Integration of Omics Approaches and Systems Biology for Clinical Applications

WILEY SERIES ON MASS SPECTROMETRY

A complete list of the titles in this series appears at the end of this volume.

Integration of Omics Approaches and Systems Biology for Clinical Applications

Edited by

Antonia Vlahou
Staff Research Scientist
Biomedical Research Foundation Academy of Athens
Athens
Greece

Harald Mischak
Robertson Chair in Biotechnology, Professor of Proteomics
University of Glasgow
Glasgow
UK

Jerome Zoidakis
Senior Research Scientist
Biomedical Research Foundation Academy of Athens
Athens
Greece

Fulvio Magni
Professor
School of Medicine and Surgery, University of Milano Bicocca
Milan
Italy

WILEY

This edition first published 2018

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication data applied for

ISBN: 9781119181149

Cover Design: Wiley
Cover Image: Designed by Theofilos Papadopoulos, PhD

Set in 10/12pt Warnock by SPi Global, Pondicherry, India

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

List of Contributors

Elena Ajello
Nephropathology Center
University of Milano-Bicocca
San Gerardo Hospital
Monza, Italy

Murat Akand
Department of Urology
School of Medicine
Selcuk University
Konya
Turkey
and
Department of Urology
School of Medicine
Katholieke Universiteit Leuven
Leuven, Belgium

Jean-Loup Bascands
Institut National de la Santé et de la Recherche Médicale (INSERM)
U1188- DéTROI- Université de La Réunion
France

Amel Bekkar
Vital-IT
SIB Swiss Institute of Bioinformatics
University of Lausanne
Lausanne
Switzerland

Maddalena Bolognesi
Nephropathology Center
University of Milano-Bicocca
San Gerardo Hospital
Monza
Italy

Daniel M. Borràs
GenomeScan B.V.
Leiden
The Netherlands

Alan Bridge
Vital-IT
SIB Swiss Institute of Bioinformatics
University of Lausanne
Lausanne
Switzerland

Giovambattista Capasso
Nephrology and Dialysis Unit
Second University of Naples
Policlinico Nuovo Napoli
Naples
Italy

Katryna Cisek
Mosaiques Diagnostics GmbH
Hannover
Germany

Riccardo Colombo
SYSBIO
Centre of Systems Biology
and Department of Informatics
Systems and Communication
University of Milano-Bicocca
Milan
Italy

Isaac Crespo
Vital-IT
SIB Swiss Institute of Bioinformatics
University of Lausanne
Lausanne
Switzerland

Elena Critselis
Proteomics Laboratory, Biotechnology Division,
Biomedical Research Foundation of the Academy of Athens
Athens
Greece

Mohammed Dakna
Mosaiques Diagnostics GmbH
Hannover, Germany

Chiara Damiani
SYSBIO
Centre of Systems Biology
and Department of Informatics
Systems and Communication
University of Milano-Bicocca
Milan, Italy

Eleni-Ioanna Delatola
Systems Biology Ireland
University College Dublin
Dublin, Ireland

Constantinos Deltas
Director, Molecular Medicine Research Center
Laboratory of Molecular and Medical Genetics
Department of Biological Sciences
University of Cyprus
Nicosia, Cyprus

Gaia De Sanctis
SYSBIO
Centre of Systems Biology
and
Department of Biotechnology and Biosciences
University of Milano-Bicocca
Milan, Italy

Julien Dorier
Vital-IT
SIB Swiss Institute of Bioinformatics
University of Lausanne
Lausanne, Switzerland

Marco Fernandes
Institute of Cardiovascular and Medical Sciences
BHF Glasgow Cardiovascular Research Centre
University of Glasgow
Glasgow, UK

Franco Ferrario
Nephropathology Center
University of Milano-Bicocca
San Gerardo Hospital
Monza, Italy

Szymon Filip
Proteomics Laboratory, Biomedical Research
Foundation
Academy of Athens
Athens, Greece

Maria Frantzi
Mosaiques Diagnostics GmbH
Hannover, Germany

Ryan B. Gil
Research Unit Analytical BioGeoChemistry
Helmholtz Zentrum München, German Research
Center for Environment Health
Neuherberg, Germany

Hiddo Lambers Heerspink
Department of Clinical Pharmacy and Pharmacology
University of Groningen, University Medical
Center Groningen
Groningen
The Netherlands

Silke Heinzmann
Research Unit Analytical BioGeoChemistry
Helmholtz Zentrum München, German Research
Center for Environment Health
Neuherberg, Germany

Holger Husi
Institute of Cardiovascular and Medical Sciences
BHF Glasgow Cardiovascular Research Centre
University of Glasgow
Glasgow, UK
and
Department of Diabetes and Cardiovascular Science
Centre for Health Science
University of the Highlands and Islands
Inverness, UK

Thomas Illig
CEO, Hannover Unified Biobank (HUB) MHH
Research Ethical Committee
Hanover
Germany

Joachim Jankowski
Institute for Molecular Cardiovascular Research
University Hospital RWTH Aachen
Aachen
Germany

Bart Janssen
GenomeScan B.V.
Leiden
The Netherlands

Julie Klein
Renal Fibrosis Laboratory
Institut National de la Santé et de la Recherche
Médicale (INSERM), U1048
Institute of Cardiovascular and Metabolic Disease

and
Renal Fibrosis Laboratory
Université Toulouse III Paul-Sabatier
Toulouse, France

Walter Kolch
Systems Biology Ireland
and
Conway Institute of Biomolecular &
Biomedical Research
and
School of Medicine
University College Dublin
Dublin, Ireland

Magdalena Krochmal
Proteomics Laboratory
Biomedical Research Foundation
Academy of Athens
Athens, Greece

Agnieszka Latosinska
Mosaiques Diagnostics GmbH
Hannover
Germany
and
Biotechnology Division
Biomedical Research Foundation
Academy of Athens
Athens, Greece

Vincenzo L'Imperio
Department of Medicine and Surgery, Pathology
University of Milano-Bicocca
San Gerardo Hospital
and
Nephropathology Center
University of Milano-Bicocca
San Gerardo Hospital
Monza, Italy

Stella Logotheti
Proteomics Laboratory
Biomedical Research Foundation
Academy of Athens
Athens
Greece

Brigitte Lohff
Institute of History, Ethics and Philosophy
of Medicine MHH
Research Ethical Committee
Hanover, Germany

Fulvio Magni
Department of Medicine and Surgery,
Proteomics and Metabolomics Unit
University of Milano-Bicocca
Monza, Italy

Manousos Makridakis
Biotechnology Division
Biomedical Research Foundation
Academy of Athens
Athens
Greece

Katerina Markoska
Medical Faculty
University "Ss. Cyril and Methodius" of Skopje
Skopje
Republic of Macedonia

Cristina Masella
Nephrology and Dialysis Unit
Second University of Naples
Policlinico Nuovo Napoli
Naples
Italy

Axel S. Merseburger
Department of Urology
University of Lübeck
Lübeck
Germany

Harald Mischak
Mosaiques Diagnostics GmbH
Hannover
Germany

Niccolò Mosele
Department of Medicine and Surgery, Proteomics and
Metabolomics Unit
University of Milano-Bicocca
Monza
Italy

Anne Niknejad
Vital-IT
SIB Swiss Institute of Bioinformatics
University of Lausanne
Lausanne
Switzerland

Alberto Ortiz
Laboratory of Nephrology
IIS-Fundacion Jimenez Diaz, School
of Medicine, UAM
and
REDinREN
and
Pathology
IIS-Fundacion Jimenez Diaz, School of Medicine, UAM
and
IRSIN, Madrid
Spain

Fabio Pagni
Department of Medicine and Surgery, Pathology
University of Milano-Bicocca
San Gerardo Hospital
and
Nephropathology Center
University of Milano-Bicocca
San Gerardo Hospital
Monza, Italy

Theofilos Papadopoulos
Renal Fibrosis Laboratory
Institut National de la Santé et de la Recherche
Médicale (INSERM), U1048
Institute of Cardiovascular and Metabolic Disease
and
Renal Fibrosis Laboratory
Université Toulouse III Paul-Sabatier
Toulouse
France

Claudia Pontillo
Mosaiques Diagnostics GmbH
Hannover
Germany

Prathibha Reddy
Institute for Molecular Cardiovascular Research
University Hospital RWTH Aachen
Aachen, Germany

Elena Sacco
SYSBIO
Centre of Systems Biology
and
Department of Biotechnology and Biosciences
University of Milano-Bicocca
Milan, Italy

Joost P. Schanstra
Renal Fibrosis Laboratory
Institut National de la Santé et de la Recherche
Médicale (INSERM), U1048
Institute of Cardiovascular and Metabolic Disease
and
Renal Fibrosis Laboratory
Université Toulouse III Paul-Sabatier
Toulouse
France

Andrew Smith
Department of Medicine and Surgery, Proteomics and
Metabolomics Unit
University of Milano-Bicocca
Monza
Italy

Goce Spasovski
Department of Nephrology
University "Ss. Cyril and Methodius," Medical Faculty
Skopje
Republic of Macedonia

Dieter Tröger
Institute for Forensic Medicine MHH
Research Ethical Committee MHH
Hanover
Germany

Marco Vanoni
SYSBIO
Centre of Systems Biology
and
Department of Biotechnology and Biosciences
University of Milano-Bicocca
Milan, Italy

Antonia Vlahou
Biotechnology Division
Biomedical Research Foundation
Academy of Athens
Athens, Greece

Ioannis Xenarios
Vital-IT
SIB Swiss Institute of Bioinformatics
University of Lausanne
Lausanne
Switzerland

Jerome Zoidakis
Proteomics Laboratory, Biomedical Research
Foundation
Academy of Athens
Athens
Greece

Preface

This book presents high-throughput analytical approaches used to investigate biological samples and omics data integration approaches that aim to offer novel solutions to clinical needs, along with two examples of their implementation in biomedical studies. Currently, there are many different experimental approaches available, and each of them provides an insight of the biological topic from a different perspective (genomics, transcriptomics, proteomics, peptidomics, metabolomics, etc.). To fully exploit the information contained in these large datasets, novel bioinformatics tools are applied. The combination of classical and computational biology has led to the development of a new discipline: systems biology. Its aim is to study biological entities globally (holistic view) rather than concentrating on their particular aspects (reductionist view).

The topics covered in this book are as follows:

a) An overview of state-of-the-art -omics techniques currently used to obtain a comprehensive molecular profile of biological specimens
b) Computational tools used for organization of these multisource data and their integration toward developing molecular models for disease pathophysiology.

As test cases the investigation of chronic kidney disease (CKD) and bladder cancer are used. These represent multifactorial, highly heterogeneous diseases and are among the most significant health issues in developed countries with a rapidly aging population. In this book, novel insights on CKD and bladder cancer obtained by "omics" data integration are presented as an example of the application of systems biology in the clinical setting.

The book is suitable for university students, researchers, and clinicians interested in clinical omics applications. The breadth of topics covered allows the reader to acquire a global view of the available omics approaches and their integration and potential for biomedical applications.

Acknowledgments

This book was conceived as a dissemination activity within the EU-funded ITN iMODE-CKD. (Clinical and system -omics for the identification of the Molecular DEterminants of established Chronic Kidney Disease, FP7-PEOPLE-2013-ITN-608332.)

The editors would like to thank all authors for their contribution.

Part I

Platforms for Molecular Data Acquisition and Analysis

1

Clinical Data Collection and Patient Phenotyping

Katerina Markoska[1] *and Goce Spasovski*[2]

[1] *Medical Faculty, University "Ss. Cyril and Methodius" of Skopje, Skopje, Republic of Macedonia*
[2] *Department of Nephrology, University "Ss. Cyril and Methodius," Medical Faculty, Skopje, Republic of Macedonia*

1.1 Clinical Data Collection

1.1.1 Data Collection for Clinical Research

The goal of clinical studies is the evaluation of interventions on clinically relevant parameters [1]. Conducting a clinical study is a major undertaking accompanied with heavy and extensive responsibilities. Good primary research calls for constant dedication by practicing physicians and patients willing to participate for the sake of knowledge and better treatment of future patients [2].

The study design is the investigator's map from which data collection follows and which enables the investigators to thoughtfully produce the necessary data forms. The formulation of a good research question, up front, informs the clinician or researcher about the most appropriate data elements to be collected [2]. Investigators often believe that collecting more data is better and that it is important to collect information on as many scientifically "interesting" factors as possible. Therefore, it is imperative to distinguish between those data elements that are essential and those that are academically "interesting" but may not be considered of interest to the key study hypothesis. This should greatly assist in narrowing down one's study questions and collecting data more efficiently [3].

1.1.2 Clinical Data Management

Clinical data management (CDM) is the process of collection, cleaning, and management of subject data in compliance with regulatory standards. The primary objective of CDM processes is to provide high-quality data by keeping the number of errors and missing data as low as possible and gather the maximum amount of data for analysis [4].

High-quality data should be absolutely accurate, have minimal or no missing points, and should be suitable for statistical analysis. The data should meet the applicable regulatory requirements specified for data quality and comply with the protocol requirements. In case of a deviation or not meeting the protocol specifications, we may think of excluding the patient from the final database [5].

Current technological developments have accelerated the rate of data collection and positively impacted the CDM process and systems by improving their quality. From the regulatory perspective, the biggest challenge would be the standardization of data management processes across organizations and development of regulations to define the procedures that has to be followed. From the industry perspective, the challenge would be the planning and implementation of data management systems in a changing operational environment. CDM is evolving to become a standard-based clinical research entity, balancing between the expectations from and constraints in the existing systems, driven by technological developments and business demands [5].

The Society for Clinical Data Management (SCDM) publishes Good Clinical Data Management Practices (GCDMP) guidelines that highlight the minimum standards and best practices, providing assistance to clinical data managers in their implementation of high-quality CDM [5]. If data have to be submitted to regulatory authorities, it should be entered and processed in accordance with the Code of Federal Regulations (CFR), Title 21, Volume 1 of Part 11, Food and Drug Administration (FDA) regulations on electronic records and electronic signatures (ERES), cited as 21CFR11.10 [6].

Many clinical data management systems (CDMS) are available for data management. Most of the CDM systems available meet these criteria, and pharmaceutical companies as well as contract research organizations

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

ensure this compliance. In multicentric trials, a CDMS has become essential to handle the huge amount of data. These CDM tools ensure the audit trail and help in the management of discrepancies [5].

One should leave sufficient time for planning and development of system and study database for the follow-up and tracking of patients throughout the study. The following issues should be defined in advance: determination of the mechanism and processes for data collection if a patient misses a scheduled appointment, implementation of quality checks, preparation and collection of patient informed consent, and institutional review board (IRB) approval. Inclusion and exclusion criteria should be defined as well as data collection elements [2].

1.1.3 Creating Data Forms

The limited focus of disease-specific consortia makes comprehensive coverage of individual areas more likely. Researchers should benefit from a clear understanding of the extensive overlap of various clinical terminologies, as well as advice regarding which standards are appropriate for a particular research context. They should also be able to address relationships between clinical research data collection standards and electronic health records (EHR) specifications, as well as the broad issue of secondary use of clinical data for research. Additional tasks could include the review of standards and their scope and relating them to needs of clinical research [7].

Item repositories can reduce the burden on new investigators to create their own items, because existing validated items or sets of items can be reused [8]. Pilot testing of data forms completed by patients allows investigators to react to suggestions from patients as well as from staff and personnel and provides more realistic estimates of data collection times [2].

1.1.3.1 Different Data Forms According to the Type of Study

Data form development is a collaborative effort among the investigators and often takes months of planning and preparation. It should be undertaken by investigators and/or stakeholders experienced in form construction and familiar with the methods of data collection, data processing, and content necessary for the study [2]. It is facilitated by review of the literature for instruments used in similar studies, also including the Clinical Data Acquisition Standards Harmonization (CDASH) recommendations, which give useful general guidance on constructing yes/no questions, scale direction, date/time formats, scope of CRF data collection, pre-populated data, and collection of calculated or derived data. Certain items (especially questionnaire-based ones) have a discrete set of permissible values (also called "responses" or "answers"), for example, the use of cigarettes (never/former/current), amount (less than 3 per day/3–10 per day/more than 10 per day), and fasting (no/yes) [7].

Study details like objectives, intervals, visits, investigators, sites, and patients should be defined in the database, and CRF layouts have to be designed for data entry [5]. In order to simplify the data collection, some answers can be coded. For example, 1 = yes and 2 = no, but these codes should be consistent throughout the CRF booklet (Table 1.1) [9].

The forms should be well designed in order to avoid variation in the responses and the site personnel can understand the format (Table 1.2) [9].

Much of the information collected in observational epidemiologic studies is collected in the form of patient/participant self-reports on standardized questionnaires that are self-administered or administered in person by an interviewer, by phone, or via mail or the Internet. The factors on which information is routinely collected in these studies include sociodemographic characteristics, lifestyle practices, medical history, and use of prescribed or over-the-counter medications [3].

Surveys are tools of great value for epidemiological research and clinical practice. They can be used as a study design, at same time serving as definitive data collection tool. On the other hand, clinical registries

Table 1.1 Coding on the case report form module.

Demography	
Date of birth (DD/MM/YYYY)	☐☐/☐☐/☐☐☐☐
Gender	Male ☐ 1 Female ☐ 2
Height (cm)	☐☐☐.☐
Weight (kg)	☐☐☐.☐☐
Smoker	Yes ☐ 1 No ☐ 2
Family history	Yes ☐ 1 No ☐ 2

Table 1.2 Well-designed and poorly designed data fields.

Poorly designed	Well designed
Date of visit:_____	Date of visit: ☐☐/☐☐/☐☐☐☐ (DD/MM/YYYY)
Blood pressure:____/____	Blood pressure: ☐☐☐/☐☐☐ (mmHg)
Pulse: _____	Pulse: ☐☐☐ (beats/min)
Temperature:_____	Temperature: ☐☐.☐ (°C)
Respiration: _____	Respiration: ☐☐ (/min)

can be also used to obtain specific data within a more comprehensive design [10].

Ideally, patient-reported outcomes are measured using standardized, validated instruments that promote the collection of high-quality data and allow meaningful comparisons across observational studies or randomized trials. National Institutes of Health Toolbox (www.nihtoolbox.org) and Patient-Reported Outcomes Measurement Information System (www.nihpromis.org) have highlighted the importance of harmonization of patient-reported outcomes data collection instruments [3].

Clinical Document Architecture templates—archetypes—are agreed-upon specifications that support rigorous computable definitions of clinical concepts like type of measure, measurement conditions, and measurement units [11].

OpenEHR, by contrast, allows the semantic model's structure to vary with the parameter being described. Clinical researchers have been specifying parameter measurement with precision long before "archetypes" were conceived [7].

CDASH addresses data collection standards through standardized CRFs. Initial CDASH standards focused on cross-specialty areas such as clinical trial safety.

The Clinical Data Interchange Standards Consortium (CDISC) (http://www.cdisc.org) is an international standards organization that aims to develop and support global, platform-independent data standards. The consortium has proposed standards valuable for general areas such as drug safety, focusing primarily on regulated studies, and does not address broader issues of clinical research. They also have CDISC Operational Data Model (ODM) for exchanging and archiving clinical study data [12]. Another clinical information model is the Health Level 7 (HL7) Reference Information Model (RIM), which use terminologies differently than the research-oriented CDISC (ODM) format. HL7 depends on mapping data elements to concepts in standard terminologies, while ODM does not support mapping of data elements themselves (e.g., serum total cholesterol, systolic BP) to terminologies and only cares that a terminology may act as a source for a data element's contents [13].

1.1.4 Case Report Form (CRF)

1.1.4.1 CRF Standards Characterization

Data collection for clinical research involves gathering variables relevant to research hypotheses. These variables ("patient parameters," "data items," "data elements," or "questions") are incorporated into data collection forms ("case report forms" (CRFs)) for study implementation [4].

CRF may exist in the form of a paper or an electronic version. The traditional method is to employ paper case report forms (pCRFs) to collect the data responses, which are translated to the database. These pCRFs are filled up by the investigator according to the completion guidelines. In the electronic case report form (eCRF)-based CDM, the investigator or a designee will be entering the data directly at the site. In the case of eCRF, the probability of erroneous data entry is lower, and the resolution of discrepancies faster [5].

A CRF is designed by the CDM team, as this is the first step in translating the protocol-specific activities into data being generated. The units in which measurements have to be made should also be mentioned next to the data field [5].

Because of the protocol-centric nature of clinical research, opportunities for shared standards at levels higher than individual items are relatively limited. Nevertheless, disease-specific CRF standardization efforts have helped identify standard pools of data items within focused research and professional communities and consequently helped achieve research efficiencies within their application areas. Of more immediate and widespread relevance are standardization efforts toward the development of section and workflow for CRF, as well as data collection and validation. The structure and content of individual CRFs/sections can be left reasonably flexible to allow adaptation to individual protocol requirements [7].

Little consensus exists on the choice and content of CRF standardization candidates. Few CRFs can be reused unchanged across all protocols. Within a specific disease domain, standard CRFs seem feasible and useful. But the segregation of data items relevant to a research protocol into individual CRFs is often based on considerations other than logical grouping and may vary with the study design. One concern about "standard" CRF use is that users should not be pressured to collect parameters defined within the CRF that are not directly related to a given protocol's research objectives. Dynamic CRF rendering offers one way out of this dilemma: protocol-specific CRF customization allows individual investigators to specify, at design time, the subset of parameters that they consider relevant. Also, web application software can read the customization metadata and render only applicable items [7].

Generally, a programmer/designer performs the CRF annotation, creates the study database, and programs the edit checks for data validation. He/she is also responsible for designing of data entry screens in the database and validating the edit checks with dummy data [5]. Databases are the clinical software applications, which are built to facilitate the CDM tasks to carry out multiple studies [14].

CRFs can be used in groups of semantically closely related parameters, which can be considered as a series

of observations. A section encompasses one or more groups. The division of CRFs into sections is often arbitrary. In paper-based data entry, CRFs consisting of a single, oversize section are sometimes used. Real-time electronic data capture (EDC) subdivision into smaller sections is generally preferred. Section headings and explanations that serve to describe the section's purpose are usually left to individual investigators [7].

1.1.4.2 Electronic and Paper CRFs

CRFs support either primary (real-time) data collection or secondarily recorded data originating elsewhere (e.g., the electronic or paper health records). Historically, CRFs were paper based. The existence of secondary EDC also influences manual workflow processes related to verification of paper-based primary data (e.g., checks for completeness, legibility, and valid codes) [7].

Although the use of validated, standardized instruments is preferred, those data collection tools are not always available. If standardized instruments do not exist for measuring a specific construct, investigators will often create "homegrown" scales, which require pilot test before using them in a formal research study. These pilot efforts ideally would involve validation of the instrument against a gold standard (e.g., clinical diagnosis) or important study outcome [3].

Collection of individual patient data on CRFs in clinical research has traditionally been done by investigators in their offices summarizing medical charts on paper forms (pCRFs), a tedious method that could result in data entry errors and wrong conclusions [15, 16].

eCRFs have improved data quality and completeness, reducing losses and transport logistics, especially for multicenter trials [17]. The choice between pCRF and eCRF is a significant step in the design of clinical studies and should be discussed with the involved stakeholders [18].

EHR and research data collection differ in that the latter records a subset of patient parameters and variables defined with the research protocol. Data are recorded in maximally structured form, avoiding narrative text, except if there is a need to record unanticipated information [7].

Le Jeannic et al. have compared the application of eCRF and pCRF and their results showed that eCRF studies were mostly used in large multicenter, national, and phase 3 clinical trials while pCRF studies were used for trials with few patients and centers. The majority of pCRFs were used in drug trials, and eCRFs were more often used in trials with a significantly higher number of patients and fewer data. The number of patients was the only explanatory variable for CRF choice. They found no difference in the average duration of recruitment. Use of eCRF and the smaller number of centers were associated with shorter study durations. The total average cost of a trial was higher with eCRFs than with pCRFs, but the mean cost per patient was lower with eCRFs. Overall, stakeholders were as satisfied with eCRFs as with pCRFs. When asked for their preference of one over the other, a majority of stakeholders chose eCRF. Preference for pCRFs is reported in monocentric trials and for eCRFs in multicentric trials. Additional advantages of eCRFs are the prevention of data entry errors by automatic checks, easier storage, and the ability of researches to oversee data collection from their offices [18].

1.1.5 Methods and Forms for Clinical Data Collection and/or Extraction from Patient's Records

In particular, a patient summary has been seen as the most appropriate way to establish eHealth interoperability. A patient summary includes patient history, allergies, active problems, test results, and medications. However, further information can be included, depending on the intended purpose of the summary and the anticipated context of use [19].

Because of the ubiquity and abundance of high-quality data embedded within medical records, they are a commonly used source of information in clinical research studies. Medical records can be important sources of information that can reliably document participants' medical history, clinical, laboratory, or physiologic profile at varying time points in a cost-efficient manner. On the other hand, the data contained in medical records can be difficult to use and, in some cases, conflicting or of questionable accuracy because of the nonstandardized manner in which this information is collected, recorded, and extracted by various healthcare professionals and members of research teams. The increasing use of electronic medical records (EMR) and their combination with administrative data have eased data extraction efforts. Moreover, the increasing use of standardized data entry sets reduced data heterogeneity [3].

1.1.5.1 Electronic Health Records (EHRs)

EHRs are basically seen as a centralized compilation of information on the patient's health [20]. The data included in paper-based patient records has provided the golden standard against which the reliability of EHRs has been assessed. The success of EHRs depends on the quality of the information available to healthcare professionals in making decisions about patient care and in the communication between healthcare professionals during patient care [19]. It has been shown that data from EHRs are reliable when compared with manual records [21, 22].

One challenge is to standardize health information systems, which also means standardization of the content and structure of EHRs [23]. EHRs have so far consisted

of unstructured narrative text but also contain structured coded data [19]. EHR contains retrospective, concurrent, and prospective information, and its primary purpose is to support continuing, efficient, and quality-integrated healthcare [24]. An EHR is used primarily for purposes of setting objectives and planning patient care, documenting the delivery of care, and assessing the outcomes of care. It includes information regarding patient needs during episodes of care provided by different healthcare professionals [25, 26]. The EHR is used by different healthcare professionals and also by administrative staff. Among the various healthcare professionals who use different components of the EHR are physicians, nurses, radiologists, pharmacists, laboratory technicians, and radiographers [19]. EHRs are used by many different healthcare professionals, and the needs and requirements of all these professionals must be taken into account in the development of the information systems. Nursing documentation, or documentation by other healthcare professionals such as physiotherapists, is an important part of the EHR and must not be excluded from medical documentation. Patients can also do parts of the documentation themselves. Patient self-documentation also reduces the workload of healthcare professionals, but it is obviously important that self-documented data components are validated by professionals [19].

Previously EHRs were classified as time oriented, problem oriented, and source oriented. Nowadays EHRs combine all three elements. In the time-oriented EMR, the data are presented in chronological order. In the problem-oriented medical record, notes are taken for each problem assigned to the patient, and each problem is described according to the subjective information, objective information, assessments, and plan. In the source-oriented record, the content of the record is arranged according to the method by which the information was obtained, for example, notes of visits, X-ray reports, and blood tests [19].

Electronic clinical records, such as conventional clinical histories, can display major shortcomings in terms of quality of information, lack of data, incomplete information, and use of multiple free terms. Before electronic clinical records can replace registries or surveys, a common terminology and set of standards must be established to encode and classify the information, and a change must be brought about in the attitude of health professionals tasked with data collection [10].

The possibility of using electronic clinical histories as a data source may depend upon the degree to which this is used within the health organization and/or system (sole data collection source, data also recorded in paper format, etc.), the completeness and coding of recorded data, and also the software available for data collection and transfer [27].

Introducing an online medical record system could play an important role in improving data collection and data quality [28].

1.1.6 Data Collection Workflow

1.1.6.1 Defining Baseline and Follow-Up Data

Before data collection begins, investigators must agree on the details of the data collection items and the process by which data collection will occur. Investigators must define the schedule according to which patients will participate in the study and outline the specific data elements to be collected each time the patient is examined. If the researcher understands office flow and can organize the follow-up process, then his or her office can map data collection in a simple and efficient manner [2].

In most cases, it is best to collect all the required initial data for a subject during a single visit at the clinic. Several steps and design features are recommended to optimize follow-up rates [2].

In order to minimize the respondent burden, follow-up questionnaires and tests should be kept to a minimum. Contact information should be collected at baseline and updated at every visit for data collection, whereas subjects with no telephone or who plan to move in the near future should be excluded. Clinicians need to plan multiple efforts at phone contact, both during and after working hours, and provide reminders for appointments. Follow-up forms should include information about treatment compliance and the exposure of patients to various operative and nonoperative treatments. Follow-up forms must also include data regarding side effects and complications of treatment (e.g., monitored events) and whether they are related to the study treatment(s). In addition to baseline and follow-up patient data, information regarding treatment must be collected [2].

1.1.6.2 Medical Coding

Pre- or coexisting illnesses are coded using the available medical dictionaries. Medical Dictionary for Regulatory Activities (MedDRA) is used for the coding of adverse events as well as other illnesses. The World Health Organization Drug Dictionary Enhanced (WHODDE) is used for coding the medications. Medical coding helps in classifying reported medical terms on the CRF to standard dictionary terms in order to achieve data consistency and avoid unnecessary duplication. The right coding and classification of adverse events and medication is crucial as an incorrect coding may lead to masking of safety issues or highlight the wrong safety concerns related to the drug [5].

It also is important to note that factors (e.g., medication use) must be defined only by clinicians, and not by study staff or study participants, in order to ensure that variables will be accurately coded [3].

1.1.6.3 Errors in Data Collection and Missing Data

If the data are inaccurate or incomplete, they will have no worth for decision-making, research, statistical, or health policy purposes [15].

One common cause of errors is not dedicating enough time in the development of data forms (i.e., in identifying the data elements and in the construction and testing of forms). An important problem is the desire of clinicians not only to create forms to meet the research goals of the study but also to provide data for routine patient care. Certain measurements needed for routine patient care are not justifiable for research forms, and vice versa. The researcher should be clear about the data necessary for assessing the primary and secondary outcomes of the study. Data items that cannot be justified should be deleted [2].

Double data entry is performed wherein the data is entered by two operators separately. The second pass entry helps in data verification by identifying the transcription errors and discrepancies caused by illegible data. Double data entry helps in getting a cleaner database compared with a single data entry. Double data entry ensures better consistency with pCRF as denoted by a lesser error rate [29, 30].

It is difficult to gather the necessary data elements at the appropriate times while avoiding missing data. It is even more difficult to collect primary data according to a very strict protocol, wherein chance, bias, and confounding factors can be addressed. No matter how sophisticated the data elements and data collection systems, human factors make or break any good research effort [2].

With increasing duration of a study, the number of participating patients usually declines, so the problem of missing data is magnified. Thus, data interpretation for long-term studies is challenging [31].

A frequently employed approach for data analysis of clinical (long-term) studies is the interpretation of missing data as therapeutic success (missing equals success (MES)) or as therapeutic failure (missing equals failure (MEF)/nonresponder imputation). A third option the exclusion of missing data (missing equals excluded (MEX)/as-treated) stands between these two extremes. Another frequently employed method for long-term studies that is criticized by statisticians is equating the last observed value with the result at the end of the study (last observation carried forward (LOCF)). The selection of the analysis method has a great impact on the results and interpretation of a study. It is recommended to combine several data analysis approaches in order to correctly interpret long-term studies and reach valid conclusions. A comparison of the characteristics of test subjects with complete as opposed to those with incomplete datasets might be helpful, in order to get indications on the possible reasons for dropping out of the study or for missing data [1].

1.1.6.4 Data Linkage, Storage, and Validation

The data management group (those responsible for data retrieval and processing) needs to link records by using a unique identifier for each patient. For example, they might use the patient hospital identification (for purposes of confidentiality, patient identification can be deleted later) plus check digits that identify the patient, the center from which the patient comes, and the type of visit [2].

The data bank must be backed up regularly. Data collection is too difficult and expensive to repeat, and patients' time is too valuable to have data lost or destroyed [2].

Data validation is the process of testing the validity of data in accordance with the protocol specifications. Discrepancy is defined as a data point that fails to pass a validation check, and it may be due to inconsistent data, missing data, range checks, and deviations from the protocol. Ongoing quality control of data processing is undertaken at regular intervals during the course of CDM. Data clarification forms (DCFs) containing queries pertaining to the discrepancies identified can be also generated [5].

In order to prevent errors from being entered, data validation rules should be implemented into the eCRF's prior to commencement of the NPC clinical trial. These data validation rules assess whether certain prespecified conditions are valid and can therefore pinpoint omissions or erroneous data [28].

The clinical trial data management system (CTDMS) prevents us from missing data or ending up with poor quality data at the end of the study, which often at that point cannot be resolved anymore [28].

Clearly, the CTDMS encourages local data managers to verify the entered data and, if necessary, ask the doctor whether the information is correct [28].

Discrepancy management helps in cleaning the data and gathers enough evidence for the deviations observed. Discrepancy management is the most critical activity in the CDM process. Being the vital activity in cleaning up the data, utmost attention must be observed while handling the discrepancies [5].

After a proper quality check and assurance, the final data validation is run. Database is locked and clean data is extracted for statistical analysis. Data extraction is done from the final database after locking. This is followed by its archival [5].

1.2 Patient Phenotyping

In clinical care settings, a wealth of longitudinal data is available through International Statistical Classification of Diseases v9 (ICD-9) codes, laboratory results,

test reports, and notes written by the physicians during multiple patient visits over several years. Essential point in improving the formation of research cohorts has been the creation of EMR-linked biobanks and enrolment of individuals from routine clinical care settings. With patient's consent and by making their data anonymous, EMRs can make a large amount of information available for research purposes, allowing studying the evolution and progression of the disease [32–34]. In this regard, using large number of patient's data from EMRs can markedly reduce the time and effort needed to identify specific phenotypes and/or markers associated with disease development, progression, and response to treatment [34–36].

1.2.1 Approaches in Defining Patient Phenotype

The Electronic Medical Records and Genomics (eMERGE) Network is a National Human Genome Research Institute (NHGRI)-funded consortium tasked with developing methods and best practices for the utilization of the EMR as a tool for genomic research. The network is developing phenotyping algorithms that are processing EMR data in order to identify cases and controls with a high degree of accuracy and confidence [37]. However, identification of particular phenotypes, especially chronic complex diseases, is challenging because of the complexity of data itself and the way in which it is recorded in EMR [38].

There has been significant debate about the optimal way to identify phenotypes in the EMR. Automated approaches using electronic phenotyping and statistical analyses are popular as compared with simpler rule-based systems. Such phenotyping algorithms are used in various applications including discovering novel genetic associations of complex diseases, tracking their natural history, isolating patients for clinical trials, and ensuring quality control in large institutions by ensuring that standard-of-care guidelines are met in these patients [38, 39].

Phenotyping algorithms are dedicated to mining biobank resources, which is essential for trial designs, and need to enable automatic identification of patients that match the research criteria. They contain keywords designed to facilitate natural language processing (NLP) and access the primarily semi-structured data fields in EHRs—procedure codes, ICD-9 codes, laboratory results, and medication data [40].

NLP content is required for enormous unstructured narrative clinical documentation that is considered to be the best resource. This is the most difficult part in phenotype algorithm construction, and although there are many NLP tools for medical domains, human involvement is still required [40].

The V-Model is a temporal model that enables visualization of textual information in a timeline, which helps in monitoring and understanding a patient's history. The model separately structures causal problems from related actions, representing apparent problems–actions (P-A) relations, and enables to extrapolate that problems occurred before actions. It enables a user to trace patient history considering semantic, temporal, and causality information in a short time. Consequently the V-Model should play a crucial role in phenotype definition and algorithm development [41].

For several relatively common conditions, such as heart failure and stroke, independently and extensively validated algorithms have been developed to ascertain the presence of these important chronic diseases [42]. eMERGE Network developed 14 robust algorithms that were extensively tested over multiple iterations (Table 1.3). The core elements of the algorithms are the administrative data (ICD-9 and CPT codes), laboratory data, and medication data (RxNorm codes), with NLP rules as an additional layer to disambiguate and refine the core data elements. Most of the developed algorithms rely on ICD-9 disease codes and use CPT procedure codes. Only two algorithms used UMLS codes (due to site-specific processing needs) [40].

1.2.2 Phenotyping CKD Patients

Clinical decision making is challenging due to variability in the rates of progression and lack of widely accepted guidelines to identify patients most at risk of progression to ESRD [43, 44].

Currently the only way of identifying CKD cases/controls is by manually reviewing laboratory values, which is cumbersome, or through ICD-9 codes. To accomplish these goals, researchers need robust phenotyping algorithms to effectively leverage disparate data sources in the EMR [38].

Nadkarni et al. developed and validated an automated algorithm for identifying diabetic/hypertensive CKD cases and controls. Their algorithm over-performed the traditional identification using ICD-9 diagnostic codes, which enabled identification of 40.1% of cases and 75.0% of controls. Their algorithm correctly identified 93.4% of cases and 95.8% of controls, indicating that it could be used for both research and clinical purposes, where rapid and accurate identification of a specific target cohort is needed [38].

Table 1.3 Distribution of codes across the 14 eMERGE phenotyping algorithms.

	Name	ICD-9[a]	CPT[b]	UMLS[c]	RxNorm[d]	Total/WC[e]	Percentage[f]
1	Alzheimer's	29	0	0	355	384/1 317	29
2	Dementia	30	0	0	20	50/634	7
3	Diabetic retinopathy	12	19	0	0	31/324	10
4	Height	156	0	0	11	167/2 101	8
5	Hypothyroidism	43	76	0	0	119/1 351	9
6	Serum lipid level	11	0	0	0	11/1 091	1
7	Low HDL[g] cholesterol level	41	10	0	0	51/2 579	2
8	Peripheral arterial disease	90	112	0	0	202/1 353	15
9	QRS duration	50	157	595	0	802/26 695	3
10	Red blood cell indices	146	141	0	0	287/2 857	10
11	Resistant hypertension	35	0	0	0	35/895	4
12	Type 2 diabetes	25	0	0	0	25/954	3
13	White blood cell indices	18	131	0	0	149/2 458	6
14	Cataract	152	20	35	0	207/3 152	6

[a] Number of ICD-9 (International Statistical Classification of Diseases, v9) codes present in the algorithm document.
[b] Number of CPT (Current Procedure Terminology) codes in document.
[c] Number of UMLS (Unified Medical Language System) codes in the document.
[d] Number of RxNorm (clinical drug) codes in the document.
[e] Total number of codes divided by the number of word tokens.
[f] Percentage of the document's word tokens that are codes.
[g] High-density lipoprotein.

1.3 Concluding Remarks

Recommendations for standardization of CRF and transition from paper to electronic health records have significantly improved and accelerated the process of clinical data collection. Technological developments in the field of CDM have enabled fast, accurate, and simplified extraction of information from enormous clinical documentations for best quality of data generated. NLP tools that are used for extracting unstructured narrative clinical data and EHR-oriented phenotyping algorithms should enable automatic selection of cases for clinical trials and other research purposes.

References

1 Boehncke, W.H., Clinical long-term studies: data collection and assessment. *J Dtsch Dermatol Ges*, 2011. 9(6): p. 479–484.
2 Weinstein, J.N. and R.A. Deyo, Clinical research: issues in data collection. *Spine*, 1976. 25(24): p. 3104–3109.
3 Saczynski, J.S., D.D. McManus, and R.J. Goldberg, Commonly used data-collection approaches in clinical research. *Am J Med*, 2013. 126(11): p. 946–950.
4 Gerritsen, M.G., et al., Data management in multi-center clinical trials and the role of a nation-wide computer network. A 5 year evaluation. *Proc Annu Symp Comput Appl Med Care*, 1993: p. 659–662.
5 Krishnankutty, B., et al., Data management in clinical research: an overview. *Indian J Pharmacol*, 2012. 44(2): p. 168–172.
6 Code of Federal Regulations Title 21. http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=11.10 (accessed August 24, 2017).
7 Richesson, R.L. and P. Nadkarni, Data standards for clinical research data collection forms: current status and challenges. *J Am Med Inform Assoc*, 2011. 18(3): p. 341–346.
8 Brandt, C.A., et al., Approaches and informatics tools to assist in the integration of similar clinical research questionnaires. *Methods Inf Med*, 2004. 43(2): p. 156–162.
9 Bellary, S., B. Krishnankutty, and M.S. Latha, Basics of case report form designing in clinical research. *Perspect Clin Res*, 2014. 5(4): p. 159–166.

10 Varela-Lema, L., A. Ruano-Ravina, and T. Cerda Mota, Observation of health technologies after their introduction into clinical practice: a systematic review on data collection instruments. *J Eval Clin Pract*, 2012. 18(6): p. 1163–1169.

11 Browne E. Archetypes for HL7 CDA Documents, 2008. https://openehr.atlassian.net/wiki/spaces/stds/pages/5373955/openEHR+Archetypes+for+HL7+CDA+Documents?preview=/5373955/5537794/Archetypes_in_CDA_4.pdf (accessed October 12, 2017).

12 CDISC. Clinical Data Acquisition Standards Harmonization: Basic Data Collection Fields for Case Report Forms. Draft version 1.0, 2008.

13 Richesson, R.L. and J. Krischer, Data standards in clinical research: gaps, overlaps, challenges and future directions. *J Am Med Inform Assoc*, 2007. 14(6): p. 687–696.

14 Fegan, G.W. and T.A. Lang, Could an open-source clinical trial data-management system be what we have all been looking for? *PLoS Med*, 2008. 5(3): p. 0050006.

15 Day, S., P. Fayers, and D. Harvey, Double data entry: what value, what price? *Control Clin Trials*, 1998. 19(1): p. 15–24.

16 Nahm, M.L., C.F. Pieper, and M.M. Cunningham, Quantifying data quality for clinical trials using electronic data capture. *PLoS One*, 2008. 3(8): p. 0003049.

17 Thwin, S.S., et al., Automated inter-rater reliability assessment and electronic data collection in a multi-center breast cancer study. *BMC Med Res Methodol*, 2007. 7: p. 23.

18 Le Jeannic, A., et al., Comparison of two data collection processes in clinical studies: electronic and paper case report forms. *BMC Med Res Methodol*, 2014. 14(7): p. 1471–2288.

19 Hayrinen, K., K. Saranto, and P. Nykanen, Definition, structure, content, use and impacts of electronic health records: a review of the research literature. *Int J Med Inform*, 2008. 77(5): p. 291–304.

20 Friedman, D.J., Assessing the potential of national strategies for electronic health records for population health monitoring and research. *Vital Health Stat*, 2006. 2(143): p. 1–83.

21 Stratmann, W.C., A.S. Goldberg, and L.D. Haugh, The utility for audit of manual and computerized problem-oriented medical record systems. *Health Serv Res*, 1982. 17(1): p. 5–26.

22 Jamison, R.N., et al., Electronic diaries for monitoring chronic pain: 1-year validation study. *Pain*, 2001. 91(3): p. 277–285.

23 European Commission. e-Health—Making Healthcare Better for European Citizens: An Action Plan for a European e-Health Area, 2004.

24 ISO/DTR 20514, Health Informatics—Electronic Health Record—Definition, Scope, and Context, 2004. https://www.iso.org/standard/39525.html (accessed October 12, 2017).

25 van Ginneken, A.M., The computerized patient record: balancing effort and benefit. *Int J Med Inform*, 2002. 65(2): p. 97–119.

26 Grimson, J., Delivering the electronic healthcare record for the 21st century. *Int J Med Inform*, 2001. 64(2–3): p. 111–127.

27 Vlug, A.E., et al., Postmarketing surveillance based on electronic patient records: the IPCI project. *Methods Inf Med*, 1999. 38(4–5): p. 339–344.

28 Wildeman, M.A., et al., Can an online clinical data management service help in improving data collection and data quality in a developing country setting? *Trials*, 2011. 12(190): p. 1745–6215.

29 Cummings, J. and J. Masten, Customized dual data entry for computerized data analysis. *Qual Assur*, 1994. 3(3): p. 300–303.

30 Reynolds-Haertle, R.A. and R. McBride, Single vs. double data entry in CAST. *Control Clin Trials*, 1992. 13(6): p. 487–494.

31 Committee for Proprietary Medicinal Products (CPMP). Points to consider on missing data: the European Agency for the Evaluation of Medicinal Products (EMEA), 2001.

32 Blumenthal, D. and M. Tavenner, The "meaningful use" regulation for electronic health records. *N Engl J Med*, 2010. 363(6): p. 501–504.

33 McCarty, C.A., et al., Informed consent and subject motivation to participate in a large, population-based genomics study: the Marshfield Clinic Personalized Medicine Research Project. *Community Genet*, 2007. 10(1): p. 2–9.

34 Denny, J.C., et al., PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, 2010. 26(9): p. 1205–1210.

35 Carroll, R.J., A.E. Eyler, and J.C. Denny, Naive Electronic Health Record phenotype identification for Rheumatoid arthritis. *AMIA Annu Symp Proc*, 2011. 2011: p. 189–196.

36 Birman-Deych, E., et al., Accuracy of ICD-9-CM codes for identifying cardiovascular and stroke risk factors. *Med Care*, 2005. 43(5): p. 480–485.

37 Gottesman, O., et al., The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genet Med*, 2013. 15(10): p. 761–771.

38 Nadkarni, G.N., et al., Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu Symp Proc*, 2014. 14: p. 907–916.

39 Shivade, C., et al., A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*, 2014. 21(2): p. 221–230.

40 Conway, M., et al., Analyzing the heterogeneity and complexity of Electronic Health Record oriented phenotyping algorithms. *AMIA Annu Symp Proc*, 2011. 2011: p. 274–283.

41 Park, H. and J. Choi, V-Model: a new perspective for EHR-based phenotyping. *BMC Med Inform Decis Mak*, 2014. 14(90): p. 1472–6947.

42 Carnahan, R.M., Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. *Pharmacoepidemiol Drug Saf*, 2012. 21 (Suppl 1): p. 90–99.

43 Keane, W.F., et al., Risk scores for predicting outcomes in patients with type 2 diabetes and nephropathy: the RENAAL study. *Clin J Am Soc Nephrol*, 2006. 1(4): p. 761–767.

44 Keith, D.S., et al., Longitudinal follow-up and outcomes among a population with chronic kidney disease in a large managed care organization. *Arch Intern Med*, 2004. 164(6): p. 659–663.

3

Nephrogenetics and Nephrodiagnostics

Contemporary Molecular Approaches in the Genomics Era

Constantinos Deltas

Director, Molecular Medicine Research Center, Laboratory of Molecular and Medical Genetics, Department of Biological Sciences, University of Cyprus, Nicosia, Cyprus

3.1 Introduction

The identification of all kinds of causative mutations by traditional and more contemporary technologies of DNA sequencing in monogenic disorders has become the gold standard in molecular genetic diagnosis during the past couple of decades. It is true though that there are still numerous occasions where, especially in smaller research or diagnostics laboratories, the indirect approach of DNA linkage analysis is a very useful alternative. This is especially the case when DNA samples are available from multiple family members and the gene or genes under investigation are very large or difficult to negotiate by direct sequence analysis, in the absence of contemporary technology such as next-generation sequencing (NGS) equipment. This practice is used in the author's laboratory. The situation is progressively changing as technology improvements enable smaller laboratories to acquire and use robust technology for DNA analysis or have access to core facilities for doing so. Also, the recent development that enables whole-genome sequencing (WGS) or whole-exome sequencing (WES) is a revolutionary improvement that is rapidly shaping the field and has already entered clinical practice and contributes to realizing a long-sought goal for personalized or precision medicine. Here, a word of caution should be mentioned as the 1000-dollar genome sequencing approach in most places is still accompanied by a greater analysis cost, let alone the complexity and the difficulty in interpreting a huge volume of data [1–3]. Even this, though, is changing rapidly. During the preparation of the last draft of this chapter, Veritas Genetics, a company in Boston, United States, announced the service of WGS for a price that broke the barrier of $1000; they offer the sequencing and the interpretation of data for $999 to participants in the Personal Genome Project (PGP) (http://www.veritasgenetics.com/).

At the same time, the advent of technologies that enable one to test simultaneously for a great number of single nucleotide polymorphisms (SNPs) or disease-causing mutations by using microchip arrays or other high-throughput technologies has contributed to the field of molecular diagnostics in a different way. This takes advantage of the known mutations that are responsible for certain monogenic disease phenotypes or for SNPs of pharmacogenetic significance and especially so in populations and for diseases where a significant percentage of patients inherit one of several known more prevalent mutations or sequence variants.

Good examples include the repertoire of mutations for cystic fibrosis, β-thalassemia, and familial Mediterranean fever. In most cases, especially for autosomal dominant or X-linked disorders, where most families have their own private mutation, one needs to take advantage of high-throughput mutation screening methods and direct automated DNA sequencing in order to achieve provision of a molecular genetic diagnostic result of high confidence and practical usefulness. Knowledge of elementary genomics and genetics as applied to the human medical genetics field is becoming indispensable for the current and next generation of medical clinicians and other medical practitioners. The understanding of the mode of inheritance of disease-causing mutations as well as the role of predisposing variants, and coming to an understanding of the new generation of laboratory genetics, is going to be a *sine qua non* in order for the laboratory geneticist and clinical geneticist to communicate. This is going to be an even more indispensable skill if one would like to be able to follow contemporary genetics and genomics literature. Of course, the revolutionary advent of, and implementation of parallel massive NGS approaches, enabled scientists and researchers in many cases to get away from the targeted search of mutations in single genes. Examples are the many multiple gene

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

panels that have been developed and used already by numerous diagnostic and research laboratories in the search for new variants and new mutations associated with specific phenotypes, aimed at better understanding of disease pathomechanisms [4–6].

Kenneth I. Berns, former editor in chief of *Genetic Testing* and director of the University of Florida's Genetics Institute, College of Medicine in Gainesville, Florida, has stated:

> As we learn more about the underlying causes of disease and link this knowledge to the emerging realization that in the not too distant future good healthcare will include information about one's genomic profile, the importance of genetic testing in clinical medicine will continue to grow.

I would like to add that in my personal view, the sacred purpose of our work as geneticists is to make a link from the dramatic picture of the affected individual—the patient as a human macroentity—to the patient as a molecular biological microentity. The next feat is the prevention or correction of the molecular malady. Therefore, deep understanding of the behavior of genetic phenomena and the mode of action of the myriads of human genetic variants is also becoming a *sine qua non* knowledge.

Genetic complexity at various levels is the norm and perhaps 20 years ago we could have predicted it. Notwithstanding this, the ability to uncover the mysteries, the interdependencies, and the cross-talk of protein coding as well as noncoding genes has led to a new level of satisfaction in regard to our achievements in providing useful information to the patients and their relatives, even though it has not become possible yet to offer genetic therapy, a dream that drove a generation of geneticists. One more word concerning complexity must be said about the emerging role of epigenetic phenomena. The regulation of gene expression by DNA methylation and histone modification is playing crucial roles in health and disease, including cancer, although we will not elaborate any further in this chapter.

3.2 Applications of Molecular Diagnostics

During the past two decades, nephrogenetics research has been extremely fruitful and productive as medical and molecular genetic investigations have led to the discovery and characterization of a great deal of disease-causing genes and mutations that result in X-linked, autosomal dominant, or autosomal recessive, in addition to mitochondrially inherited renal conditions. This has undeniably improved substantially the understanding of their clinical and molecular pathology, the diagnosis and/or prognosis, and the genetic counseling accompanying these feats and naturally the clinical intervention for a targeted design of better treatments. Equally important is the fact that more and more human maladies are recognized to have a genetic component. It is worth mentioning that an older survey using the British Columbia Health Surveillance Registry, which included more than one million consecutive births, showed that the frequency of individuals younger than 25 years of age who develop a disease with an important genetic component was 5.3% [7]. Of the 5.3%, single gene disorders represented 0.36% and multifactorial conditions such as cleft lip or diabetes represented 4.6%. Clearly, when considering genetic disorders as a whole, they become relatively common, although each one of them alone can be extremely rare. Rare or orphan diseases (population prevalence of $<5:10\,000$) have attracted special attention in recent years, with both the United States and the European Union announcing dedicated calls for funding research aimed at better diagnosis, treatment, and drug discovery for these conditions. In Europe, a disease is defined as rare if it has a prevalence of fewer than $5:10\,000$, while in the United States a disorder is defined as rare when it affects fewer than 200 000 Americans at any given time (Orphanet: http://www.orpha.net/consor/cgi-bin/index.php). Most inherited monogenic renal disorders satisfy this definition. Perhaps the only two exceptions are the autosomal dominant form of polycystic kidney disease (ADPKD), which is reported to have a prevalence of $1:400$ to $1:1000$ [8], and the thin basement membrane nephropathy (TBMN), which is estimated to have a prevalence as high as 1% in the general population ([9] and references therein).

It is not always obvious or easy to verify that a condition has a familial nature. How does one validate the existence of a heritable disease component in a family or in the index patient proband? I am sure most if not all geneticists agree and know firsthand that specifically as regards the genetic investigation of the index patient and their family, the most effective, cheapest, painless, noninvasive, and most informative genetic investigation one can do is the accurate drawing of a detailed family pedigree. It is not unusual for people to reveal the existence of family members, either close or distantly related, that have not previously been considered to be affected by the condition under investigation. Of course this approach does not always enable one to identify the genetic and heritable nature of the disease, especially in sporadic cases, which in many instances represent autosomal recessive or X-linked diseases with no known positive family history. The fact that modern kindred are usually small with only 1–3 offspring, compared with 5–10 in older generations, adds to the difficulty in identifying other affected first- or second-degree relatives.

3.3 Aims of Present-Day Molecular Genetic Investigations

What are usually the aims of contemporary molecular genetic investigation?

1) Facilitation of narrowing the clinical differential diagnosis and confirmation or exclusion of the clinical diagnosis or suspected disease entity
2) Presymptomatic genetic diagnosis
3) Investigation of the inheritance of a genetic predisposition for a certain condition
4) Pharmacogenetic application (many are in place and more are being developed)
5) Prenatal or preimplantation genetic diagnosis (PGD)
6) Investigation for basic science research purposes within the framework of an approved clinical protocol

For 1–4 above, molecular genetic investigation and provision of genetic results should aim at assisting clinicians in reaching a defined diagnosis during their differential approach and dictate the best possible therapeutic intervention. It is equally important to be able to use the produced data for interrupting an equivocal therapy that had been instituted based on wrong assumptions or for the reduction of the general morbidity and mortality or the risk for the specific patient. Occasionally genetic results can help to reduce the likely adverse reactions and avoid the suffering owing to uncertainty and to avoid drug trials and interventions that are inadequate and destined to fail. The pharmacogenetic application aims at preventing the under- or overdose of drugs, the acting ingredient of which is susceptible to a metabolic pathway by a gene product that is highly polymorphic or of a variable enzyme genetics. Consider, for example, the toxicity ensued when a patient is overtreated with a medication, where its metabolism is unknowingly slowed due to a genetic variant. There are numerous examples of enzymes encoded by the P450 family of genes that play a significant role for different classes of (potentially toxic) medical substances that are administered for treating renal conditions or for immunosuppression aimed at preventing rejection after kidney transplantation. A prime example is the enzyme encoded by gene P450 3A5 (CYP3A5), variants of which affect substantially the kinetics of metabolism of tacrolimus, a frequently used calcineurin inhibitor [10–12].

Certainly, it seems likely that molecular tests will be implemented, which, in conjunction with collecting a doctor's prescription from the pharmacy store, are going to allow personalized dose adjustments [13, 14]. These adjustments are going to be based on genotyping related to adverse drug reactions and side effects and will allow the most effective class or brand of medicine among several similar candidates to be dispensed.

3.4 Material Used for Genetic Testing

Which biological material should be used for molecular genetic diagnostics? For purely clinical routine diagnostic purposes, the overwhelming majority of cases require genomic DNA that is isolated from various sources. Most frequently, DNA is isolated from peripheral blood mononuclear cells (PBMCs), from a whole blood sample (typically 3–5 ml), collected in the presence of EDTA as anticoagulant. Saliva or a mouth wash, or material from a tissue biopsy, or cells from a tissue culture that has been established in the context of the investigation of the patient, can be alternative sources of genetic material. Obtaining genomic DNA could be part of a routine clinical investigation of the patient or under an ethically approved research program, following signed informed consent. Once obtained, anticoagulated whole blood is processed for isolation of genomic DNA using routine protocols after selective lysis of the red blood cells and removal of the proteins. Subsequent lysis of the PBMCs allows extraction of the genomic material by a popular salting-out procedure [15] or by older phenol/chloroform extraction methods [16]. The use of these organic solvents is presently avoided in most applications unless specific subsequent research protocols demand for it. In present-day procedures, commercial kits containing columns and requiring a series of centrifugations allow easy isolation of good quality DNA. Final washes with chilled 70% ethanol ensure removal of excess salts and other low molecular weight inorganic molecules. Standardization of downstream procedures, most probably polymerase chain reaction (PCR) amplification of DNA for diagnostic purposes, requires the examination of the concentration and the quality of the DNA isolated by quick agarose gel electrophoresis and/or spectrophotometric analysis at 260 and 280 nm. The ratio of the 260/280 absorbencies should be 1.8–2.0. DNA isolation procedures that require forcing the biological material through the columns perhaps result in obtaining cleaner but fragmented DNA with relatively smaller molecular size, on average on the order of 30000–50000 Da. This might not be an option if one wanted to proceed with other analytical techniques, like normal Southern blotting or pulsed-field gel electrophoresis, for detecting higher molecular weight fragments, searching for large deletions and insertions or DNA rearrangements. Fortunately, nowadays the PCR technique, which can work and amplify the sequence of interest using very fragmented DNA as template, is used in the overwhelming majority of molecular diagnostics investigations, very frequently followed up by restriction enzyme digests and size fractionation by agarose or polyacrylamide gel electrophoresis. Selected applications require long-range PCR amplifications during which DNA fragments on the

order of a few thousand base pairs (e.g., 10 kb) can be obtained, followed by nested PCR amplifications of smaller fragments. An excellent example of such an application concerns the *PKD1* gene (coding for polycystin 1), which because of existing additional homologous pseudogenes (with up to 90% homology to its nearly 75% 5′ region at the nucleotide level) long-range PCR is a successful approach, with the use of primers carefully designed within regions of least homology [17–20].

In some cases, depending on the examined gene (e.g., very large genes with large numbers of exons) or disease, it is probable that the geneticist might prefer to isolate and analyze mRNA instead of DNA, provided an easy source of expressed mRNA is available. Unfortunately not all genes are expressed in peripheral blood leukocytes. Where mRNA is more desirable, PCR follows after a prior step of converting mRNA into complementary DNA (cDNA) with the use of a viral reverse transcriptase enzyme (RT-PCR). There is one major advantage as well as one major pitfall when analyzing cDNA for mutations. The advantage is that one limits the analysis to the spliced exonic sequence and one is able to cover these large sequence genes in a much smaller number of PCR reactions, crossing many exon–exon junctions simultaneously. This saves time and reagents, especially when dealing with genes of many exons, examples of which are the autosomal dominant *PKD1* (47 exons), the autosomal recessive *PKHD1* (87 exons, coding for fibrocystin/polyductin), and the collagen IV genes of basement membranes (48–52 exons). The disadvantage is that this approach may miss larger genomic aberrations such as insertions or deletions and rearrangements. For example, if one entire allele or a large part of it is heterozygously deleted, the patient will be falsely diagnosed as homozygous normal (in reality being hemizygous), because only the normal allele will be amplified and analyzed. Even with the ability nowadays to use more readily the gold standard method for mutation detection, which is direct DNA re-sequencing, the use of the Sanger sequencing method may miss them. Other techniques such as comparative genomic hybridization (CGH) and multiplex ligation-dependent probe amplification (MLPA) [21] are better suited to detect such genomic aberrations. A recent example of a renal monogenic disorder with a genomic aberration was the identification of an exon 2–3 duplication in the *CFHR5* gene in patients with an autosomal dominant form of C3 glomerulopathy [22, 23].

3.5 Clinical, Genetic, and Allelic Heterogeneity

When the gene or genes at fault are known, the genetic investigation of a patient can be a relatively simple procedure of collecting peripheral blood samples, isolating genomic DNA, and performing simple PCR-based genetic tests. It is even simpler when the individual has a known family history, for example, of ADPKD (an inherited polycystic nephropathy with severe symptoms usually after the fourth or fifth decade) and belongs to a family with a previously identified mutation, for example, in the *PKD1* or the *PKD2* gene. Under these circumstances, by screening for the identical genetic change in the previously identified exon, a molecular genetic diagnosis can be produced within a few hours. A negative result can relieve existing anxiety and spare the patient from unnecessary frequent doctor visits, while a positive result will direct the doctor to the correct decisions for close follow-up and perhaps intervention, either through the administration of proper medication or otherwise. Another especially useful application of molecular diagnostics for late-onset diseases is when testing a living-related potential kidney donor for a mutation that segregates in the family. Sometimes, in the absence of confident pathognomonic features that would enable a clear clinical diagnosis, only genetics can provide the answer and give the green light for the donor to donate his/her kidney.

Unfortunately in many cases things are more complicated, and it is required to have a close collaboration and exchange of information between the nephrologist and the geneticist, who needs to be an expert in his/her field if he/she is to be of help to the clinician. As in every other scientific field, the good geneticist will help evaluate the difficult cases that require multiple approaches.

The performance of a molecular test and the reporting of a test result, even a negative one, is an act that demonstrates to the doctor and the person/family under investigation that certain action has been taken up and a worthwhile result has been produced. The issued report, however, may give an erroneous impression especially if it is not accompanied by adequate interpretation of the result or if the clinician is unable to comprehend the result and he does not seek further detailed explanation. When a mutation is found that was previously verified and reported by others or when a stop codon mutation or a frameshift or splice site defect is detected, it is usually straightforward. However, if a sporadic patient with a suspicion for cystinuria, for example, without a family history, is investigated by DNA sequencing of the *SLC3A1* gene alone and no mutation is detected, the diagnosis for cystinuria is still not excluded. This is because cystinuria, as with many other inherited conditions, is genetically heterogeneous and may be caused by mutations in either of two genes—*SLC3A1* or *SLC7A9*—resulting in similar phenotypes, even though the cystine concentration in the urine of heterozygote individuals can produce suspicion in favor of one or the other gene (http://omim.org/entry/220100). Situations

like this represent phenocopies because of clinical heterogeneity, that is, mutations in more than one gene producing the same heritable monogenic disorder, a phenomenon which is common. Other examples include autosomal dominant polycystic kidney disease (PKD), autosomal dominant medullary cystic kidney disease, nephronophthisis, renal tubular acidosis, Alport syndrome, Bartter's syndrome, and steroid-resistant and congenital nephrotic syndromes. On clinical grounds it is often impossible to dissect out which one of the candidate genes is at fault, even though there may be reasonable suspicions. In ADPKD, for example, it was shown that if there is a family history of all patients reaching end-stage renal disease (ESRD) after the age of 70 years, this is strong indication for involvement of the *PKD2* gene [24]. Torra et al. [25] had earlier shown that there was increased prevalence of PKD type 2 among elderly PKD patients [25]. However, if one is to offer a presymptomatic molecular diagnosis for a potential living-related kidney donor, testing of both genes is absolutely advised for reaching a firm conclusion.

Usually the screening for mutations is focused within the exonic coding regions and 10–20 bp of the flanking sequences that contain the invariably conserved splicing signals. Consequently, a negative molecular test may be attributed to other factors such as the presence of mutations in gene regions not routinely searched for, such as the promoter region, the 3′UTR or deep intronic regions, distant from the exon/intron splice site junctions, perhaps leading to activation of cryptic splice sites. Indicative examples, among many, of mutations that on first sight seem benign but proved to be deleterious, are those found in intronic sequences: (i) the IVS1-110 single nucleotide substitution at position 110 of the first intron of the β-globin gene, resulting in partly aberrant splicing that includes additional sequences and partly normally spliced β-globin mRNA; (ii) the 3849 + 10 kb C > T transition mutation deep in intron 19 of the *CFTR* gene, resulting in inclusion of a cryptic exon of 84 bp, responsible for cystic fibrosis; and (iii) the exceptionally frequent recurrent synonymous mutation in an exon of the LMNA/C gene of lamin A/C, which creates a cryptic splice site and results in severe Hutchinson–Gilford progeria syndrome [26–28].

Similar scenarios apply to numerous other gene systems and heritable conditions, including genes that are implicated in several inherited renal diseases. Examples include mutations in the *PKD1* gene where mutations in noncoding regions cause aberrant splicing of a small 75-bp intron [29], intronic mutations in the *SLC12A1* gene of type 1 Bartter syndrome [30], a missense mutation altering the first nucleotide of *PKD2* exon 6 and resulting in aberrant splicing [31], and many others.

A particularly challenging situation concerns the highly heterogeneous group of childhood and adolescent syndromes of focal segmental glomerulosclerosis (FSGS) that result in steroid-resistant or steroid-sensitive nephrotic syndromes. The number of genes involved is already large and it is more than certain that more will be identified. It is hoped that the specific clinical or histological characteristics, age of presentation or yet other biomarkers of each one, will direct the investigations of the geneticist, a situation nevertheless that emphasizes again the necessity for close collaboration between the genetics laboratory and the clinic. It is fascinating that causative mutations have been identified in genes that encode proteins located in the slit diaphragm, or the podocyte membrane, the podocyte cytoskeleton, or even the nucleus and the mitochondrion, a situation that highlights the complexity and the interdependence of biological processes in the glomerulus [32]. Recent investigations of large cohorts of families with steroid-resistant nephrotic syndrome (SRNS) have been revealing. Panels of 27 genes that are implicated in this condition have been examined by NGS in 1783 unrelated international families, and a single gene cause was determined in 29.5% [33]. Others showed that *NPHS2* mutations account for only 15% of nephrotic syndrome, emphasizing again that more genes are expected to be found in monogenic SRNS patients [34]. Here it should be mentioned that in addition to the more traditional distinct glomerulopathy genes that are being investigated for SRNS and FSGS, mutations have been reported initially by our group in the *COL4A3*/*COL4A4* genes in a large Greek–Cypriot cohort of patients who presented with familial microscopic hematuria (MH) and the dual diagnosis of TBMN and FSGS [35]. This fact allows the hypothesis that the collagen IV mutations link familial hematuria and FSGS while providing insight relating glomerular epithelium destruction via basement membrane thinning [9, 35–37]. Several other groups supported and corroborated these results [38, 39].

The role of putative modifier genes and perhaps of environmental factors certainly cannot be excluded (see following text). As regards the discrimination between steroid-sensitive and steroid-resistant patients, a recent publication showed that neutrophil gelatinase-associated lipocalin (NGAL) levels in urine could differentiate the two forms of nephrotic syndrome with higher levels correlating with disease severity in SRNS [40].

In addition to the multiplicity of genes mutated and causing the same phenotype, allelic heterogeneity is another level of complexity for all monogenic disorders regardless of mode of inheritance, although there are notable exceptions for recessive conditions with a small number of mutations accounting for the great majority of mutant alleles (e.g., the large deletion in the *NPHS1* gene causing the Finnish-type nephrotic syndrome) [41]. Allelic heterogeneity has been invoked to explain the variable expressivity, and many researchers have tried to

reach a genotype/phenotype correlation algorithm for each gene. Despite the initial hopes by the scientific community that every mutation or classes of mutations would enable us easily to predict disease outcome and prognosis and act accordingly (allelic heterogeneity), it turned out that we can only partially trust this line of evaluation. It is well known, for instance, that mutations in the *PKD2* gene cause a milder form of ADPKD compared with mutations in the *PKD1* gene, as evidenced by later age of onset of ESRD and a smaller overall percentage of patients reaching ESRD [42, 43]. This explains, to some extent, the significant interfamilial clinical variability observed among the population of patients with ADPKD but certainly cannot explain the variability among patients with type 1 or type 2 ADPKD and, even worse, among patients within the same family who share the exact same mutation. Several works attempted to address these important issues. Magistroni et al. showed that the site of the *PKD2* mutation does not play a role, while surprisingly patients with splice site mutations appeared to have milder renal symptoms compared with patients with other types of mutations [44]. The group of Peter Harris at Mayo Clinic, MN, United States, showed that mutations in the 5′ half of the *PKD1* gene confer more severe disease compared with mutations in the 3′ half. Importantly, they also showed that mutations located further 5′ of the gene are associated with significant risk for developing intracranial aneurysms [45]. Hateboer et al. had shown that different groups of mutations and the location of the mutation within the *PKD2* gene do influence clinical outcome [46, 47]. Most recent work by Hwang et al. showed in a large cohort of ADPKD patients that *PKD1* in-frame insertion/deletion or non-truncating *PKD1* mutations or mutations in the *PKD2* gene have smaller height-adjusted total kidney volume and reduced risks for ESRD and death [48].

Promising results for genotype/phenotype correlation have been published for other gene systems, including the collagen IV genes, mutations in which cause X-linked (*COL4A5*) or autosomal recessive Alport syndrome (*COL4A3*/*COL4A4*). Large deletions/insertions and gene rearrangements have been associated with earlier age at onset of ESRD and/or with more frequent establishment of hearing loss, while glycine substitutions in the collagenous domain have been associated with variable expressivity depending on exact position along the triple helical domain or relative to the position of natural interruptions of the collagenous domain. However, it is not always possible to predict the disease outcome for the patient before you with absolute certainty. Especially in regard to TBMN, the experience by several researchers has been that even within families with the same mutation, different patients progress to kidney function decline with different rate [49–52].

3.6 Oligogenic Inheritance

To make things more complex, digenic inheritance or even triallelic inheritance, although rare, cannot be excluded. In cases where the clinical presentation is the result of digenic inheritance, there is indeed an interesting phenomenon that attracts added attention. This has been documented to be the case in Bardet–Biedl syndrome (BBS) [53] and more recently in patients with SRNS and Alport Syndrome where heterozygous mutations in different genes account for the phenotype. More specifically, Löwik et al. [54], in a cohort of 19 nonfamilial childhood-onset steroid-resistant FSGS patients, reported that two patients showed mutations in the *CD2AP* gene, one combined with an *NPHS2* (podocin) mutation [54]. Another patient carried three mutations, as the patient was compound heterozygous for *NPHS2* mutations and heterozygous for a *NPHS1* (nephrin) mutation. Yet another patient carried a de novo *WT-1* (Wilms' tumor 1) mutation that was combined with a heterozygous *NPHS1* mutation, while two other patients showed three heterozygous *PLCE1* (phospholipase C epsilon 1) mutations. All aforementioned mutated genes are expressed in the podocytes that are crucial cells for the maintenance of the glomerular filtration barrier and especially the slit diaphragm between interdigitating podocyte processes. These findings, therefore, emphasize that combined gene defects are capable of causing FSGS and consequently complicating things for the genetics laboratory.

Meckel syndrome (MKS) is a genetically heterogeneous ciliopathy where mutations in six genes have been described. It is embryonic lethal and characterized by polycystic kidneys, central nervous system defects, polydactyly, and liver fibrosis. Genetic analysis of additional ciliopathy candidates by exon-enriched NGS revealed a splice donor mutation in one allele of the *B9D1* gene that abolished exon 4 and a large genomic deletion that removed the entire second *B9D1* allele. In the same patient a substitution mutation (p.R2210C) was found in another *MKS* gene, *CEP290*, thereby suggesting oligogenic inheritance [55].

Chen et al. screened 20 genes (15 BBS plus *RPGRIP1L*, *CC2D2A*, *NPHP3*, *TMEM67*, and *INPP5E*) for mutations in BBS patients and found causative and probably causative mutations in ten of the genes in 46/55 families studied (84%) [56]. Importantly, once again the authors reported triallelic inheritance suggesting oligogenic inheritance in Caucasian and Arabian families, thereby complicating approaches for molecular genetic testing for diagnostic purposes. The role of a third allele in modifying the phenotype rests in the borderline between the requirement for oligogenic inheritance for disease expression and genetic modifiers that add to the severity of already clinically expressed diseases (see next section).

As regards ADPKD, normally inherited as an autosomal dominant nephropathy, hypomorphic alleles with incomplete penetrance were previously shown to exert a dose effect where two alleles were needed to cause a phenotype. Also, it has been known for many years that a small subset of about 1% of ADPKD patients present with an early and severe phenotype more reminiscent to the autosomal recessive form of PKD that is normally caused by mutations in the *PKHD1* gene. A few years back it was shown that at the cellular level ADPKD might behave as a recessive disease because several groups documented that the *PKD1* or *PKD2* heterozygous germinal mutations were necessary but not sufficient to cause cystogenesis. Instead, acquired somatic second hits, producing trans-heterozygosity with a mutation on the other allele of the same gene, or even in one allele of the other *PKD* gene, are necessary for cystogenesis. The timing, the nature, and the multiplicity of these second hits might clearly be important factors in determining the age at onset and the overall severity of disease [57–60]. More recently Bergmann et al. showed that at least part of the variable expressivity in patients with the ADPKD can be attributed to the co-inheritance of mutations in multiple *PKD* genes, including *PKD1*, *PKD2*, *PKHD1* (mutated in autosomal recessive PKD), and *HNF-1β* (mutated in the renal cysts and diabetes syndrome (RCAD)) [61]. They showed that co-inheritance of two or even three genetic variants is another mechanism that could explain early onset and severe phenotypes, thereby supporting a pattern of oligogenic inheritance. Patients with *HNF-1β* mutations do develop cysts, among other symptoms. Finally, digenic inheritance was shown for Alport syndrome. Specifically, co-inheritance of two heterozygous mutations in the *COL4A3* and *COL4A4* genes was documented in patients who had an intermediate phenotype with respect to the autosomal dominant form and the autosomal recessive one, thus reflecting the dose of the final wild-type collagen IV triple helix [62].

3.7 ADPKD, Phenotypic Heterogeneity, and Genetic Modifiers

Phenotypic heterogeneity exemplified as disparate spectrum of symptoms is the norm, and it can be based on allelic heterogeneity, nature, and position of a mutation in a given gene, while more recently an additive role is attributed to mostly unknown modifier genes. The implication of modifier genes that somehow affect the function or the outcome of the primary genes through a cross-talk mechanism, which is not always clear or apparent, has become a necessary prerequisite for interpreting the intrafamilial variable expressivity or phenotypic heterogeneity of many diseases. In these cases, it is obvious that the disease inheritance is dependent on a single gene with a defined mode of inheritance, yet one or more genes, perhaps in concert with environmental factors, have an effect on the severity of symptoms or the age at onset of the disease. In a way, even classical monogenic disorders have an element of multifactorial or polygenic inheritance as regards the full spectrum of clinical presentation. Fain et al. demonstrated that up to 18–59% of the phenotypic expression of *PKD1* mutations can be attributed to genetic modifiers [63]. A similar study by Paterson et al. showed that for patients before reaching ESRD the heritability of phenotypic variation was 42%, whereas for patients with ESRD it was estimated to 78% [64]. Also, Persu et al. first and subsequently Lamnisou et al. suggested a modifier role for an *ENOS* (endothelial nitric oxide synthase) polymorphism on the age at onset or rate of progression of ADPKD [65, 66] (Table 3.1). However other publications had mixed results. Finally, a meta-analysis for the role of the ACE I/D (angiotensin-converting enzyme, insertion/deletion) polymorphism failed to confirm any association with ADPKD phenotype [76].

In the largest study thus far, in search for genetic modifiers, Liu et al. studied two separate cohorts of *PKD1* patients, investigating the potential role of 173 biological candidate genes [69]. The first cohort included 794 white patients from 227 families and the second 472 white patients from 229 families. They found statistical significance for three SNPs in the *DKK3* gene that is supposed to play a role in regulating the Wnt/β-catenin signaling, thereby modulating renal cyst growth. The modification was significant for eGFR as a primary outcome but not for ESRD. The SNP with the strongest association was rs3750940 ($p = 4.6 \times 10^{-5}$) and accounts for only 1.4% of the total variance of eGFR. This SNP is in partial linkage disequilibrium (LD) with the other two, all being intronic, thereby raising the possibility that they are in LD with another yet unknown functional variant. These data indicate that more genetic modifiers of eGFR and renal survival should exist with variable contribution, which can only be detected using large enough cohorts (Table 3.1).

The search for identifying true modifier genes has added difficulties, similar to the ones encountered when searching for genes implicated in polygenic conditions. In such studies there is a need for large patient cohorts and accurate detailed diagnosis of patients who, after all, may develop some of the more severe symptoms at older ages, because age-dependent penetrance is the norm in many late-onset heritable conditions.

Table 3.1 Genes with evidence that they modify the disease outcome of monogenic renal conditions.

Gene (protein)	Renal disease	Modifying effect	Significance	Reference
RPGRIP1L (retinitis pigmentosa GTPase regulator interacting protein-1 like)	Ciliopathies: Meckel–Gruber syndrome, Joubert syndrome, Bardet–Biedl syndrome, Senior–Løken syndrome, nephronophthisis	Retinal degeneration	$p = 7.35E{-}05$	[67]
AHI1	Nephronophthisis	Retinal degeneration	$p = 5.36E{-}06$, RR = 7.5	[68]
DKK3 (Dickkopf 3)	Autosomal dominant polycystic kidney disease	Lowers age of ESRD onset	$p = 4.6 \times 10^{-5}$	[69]
ENOS (endothelial nitric oxide synthase)	Autosomal dominant polycystic kidney disease	Lowers age of ESRD onset	$p = 0.006$ $p = 0.048$ $p = 0.018$	[65, 66, 70]
ACE (angiotensin-converting enzyme)	IgA nephropathy	Progression to CKD	$p < 0.001$	[71] (*Confirmed by at least 7 more publications*)
NPHS2 (podocin)	TBMN, familial hematuria	Proteinuria, CKD	$p < 0.05$	[72–74]
HBEGF 3′ UTR target for miRNA **miR-1207-5p**	CFHR5 nephropathy	Proteinuria, CKD	$p = 0.038$	[75]

IgA nephropathy is not a monogenic condition, but it is included based on the strong association found by several studies.

3.8 Collagen IV Nephropathies, Genetic and Phenotypic Heterogeneity, and Genetic Modifiers

Mutations in *COL4A5* are known to cause the classical X-linked form of Alport Syndrome, a hereditary progressive glomerulopathy, usually with onset in childhood and adolescence. In addition to characteristic histological findings, 82.5% of male patients develop sensorineural hearing loss, and a minority of 44% of patients develop ocular abnormalities, in the form of dot and fleck retinopathy and anterior lenticonus. Interestingly though, a subset of patients follow a more benign course, more reminiscent to TBMN rather than Alport. These patients develop ESRD at ages after 40 or even 50 years and may or may not develop hearing loss or ocular defects. Several studies attempted to explain this heterogeneity by attributing the variable phenotypes to the nature and position of the *COL4A5* mutation [51, 77–81].

Gross et al. [50] proposed the following classification of phenotypes of X-linked Alport patients:

a) Type S (severe), characterized by juvenile-onset ESRD (~20 years of age), 80% incidence of hearing loss and 40% incidence of ocular lesions. This classical picture is caused by large rearrangements, premature stop, frameshift, donor splice site, and mutations in the carboxy-terminal non-collagenous globular NC1 domain.
b) Type MS (moderate–severe), including patients that progress to ESRD at age ~26 and present lower frequencies of the extrarenal manifestations, implicating non-glycine missense mutations, glycine substitutions in exons 21–47, and in-frame and acceptor splice site mutations.
c) Type M (moderate), associated with glycine substitutions in exons 1–20 and characterized by late-onset ESRD (after the age of 30), 70% hearing loss, and less than 30% ocular lesions.

Bekheirnia et al. [49], in a large cohort of US patients with X-linked Alport, support the proposition of Gross et al. [50] that the most aggressive phenotypes are caused by truncating mutations, large and small deletions, and splice mutations; however they make the point that the position of Gly-X-Y mutations may not always predict the age of onset of ESRD [49]. Additionally, in a work where we evaluated male patients with *COL4A5* glycine substitution mutations in the collagenous domain, we showed that the age at onset of ESRD decreased with increasing number of side-chain carbon atoms in the substituting residue (r^2: 0.1362; p: 0.0017) [52].

Notwithstanding the aforementioned findings, there are reports, including recent work from our laboratory, describing mutations in *COL4A5* that are associated with milder phenotypes that are not easily recognized as X-linked Alport syndrome, even when the inheritance is suggestive of this diagnosis [82–87]. Prime examples among several are the mutations p.G624D and p.P628L

in *COL4A5* that were described in Caucasian patients of several centers. In our setting, onset of ESRD varied from 30 to 57 years in seven males. In a Hellenic family, GR4209, a male patient reached ESRD at the age of 39 with sensorineural hearing loss. Another affected male showed hematuria and proteinuria accompanied by sensorineural hearing loss, but his serum creatinine was normal at 1.03 mg/dl, at the age of 30 years. In another family GR4211, two brothers carry the mutation, one of which reached ESRD at 50 years without hearing loss, with FSGS on biopsy and uniform thinning of the glomerular basement membrane (GBM) that included focal splitting. His brother is proteinuric with mildly reduced GFR and no hearing loss at 55 years. No EM results are available. All findings in these families differed from classic, adolescence onset, and adult type X-linked Alport. Both mutations are near the 12th natural interruption of the collagenous domain of *COL4A5*, which is of G1G type and is converted to G4G type, a fact we hypothesized may explain the milder course of disease, as it may affect less drastically the triple helix formation and the structural function of the collagen IV mature trimer [83, 84]. It is interesting to note that mutation p.G624D may represent an old founder effect because it has been reported in several populations of Caucasian origin, including seven Greek families ([84, 88] and unpublished results).

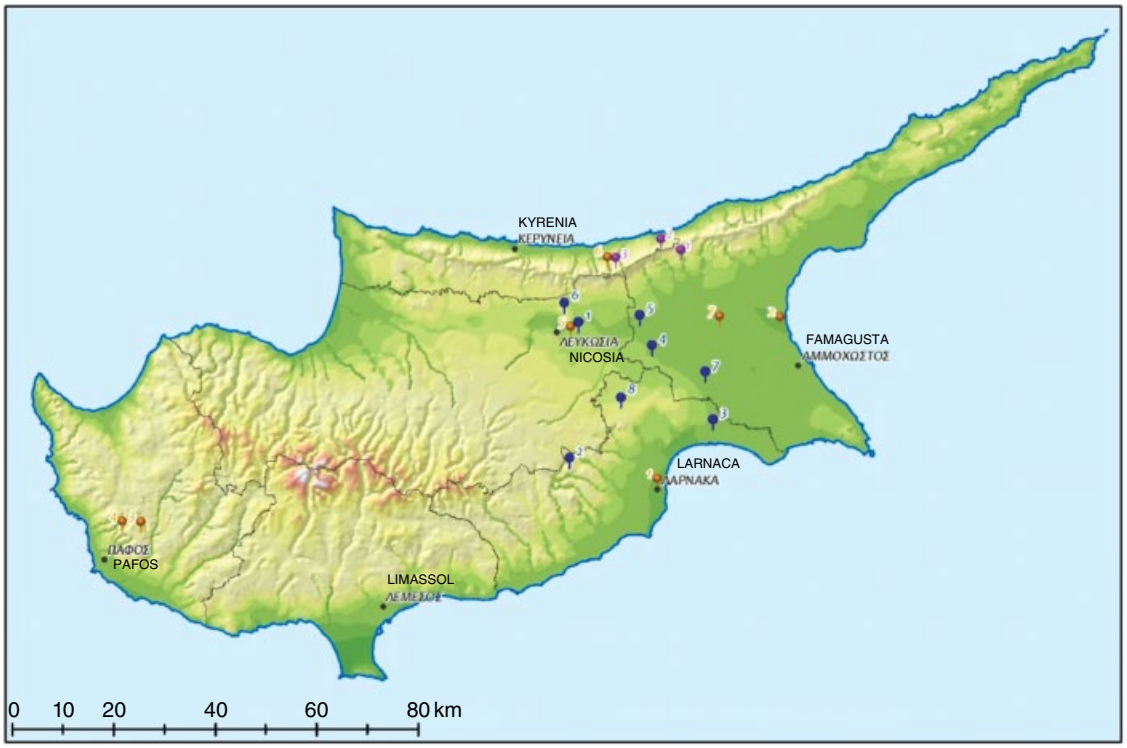
Another striking example of the wide spectrum of symptoms caused by mutations in the same gene is the one involving mutations in genes *COL4A3*/*COL4A4* on chromosome 2q36-37 that encode the alpha 3 and alpha 4 chains of GBM collagen IV molecules. Homozygous or compound heterozygous mutations in either of these genes cause classical full-blown autosomal recessive Alport syndrome whereas heterozygous mutations were shown and considered to be responsible for benign familial MH, most of the times an isolated symptom, as a consequence of TBMN [89–92]. Apart from occasional episodes of macroscopic hematuria and low-grade proteinuria, heterozygous patients have been considered to follow a benign course with excellent prognosis. Earlier reports had alluded to patients who had developed more severe glomerulopathy in the presence of TBMN, but no molecular testing had been performed to document the cause of TBMN [93]. More recent work in our laboratory described a large cohort of Greek–Cypriot families with TBMN that segregated heterozygous mutations in the *COL4A3*/*COL4A4* genes. One mutation, *COL4A3*-p.G1334E, is a founder with more than 150 carriers. Among 228 Greek–Cypriot patients, nearly half of patients will experience a reduction in their glomerular filtration rate of variable degree by the age of 70 years, while 30% of them will progress to ESRD, a fact that clearly challenges the formerly thought benign nature of the disease, at least in this cohort [35, 37, 94, 95].

Based on selected renal biopsies from patients who had developed proteinuria and kidney function decline in the studied families, the cause of the severe kidney function impairment was focal and segmental glomerulosclerosis, thereby satisfying a dual diagnosis [35]. This obscure finding cannot be explained by the kind of the mutation as 134 of the patients share the same exact heterozygous mutation in the *COL4A3* gene, p.Gly1334Glu (substitution of glycine by glutamate), as a result of a strong founder effect (Figure 3.1a). Consequently, it was reasonable to hypothesize the putative effect of a modifier gene (or genes) that when co-inherited with the aforementioned mutation predisposes a subset of these patients to a more severe clinical outcome and chronic kidney function decline. It was also reasonable to assume that the putative modifier might be a gene playing a role in the glomerular filtration barrier environment, affecting the podocyte supporting function. Without knowledge about the identity and the exact role of putative genetic modifiers that somehow predispose to more severe phenotype, it is impossible to make predictions of the phenotype of a monogenic disorder, which is primarily caused by a causative mutation in a responsible gene.

To this end we set out to investigate the putative role of a previous suspect, the p.Arg229Gln variant in the podocin (*NPHS2* gene) that was shown to predispose to proteinuria on the background of TBMN and to microalbuminuria in the general population [72, 96]. We should also mention that Kottgen et al. [97] did not corroborate any significant association between p.Arg229Gln and eGFR in either white or black individuals, while Franceschini et al. [98] had concluded that the same variant confers a nonsignificant increased risk for FSGS by 20–70%, in European descent populations, based on a very detailed review and meta-analysis.

We used a cohort of 147 patients with a familial form of MH, of whom 102 had TBMN and 45 had C3 glomerulopathy as a result of CFHR5 nephropathy (see next section), all with known mutations. The patients were categorized as "mild" or "severe," based on the presence of microhematuria only, or proteinuria and renal impairment. Nine p.Arg229Gln carriers were found in the "severe" category and none in the "mild" ($p = 0.010$ for genotypic association; $p = 0.043$ for allelic association, adjusted for patients' relatedness), thus supporting the possible contribution of 229Gln allele in disease progress. These results offer more evidence that in patients with familial hematuria, *NPHS2*- p.Arg229Gln predisposes to proteinuria and ESRD. Subsequent work in our lab corroborated this finding and implicated the role of another *NPHS2* variant, p.Glu237Gln, as predisposing to more severe disease when co-inherited on the background of TBMN in our cohort. This series of experiments was supported by cell culture functional studies

(a)



(b)

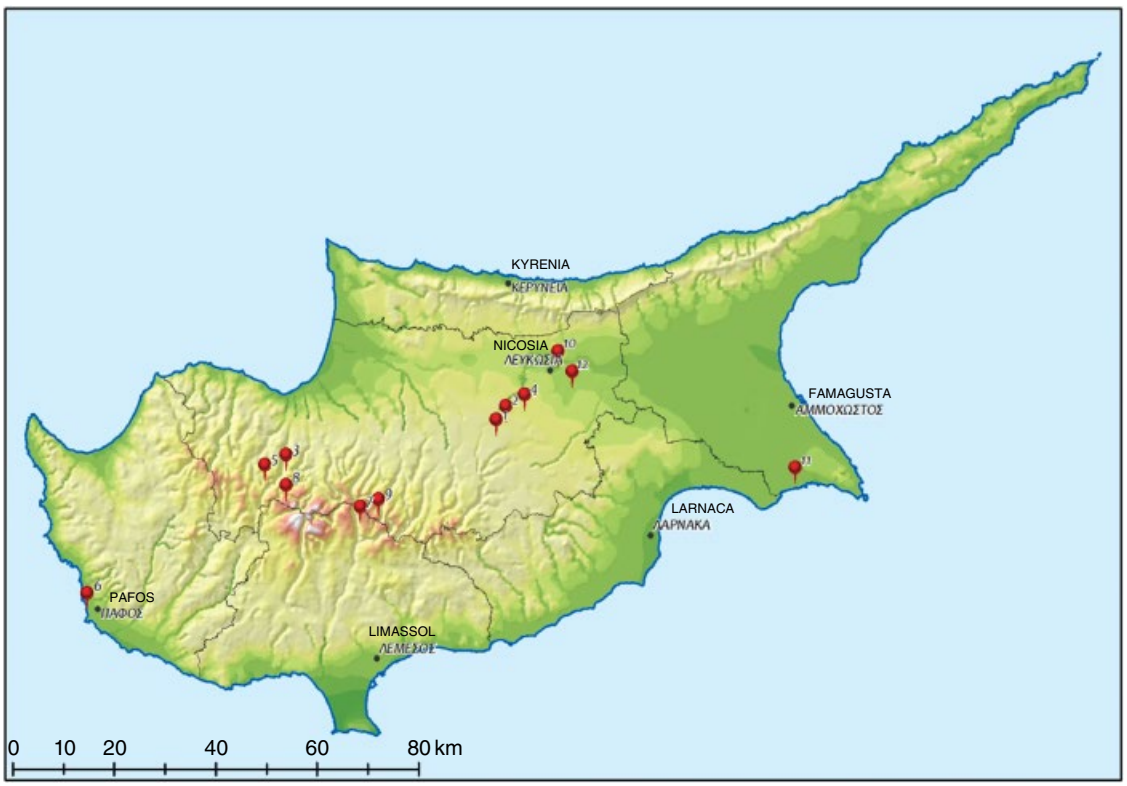


Figure 3.1 (a) The thin basement membrane nephropathy (TBMN) genetic map of Cyprus. Shown are villages and cities where 26 families of TBMN have been detected so far. All patients carry founder mutations in the *COL4A3*/*COL4A4* genes. Families in villages shown with blue dots carry the founder mutation in the *COL4A3* gene, p.G1334E. (b) The CFHR5 genetic map of Cyprus. Shown are 12 villages where 23 families of CFHR5 nephropathy have been detected. All patients carry the same *CFHR5* exon 2–3 duplication, as a result of a founder effect.

that showed that these two *NPHS2* variants interfered with normal trafficking of podocin and nephrin, demonstrating perinuclear staining. Immunoprecipitation experiments showed stronger binding of mutant podocin to WT podocin or nephrin, thus adding support for their negative effect when co-inherited with *COL4A3*/*A4* mutations [99].

In conclusion, these two variants, p.Arg229Gln and p.Glu237Gln, may be good prognostic markers for young hematuric patients, predicting future progressive kidney function decline on long follow-up, through the development of FSGS [9]. These findings, however, make only one piece of a larger puzzle, because we found these variants in only 10.6 or 8.5% of the severely affected patients [73, 74]. Clearly, therefore, there are more gene modifiers expected to be found, likely genes encoding for components or regulators of the glomerular filtration barrier, the damage of which creates the preconditions for loss of protein, admittedly a grave development.

3.9 CFHR5 Nephropathy, Phenotypic Heterogeneity, and Genetic Modifiers

As a first line of defense against pathogens, the complement system constitutes a significant part of innate immunity in humans. While the classical complement system requires immune complexes that act as the triggers for its activation, the alternative complement pathway requires no immune complexes, is independent of the presence of antibodies, and may be active on a permanent basis. Recognition that in CFHR5 nephropathy a mutation in the *CFHR5* gene is responsible for familial C3 glomerulonephritis has suggested that this gene plays an important role in regulation of the alternative complement pathway in the kidney [23, 100].

This new hereditary form of MH had been described before, but its hereditary nature and pathophysiological connection to the complement alternate pathway system had not been recognized [101–103]. Similarly, in regard to its histology, mesangial C3 glomerulopathy has been known at least since 1980 from published reports in Europe and Japan, while more recently, loss-of-function mutations in important regulatory proteins such as CFH, complement factor I, and membrane cofactor protein have been detected in patients with inherited nephropathies characterized by isolated C3 mesangial and subendothelial deposits [104–108]. There may be also mild mesangial proliferation, and it may or may not be accompanied by mild membranoproliferative glomerulonephritis (MPGN). In fact, Abrera-Abeleda et al. first tried to associate DNA variations in the *CFH* and the *CFHR5* genes with MPGN type II (also known as dense deposit disease (DDD)) [104]. No clearly pathogenic mutations were reported.

The association of *CFHR5* mutations with this familial form of MH was the occurrence of two parallel events. The first was that D. Gale, then a research fellow taking off from his residency in Nephrology at UCL, embarked on the study of a few patients of Cypriot origin living in London whom he encountered in clinical practice and who were observed by the histopathologist T. Cook to manifest a highly unusual form of familial glomerulonephritis. The second event has been the ongoing collection of samples and preparation of a DNA biobank from Cypriot families with inherited kidney disorders, at the Laboratory of Molecular and Medical Genetics and more recently at the Molecular Medicine Research Center of the University of Cyprus, by our group [22, 23].

The index patient and affected relatives with the disease presented with MH as well as episodes of macroscopic hematuria following upper respiratory tract infections (a pattern termed "synpharyngitic macroscopic hematuria"). Based on this presentation it was reasonable to suspect IgA nephropathy that however was rejected based on renal biopsy results, which did not show glomerular deposition of IgA. Additionally, IgA nephropathy is sporadic in the overwhelming majority of cases. Some familial cases have been reported and a locus has been mapped, although no gene has been cloned as yet [109].

Histology showed that the biopsies were highly unusual: there was mild MPGN, also referred to as mesangiocapillary glomerulonephritis, with slight increase in mesangial cells and matrix. Some cells had slight capillary wall thickening. The EM showed subendothelial GBM electron-dense C3 deposition. Importantly, there was no deposition of immune complexes. It is worth mentioning that not all the biopsies in CFHR5 nephropathy actually show MPGN (Figure 3.2) (see also [110]).

These appearances, which are now termed C3 glomerulonephritis, implicated dysregulation of the complement alternative pathway [23]. Molecular investigation of this potentially monogenic disorder was commenced in the initial two families in whom there was autosomal dominant inheritance of the disease (Figure 3.3). Initial molecular studies resulted in identifying a copy number variation in the *CFHR5* gene, where exons 2–3 were duplicated in all patients and in none of healthy family members or a number of other healthy subjects. Further extensive investigation in London and Cyprus identified 23 families in total, thus far, with >150 affected subjects and clear autosomal dominant segregation of the condition. This specific mutation appears to be endemic to Cyprus [22]. For the sake of completeness, only very few other hereditary C3 glomerulopathies have been described in other populations, in small nuclear families, some with a mutation involving the *CFHR5* [111].

The largest family is CY5308 with 37 mutation carriers, and this mutation represents yet another strong founder

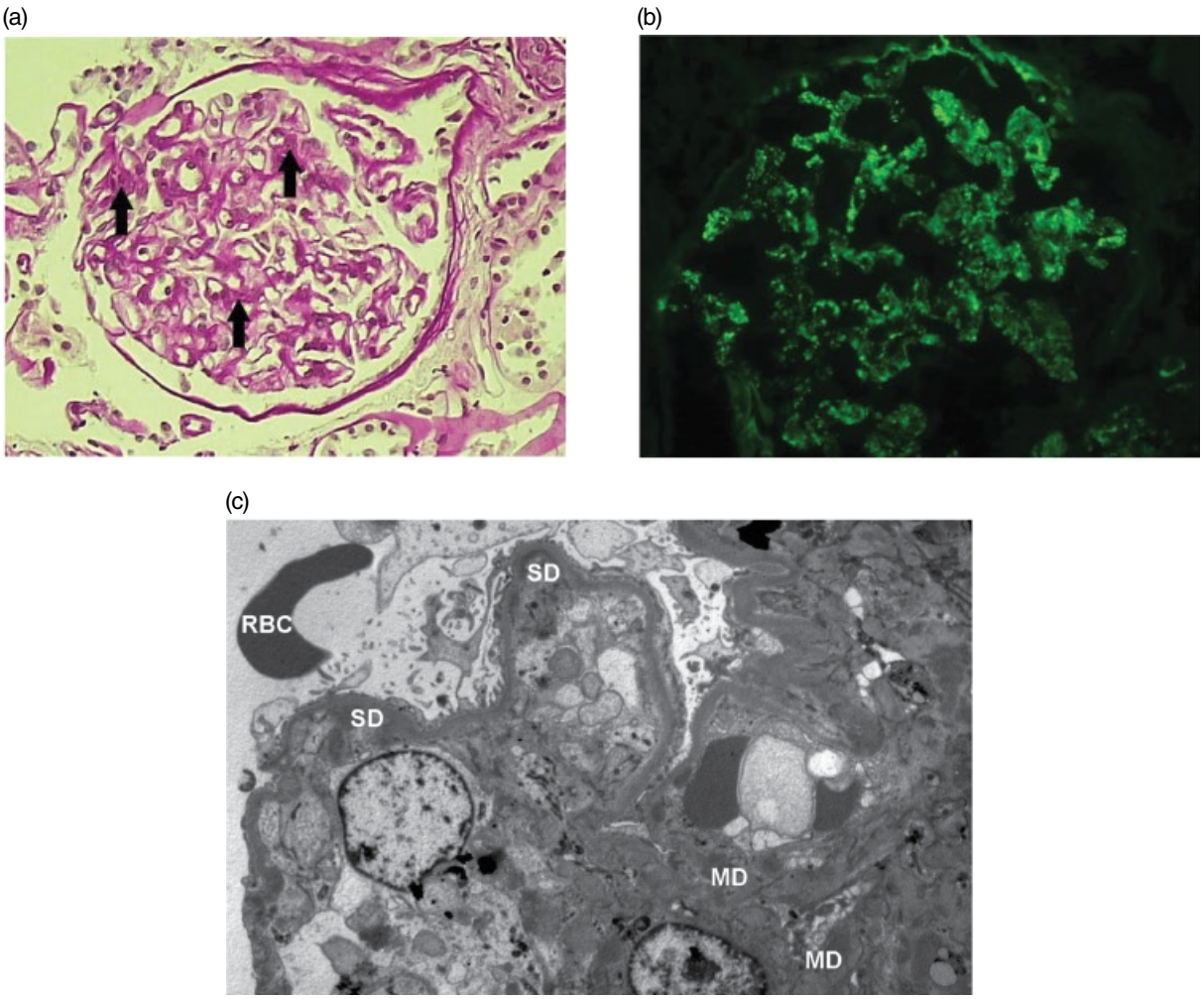


Figure 3.2 (**a**) Patient with CFHR5 nephropathy. Light micrograph from a representative kidney biopsy in family CY5308 stained with PAS showing mesangial hypercellularity (arrows) (×400). (**b**) Immunofluorescence from a representative kidney biopsy in family CY5308 showing diffuse granular C3 staining on the mesangium and capillary wall. (**c**) Electron micrograph from a representative kidney biopsy in family CY5308 showing the presence of subendothelial deposits (SD) as well as mesangial deposits (MD). A red blood cell (RBC) is seen in Bowman's capsule (×8000). *Source*: Athanasiou et al. [22]. Reproduced with permission of American Society of Nephrology.

phenomenon in Cyprus (Figure 3.1b). This inherited glomerulopathy is characterized by impressive variable expressivity, exemplified as a broad spectrum of symptoms in the cohort of the patients. There is clear evidence for reduced penetrance, 90%, as 14 of 136 patients tested are negative on urine findings. 59% of the patients have hematuria only, 4% hematuria and proteinuria but no kidney function decline, and 50% (30/61, 24 M) of patients over 50 years have progressed to renal impairment. In total, 19 patients reached ESRD (14%), 16 of them, all males, after the age of 50 years. It is of great interest that of all 19 patients who progressed to ESRD, only 3 (16%) are females. It is also interesting that all males who progressed to ESRD had demonstrated episodes of macroscopic hematuria during childhood and adolescence after upper respiratory tract infections. This was indeed a finding that complicated the clinical diagnosis as it was occasionally confused with IgA nephropathy that also presents with episodes of synpharyngitic macroscopic hematuria, associated with infection and pyrexia [22].

Similarly to TBMN, CFHR5 nephropathy is a progressive later-onset disease, as patients who carry this mutation in *CFHR5* are nearly asymptomatic until the age of about 30 years, with only isolated MH or negative urine findings. After the age of 30, proteinuria develops in most patients, and once this occurs, patients may progress to kidney function decline [22, 23]. It is presently unknown what protects women from reaching a more severe phenotype. This great phenotypic heterogeneity observed among patients within same families is hypothesized to be attributed to unknown modifier genes and perhaps environmental factors. A couple of candidate

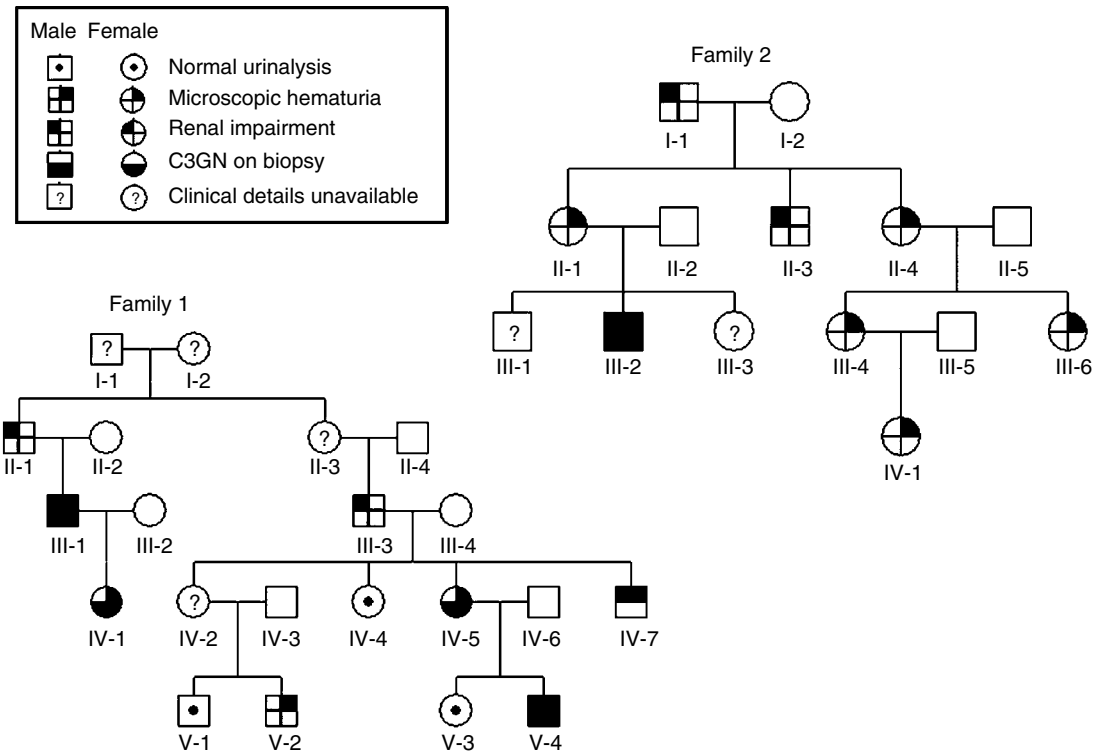


Figure 3.3 Pedigrees of the two original Greek–Cypriot families segregating CFHR5 nephropathy that were described in London. *Source:* Gale et al. [23]. Reproduced with permission of Elsevier.

modifier genes are under investigation in our laboratory in Cyprus at the Molecular Medicine Research Center. We recently reported our evidence that the glutamine variant of the p.Arg229Gln podocin (*NPHS2*) mutation may act as a high-risk genetic factor predisposing patients with familial hematuria (TBMN or CFHR5 nephropathy) to a more severe disease [74]. We had similar findings for another genetic modifier, this time a polymorphism in the target sequence of miRNA gene hsa-miR-1207-5p, in the 3′ UTR of *HBEGF* (miRSNP C1936T, rs13385) [75]. Importantly, its significance was demonstrated by functional studies in undifferentiated cultured podocytes and by association studies in two cohorts of patients. Specifically, in the presence of a mimic for miRNA hsa-miR-1207-5p, there was down-regulation of the *HBEGF* expression, judged by Western blot analysis. This was corroborated by the use of luciferase sensor constructs of both alleles, where the 1936T allele demonstrated abrogation of miRNA binding. Most interestingly, the 1936T allele was shown to act as a genetic modifier, as it was genetically associated with a higher risk for progression to severe renal disease in the presence of a primary glomerulopathy, C3 glomerulonephritis. The exact mechanism by which *HBEGF* can alter disease phenotype is currently unknown. However, we hypothesize that the role of *HBEGF* in proliferation and fibrosis of mesangial cells is very critical to this end [75].

3.10 Unilocus Mutational and Phenotypic Diversity (UMPD)

We first referred to this phenomenon in 1993, in an attempt to describe the situation where allelic mutations in the same gene resulted in so diverse and distinct phenotypes that they were described with a different clinical diagnosis, if as though they represented diverse disease entities [112, 113]. An overlap in symptomatology is not entirely excluded.

Based on the previous sections, it should be no surprise to have patients with mutations in the same gene presenting with same diagnosis and symptoms fitting on a broad spectrum of diverse phenotypes. However, it is intriguing to have patients with mutations in the same gene presenting with overlapping phenotypes and disparate clinical diagnoses. Put another way, the phenomenon of unilocus mutational and phenotypic diversity (UMPD) is the extreme exhibition of variable expressivity.

A good example of this phenomenon in nephrogenetics is the *MCKD2* gene (coding for uromodulin), where patients carrying mutations can present with one of three rare renal autosomal dominant conditions, those being medullary cystic kidney disease type II [114], familial juvenile hyperuricemic nephropathy (HNFJ1) [115], or glomerulocystic kidney disease [116]. In HNFJ1,

Bowman's space is dilated forming cysts and glomerular tuft collapses. Predominant features of HNFJ1 resemble MCKD2, with hyperuricemia-associated gouty arthritis and early progression to renal impairment [117]. The major clinical manifestations of glomerulocystic kidney disease are reduced fractional excretion of uric acid, while there is impaired urine concentration ability of the kidneys, which either appear hypoplastic or have a normal size [118]. Based on overlapping features though, Scolari et al. suggested referring to these conditions as "uromodulin storage diseases," as uromodulin is shown to be retained in the endoplasmic reticulum and perhaps demonstrating a major role in determining tubulointerstitial fibrosis and renal impairment [119]. A similar suggestion had been done earlier by Hart et al., who based on the fact that MCKD2 and HNFJ1 are allelic disorders they designated them as "uromodulin-associated kidney diseases" [115]. More recently, a Kidney Disease Improving Global Outcomes (KDIGO) consensus report grouped these conditions under a common heading that includes additional diseases, as "Autosomal Dominant Tubulointerstitial Kidney Disease (ADTKD)" [120].

Another, even more radical example is the complement factor H (*CFH*) gene, a major regulator of the alternative pathway of complement [121]. Depending on the location and nature of the mutation, the phenotype can range from DDD (formerly considered as a form of MPGN II) to atypical hemolytic uremic syndrome (aHUS) or basal laminar drusen. Also, mutations in the same gene have been documented to confer a significantly increased predisposition and high risk for an ocular condition of late onset, the age-related macular degeneration (AMD). High risk for AMD is specifically conferred by a substitution at amino acid residue 402 from histidine to tyrosine (His402Tyr) [122, 123]. DDD and aHUS certainly have overlapping features; however the clinical differences are striking, the cardinal symptom of DDD being the deposition of complement C3 in the GBM and the cardinal symptoms of aHUS being thrombotic microangiopathy and thrombocytopenia. Most mutations resulting in aHUS are crowded in the short consensus repeats toward the carboxy-terminal end of the protein [124].

Collagen IV nephropathies are caused by mutations in the collagen IV genes (see previous section). Specific mutations in *COL4A3*/*COL4A4* result in TBMN and MH since childhood, with a varying likelihood for progressing to kidney function decline and even ESRD. In fact some authors refer to these severe heterozygous cases as autosomal dominant Alport syndrome, mostly without extrarenal manifestations. Also, a large series of patients from Cyprus was reported where there was a dual diagnosis of TBMN and FSGS, as a cardinal histological feature. Equally important is the finding of male patients with hypomorphic *COL4A5* mutations who develop a much milder form of Alport syndrome and very late age at onset of ESRD, mostly presenting as a phenocopy of TBMN.

On the opposite end, Becknell et al. [125] reported on a large American family with mutation p.Phe222Cys in the *COL4A5* gene, where male patients presented with a novel severe glomerulopathy that rapidly progressed to ESRD at 10–22 years old. The mutation was within a G4G interruption of the collagenous domain, substituting a conserved phenylalanine residue, thereby attributing an unknown functional role. Interestingly, the authors emphasize the absence of typical Alport syndrome clinical and biopsy findings. The symptoms included proteinuria and variable hematuria, while the biopsy showed global and segmental glomerulosclerosis, mesangial hypercellularity, and GBM immune complex deposition, quite unusual for Alport. There were no typical Alport biopsy findings such as GBM alternate thinning and thickening nor any GBM splitting and lamellation.

Possible explanations for these extreme phenotypes that support the UMPD as a phenomenon arising from allelic mutations could be that responsible mutations occur in distinct functional domains with disparate interacting partners, perhaps relating to posttranslational modifications of the domains harboring the mutation, or contribution of confounding defects and genetic modifiers in different patients. Take into consideration, for example, a recent review publication by Parkin et al., who presented the interactome of collagen IV, where they described the various domains and the multiple interacting partners of the protein moieties [126].

Finally, an exemplified system of UMPD is the one related to mutations in the lamin A/C gene. Here, different mutations have been found in patients with a clinical diagnosis as disparate as Hutchinson–Gilford progeria syndrome and several neuropathies or lipodystrophies or even cardiac defects, collectively referred to as laminopathies, not to be discussed in this chapter [127].

3.11 Next-Generation Sequencing (NGS)

During the past few years, there has been a great competition and a race for achieving what until recently was totally unimaginable, that is, the $1000-genome sequencing. The objective has been to enable the sequencing of one's entire human genome with a cost under $1000 within a few days, with trusty worthwhile results [3]. In addition to its usefulness in numerous research settings, this technology has already revolutionized clinical diagnostics regimens. Useful applications are going to be to determine one's overall morbid genetic load at birth or search for new mutations in unknown genes or search for mutations in one of several

candidate genes owing to extreme genetic heterogeneity, where tens or hundreds of coding exons will need to be analyzed. Consider, for example, a case of nephronophthisis or BBS or inherited nephrotic syndromes, with more than 10–30 genes implicated in each. This massive and robust approach will clearly be beneficial although in cases of WGS or even WES the complexity of the derived data requires time-consuming and expensive software for analysis. In most diagnostic scenarios, the genes at fault are known beforehand, and in many frequent autosomal recessive conditions, the repertoire of existing causative mutations is also known in the populations under study. In fact, in the overwhelming majority of such cases, there are well-established protocols for mutation detection. Practically speaking therefore, knowing one's population under study and the previous determination of the mutation repertoire is of clinical utility in the clinical routine environment, making NGS more of a research, at least for the present, that the logistics are still complicated [3].

There are, however, numerous occasions where specific panels of genes, implicated, for example, in inherited kidney disorders, have been verified and are in use for diagnostic purposes. Such panels concern Alport syndrome and collagen IV nephropathies [5, 6, 128], nephrotic syndrome [33, 129], congenital anomalies of the kidney and urinary tract (CAKUT) [130], and others. The same reservations are applied here in regard to identifying and verifying the disease-causing variants, distinguishing them from neutral variants with the help of comprehensive DNA variant databases, robust bioinformatics approaches, and deep knowledge of human genetics.

In this respect, WGS will certainly reveal thousands of DNA variants of unknown functional significance. In the effort to document the pathogenic nature of a DNA variant in the absence of trusty robust functional tests, certain axioms need to be satisfied: the presence of the variant only in affected members in a family and in no healthy subjects (occasions of incomplete penetrance are exceptions) and the absence of the variant in 50–100, if not more, subjects of the general population. It is well known though that not all patients belong to large families with additional patients. Especially for rare monogenic recessive conditions, most patients represent sporadic cases, making things even more intricate. In the case of pathogenic variants linked to recessive conditions, things are not clear because they may be of high population frequency in healthy heterozygous carriers. Several software are useful and supportive in assessing the functional significance when considering the nature of a single nucleotide substitution, but one can never place absolute confidence in them because there are many exceptions. Such software in use are the following:

SNPs3D: http://www.ncbi.nlm.nih.gov/pubmed/16551372
CLC Bioinformatics Database: (http://www.clcbio.com/index.php?id=1243)
SIFT: www.ncbi.nlm.nih.gov/pmc/articles/PMC168916/
MutationTaster: http://www.mutationtaster.org/

A word of caution deserves to be mentioned on the fact that even though the actual determination of the DNA sequence has become or is becoming available with acceptable cost, the analysis of the results is still a complex, expensive, and highly demanding task by a number of bioinformatics experts. Hopefully, new electronic and software tools will make it easier and allow it to enter clinical practice sooner than anticipated. In regard to research, the impact upon it is enormous, and recent publications described elegant work that included a combination of technologies for studying clear cell renal cell carcinoma, including single-cell WES [131, 132]. The complexity of tumor DNA variation at single cell level could only be effectively investigated with technologies such as NGS, thereby deriving useful and decisive new information pertaining to tumor initiation and progression. The scientific society should not disregard that NGS results and interpretation will be accompanied by important social, ethical, and legal dilemmas, and therefore appropriate measures and studies will need to be undertaken when implementing this revolutionary technology as a routine diagnostic or prognostic tool. Associated with this development is the issue of incidental findings whereby laboratory people who are searching for mutations in specific genes find important actionable mutations in other genes that are responsible for severe phenotypes. Several papers have been published that offer solutions to these dilemmas, but apparently one safe approach is informing all concerned beforehand and signing an informed consent, stating whether the interested party wish to know of such findings (see, e.g., the relevant position of the American College of Medical Genetics and Genomics) [133].

3.12 Conclusions

The molecular genetic approach is a very strong tool in the armamentarium of modern science; in many cases it can work on its own and provide confirmation of the clinical diagnosis or dispute and even reject the clinical diagnosis. However, in most cases the combination of the clinical data and the guidance the geneticist can have from the clinician will save money, time, and effort. An early biopsy, when indicated, could be of paramount importance in designating candidate genes to analyze or preclude others. At the same time, either in the presence or the absence of a biopsy, from the

practical point of view, once a causative mutation is found, other members of the family will be readily tested in a targeted manner, thereby avoiding additional invasive procedures for diagnostic purposes. And let us not forget that a negative molecular outcome upon testing a subject at risk of developing a heritable disorder later in life is of equal significance to identifying a positive result, both from the medical and the ethical perspective.

The previous work of others as well as our work proves once again the benefit a whole society can derive from genetics studies that identify likely common mutations in the respective populations we serve. Knowing the repertoire of mutations in the gene pool under study facilitates tremendously the genetic diagnosis, while the identification of a causative DNA defect may obviate the need for another biopsy.

The last word of this chapter is dedicated to the genetic modifiers that are arising as new promising factors to explain part of the variability observed in nearly all genetic disorders, so much for the monogenic as well as the multifactorial. Despite all the problems and the complexity of biological systems, one wishes that at least in some cases common variants will be identified that will prove to confer a strong modifying effect and hopefully better personalized treatments will be designed or discovered in the near future. The identification of such likely modifiers is now becoming easier based on newer more robust technologies that make research discovery faster, cheaper, and more effective. We anticipate that in the near future all this information on potential modifiers as well as the knowledge on pharmacogenetics applications will find their way into the general medical practice for the benefit of patients.

It is unfortunate that in most cases of genetic modifiers, no gene is found to exert a major modifying effect and also being of high frequency, thereby satisfying the desired original common variant-strong effect hypothesis. On the contrary, we keep finding rare variants with small effects, with exemplary example being variants found to predispose to chronic kidney disease and the ones we described here for the familial hematurias, among others. Occasionally, rare variants are found that exert strong effects. One cannot exclude, however, that more targeted approaches in the future may identify such more frequent modifiers of disease outcome.

Acknowledgments

We wish to thank all patients and their families who participated in the research work performed in our laboratory and we describe here. We warmly thank all clinicians of the public and private sector in Cyprus and elsewhere who collaborated with our group and contributed biological material for this research and especially Dr. A. Pierides, who coordinated most of the clinical work. Most of our work described here was funded by the following research grants, from the Cyprus Research Promotion Foundation: ΠΕΝΕΚ ΕΝΙΣΧ/0505/02, ΠΕΝΕΚ/ΕΝΙΣΧ/0308/08, New Infrastructure/ΣΤΡΑΤΗ/0308/24 (a grant co-funded by the European Regional Development Fund and the Republic of Cyprus through the Research Promotion Foundation). Partial funding was also provided by the Cyprus Kidney Association, the George & Maria Tyrimos endowment through a grant by the Pancyprian Gymnasium, Nicosia, as a scholarship to support a PhD student in my laboratory and by the University of Cyprus.

References

1 Harrison, R. J. 2012. Understanding genetic variation and function—the applications of next generation sequencing. *Semin Cell Dev Biol*, 23, 230–236.

2 Mardis, E. R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet*, 24, 133–141.

3 Mardis, E. R. 2010. The $1,000 genome, the $100,000 analysis? *Genome Med*, 2, 84.

4 Bullich, G., Trujillano, D., Santin, S., Ossowski, S., Mendizabal, S., Fraga, G., Madrid, A., Ariceta, G., Ballarin, J., Torra, R., Estivill, X. & Ars, E. 2015. Targeted next-generation sequencing in steroid-resistant nephrotic syndrome: mutations in multiple glomerular genes may influence disease severity. *Eur J Hum Genet*, 23, 1192–1199.

5 Moriniere, V., Dahan, K., Hilbert, P., Lison, M., Lebbah, S., Topa, A., Bole-Feysot, C., Pruvost, S., Nitschke, P., Plaisier, E., Knebelmann, B., Macher, M. A., Noel, L. H., Gubler, M. C., Antignac, C. & Heidet, L. 2014. Improving mutation screening in familial hematuric nephropathies through next generation sequencing. *J Am Soc Nephrol*, 25, 2740–2751.

6 Papazachariou, L., Demosthenous, P., Pieri, M., Papagregoriou, G., Savva, I., Stavrou, C., Zavros, M., Athanasiou, Y., Ioannou, K., Patsias, C., Panagides, A., Potamitis, C., Demetriou, K., Prikis, M., Hadjigavriel, M., Kkolou, M., Loukaidou, P., Pastelli, A., Michael, A., Lazarou, A., Arsali, M., Damianou, L., Goutziamani, I., Soloukides, A., Yioukas, L., Elia, A., Zouvani, I., Polycarpou, P., Pierides, A., Voskarides, K. & Deltas, C. 2014.

Frequency of COL4A3/COL4A4 mutations amongst families segregating glomerular microscopic hematuria and evidence for activation of the unfolded protein response. Focal and segmental glomerulosclerosis is a frequent development during ageing. *PLoS One*, 9, e115015.

7 Baird, P. A., Anderson, T. W., Newcombe, H. B. & Lowry, R. B. 1988. Genetic disorders in children and young adults: a population study. *Am J Hum Genet*, 42, 677–693.

8 Gabow, P. A. 1993. Autosomal dominant polycystic kidney disease. *N Engl J Med*, 329, 332–342.

9 Deltas, C., Savva, I., Voskarides, K., Papazachariou, L. & Pierides, A. 2015. Carriers of autosomal recessive alport syndrome with thin basement membrane nephropathy presenting as focal segmental glomerulosclerosis in later life. *Nephron*, 130, 271–280.

10 Haufroid, V., Mourad, M., Van Kerckhove, V., Wawrzyniak, J., De Meyer, M., Eddour, D. C., Malaise, J., Lison, D., Squifflet, J. P. & Wallemacq, P. 2004. The effect of CYP3A5 and MDR1 (ABCB1) polymorphisms on cyclosporine and tacrolimus dose requirements and trough blood levels in stable renal transplant patients. *Pharmacogenetics*, 14, 147–154.

11 Haufroid, V., Wallemacq, P., Vankerckhove, V., Elens, L., De Meyer, M., Eddour, D. C., Malaise, J., Lison, D. & Mourad, M. 2006. CYP3A5 and ABCB1 polymorphisms and tacrolimus pharmacokinetics in renal transplant candidates: guidelines from an experimental study. *Am J Transplant*, 6, 2706–2713.

12 Thummel, K. E. 2004. A genetic test for immunosuppressant dose selection? *Pharmacogenetics*, 14, 145–146.

13 Macphee, I. A. 2010. Use of pharmacogenetics to optimize immunosuppressive therapy. *Ther Drug Monit*, 32, 261–264.

14 Quteineh, L. & Verstuyft, C. 2010. Pharmacogenetics in immunosuppressants: impact on dose requirement of calcineurin inhibitors in renal and liver pediatric transplant recipients. *Curr Opin Organ Transplant*, 15, 601–607.

15 Miller, S. A., Dykes, D. D. & Polesky, H. F. 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*, 16, 1215.

16 Maniatis, T., Fritsch, E. & Sambrook, J. 1982. *Molecular Cloning, A Laboratory Manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

17 Koptides, M., Constantinides, R., Kyriakides, G., Hadjigavriel, M., Patsalis, P. C., Pierides, A. & Deltas, C. C. 1998. Loss of heterozygosity in polycystic kidney disease with a missense mutation in the repeated region of PKD1. *Hum Genet*, 103, 709–717.

18 Rossetti, S., Strmecki, L., Gamble, V., Burton, S., Sneddon, V., Peral, B., Roy, S., Bakkaloglu, A., Komel, R., Winearls, C. G. & Harris, P. C. 2001. Mutation analysis of the entire PKD1 gene: genetic and diagnostic implications. *Am J Hum Genet*, 68, 46–63.

19 Thomas, R., Mcconnell, R., Whittacker, J., Kirkpatrick, P., Bradley, J. & Sandford, R. 1999. Identification of mutations in the repeated part of the autosomal dominant polycystic kidney disease type 1 gene, PKD1, by long-range PCR. *Am J Hum Genet*, 65, 39–49.

20 Watnick, T. J., Piontek, K. B., Cordal, T. M., Weber, H., Gandolph, M. A., Qian, F., Lens, X. M., Neumann, H. P. & Germino, G. G. 1997. An unusual pattern of mutation in the duplicated portion of PKD1 is revealed by use of a novel strategy for mutation detection. *Hum Mol Genet*, 6, 1473–1481.

21 Schouten, J. P., Mcelgunn, C. J., Waaijer, R., Zwijnenburg, D., Diepvens, F. & Pals, G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res*, 30, e57.

22 Athanasiou, Y., Voskarides, K., Gale, D. P., Damianou, L., Patsias, C., Zavros, M., Maxwell, P. H., Cook, H. T., Demosthenous, P., Hadjisavvas, A., Kyriacou, K., Zouvani, I., Pierides, A. & Deltas, C. 2011. Familial C3 glomerulopathy associated with CFHR5 mutations: clinical characteristics of 91 patients in 16 pedigrees. *Clin J Am Soc Nephrol*, 6, 1436–1446.

23 Gale, D. P., De Jorge, E. G., Cook, H. T., Martinez-Barricarte, R., Hadjisavvas, A., Mclean, A. G., Pusey, C. D., Pierides, A., Kyriacou, K., Athanasiou, Y., Voskarides, K., Deltas, C., Palmer, A., Fremeaux-Bacchi, V., De Cordoba, S. R., Maxwell, P. H. & Pickering, M. C. 2010. Identification of a mutation in complement factor H-related protein 5 in patients of Cypriot origin with glomerulonephritis. *Lancet*, 376, 794–801.

24 Barua, M., Cil, O., Paterson, A. D., Wang, K., He, N., Dicks, E., Parfrey, P. & Pei, Y. 2009. Family history of renal disease severity predicts the mutated gene in ADPKD. *J Am Soc Nephrol*, 20, 1833–1838.

25 Torra, R., Badenas, C., Perez-Oller, L., Luis, J., Millan, S., Nicolau, C., Oppenheimer, F., Mila, M. & Darnell, A. 2000. Increased prevalence of polycystic kidney disease type 2 among elderly polycystic patients. *Am J Kidney Dis*, 36, 728–734.

26 Augarten, A., Kerem, B. S., Yahav, Y., Noiman, S., Rivlin, Y., Tal, A., Blau, H., Ben-Tur, L., Szeinberg, A. & Kerem, E., 1993. Mild cystic fibrosis and normal or borderline sweat test in patients with the 3849 + 10 kb C-->T mutation. *Lancet*, 342, 25–26.

27 Eriksson, M., Brown, W. T., Gordon, L. B., Glynn, M. W., Singer, J., Scott, L., Erdos, M. R., Robbins, C. M., Moses, T. Y., Berglund, P., Dutra, A., Pak, E., Durkin, S., Csoka, A. B., Boehnke, M., Glover, T. W. & Collins, F. S. 2003. Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome. *Nature*, 423, 293–298.

28 Kyrri, A. R., Felekis, X., Kalogerou, E., Wild, B. J., Kythreotis, L., Phylactides, M. & Kleanthous, M. 2009. Hemoglobin variants in Cyprus. *Hemoglobin*, 33, 81–94.

29 Peral, B., Gamble, V., San Millan, J. L., Strong, C., Sloane-Stanley, J., Moreno, F. & Harris, P. C. 1995. Splicing mutations of the polycystic kidney disease 1 (PKD1) gene induced by intronic deletion. *Hum Mol Genet*, 4, 569–574.

30 Nozu, K., Iijima, K., Kawai, K., Nozu, Y., Nishida, A., Takeshima, Y., Fu, X. J., Hashimura, Y., Kaito, H., Nakanishi, K., Yoshikawa, N. & Matsuo, M. 2009. In vivo and in vitro splicing assay of SLC12A1 in an antenatal salt-losing tubulopathy patient with an intronic mutation. *Hum Genet*, 126, 533–538.

31 Tan, Y. C., Blumenfeld, J., Michaeel, A., Donahue, S., Balina, M., Parker, T., Levine, D. & Rennert, H. 2011. Aberrant PKD2 splicing due to a presumed novel missense mutation in autosomal-dominant polycystic kidney disease. *Clin Genet*, 80, 287–292.

32 D'Agati, V. D., Kaskel, F. J. & Falk, R. J. 2011. Focal segmental glomerulosclerosis. *N Engl J Med*, 365, 2398–2411.

33 Sadowski, C. E., Lovric, S., Ashraf, S., Pabst, W. L., Gee, H. Y., Kohl, S., Engelmann, S., Vega-Warner, V., Fang, H., Halbritter, J., Somers, M. J., Tan, W., Shril, S., Fessi, I., Lifton, R. P., Bockenhauer, D., El-Desoky, S., Kari, J. A., Zenker, M., Kemper, M. J., Mueller, D., Fathy, H. M., Soliman, N. A., Group, S. S. & Hildebrandt, F. 2015. A single-gene cause in 29.5% of cases of steroid-resistant nephrotic syndrome. *J Am Soc Nephrol*, 26, 1279–1289.

34 Guaragna, M. S., Lutaif, A. C., Piveta, C. S., Souza, M. L., De Souza, S. R., Henriques, T. B., Maciel-Guerra, A. T., Belangero, V. M., Guerra-Junior, G. & De Mello, M. P. 2015. NPHS2 mutations account for only 15% of nephrotic syndrome cases. *BMC Med Genet*, 16, 88.

35 Voskarides, K., Damianou, L., Neocleous, V., Zouvani, I., Christodoulidou, S., Hadjiconstantinou, V., Ioannou, K., Athanasiou, Y., Patsias, C., Alexopoulos, E., Pierides, A., Kyriacou, K. & Deltas, C. 2007. COL4A3/COL4A4 mutations producing focal segmental glomerulosclerosis and renal failure in thin basement membrane nephropathy. *J Am Soc Nephrol*, 18, 3004–3016.

36 Deltas, C., Pierides, A. & Voskarides, K. 2013. Molecular genetics of familial hematuric diseases. *Nephrol Dial Transplant*, 28, 2946–2960.

37 Voskarides, K., Pierides, A. & Deltas, C. 2008. COL4A3/COL4A4 mutations link familial hematuria and focal segmental glomerulosclerosis. glomerular epithelium destruction via basement membrane thinning? *Connect Tissue Res*, 49, 283–288.

38 Gast, C., Pengelly, R. J., Lyon, M., Bunyan, D. J., Seaby, E. G., Graham, N., Venkat-Raman, G. & Ennis, S. 2016. Collagen (COL4A) mutations are the most frequent mutations underlying adult focal segmental glomerulosclerosis. *Nephrol Dial Transplant*, 31(6), 961–970.

39 Malone, A. F., Phelan, P. J., Hall, G., Cetincelik, U., Homstad, A., Alonso, A. S., Jiang, R., Lindsey, T. B., Wu, G., Sparks, M. A., Smith, S. R., Webb, N. J., Kalra, P. A., Adeyemo, A. A., Shaw, A. S., Conlon, P. J., Jennette, J. C., Howell, D. N., Winn, M. P. & Gbadegesin, R. A. 2014. Rare hereditary COL4A3/COL4A4 variants may be mistaken for familial focal segmental glomerulosclerosis. *Kidney Int*, 86, 1253–1259.

40 Bennett, M. R., Piyaphanee, N., Czech, K., Mitsnefes, M. & Devarajan, P. 2012. NGAL distinguishes steroid sensitivity in idiopathic nephrotic syndrome. *Pediatr Nephrol*, 27, 807–812.

41 Kestila, M., Lenkkeri, U., Mannikko, M., Lamerdin, J., Mccready, P., Putaala, H., Ruotsalainen, V., Morita, T., Nissinen, M., Herva, R., Kashtan, C. E., Peltonen, L., Holmberg, C., Olsen, A. & Tryggvason, K. 1998. Positionally cloned gene for a novel glomerular protein--nephrin—is mutated in congenital nephrotic syndrome. *Mol Cell*, 1, 575–582.

42 Demetriou, K., Tziakouri, C., Anninou, K., Eleftheriou, A., Koptides, M., Nicolaou, A., Deltas, C. C. & Pierides, A. 2000. Autosomal dominant polycystic kidney disease-type 2. Ultrasound, genetic and clinical correlations. *Nephrol Dial Transplant*, 15, 205–211.

43 Hateboer, N., Van Dijk, M. A., Bogdanova, N., Coto, E., Saggar-Malik, A. K., San Millan, J. L., Torra, R., Breuning, M. & Ravine, D. 1999. Comparison of phenotypes of polycystic kidney disease types 1 and 2. European PKD1-PKD2 Study Group. *Lancet*, 353, 103–107.

44 Magistroni, R., He, N., Wang, K., Andrew, R., Johnson, A., Gabow, P., Dicks, E., Parfrey, P., Torra, R., San-Millan, J. L., Coto, E., Van Dijk, M., Breuning, M., Peters, D., Bogdanova, N., Ligabue, G., Albertazzi, A., Hateboer, N., Demetriou, K., Pierides, A., Deltas, C., St George-Hyslop, P., Ravine, D. & Pei, Y. 2003. Genotype-renal function correlation in type 2 autosomal dominant polycystic kidney disease. *J Am Soc Nephrol*, 14, 1164–1174.

45 Rossetti, S., Chauveau, D., Kubly, V., Slezak, J. M., Saggar-Malik, A. K., Pei, Y., Ong, A. C., Stewart, F., Watson, M. L., Bergstralh, E. J., Winearls, C. G., Torres, V. E. & Harris, P. C. 2003. Association of mutation position in polycystic kidney disease 1 (PKD1) gene and development of a vascular phenotype. *Lancet*, 361, 2196–2201.

46 Hateboer, N., Veldhuisen, B., Peters, D., Breuning, M. H., San-Millan, J. L., Bogdanova, N., Coto, E., Van Dijk, M. A., Afzal, A. R., Jeffery, S., Saggar-Malik, A. K.,

Torra, R., Dimitrakov, D., Martinez, I., De Castro, S. S., Krawczak, M. & Ravine, D. 2000. Location of mutations within the PKD2 gene influences clinical outcome. *Kidney Int*, 57, 1444–1451.

47 Rossetti, S., Burton, S., Strmecki, L., Pond, G. R., San Millan, J. L., Zerres, K., Barratt, T. M., Ozen, S., Torres, V. E., Bergstralh, E. J., Winearls, C. G. & Harris, P. C. 2002. The position of the polycystic kidney disease 1 (PKD1) gene mutation correlates with the severity of renal disease. *J Am Soc Nephrol*, 13, 1230–1237.

48 Hwang, Y. H., Conklin, J., Chan, W., Roslin, N. M., Liu, J., He, N., Wang, K., Sundsbak, J. L., Heyer, C. M., Haider, M., Paterson, A. D., Harris, P. C. & Pei, Y. 2016. Refining genotype-phenotype correlation in autosomal dominant polycystic kidney disease. *J Am Soc Nephrol*, 27(6), 1861–1868.

49 Bekheirnia, M. R., Reed, B., Gregory, M. C., Mcfann, K., Shamshirsaz, A. A., Masoumi, A. & Schrier, R. W. 2010. Genotype-phenotype correlation in X-linked Alport syndrome. *J Am Soc Nephrol*, 21, 876–883.

50 Gross, O., Netzer, K. O., Lambrecht, R., Seibold, S. & Weber, M. 2002. Meta-analysis of genotype-phenotype correlation in X-linked Alport syndrome: impact on clinical counselling. *Nephrol Dial Transplant*, 17, 1218–1227.

51 Jais, J. P., Knebelmann, B., Giatras, I., de Marchi, M., Rizzoni, G., Renieri, A., Weber, M., Gross, O., Netzer, K. O., Flinter, F., Pirson, Y., Verellen, C., Wieslander, J., Persson, U., Tryggvason, K., Martin, P., Hertz, J. M., Schroder, C., Sanak, M., Krejcova, S., Carvalho, M. F., Saus, J., Antignac, C., Smeets, H. & Gubler, M. C. 2000. X-linked Alport syndrome: natural history in 195 families and genotype- phenotype correlations in males. *J Am Soc Nephrol*, 11, 649–657.

52 Tsiakkis, D., Pieri, M., Koupepidou, P., Demosthenous, P., Panayidou, K. & Deltas, C. 2012. Genotype-phenotype correlation in X-linked Alport syndrome patients carrying missense mutations in the collagenous domain of COL4A5. *Clin Genet*, 82, 297–299.

53 Gropman, A. L. & Adams, D. R. 2007. Atypical patterns of inheritance. *Semin Pediatr Neurol*, 14, 34–45.

54 Lowik, M., Levtchenko, E., Westra, D., groenen, P., Steenbergen, E., Weening, J., Lilien, M., Monnens, L. & Van Den Heuvel, L. 2008. Bigenic heterozygosity and the development of steroid-resistant focal segmental glomerulosclerosis. *Nephrol Dial Transplant*, 23, 3146–3151.

55 Hopp, K., Heyer, C. M., Hommerding, C. J., Henke, S. A., Sundsbak, J. L., Patel, S., Patel, P., Consugar, M. B., Czarnecki, P. G., Gliem, T. J., Torres, V. E., Rossetti, S. & Harris, P. C. 2011. B9D1 is revealed as a novel Meckel syndrome (MKS) gene by targeted exon-enriched next-generation sequencing and deletion analysis. *Hum Mol Genet*, 20, 2524–2534.

56 Chen, J., Smaoui, N., Hammer, M. B., Jiao, X., Riazuddin, S. A., Harper, S., Katsanis, N., Riazuddin, S., Chaabouni, H., Berson, E. L. & Hejtmancik, J. F. 2011. Molecular analysis of Bardet-Biedl syndrome families: report of 21 novel mutations in 10 genes. *Invest Ophthalmol Vis Sci*, 52, 5317–5324.

57 Koptides, M., Hadjimichael, C., Koupepidou, P., Pierides, A. & Deltas, C. C. 1999. Germinal and somatic mutations in the PKD2 gene of renal cysts in autosomal dominant polycystic kidney disease. *Hum Mol Genet*, 8, 509–513.

58 Koptides, M., Mean, R., Demetriou, K., Pierides, A. & Deltas, C. C. 2000. Genetic evidence for a trans-heterozygous model for cystogenesis in autosomal dominant polycystic kidney disease. *Hum Mol Genet*, 9, 447–452.

59 Pei, Y. 2001. A "two-hit" model of cystogenesis in autosomal dominant polycystic kidney disease? *Trends Mol Med*, 7, 151–156.

60 Wu, G., Tian, X., Nishimura, S., Markowitz, G. S., D'Agati, V., Park, J. H., Yao, L., Li, L., Geng, L., Zhao, H., Edelmann, W. & Somlo, S. 2002. Trans-heterozygous Pkd1 and Pkd2 mutations modify expression of polycystic kidney disease. *Hum Mol Genet*, 11, 1845–1854.

61 Bergmann, C., Von Bothmer, J., Ortiz Bruchle, N., Venghaus, A., Frank, V., Fehrenbach, H., Hampel, T., Pape, L., Buske, A., Jonsson, J., Sarioglu, N., Santos, A., Ferreira, J. C., Becker, J. U., Cremer, R., Hoefele, J., Benz, M. R., Weber, L. T., Buettner, R. & Zerres, K. 2011. Mutations in multiple PKD genes may explain early and severe polycystic kidney disease. *J Am Soc Nephrol*, 22, 2047–2056.

62 Mencarelli, M. A., Heidet, L., Storey, H., Van Geel, M., Knebelmann, B., Fallerini, C., Miglietti, N., Antonucci, M. F., Cetta, F., Sayer, J. A., van den Wijngaard, A., Yau, S., Mari, F., Bruttini, M., Ariani, F., Dahan, K., Smeets, B., Antignac, C., Flinter, F. & Renieri, A. 2015. Evidence of digenic inheritance in Alport syndrome. *J Med Genet*, 52(3), 163–174.

63 Fain, P. R., Mcfann, K. K., Taylor, M. R., Tison, M., Johnson, A. M., Reed, B. & Schrier, R. W. 2005. Modifier genes play a significant role in the phenotypic expression of PKD1. *Kidney Int*, 67, 1256–1267.

64 Paterson, A. D., Magistroni, R., He, N., Wang, K., Johnson, A., Fain, P. R., Dicks, E., Parfrey, P., St George-Hyslop, P., & Pei, Y. 2005. Progressive loss of renal function is an age-dependent heritable trait in type 1 autosomal dominant polycystic kidney disease. *J Am Soc Nephrol*, 16, 755–762.

65 Lamnissou, K., Zirogiannis, P., Trygonis, S., Demetriou, K., Pierides, A., Koptides, M. & Deltas, C. C. 2004. Evidence for association of endothelial cell nitric oxide synthase gene polymorphism with earlier progression to end-stage renal disease in a cohort of Hellens from Greece and Cyprus. *Genet Test*, 8, 319–324.

66 Persu, A., Stoenoiu, M. S., Messiaen, T., Davila, S., Robino, C., El-Khattabi, O., Mourad, M., Horie, S., Feron, O., Balligand, J. L., Wattiez, R., Pirson, Y., Chauveau, D., Lens, X. M. & Devuyst, O. 2002. Modifier effect of ENOS in autosomal dominant polycystic kidney disease. *Hum Mol Genet*, 11, 229–241.

67 Khanna, H., Davis, E. E., Murga-Zamalloa, C. A., Estrada-Cuzcano, A., Lopez, I., Den Hollander, A. I., Zonneveld, M. N., Othman, M. I., Waseem, N., Chakarova, C. F., Maubaret, C., Diaz-Font, A., Macdonald, I., Muzny, D. M., Wheeler, D. A., Morgan, M., Lewis, L. R., Logan, C. V., Tan, P. L., Beer, M. A., Inglehearn, C. F., Lewis, R. A., Jacobson, S. G., Bergmann, C., Beales, P. L., Attie-Bitach, T., Johnson, C. A., Otto, E. A., Bhattacharya, S. S., Hildebrandt, F., Gibbs, R. A., Koenekoop, R. K., Swaroop, A. & Katsanis, N. 2009. A common allele in RPGRIP1L is a modifier of retinal degeneration in ciliopathies. *Nat Genet*, 41, 739–745.

68 Louie, C. M., Caridi, G., Lopes, V. S., Brancati, F., Kispert, A., Lancaster, M. A., Schlossman, A. M., Otto, E. A., Leitges, M., Grone, H. J., Lopez, I., Gudiseva, H. V., O'Toole, J. F., Vallespin, E., Ayyagari, R., Ayuso, C., Cremers, F. P., Den Hollander, A. I., Koenekoop, R. K., Dallapiccola, B., Ghiggeri, G. M., Hildebrandt, F., Valente, E. M., Williams, D. S. & Gleeson, J. G. 2010. AHI1 is required for photoreceptor outer segment development and is a modifier for retinal degeneration in nephronophthisis. *Nat Genet*, 42, 175–180.

69 Liu, M., Shi, S., Senthilnathan, S., Yu, J., Wu, E., Bergmann, C., Zerres, K., Bogdanova, N., Coto, E., Deltas, C., Pierides, A., Demetriou, K., Devuyst, O., Gitomer, B., Laakso, M., Lumiaho, A., Lamnissou, K., Magistroni, R., Parfrey, P., Breuning, M., Peters, D. J., Torra, R., Winearls, C. G., Torres, V. E., Harris, P. C., Paterson, A. D. & Pei, Y. 2010. Genetic variation of DKK3 may modify renal disease severity in ADPKD. *J Am Soc Nephrol*, 21, 1510–1520.

70 Stefanakis, N., Ziroyiannis, P., Trygonis, S. & Lamnissou, K. 2008. Modifier effect of the Glu298Asp polymorphism of endothelial nitric oxide synthase gene in autosomal-dominant polycystic kidney disease. *Nephron Clin Pract*, 110, c101–c106.

71 Yoshida, H., Mitarai, T., Kawamura, T., Kitajima, T., Miyazaki, Y., Nagasawa, R., Kawaguchi, Y., Kubo, H., Ichikawa, I. & Sakai, O. 1995. Role of the deletion of polymorphism of the angiotensin converting enzyme gene in the progression and therapeutic responsiveness of IgA nephropathy. *J Clin Invest*, 96, 2162–2169.

72 Tonna, S., Wang, Y. Y., Wilson, D., Rigby, L., Tabone, T., Cotton, R. & Savige, J. 2008. The R229Q mutation in NPHS2 may predispose to proteinuria in thin-basement-membrane nephropathy. *Pediatr Nephrol*, 23, 2201–2207.

73 Stefanou, C., Pieri, M., Savva, I., Georgiou, G., Pierides, A., Voskarides, K. & Deltas, C. 2015. Co-inheritance of functional podocin variants with heterozygous collagen IV mutations predisposes to renal failure. *Nephron*, 130, 200–212.

74 Voskarides, K., Arsali, M., Athanasiou, Y., Elia, A., Pierides, A. & Deltas, C. 2012. Evidence that NPHS2-R229Q predisposes to proteinuria and renal failure in familial hematuria. *Pediatr Nephrol*, 27, 675–679.

75 Papagregoriou, G., Erguler, K., Dweep, H., Voskarides, K., Koupepidou, P., Athanasiou, Y., Pierides, A., Gretz, N., Felekkis, K. N. & Deltas, C. 2012. A miR-1207-5p binding site polymorphism abolishes regulation of HBEGF and is associated with disease severity in CFHR5 nephropathy. *PLoS One*, 7, e31021.

76 Pereira, T. V., Nunes, A. C., Rudnicki, M., Magistroni, R., Albertazzi, A., Pereira, A. C. & Krieger, J. E. 2006. Influence of ACE I/D gene polymorphism in the progression of renal failure in autosomal dominant polycystic kidney disease: a meta-analysis. *Nephrol Dial Transplant*, 21, 3155–3163.

77 Alport, A. C. 1927. Hereditary familial congenital haemorrhagic nephritis. *Br Med J*, 1, 504–506.

78 Barker, D. F., Hostikka, S. L., Zhou, J., Chow, L. T., Oliphant, A. R., Gerken, S. C., Gregory, M. C., Skolnick, M. H., Atkin, C. L. & Tryggvason, K. 1990. Identification of mutations in the COL4A5 collagen gene in Alport syndrome. *Science*, 248, 1224–1227.

79 Feingold, J., Bois, E., Chompret, A., Broyer, M., Gubler, M. C. & Grunfeld, J. P. 1985. Genetic heterogeneity of Alport syndrome. *Kidney Int*, 27, 672–677.

80 Jais, J. P., Knebelmann, B., Giatras, I., de Marchi, M., Rizzoni, G., Renieri, A., Weber, M., Gross, O., Netzer, K. O., Flinter, F., Pirson, Y., Dahan, K., Wieslander, J., Persson, U., Tryggvason, K., Martin, P., Hertz, J. M., Schroder, C., Sanak, M., Carvalho, M. F., Saus, J., Antignac, C., Smeets, H. & Gubler, M. C. 2003. X-linked Alport syndrome: natural history and genotype-phenotype correlations in girls and women belonging to 195 families: a "European Community Alport Syndrome Concerted Action" study. *J Am Soc Nephrol*, 14, 2603–2610.

81 Williamson, D. A. 1961. Alport's syndrome of hereditary nephritis with deafness. *Lancet*, 2, 1321–1323.

82 Barker, D. F., Denison, J. C., Atkin, C. L. & Gregory, M. C. 2001. Efficient detection of Alport syndrome COL4A5 mutations with multiplex genomic PCR-SSCP. *Am J Med Genet*, 98, 148–160.

83 Deltas, C., Voskarides, K., Demosthenous, P., Papazachariou, L., Zirogiannis, P. & Pierides, A. 2012. The power of molecular genetics in establishing the diagnosis and offering prenatal testing: the case for Alport Syndrome. In Sahay, M. (ed.) *Diseases of Renal Parenchyma*, InTech Publishing, Rijeka.

84 Demosthenous, P., Voskarides, K., Stylianou, K., Hadjigavriel, M., Arsali, M., Patsias, C., Georgaki, E., Zirogiannis, P., Stavrou, C., Daphnis, E., Pierides, A. & Deltas, C. 2012. X-linked Alport syndrome in Hellenic families: phenotypic heterogeneity and mutations near interruptions of the collagen domain in COL4A5. *Clin Genet*, 81, 240–248.

85 Martin, P., Heiskari, N., Zhou, J., Leinonen, A., Tumelius, T., Hertz, J. M., Barker, D., Gregory, M., Atkin, C., Styrkarsdottir, U., Neumann, H., Springate, J., Shows, T., Pettersson, E. & Tryggvason, K. 1998. High mutation detection rate in the COL4A5 collagen gene in suspected Alport syndrome using PCR and direct DNA sequencing. *J Am Soc Nephrol*, 9, 2291–2301.

86 Pierides, A., Voskarides, K., Kkolou, M., Hadjigavriel, M. & Deltas, C. 2013. X-linked, COL4A5 hypomorphic Alport mutations such as G624D and P628L may only exhibit thin basement membrane nephropathy with microhematuria and late onset kidney failure. *Hippokratia*, 17, 207–213.

87 Slajpah, M., Gorinsek, B., Berginc, G., Vizjak, A., Ferluga, D., Hvala, A., Meglic, A., Jaksa, I., Furlan, P., Gregoric, A., Kaplan-Pavlovcic, S., Ravnik-Glavac, M. & Glavac, D. 2007. Sixteen novel mutations identified in COL4A3, COL4A4, and COL4A5 genes in Slovenian families with Alport syndrome and benign familial hematuria. *Kidney Int*, 71, 1287–1295.

88 Pierides, A., Voskarides, K., Kkolou, M., Hadjigavriel, M. & Deltas, C. 2013. X-linked, COL4A5 hypomorphic Alport mutations such as G624D and P628L may only exhibit thin basement membrane nephropathy with microhematuria and late onset kidney failure. *Hippokratia*, 17, 7.

89 Kashtan, C. E. 2005. Familial hematurias: what we know and what we don't. *Pediatr Nephrol*, 20, 1027–1035.

90 Lemmink, H. H., Nillesen, W. N., Mochizuki, T., Schroder, C. H., Brunner, H. G., Van Oost, B. A., Monnens, L. A. & Smeets, H. J. 1996. Benign familial hematuria due to mutation of the type IV collagen alpha4 gene. *J Clin Invest*, 98, 1114–1118.

91 Mochizuki, T., Lemmink, H. H., Mariyama, M., Antignac, C., Gubler, M. C., Pirson, Y., Verellen-Dumoulin, C., Chan, B., Schroder, C. H., Smeets, H. J., & Reeders, S. T. 1994. Identification of mutations in the alpha 3(IV) and alpha 4(IV) collagen genes in autosomal recessive Alport syndrome. *Nat Genet*, 8, 77–81.

92 Rana, K., Wang, Y. Y., Buzza, M., Tonna, S., Zhang, K. W., Lin, T., Sin, L., Padavarat, S. & Savige, J. 2005. The genetics of thin basement membrane nephropathy. *Semin Nephrol*, 25, 163–170.

93 Deltas, C. 2009. Thin basement membrane nephropathy: is there genetic predisposition to more severe disease? *Pediatr Nephrol*, 24, 877–879.

94 Deltas, C., Pierides, A. & Voskarides, K. 2012. The role of molecular genetics in diagnosing familial hematuria(s). *Pediatr Nephrol*, 27(8), 1221–1231.

95 Pierides, A., Voskarides, K., Athanasiou, Y., Ioannou, K., Damianou, L., Arsali, M., Zavros, M., Pierides, M., Vargemezis, V., Patsias, C., Zouvani, I., Elia, A., Kyriacou, K. & Deltas, C. 2009. Clinico-pathological correlations in 127 patients in 11 large pedigrees, segregating one of three heterozygous mutations in the COL4A3/COL4A4 genes associated with familial haematuria and significant late progression to proteinuria and chronic kidney disease from focal segmental glomerulosclerosis. *Nephrol Dial Transplant*, 24, 2721–2729.

96 Pereira, A. C., Pereira, A. B., Mota, G. F., Cunha, R. S., Herkenhoff, F. L., Pollak, M. R., Mill, J. G. & Krieger, J. E. 2004. NPHS2 R229Q functional variant is associated with microalbuminuria in the general population. *Kidney Int*, 65, 1026–1030.

97 Kottgen, A., Hsu, C. C., Coresh, J., Shuldiner, A. R., Berthier-Schaad, Y., Gambhir, T. R., Smith, M. W., Boerwinkle, E. & Kao, W. H. 2008. The association of podocin R229Q polymorphism with increased albuminuria or reduced estimated GFR in a large population-based sample of US adults. *Am J Kidney Dis*, 52, 868–875.

98 Franceschini, N., North, K. E., Kopp, J. B., Mckenzie, L. & Winkler, C. 2006. NPHS2 gene, nephrotic syndrome and focal segmental glomerulosclerosis: a HuGE review. *Genet Med*, 8, 63–75.

99 Stefanou, C., Pieri, M., Savva, I., Georgiou, G., Pierides, A., Voskarides, K. & Deltas, C. 2015. Co-inheritance of functional podocin variants with heterozygous collagen IV mutations predisposes to renal failure. *Nephron-Experimental Nephrology and Genetics*, 130(3), 200–212.

100 Mcrae, J. L., Cowan, P. J., Power, D. A., Mitchelhill, K. I., Kemp, B. E., Morgan, B. P. & Murphy, B. F. 2001. Human factor H-related protein 5 (FHR-5). A new complement-associated protein. *J Biol Chem*, 276, 6747–6754.

101 Grekas, D., Morley, A. R., Wilkinson, R. & Kerr, D. N. 1984. Isolated C3 deposition in patients without systemic disease. *Clin Nephrol*, 21, 270–274.

102 Manno, C., Proscia, A. R., Laraia, E., Giangrande, M., Di Carlo, M., Salvatore, C., Tasco, A. & Schena, F. P. 1990. Clinicopathological features in patients with isolated C3 mesangial proliferative glomerulonephritis. *Nephrol Dial Transplant*, 5 Suppl 1, 78–80.

103 Sirbat, D., Saudax, E., Hurault De Ligny, B., Bene, M. C. & Raspiller, A. 1983. A new etiology of episcleritis: nephropathies with IgA and/or isolated C3 deposits. *J Fr Ophtalmol*, 6, 921–925.

104 Abrera-Abeleda, M. A., Nishimura, C., Smith, J. L., Sethi, S., Mcrae, J. L., Murphy, B. F., Silvestri, G., Skerka, C., Jozsi, M., Zipfel, P. F., Hageman, G. S. & Smith, R. J. 2006. Variations in the complement regulatory genes factor H (CFH) and factor H related 5 (CFHR5) are associated with membranoproliferative glomerulonephritis type II (dense deposit disease). *J Med Genet*, 43, 582–589.

105 Calls Ginesta, J., Almirall, J., Torras, A., Darnell, A. & Revert, L. 1995. Long-term evolution of patients with isolated C3 mesangial glomerulonephritis. *Clin Nephrol*, 43, 221–225.

106 Gallego, N., Teruel, J. L., Mampaso, F., Gonzalo, A. & Ortuno, J. 1991. Acute interstitial nephritis superimposed on glomerulonephritis: report of a case. *Pediatr Nephrol*, 5, 229–231.

107 Habbig, S., Mihatsch, M. J., Heinen, S., Beck, B., Emmel, M., Skerka, C., Kirschfink, M., Hoppe, B., Zipfel, P. F. & Licht, C. 2009. C3 deposition glomerulopathy due to a functional factor H defect. *Kidney Int*, 75, 1230–1234.

108 Servais, A., Fremeaux-Bacchi, V., Lequintrec, M., Salomon, R., Blouin, J., Knebelmann, B., Grunfeld, J. P., Lesavre, P., Noel, L. H. & Fakhouri, F. 2007. Primary glomerulonephritis with isolated C3 deposits: a new entity which shares common genetic risk factors with haemolytic uraemic syndrome. *J Med Genet*, 44, 193–199.

109 Paterson, A. D., Liu, X. Q., Wang, K., Magistroni, R., Song, X., Kappel, J., Klassen, J., Cattran, D., St George-Hyslop, P. & Pei, Y. 2007. Genome-wide linkage scan of a large family with IgA nephropathy localizes a novel susceptibility locus to chromosome 2q36. *J Am Soc Nephrol*, 18, 2408–2415.

110 Pickering, M. C., D'agati, V. D., Nester, C. M., Smith, R. J., Haas, M., Appel, G. B., Alpers, C. E., Bajema, I. M., Bedrosian, C., Braun, M., Doyle, M., Fakhouri, F., Fervenza, F. C., Fogo, A. B., Fremeaux-Bacchi, V., Gale, D. P., Goicoechea de Jorge, E., Griffin, G., Harris, C. L., Holers, V. M., Johnson, S., Lavin, P. J., Medjeral-Thomas, N., Paul Morgan, B., Nast, C. C., Noel, L. H., Peters, D. K., Rodriguez de Cordoba, S., Servais, A., Sethi, S., Song, W. C., Tamburini, P., Thurman, J. M., Zavros, M. & Cook, H. T. 2013. C3 glomerulopathy: consensus report. *Kidney Int*, 84, 1079–1089.

111 Medjeral-Thomas, N., Malik, T. H., Patel, M. P., Toth, T., Cook, H. T., Tomson, C. & Pickering, M. C. 2014. A novel CFHR5 fusion protein causes C3 glomerulopathy in a family without Cypriot ancestry. *Kidney Int*, 85, 933–937.

112 Constantinou-Deltas, C. D., Ladda, R. L. & Prockop, D. J. 1993. Somatic cell mosaicism: another source of phenotypic heterogeneity in nuclear families with osteogenesis imperfecta. *Am J Med Genet*, 45, 246–251.

113 Deltas, C. & Papagregoriou, G. 2010. Cystic diseases of the kidney: molecular biology and genetics. *Arch Pathol Lab Med*, 134, 569–582.

114 Scolari, F., Puzzer, D., Amoroso, A., Caridi, G., Ghiggeri, G. M., Maiorca, R., Aridon, P., De Fusco, M., Ballabio, A. & Casari, G. 1999. Identification of a new locus for medullary cystic disease, on chromosome 16p12. *Am J Hum Genet*, 64, 1655–1660.

115 Hart, T. C., Gorry, M. C., Hart, P. S., Woodard, A. S., Shihabi, Z., Sandhu, J., Shirts, B., Xu, L., Zhu, H., Barmada, M. M. & Bleyer, A. J. 2002. Mutations of the UMOD gene are responsible for medullary cystic kidney disease 2 and familial juvenile hyperuricaemic nephropathy. *J Med Genet*, 39, 882–892.

116 Rampoldi, L., Caridi, G., Santon, D., Boaretto, F., Bernascone, I., Lamorte, G., Tardanico, R., Dagnino, M., Colussi, G., Scolari, F., Ghiggeri, G. M., Amoroso, A. & Casari, G. 2003. Allelism of MCKD, FJHN and GCKD caused by impairment of uromodulin export dynamics. *Hum Mol Genet*, 12, 3369–3384.

117 Stiburkova, B., Majewski, J., Hodanova, K., Orova, L., Jebkova, M., Zikanova, M., Vylet'Al, P., Sebesta, I., Marinaki, A., Simmonds, A., Matthijs, G., Fryns, J. P., Torres, R., Puig, J. G., Ott, J. & Kmoch, S. 2003. Familial juvenile hyperuricaemic nephropathy (FJHN): linkage analysis in 15 families, physical and transcriptional characterisation of the FJHN critical region on chromosome 16p11.2 and the analysis of seven candidate genes. *Eur J Hum Genet*, 11, 145–154.

118 Gusmano, R., Caridi, G., Marini, M., Perfumo, F., Ghiggeri, G. M., Piaggio, G., Ceccherini, I. & Seri, M. 2002. Glomerulocystic kidney disease in a family. *Nephrol Dial Transplant*, 17, 813–818.

119 Scolari, F., Caridi, G., Rampoldi, L., Tardanico, R., Izzi, C., Pirulli, D., Amoroso, A., Casari, G. & Ghiggeri, G. M. 2004. Uromodulin storage diseases: clinical aspects and mechanisms. *Am J Kidney Dis*, 44, 987–999.

120 Eckardt, K. U., Alper, S. L., Antignac, C., Bleyer, A. J., Chauveau, D., Dahan, K., Deltas, C., Hosking, A., Kmoch, S., Rampoldi, L., Wiesener, M., Wolf, M. T. & Devuyst, O. 2015. Autosomal dominant tubulointerstitial kidney disease: diagnosis, classification, and management-A KDIGO consensus report. *Kidney Int*, 88(4), 676–683.

121 Pickering, M. & Cook, H. T. 2011. Complement and glomerular disease: new insights. *Curr Opin Nephrol Hypertens*, 20, 271–277.

122 Edwards, A. O., Ritter, R., 3rd, Abel, K. J., Manning, A., Panhuysen, C. & Farrer, L. A. 2005. Complement factor H polymorphism and age-related macular degeneration. *Science*, 308, 421–424.

123 Haines, J. L., Hauser, M. A., Schmidt, S., Scott, W. K., Olson, L. M., Gallins, P., Spencer, K. L., Kwan, S. Y., Noureddine, M., Gilbert, J. R., Schnetz-Boutaud, N.,

Agarwal, A., Postel, E. A. & Pericak-Vance, M. A. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science*, 308, 419–421.

124 Hirt-Minkowski, P., Dickenmann, M. & Schifferli, J. A. 2010. Atypical hemolytic uremic syndrome: update on the complement system and what is new. *Nephron Clin Pract*, 114, c219–c235.

125 Becknell, B., Zender, G. A., Houston, R., Baker, P. B., Mcbride, K. L., Luo, W., Hains, D. S., Borza, D. B. & Schwaderer, A. L. 2011. Novel X-linked glomerulopathy is associated with a COL4A5 missense mutation in a non-collagenous interruption. *Kidney Int*, 79, 120–127.

126 Parkin, J. D., San Antonio, J. D., Pedchenko, V., Hudson, B., Jensen, S. T. & Savige, J. 2011. Mapping structural landmarks, ligand binding sites, and missense mutations to the collagen IV heterotrimers predicts major functional domains, novel interactions, and variation in phenotypes in inherited diseases affecting basement membranes. *Hum Mutat*, 32, 127–143.

127 Carboni, N., Mateddu, A., Marrosu, G., Cocco, E. & Marrosu, M. G. 2013. Genetic and clinical characteristics of skeletal and cardiac muscle in patients with lamin A/C gene mutations. *Muscle Nerve*, 48, 161–170.

128 Fallerini, C., Dosa, L., Tita, R., Del Prete, D., Feriozzi, S., Gai, G., Clementi, M., La Manna, A., Miglietti, N., Mancini, R., Mandrile, G., Ghiggeri, G. M., Piaggio, G., Brancati, F., Diano, L., Frate, E., Pinciaroli, A. R., Giani, M., Castorina, P., Bresin, E., Giachino, D., De Marchi, M., Mari, F., Bruttini, M., Renieri, A. & Ariani, F. 2014. Unbiased next generation sequencing analysis confirms the existence of autosomal dominant Alport syndrome in a relevant fraction of cases. *Clin Genet*, 86, 252–257.

129 Mccarthy, H. J., Bierzynska, A., Wherlock, M., Ognjanovic, M., Kerecuk, L., Hegde, S., Feather, S., Gilbert, R. D., Krischock, L., Jones, C., Sinha, M. D., Webb, N. J., Christian, M., Williams, M. M., Marks, S., Koziell, A., Welsh, G. I., Saleem, M. A. & Group, R. T. U. S. S. 2013. Simultaneous sequencing of 24 genes associated with steroid-resistant nephrotic syndrome. *Clin J Am Soc Nephrol*, 8, 637–648.

130 Nicolaou, N., Pulit, S. L., Nijman, I. J., Monroe, G. R., Feitz, W. F., Schreuder, M. F., Van Eerde, A. M., de Jong, T. P., Giltay, J. C., van der Zwaag, B., Havenith, M. R., Zwakenberg, S., van der Zanden, L. F., Poelmans, G., Cornelissen, E. A., Lilien, M. R., Franke, B., Roeleveld, N., Van Rooij, I. A., Cuppen, E., Bongers, E. M., Giles, R. H., Knoers, N. V. & Renkema, K. Y. 2016. Prioritization and burden analysis of rare variants in 208 candidate genes suggest they do not play a major role in CAKUT. *Kidney Int*, 89(2), 476–486.

131 Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., Mcindoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., Mccombie, W. R., Hicks, J. & Wigler, M. 2011. Tumour evolution inferred by single-cell sequencing. *Nature*, 472, 90–94.

132 Xu, X., Hou, Y., Yin, X., Bao, L., Tang, A., Song, L., Li, F., Tsang, S., Wu, K., Wu, H., He, W., Zeng, L., Xing, M., Wu, R., Jiang, H., Liu, X., Cao, D., Guo, G., Hu, X., Gui, Y., Li, Z., Xie, W., Sun, X., Shi, M., Cai, Z., Wang, B., Zhong, M., Li, J., Lu, Z., Gu, N., Zhang, X., Goodman, L., Bolund, L., Wang, J., Yang, H., Kristiansen, K., Dean, M. & Li, Y. 2012. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, 148, 886–895.

133 Green, R. C., Berg, J. S., Grody, W. W., Kalia, S. S., Korf, B. R., Martin, C. L., Mcguire, A. L., Nussbaum, R. L., O'Daniel, J. M., Ormond, K. E., Rehm, H. L., Watson, M. S., Williams, M. S., Biesecker, L. G., American College of Medical, G. & Genomics 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med*, 15, 565–574.

2

Biobanking, Ethics, and Relevant Legal Issues

Brigitte Lohff[1], *Thomas Illig*[2], *and Dieter Tröger*[3,†]

[1] *Institute of History, Ethics and Philosophy of Medicine MHH, Research Ethical Committee, Hanover, Germany*
[2] *CEO, Hannover Unified Biobank (HUB) MHH, Research Ethical Committee, Hanover, Germany*
[3] *Institute for Forensic Medicine MHH, Research Ethical Committee MHH, Hanover, Germany*

2.1 Introduction

Basic medical research allows development of new therapeutic modalities and requires the use of human biomaterials. Since the development of new therapeutics involves clinical trials, an issue of moral limits and ethical norms appear, which need to be considered, when experiments with patients and healthy individuals are planned.

Clinical trials involve not only research-oriented physicians but also experts from the laboratories, the pharmaceutical industry, and developers of medical devices. There is a consensus among all these groups involved in medical research about the necessity of development of new diagnostics and therapeutics; however the opinions, whether a study is still ethically acceptable, may differ. The knowledge of the benefits versus risks, to which the enrolled study individuals are exposed, plays a big role in shaping the opinions on ethical issues. Although it may suffer from lack of objectivity, each opinion is relevant and shall be respected, as it represents moral values of a person stating it.

Three aspects of medical research will be discussed hereby:

1) Brief historical derivation to the ethical guidelines in medical research
2) Biobanking: definition, role, and guidelines of national and European biobanks
3) Fundamentals and tasks of ethics committees in research with biobank materials

2.2 Brief Historical Derivation to the Ethical Guidelines in Medical Research

Brigitte Lohff

Over the past decades it has become increasingly common that an approval by an operationally independent ethics commission must be obtained before starting any clinical study. Many clinical trial applications, which are submitted to the competent ethics commission, refer often to the Declaration of Helsinki. Indeed this declaration was adopted at the 18th General Assembly of the World Medical Association (WMA) in June 1964 in Helsinki, Finland. In the last 50 years it has been revised several times. The additions and adaptations aimed to adjust it to the current state of medical research [2]. The goal was to consider the development of research-based medicine and to unify the different national and international standards. Currently any clinical research must agree with the guidelines of the 7th revision of the Declaration of Helsinki, signed in Fortaleza, Brazil, in 2013 [3].

Meanwhile gaining an approval by an ethics commission to perform a clinical study has become standard. Yet, it is often unclear why an approval of the respective ethics commission needs to be obtained for studies involving human materials from biobanks and how long obtaining a permission to perform such studies takes place and why doctors have agreed worldwide in June 1964 to evaluate clinical research not solely based on scientific criteria but also on ethical criteria. This can be understood only when considering the recent history of tests on humans, having taken into account Germany in particular [4].

† Died 15 October 2016.

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

2.2.1 1900: Directive to the Head of the Hospitals, Polyclinics, and Other Hospitals

Since the end of the nineteenth century, a public controversy had begun whether a physician without any restriction and alone can decide on carrying a study with patients to satisfy his scientific interests or curiosity [5]. Initially medical doctors sought to find answers to such questions like the route of infection in infectious diseases or whether cancer can be transmitted by cells with oncogenic phenotype. On the one hand, the doctors' freedom allowed promising innovative experiments, such as with "immunization" of sick and healthy people. On the other hand, it resulted in many futile experiments, including those of the Breslauer dermatologists Albrecht Neisser, who infected children and prostitutes with syphilis, in order to test on them blood serum used previously against diphtheria.

In 1900 a collection of essays and newspaper articles appeared under the title "Poor people in the hospital" about the unimaginable inhuman experiments in hospitals. The subsequent public debate led to a first "Instruction to the head of the hospitals, polyclinics and other hospitals" (1900 12. 29.) by the Prussian Ministry of Culture [1]. The following rules were established:

- Medical investigations are permitted only for diagnostic and immunization purposes.
- Minors or not qualified for legal acts persons have to be excluded from experimental medical procedures.
- Patients who do not declare consent cannot be included.
- The same applies if no proper instructions on adverse consequences were given to the patients.
- Such interventions may be only performed by the head of a clinic or authorized by him/her.

2.2.2 1931: Guidelines for Novel Medical Treatments and Scientific Experimentation

The 1900 issued instructions were ineffective and not implemented by the medical professionals. The experiments with mentally and physically disabled children and socially disadvantaged and excluded people were not stopped. So finally, on February 28, 1931, the Reich Ministry of the Interior formulated "Guidelines for novel medical treatments and scientific experimentation." These guidelines were published in the widely read by medical doctors' journal *Deutsche Medizinische Wochenschrift* [12]. Since these guidelines were—as Gabriele Moser could prove—published in 1935, 1938, and 1942 in the so-called doctor's etiquette, they were supposed to apply over the period of Nazi dictatorship [8, p. 193]!

In these guidelines, a clear distinction between therapeutic trial and medical test was made for the first time. It was also made clear that medical tests are necessary. In addition, the directive from 1900 was supplemented with the following points:

- Tests must always be carried out according to the principles of medical ethics.
- New therapies must have been previously tested in animal experiments.
- There must be the consent of the person concerned.
- Social hardship must not be exploited.
- Records of the tests must be prepared.
- The publication of the results should not infringe upon the rights of the sick.
- Experiments on children and adolescents as well as the dying are not permitted.

2.2.3 1947: The Nuremberg Code

It is clear that the guidelines from 1931, which aimed to strengthen the rights of patients and to instruct on ethical standards on the part of the doctors, were not followed in the years 1933–1945; on the contrary these guidelines were repeatedly and intentionally violated [7]. The Nuremberg doctors' trials (October 25, 1946—August 20, 1947) revealed that various experiments were planned and carried out with inhuman cruelty by hundreds of doctors, nurses, and assistants at the concentration camps and children's and psychiatric hospitals. Parts of the verdicts of Nuremberg trials 1947 comprised criteria for "permitted medical experiments." This formed the basis of medical discourse in the postwar period.

The Nuremberg Code [9, 15] introduced, in addition to those named already in the former directives, new points to prevent the permanent violation of fundamental rights of the human subjects or patients:

- The necessity of a voluntary consent of the person before enrollment in the study.
- Clear definition of the nature, length, and purpose of the trial.
- All measures to avoid unnecessary physical and mental suffering/injury of the subjects must be undertaken.
- A subject can stop participating in the trial at any time.
- Tests shall be conducted only by scientifically trained staff.

2.2.4 1964: The Declaration of Helsinki

It took a few more years with occurrences of violations of human rights in the medical context until an international agreement was signed to adopt binding guidelines for the doctors and all personnel involved in biomedical research.

Looking back, the historical experience had shown that a commitment of the medical profession to abide by basic ethical values in research involving human subjects

is not an obvious matter [6]. The Declaration of Helsinki 1964 became the first internationally recognized regulatory framework with ethical principles for medical research involving human subjects [14]. Based on the Nuremberg Code of 1947, in 1964—in addition to the principles laid down already in the older ethical guidelines for research in medicine—new guidelines have been stated:

- Experiments should be terminated when the risk outweighs the benefit for research interests.
- Studies should not be performed if confirmed results are already available from other studies.
- Even if an informed consent of a person is there, the responsibility lies always with the doctor.

Until this declaration became part of the self-evident basis for all categories of people involved in medical research, it took some time and some revisions of these guidelines. In the first revision in 1975 of the Declaration of Helsinki in Tokyo, it was recommended and adopted that the approval of an independent ethics committees for clinical trial (§23) must be obtained before the start of the studies. In the course of a few years, this body has become gradually necessary for everyone involved in clinical trials.

2.2.5 The Declaration of Helsinki and Research on Human Materials and Data

Some of the established rules are of importance also in terms of research with materials from biobanks. At the General Assembly of WMA in October 2000 in Edinburgh, it was decided that:

- Experiments with persons unable to provide consent may only take place if there are no other adequate subjects with "informed consent" and the research would bring benefits to the affected group (§30) [16].
- It is necessary for the progress of the research and also for the protection of patients that negative results are published (§36); the freedom of the researcher to publish the results is of higher value than the economic interests of contracting entity.

The continuously improved guidelines were expanded to research on identifiable human material and data (§25) in 2008. It was fixed from the perspective of the WMA that the doctor/researcher must inform patients completely on the research-related aspects of the treatment. This applied also the use and the further handling of human biomaterials and the resulting data. The refusal of a patient to participate in a study or the decision of the patient to withdraw from the study must never adversely affect the relationship between patient and physician (§31).

2.2.6 2013: Current Valid Declaration of Helsinki in the 7th Revision

The new version from 2013 for the first time includes binding regulations in respect to biobanks:

- §32 specifies that the same principles, as the ones used for the collection of materials and data from human subjects, should apply also to human materials from biobanks.
- Patients must give their informed consent for "collection, storage, and/or reuse."
- §32 provides further that the informed consent can be avoided, if impossible or impractical to obtain.
- In these situations, it requires the consent of an ethics commission!
- §35 requires registration of all studies involving humans or human material from biobanks, not only—as previously—clinical trials.

There are good reasons for significant extension of registration of studies—which also applies to nonclinical studies. An important argument is that unnecessary research and thus unnecessary risks to study participants can be avoided by this registration.

All individuals involved in the research should regard it as something meaningful and necessary in order to adhere to the guidelines for biomedical research derived from the Declaration of Helsinki.

Research today is based on the implementation of the Declaration of Helsinki at the European level and the respective national legal regulations. After long discussions, the principle "a clinical trial cannot be carried out without an affirmative vote of an ethics committee" stated by the European guidelines for the ethics committee from 2013 applies also to research with biobank materials [17]. This principle should apply to multicenter studies, that is, studies where several centers from different countries are involved.

References

1 Anweisung an die Vorsteher der Kliniken, Polikliniken und sonstigen Krankenanstalten. Centralblatt für die gesamte Unterrichts-Verwaltung in Preußen 2 (1901) Hrsg. von dem Ministerium der geistlichen, Unterrichts- und Medizinalangelegenheiten. Berlin: Cotta, S. 188–189. http://goobiweb.bbf.dipf.de/viewer/image/ZDB985843438_0043/1/ (accessed August 26, 2017).

2 Declaration of Helsinki 1964–2013. https://en.wikipedia.org/wiki/Declaration_of_Helsinki (accessed August 26, 2017).
3 Declaration of Helsinki 7. Revision 2013. https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/ (accessed October 16, 2017).
4 Eckhart, W.U.: *Man, Medicine, and the State—The Human Body as an Object of Government Sponsored Medical Research in the 20th Century*, Stuttgart: Steiner, 2006.
5 Elkeles, B.: Medizinische Menschenversuche gegen Ende des 19. Jahrhunderts und der Fall Neisser. Kritik und Rechtfertigung einer wissenschaftlichen Methode. *Medizinhistorisches Journal* 20 (1985): 135–148.
6 Frewer, A.; Schmidt, U. (Hrsg.): Standards der Forschung Historische Entwicklung und ethische Grundlagen klinischer Studien Reihe: Klinische Ethik. Biomedizin in Forschung und Praxis/Clinical Ethics Biomedicine in Research and Practice—Band 1. Frankfurt: Peter Lang, 2007.
7 Mitscherlich, A.; Mielke, F. (Eds.): Medizin ohne Menschlichkeit: Dokumente des Nürnberger Ärzteprozesses 18. Aufl. 2012.
8 Moser, G.: Die Deutsche Forschungsgemeinschaft und die Krebsforschung 1920-1970 (Studien zur Geschichte der Deutschen Forschungsgemeinschaft 7). Stuttgart: Steiner, 2011.
9 Nürnberger Kodex. http://www.ippnw-nuernberg.de/aktivitaet2_1.html (accessed August 26, 2017).
10 Percival, T.: *Medical Ethics a Code of Institutes and Precepts*, 3rd Edition. Chicago: American Medical Association Press, 1847.
11 Reuland A.J.: Menschenversuche in der Weimarer Republik. Books on Demand GmbH, Norderstedt, 2004.
12 Richtlinien für neuartige Heilbehandlungen und wissenschaftliche Versuche. DMW, 1931, S. 509. http://dg-pflegewissenschaft.de/wp-content/uploads/2017/05/ForschungsrichtlinienReichsinnenministeriums.pdf (accessed October 16, 2017).
13 Statement Council of Europe, Steering Committee on Bioethics. http://www.coe.int/en/web/bioethics/home (accessed August 26, 2017).
14 Weindling, P.J.: *Nazi Medicine and the Nuremberg Trials—From Medical War Crimes to In-formed Consent*, Basingstoke: Palgrave Macmillan, 2004.
15 Weindling, P.J.: The Origins of Informed Consent: The International Scientific Commission on Medical War Crimes, and the Nuremberg Code. *Bulletin of the History of Medicine* 75.1 (2001): 37–71.
16 Wenz, V.: *Forschung mit einwilligungsunfähigen Personen aus der Perspektive des deut-schen und des englischen Rechtes*. Göttingen: Cuvillier, 2006.
17 Ethik-Kommissionen bleiben Pflicht. http://www.pharmazeutische-zeitung.de/index.php?id=50248 (accessed August 26, 2017).

2.3 Biobanking: Definition, Role, and Guidelines of National and International Biobanks

Thomas Illig

2.3.1 Introduction

Biobanks were presented in a 2009 issue of the American *Time Magazine* as one of ten ideas that can substantially change the world [1]. This illustrates the enormous potential of biobanks in the future to impact medical research, as well as diagnostic and therapeutic approaches, especially taking into consideration the significant development of molecular methods that enable analysis of thousands of molecules in parallel (OMICS approaches). Understanding the molecular and environmental foundations of human diseases, in order to improve diagnosis and treatment, is a top priority both for biomedical research and for society. A very important prerequisite for this project is the development of infrastructures, which are crucial for biomedical research. One of these central infrastructures is biobanks, which store biomaterials and associated data. Therefore, they form the basis for a large part of biomedical research.

New methods for biomarkers and therapeutic research offer a great potential for individual preventative and therapeutic measures (personalized medicine). In recent years, new common DNA risk variants for numerous widespread diseases have been discovered and could be at least partially functionally characterized. Currently, mutations for rare diseases are being identified by exome and whole-genome sequencing approaches. The metabolome and proteome are investigated in order to discover novel biomarkers. Due to these new partially very sensitive molecular analyses, there is a necessity for large studies with high number of individuals and improved cooperation between existing biobanks that allow smooth sample and data exchange. The lack of sufficient high-quality biomaterials within the Cancer Genome Atlas project constitutes a serious problem [2]. There are significant efforts in several countries to standardize biobanks (e.g., International Organization for Standardization (ISO) 9001, OECD Good Practices) and to implement policies for biomaterials used in research. The German Institute for Standardization (DIN) has in

the ISO successfully sought to provide internationally accepted standards for the accreditation and certification of biobanks. The Technical Committee (ISO TC 276) has already begun its work to create an international standard (IS) since 2016.

In recent years biobank projects in the focus of public attention are those, which have been most comprehensively established (have gathered particularly large number of data and samples from subjects or donors) or are focusing on genetic issues, with emphasis on future predictions with probability of disease occurrence and/or particularly wide-ranging (new) formulated medical uses as targets. Such biobanks serve mainly as a basis for research on common diseases. With the development, the construction, and operation of biobanks, a variety of questions arise, which are mainly relative to the collection, storage, use, and transfer of samples and data, as well as the social aspects of these processes. In numerous countries (including Germany) and institutions, biobanks are attracting increasing attention from politics, science, economics, and advisory commissions.

In addition to the definition of human biobanks, various types of human biobanks, the quality of the samples, IT aspects, type of harmonization efforts within biobanks, cooperation and the access rights, and financial aspects and sustainability of biobanks will be described in this chapter.

Legal privacy and ethical factors also play a very important role in biobanks. It cannot be ruled out with the new sequencing methods (next-generation sequencing (NGS)) that persons can be clearly identified solely on the basis of genetic information. This part will be discussed in a separate section of this chapter.

2.3.2 Definition of Biobanks

The word "biobank" is only a little less than two decades old [3] and has recently been defined by the OECD [4] as "a collection of biological material and associated data and information." However, the definition remains controversial.

There is a consensus that the term biobank refers to biological collections of human, animal, plant, or microbial samples and that associated sample data must be available. Moreover, biobanks must operate according to professional standards. But there is so far no consensus whether the purpose of the collection, size, or access rights determine the concept of a biobank. It is probably appropriate that a general, broad definition of a biobank will be accepted, and the next step will be to pay attention to a widely accepted universal classification of different types of biobanks (organism, type of material (tissue, liquid, clinical, epidemiological)) [5]. The establishment and operation of biobanks in Germany are not subjected to a general approval requirement. It is also stated by the National Ethics Council that "The collection and use of human bodily substances and personal data is part of the normal medical research. Usually it harbours no special risks for donors and is recognized by the established standards of medical research. Therefore it does not require widespread official preliminary control."

2.3.3 Human Biobank Types

In Germany, as well as in many European and non-European countries, there are large numbers of biobanks, each with different characteristics regarding their organizational and legal form and their research practice. Biobanks may be designed either for a specific disease or without disease focus, take into account environmental factors, store liquid or tissue samples, be population representative or not, and include samples from children or adults.

Biobanks have been established by research institutions, pharmaceutical industrial enterprises, commercial companies (e.g., Vita 34, Indivumed), and other authorities. Clinical biobanks had in the past often a disease focus. Nearly every university hospital in the Western world has established several different biomaterial collections. There are significant differences in quality, size, and other important variables between biobanks. In recent years, many universities have decided to establish central biobanks in order to increase the quality of biomaterial collection and adapt to the IS. So in 2011 five (Aachen, Berlin, Heidelberg, Kiel, and Würzburg) central biomaterials banks (cBMBs) were selected by the Federal Ministry of Education and Research (BMBF)-funded initiative and were provided for 5 years with a total budget of €18 million. A central objective of the BMBF was that these five cBMBs provide biomaterials and clinical data for scientists outside the respective university. Central biobanks have been established also in other universities (e.g., Hannover, Jena, Leipzig, Mannheim, and Munich). Many other medical faculties in Germany also plan to set up central biobanks.

2.3.4 Clinical Biobanks

Different biomaterials are stored in clinical biobanks. In addition to the tissues, body fluids (e.g., blood and blood derivatives, urine, lung lavage, cerebrospinal fluid, saliva), stool, hair, or various swabs (e.g., skin, mouth, and nose) are collected and stored. Hannover Unified Biobank (HUB) of Hannover Medical School (MHH) is an example of a clinical centralized biobank. Scientific and clinical projects at the MHH generate large amounts of samples, including tissue, blood, living cells, cell cultures, urine,

swabs, stool, bile, bronchoalveolar lavage (BAL), and other biological samples. To ensure maximum comparability of the samples, it is necessary to ensure maximum harmonization in the pre-analysis, archiving, and publication of the samples. Therefore, the samples are in accordance with strict instructions processed and stored (standard operation procedures (SOPs)). In order to achieve an optimization of the sample quality, the HUB works with the following principles:

- Accurate, detailed, and harmonized SOPs to the highest quality of biomaterials
- High degree of automation in the pre-analysis, storage, and retrieval of samples and complete sample tracking through laboratory and biobank information management systems (LIMS/BIMS) to exclude contamination or incorrect labeling of samples
- Storage of most samples in the gas phase of liquid nitrogen in order to ensure high quality of biological samples even after prolonged storage
- Pseudonymization of samples and high data security to protect patients
- Harmonization of patient consent forms
- Further development of the biobank software/database and link with relevant clinical data

The transfer of all sample collections from MHH to the central biobank HUB is planned.

HUB acts as administrator of the samples; however the access rights remain with the principal investigators (PIs) of the studies, the so-called gatekeepers. The storage, management, and release of the samples will be provided to biobank users invoiced according to a defined cost key. The relevant data for the biomaterials are managed by a biobank software. HUB manages and supports different types of samples (e.g., tissue, cells, cell lines, microorganisms, body fluids). It is possible to apply for biomaterials research projects. However, a prerequisite for obtaining samples is the agreement of an access committee and of the gatekeeper.

The gatekeepers can request "their own" samples at any time without the consent of the access committee. In addition to the storage of biomaterials, the HUB services provide DNA and RNA isolations.

2.3.5 Governance in HUB

The gatekeepers choose a governing body (steering committee), consisting of several members of the MHH. The steering committee may establish working groups to support the HUB staff with scientific expertise in the SOP development, ethical issues, and data security concepts. It shall appoint an access committee that verifies the sample and data for their scientific merit and feasibility and recommends certain sample release. The external scientific advisory board of the HUB consists of national and international experts (e.g., data security experts, experts in bioethics, experts in quality of biomaterials). Other examples of large clinical biobanks are EuroBioBank (rare diseases), Biobank Graz University (mainly pathological samples), Biobank of Pathology, Charité (mainly pathological samples), and many more.

2.3.6 Epidemiological Biobanks

Some European countries have or will establish very large epidemiological biobanks. The first recruitment phase has recently been completed in England. The UK Biobank comprises 500 000 adult participants [6]. Similar activities have been initiated in Germany. The National Cohort was initiated by a network of German research institutions of the Helmholtz Association, numerous universities, the Leibniz Association, and other partners. The goal is to build a large-scale long-term population study focused on the causes of common diseases such as cardiovascular disease, cancer, diabetes, and dementia and infectious diseases in order to identify risk factors, discover effective preventive actions, and develop approaches for early detection of diseases. In this (cohort) study of 200 000 people, aged 20–69 years, participants from all the regions of Germany will be medically examined and interviewed about life habits (e.g., physical activity, smoking, diet, occupation). In addition, the blood samples and other biomaterials from participants will be stored at a central "biorepository," which is located at the Helmholtz Centre in Munich (HMGU). After 5 years all participants will be invited again for an examination and second survey in the 18 study centers. Over 10–20 years, some participants will naturally experience diseases, which can then be correlated with the data collected. The study thus offers a unique potential for a variety of scientific studies. The researchers will obtain information on genetic factors, environmental conditions, social environment, and lifestyle that are all interconnected and play a role in the development of diseases. Based on this information the researchers can develop strategies for improved prevention and treatment of major public diseases [7]. The UK Biobank and the National Cohort were confronted with ethical and data protection issues. In particular, guidelines or standards were formulated in regard to participation, consent, confidentiality, data and sample access and ownership, access policy, management of biobanks, accountability, dissemination, and patenting. Both biobanks agree that there should be regular information on scientific results through newsletters, websites, helplines, and events. Scientists who want to use the database are invited to publish all data, both positive and negative. The publication should go through a peer

review process. The participants should be tracked over a longer period in both studies. Such very large, long-running studies with well-characterized phenotypes offer numerous advantages. Since the cohorts are established prospectively, predictive biomarkers can be discovered and in the future employed to identify high-risk persons. Furthermore, subtle molecular signatures associated with diseases can be uncovered. These molecular features often characterize lifestyle diseases, such as type 2 diabetes, heart disease, asthma, and Alzheimer's disease. Additionally, due to the enormous size of the studies, the specific environment and lifestyle factors can be examined. However, large epidemiological studies generally collect only body fluids, smears, stools, and no tissues. Other major epidemiological biobank projects are European Prospective Investigation into Cancer and Nutrition (EPIC), MORGAM (cardiovascular disease), National Biobank Estonia, Iceland Biobank, BioBank Japan, CARTaGENE (Canada), and SAPALDIA (Switzerland), to name just a few.

2.3.7 Quality of Samples

Biomaterials form the basis for biomedical research. They can be significantly altered during the extraction, processing, long-term storage, and retrieval. It is therefore necessary to optimize these processes and standardize them among different biobanks by establishing SOPs. It is also important in the process of sample collection, processing, and storage to document every step. In the processing (pre-analytics) of whole blood to serum (obtained from clotted blood) or plasma (obtained from blood treated with anticoagulant substances), many factors play an important role. Samples should be taken, whenever possible, always in the same position (seated). During further processing the cellular blood constituents shall be quickly separated from the plasma or serum (within hours). In tissue samples, ischemia plays a central role for the sample quality. The word ischemia refers to the time period in which a transplanted organ or tissue is cut from the normal blood supply and is therefore no longer supplied with oxygen. A distinction is made between cold ischemia time, in which the blood-free organ is kept on ice, and warm ischemia, in which the organ remains at room temperature, but is not supplied with blood. The plasma, serum, or tissue should then be frozen as quickly as possible at −80°C or in the gas phase of liquid nitrogen (within minutes or hours). The long-term storage should be at −80°C or colder. Recently, fully automated freeze and sorting robots are available for this purpose, which leads to a significant improvement in the quality of the samples.

The pre-analytical variability of the sample preparation causes substantial changes of various molecules such as peptides, proteins, metabolites, or enzymes [8, 9]. These modifications may not only affect [10] diagnostic tests but also have an effect on multiparametric tests (omics) and can interfere with biomarker identification for diseases or even make it impossible for a meaningful interpretation of data [11, 12]. If the quality of the samples is unclear, there are certain markers that can indicate the quality of the samples. There are different quality markers for different tissues or body fluids [9, 13, 14]. The control of pre-analytical variables, however, is highly complex because it is dependent not only on the influence of the sample quality and on the class of the biomolecules (DNA, RNA, protein, peptide, metabolite) but also on the nature of the analytical method and its specificity, sensitivity, and robustness.

Additional drawbacks are the natural variations in markers within a study group, which can be difficult to distinguish from variations caused by the quality of the biomaterial. Thus, one should attempt to maintain the optimal biomaterial quality. Yet, this turns out quite often to be extremely difficult in clinical settings. The samples should be tested with quality markers. Recently, a method has been presented, in which an addition of a synthetic peptide reporter to blood samples is followed by relative quantification of the produced proteolytic peptide fragments. Thus, pre-analytical variability can be recorded. The method enables assessment of serum and plasma quality. This method, however, requires availability of synthetic reporter peptides, in a laboratory involved in a clinical or epidemiological study, and is associated with huge additional effort [9]. In the future, blood tubes could be preloaded by manufacturers with these peptide markers.

2.3.8 Harmonization and Cooperation of Biobanks

The collaboration among biobanks is becoming increasingly important because initial research results obtained using samples from a biobank must be validated in an independent set of samples obtained from another biobank in order to verify the accuracy, validity, and transferability of the results. Furthermore, the funding bodies in medical research support large multicentric projects. For better comparability of the results among different biobanks, similar standard procedures should be followed. In addition, the following factors require harmonization:

- The analytical molecular high-throughput technologies
- The procedures of data and sample collection
- The rules for the protection and safeguarding of the rights of patients and volunteers

For this reason, various national and international consortia have been formed to define those standards and to facilitate data exchange between biobanks.

2.3.9 Situation in Germany

In Germany, the Working Group Biomaterial has taken a leading role in the harmonization and standardization of biobanks and biobank processes in the frame of Technology, Methods, and Infrastructure for Networked Medical Research (TMF). It has been already recognized early within the TMF that the harmonization and standardization of biobanks is a key challenge for the networked medical research and thus an important common field of action. After the first inventory from 2003/2004 to 2005, a large-scale project to clarify the legal, ethical, and organizational framework for the establishment and operation has been carried out. Action guidelines and templates are available in the form of advice for the researchers on various aspects:

- Categorization and modeling of biobank projects
- Regulatory frameworks
- Data protection concepts
- Patient consent forms
- Quality assurance of biomaterial

The medical collaborative research projects go beyond national borders. The Working Group Biomaterial within the TMF has therefore launched a project to clarify rules for the disclosure of materials from German donors to cooperating biobanks—the legal basis for European partner countries. Another current project aims to create a German register of medical biobanks.

In another research project high-dimensional molecular data are linked to clinical data from patients, type of studies, and centers. Under the umbrella of the TMF, the researchers are working together to ensure the quality control of molecular data on the stages of the production, interpretation, storage, and validation.

The specific data protection requirements, which arise in connection with the storage, management, and dissemination of molecular, in particular genetic data, are the subject of cross-border activities [15]. The German Biobank Registry (DBR) was established in the frame of TMF, in which all German biobanks can register. Thus, the DBR provides good overview of the majority of biobanks in Germany and about what materials and data they contain [16].

2.3.10 Situation in Europe and Worldwide

Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) has been promoted as one of the first EU infrastructure projects by the European Commission. Since 2011 it has grown to 54 members from over 30 countries, with more than 225 associated organizations, mainly biobanks. Thus it represents one of the largest EU-funded infrastructure projects. This project aims to build a coordinated, large-scale pan-European infrastructure for biobanks to improve the treatment and prevention of common and rare diseases [17]. In addition to instructions for the harmonization and standardization of biobanks, a register has been set up, in which more than 200 biobanks across Europe are already registered.

Under the 6th EU Research Framework Program, a cooperation project of the P3G consortium leading 18 European and Canadian research institutions had been funded. The project entitled "Harmonising population-based biobanks and cohort studies to strengthen the foundation of European biomedical science in the post-genome era" has the following objectives [18]:

- To categorize population-based biobanks and cohort studies in Europe systematically. Particular attention should therefore fall on studies that contribute to research on the genetic and environmental causes of complex diseases.
- To identify genetically isolated populations with special consideration of new possibilities for the construction of biobanks in Europe.
- To establish standardized criteria for the selection and collection of sample collections in genetically isolated populations.
- To build an infrastructure for the exchange of methods of genotyping in large cohorts.
- To prepare a communication forum and to discuss the selection of markers, quality control, database structure, and analysis.
- To develop a standard for the determination of complex phenotypes and lifestyle factors.
- To work on the solution of the statistical–methodological problems with the study design and analysis and the merging of data from different studies. An expert platform will be built to develop mathematical models integrating genetic epidemiology and statistics.

Apart from the aforementioned, there are many other projects that deal with the harmonization of biobanks, such as regional and international organizations (IARC, FIBO, ESBB, ISBER), further research and infrastructure initiatives (caHUB, caBIG, Biomarkers Consortium, BioSHaRE-EU, HuGENet, PHOEBE), and Internet databases (dbGaP, datSHaPER, DataShield, HapMap, HumGen, OBiBa, and OBO).

2.3.11 Definition of Ownership, Access Rights, and Governance of Biobanks

The access rights are different among biobanks. While certain biobanks do not even provide insight into the existing samples and data—not to mention to provide material for the research of other research groups—there are efforts in Germany to make those biomaterials and their associated data accessible for scientists.

Prerequisites therefore are as follows:

- A request of the potential sample user, which describes the planned project, the scientific background, and the number and type of samples and data (material and data transfer agreement) required
- The patients/volunteers' informed consent
- Permission of the access committee and biobank PIs

The "gatekeeper" (scientist/clinician) has a right of veto over the samples or data transfer.

Assessing the compliance of cBMBs, a prerequisite of the funding agencies was that the biomaterials, which will be stored, will be available for the broad scientific community. A similar approach is envisaged in the National Cohort. Currently it is discussed how to deal with biomaterials, which are very valuable and limited in amount (e.g., fresh tissues, cells, serum, plasma). A final consensus could not be achieved here. Furthermore, a patient consent text was drafted recently by the working group of medical ethics committees that gives the consent for the broadest possible medical research and ensures that as many researchers as possible can obtain access to these data and samples.

This so-called broad consent or open consent is currently being discussed or implemented in several university hospitals or in large research associations (German Centres for Health Research) or large epidemiological studies (National Cohort).

The ownership of biomaterial samples in a medical context is legally not uncontroversial (at least when there is no explicit transfer of ownership from the donor to the collector of the samples or the biobank). Also it is important to distinguish between those samples that fall in context of pure treatment and those that are already procured for the purpose of research. The former enter the laboratory (pathology) for treating or providing diagnosis, where they are held and used for the institution's own research without the need for consent. If samples are intended to be used in research, the patient or clinical trial participant consent is required; but this refers to date in most cases only on the right of use for research, but not on the transfer of property.

2.3.12 IT in Biobanks

The value of biobanks increases enormously by the presence of a professional, high-quality IT system. The IT system, by connecting the sample data with clinical or laboratory data, has to cope with various tasks that range from the "consent management" (management of informed consent), specimen collection, processing, storage, and distribution to quality control [19]. Factors like data security, data access, and reporting play a central role. Different biobank information management systems, such as BIMS or LIMS, were developed for this purpose from different companies or by biobanks.

These systems play an absolutely critical role in monitoring the sample quality for different processes of sample collection, sample processing, storage, and utilization. To ensure complete documentation of samples, all equipment involved in the sample processing and storage must be integrated into the BIMS interfaces. The use of 2D barcodes is recommended during sample handling and reduces the mismanagement of samples dramatically. In addition, such management systems can cope with the huge amounts of data that can be generated with the samples. This is of increasing importance, facing the availability of the new "omics" approaches, such as the whole-genome sequencing. A well-functioning IT system is able to integrate large amounts of data from different sources (clinical data, sample quality, sequencing, and other laboratory data). It is increasingly important to harmonize database structures and use standard formats for exchange between biobanks and different institutions, as well as for joint analysis [20]. Another important aspect is the IT data protection compliance. In Germany a generic data protection concept for biomaterial was developed by TMF e.V. that serves as the basis for numerous data protection concepts [15].

2.3.13 Financial Aspects and Sustainability

The establishment and maintenance of biobanks is very costly. Main costs are due to staff, infrastructure (automated freeze storage robots, manual liquid nitrogen tanks, manual −80°C storage, and transport vessels), biobank IT structure (BIMS, LIMS), and consumables (2D barcode tubes, electricity, and liquid nitrogen). Biobanks are often set up without elaborate long-term plans for operational sustainability [21, 22]. Whereas a large part of the costs is associated with building the entire infrastructure at the beginning, significant costs arise from sample storage and establishment of procedures allowing access to the data and to the biological samples. Currently, full-cost models are created for different biobanks but also at the level of the TMF. Specific prices for services by biobanks are not yet published. Different models are pursued for biobank operations, ranging from institutional support to third-party funding and user fees. Most of the time, mixed financing takes place [22].

2.3.14 Conclusion

In the age of personalized medicine, high-quality biobanks can support the medical/molecular biological

research and operate to optimize the personal protection of the sample donor. Studies using biobank materials and data can help to better understand diseases, to optimize biomarker discovery and validation, and to identify new treatment options, including tailored therapies specific for a patient group.

Currently, large, harmonized, professional, and modern biobanks are created in many areas. Numerous universities are in the process of setting up centralized biobanks. Even larger multicenter projects of biobanks in Germany harmonize their activities, as it is done in the whole world. Despite these positive developments, there are still many tasks to be solved, such as the sustainable funding of biobanks, the challenge of different legal systems for biobanks in cooperating countries, and tightening of data protection in Europe, which complicates the cooperation between biobanks. Overall, however, most funding agencies came to the conclusion that clinical projects involving biomaterials should be supported only if professional biobanks are involved. It is time to integrate the numerous small, less organized biobanks in larger structures to obtain reliable results in biomedical research in the future.

References

1 Park, A. Biobanks. TIME. 2009;173(12). http://content.time.com/time/specials/packages/article/0,28804,1884779_1884782_1884766,00.html (accessed October 16, 2017).
2 Waltz, E. Pricey cancer genome project struggles with sample shortage. *Nat Med*. 2007;13(4):391.
3 Loft, S., Poulsen, H.E. Cancer risk and oxidative DNA damage in man. *J Mol Med (Berl)*. 1996;74(6):297–312.
4 Creation and Governance of Human Genetic Research Databases. Paris: OECD publishing, October 25, 2006.
5 Hewitt, R., Watson, P. Defining biobank. *Biopreserv Biobank*. 2013;11(5):309–315.
6 Ollier, W., Sprosen, T., Peakman, T. UK biobank: from concept to reality. *Pharmacogenomics*. 2005;6(6):639–646.
7 Gemeinsam forschen für eine gesündere Zukunft. Www.nationale-kohorte.de (accessed August 26, 2017).
8 Yi, J., Liu, Z., Gelfand, C.A., Craft, D. Investigation of peptide biomarker stability in plasma samples using time-course MS analysis. *Methods Mol Biol.* 2011;728:161–175.
9 Findeisen, P., Sismanidis, D., Riedl, M., Costina, V., Neumaier, M. Preanalytical impact of sample handling on proteome profiling experiments with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Clin Chem*. 2005;51(12):2409–2411.
10 Guder, W.G. History of the preanalytical phase: a personal view. *Biochem Med (Zagreb)*. 2014;24(1):25–30.
11 Karsan, A., Eigl, B.J., Flibotte, S., et al. Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis. *Clin Chem*. 2005;51(8):1525–1528.
12 McLerran, D., Grizzle, W.E., Feng, Z., et al. Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias. *Clin Chem*. 2008;54(1):44–52.
13 Betsou, F., Gunter, E., Clements, J., et al. Identification of evidence-based biospecimen quality-control tools: a report of the international society for biological and environmental repositories (IS-BER) biospecimen science working group. *J Mol Diagn*. 2013;15(1):3–16.
14 Govorukhina, N.I., de Vries, M., Reijmers, T.H., Horvatovich, P., van der Zee, A.G., Bischoff, R. Influence of clotting time on the protein composition of serum samples based on LC-MS data. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2009;877(13):1281–1291.
15 Pommerening, K., Drepper, J., Helbing, K., Ganslandt, T. Leitfaden zum Datenschutz in medizinischen Forschungsprojekten Generische Lösungen der TMF 2.0. TMF. 2014. http://www.tmf-ev.de/EnglishSite/TMFBookSeries.aspx (accessed August 26, 2017).
16 http://dbr.biobanken.de/en/bdb (accessed August 26, 2017).
17 www.bbmri-eric.eu (accessed August 26, 2017).
18 www.p3gconsortium.org (accessed August 26, 2017).
19 National cancer institute best practices for biospecimen resources, 2011. https://biospecimens.cancer.gov/bestpractices/2011-ncibestpractices.pdf (accessed October 16, 2017).
20 Olson, J.E., Bielinski, S.J., Ryu, E., et al. Biobanks and personalized medicine. *Clin Genet*. 2014;86(1):50–55.
21 Hewitt, R.E. Biobanking: the foundation of personalized medicine. *Curr Opin Oncol*. 2011;23(1):112–119.
22 Riegman, P.H., Morente, M.M., Betsou, F., de Blasio, P., Geary, P. Marble Arch International Working Group on Biobanking for Biomedical Research. Biobanking for better healthcare. *Mol Oncol*. 2008;2(3):213–222.

2.4 Tasks of Ethics Committees in Research with Biobank Materials

Dieter Tröger

2.4.1 General Basic Concept

The ethics committee is an independent body made up of healthcare professionals and those involved in nonmedical fields. It basically ensures that the ethics committee independent judges are not bound by instructions from the underlying organization (z. B. University, faculty). The tasks of the ethics committee are as follows:

- To protect the rights, safety, and well-being of individuals participating in a clinical trial
- To establish the confidence of the public in medical research involving human subjects
- To evaluate a medical research project according to an ethically legal point of view
- To provide opinions on the inspection/quality control plan, the researchers qualifications, the suitability of the facilities, and the methods and information material by which research participants were instructed and informed to get their consent

In addition to the protection of the patients and study participants, the commission is also responsible to prevent scientists, as well as institutions (e.g., universities, hospitals), from unjustified and dangerous research.

To protect the patients, the ethics commission evaluates if the project is medically justifiable, regarding the risks for the patient on the one hand and the significance for medical purposes on the other hand. Nevertheless, the benefits for the patient and his/her safety always outweigh the latter, which means that the well-being of the patient always comes first and against possible benefits for society and state.

In Germany, the objective of an ethics commission under public law is put down in its respective statute. In the statute of the ethics commission of the medical university of Hannover (MHH), for example, it is agreed that:

1) The commission has to consist of not less than seven members, of which at least four should be doctors and one a lawyer with the qualification of judgeship. Two doctors should be experienced interns, one pediatrist and one from the field of theoretical medicine. Also one of the members should be acquainted with statistics and test planning.
2) In the statutes it is also generally stated that official experts and specialists could be consulted for advice.

2.4.1.1 The Application Procedure

1) The ethics commission must decide promptly about an application, because researchers and employers have the right to receive a decision within 60 days.
2) The ethics commission has a quorum when at least half of the members contribute to the decision. The decision is taken by the simple majority.
3) The commission can either accept or turn down the application.
4) A rejection is only valid after an official hearing with the applicant and can be withdrawn if prearranged requirements have been fulfilled.
5) The committee should follow the rules imposed by the laws of each country. For example, in the case of applications that are subject to the German Pharmaceutical Act (AMG) or the Medical Devices Act (MPG), the ethics commission is requested to follow §42 of the AMG and §22 of the MPG and therefore enacted acts (GCP-V, MPG-V).

In general, every doctor/researcher, planning to perform research on humans or human material, is obligated, according to both the Declaration of Helsinki of 2013 and the respective code of medical ethics of the responsible medical association (e.g., Landesärztekammer), to consult the ethics commission about his intent and discuss the ethical and legal questions involved.

2.4.2 About the Respective Ethics Commissions

For doctors, as well as scientists, who are employed at universities or teaching hospitals, the ethics commission of the university is responsible. For all other doctors the ethics commission of the "Landesärztekammer" (State Medical Council) is responsible of this task.

All ethics commissions in Germany impose a fee in order to cover the costs associated with their meetings. The amount is preassigned to the respective statutes and sometimes differs greatly from one to another, which is critically perceived by the clinics and the pharmaceutical industry. There are particular rules of ethics commissions, for example, in MHH, where only applications that follow the AMG and MPG require a fee. Every other application is processed without a fee.

Multicenter studies also fall under the act about the exercise of the "Gute Klinische Praxis" in case of clinical examinations with medicaments used for humans (= GCP-V), which controls the advisory process by the regional ethics commissions (§8 Abs. 5 GCP-V).

The ethics commission in charge evaluates the clinical examination together with other ethics commissions who are responsible for their researchers. They verify the qualifications of the researchers and the research facilities

in their jurisdiction. Their evaluation must be passed to the ethics commission in charge within 30 days after successful application submission.

The decision of the ethics commission will be conveyed to the applicant in writing, in the form of a so-called "Votum," where it is written, if no ethical or legal concerns were found against the project or if there have to be made changes to the project.

In the case of non-AMG and non-MPG applications, the applicant is not forced by law to follow the "Votum" or the recommended changes, because the ethics commission only fulfills an advisory function, but if the applicant still decides to pursue his project and it fails, he could be prosecuted by civil law.

When it comes to applications for the ethics commission that fall under the AMG or MPG and include the collection or use of assembled biomaterials (e.g., blood, urine, saliva, tissue, liquor), these have to be dealt with as §42 AMG and §22 MPG dictate.

For Non-AMG or MPG studies with up-to-date, project-based biomaterial collections, an application is needed in which the project and possible ethical problems are briefly described. The specifications on what evidence has to be provided depends on the respective ethics commission. They can be normally found on their homepage.

Regarding the collection of biomaterials, it has to be settled:

- Whether the materials will be destroyed after the project is finished
- Or if they should be used for a latter project from the same or different field
- Or if they should be given to cooperating institutes
- Or if a genetic examination or even a whole-genome sequencing had been performed or has been planned.

When writing a report, it is recommended to use the specimen description for the patient information and the declaration of consent for adults, as well as youths and children.

2.4.3 The Establishment of Biobanks

When it comes to the special case of establishing a clinical biobank, which follows a professional standard like the "Hannover Unified Biobank" (HUB), a detailed application is necessary and should especially include information about the quality of storage and the organizational structure. To evaluate an application for the establishment of a biobank, the ethics commission needs in particular

- Specified information about the purpose, the organization, the courses of procedure, the documentation, and the financing concept of the biobank
- Information on the biomaterial and data, for example, acquisition, storage, quality assurance, utilization, and safeguarding
- Documents about the donors and their declaration of consent

In conclusion it is essential to point out that researchers and doctors are strictly obliged by the code of medical ethics to inform the respective ethics commission before beginning a project. This rule should be applied if the used biomaterial originates from a small internal collection or from a big clinical biobank like HUB.

Further Reading

Act on the Transport of Medicinal Products. http://www.gesetze-im-internet.de/amg_1976/BJNR024480976.html (accessed August 26, 2017).

Bundesrecht konsolidiert: Gesamte Rechtsvorschrift für Medizinproduktegesetz, Fassung vom 26.08.2017. https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10011003 (accessed August 26, 2017).

Ethics committee Hannover Medical School. https://www.mh-hannover.de/3371.html (accessed August 26, 2017).

German Medical Association. http://www.bundesaerztekammer.de/weitere-sprachen/english/german-medical-association/ (accessed August 26, 2017).

Gesetz über den Verkehr mit Arzneimitteln (Arzneimittelgesetz—AMG) § 42 Verfahren bei der Ethik-Kommission Genehmigungsverfahren bei der Bundesoberbehörde. http://www.gesetze-im-internet.de/amg_1976/__42.html (accessed August 26, 2017).

Hrsg. Kiehntopf, M., Böer, K. Biomaterialbanken—Checkliste zur Qualitätssicherung. Berlin: Medizinisch Wissenschaftliche Verlagsgesellschaft. TMF Band 5, 2008.

Hrsg. Deutsch, E., Lippert, H.-D. Kommentar zum Medizinproduktegesetz (MPG). Berlin: Springer, 2. Auflage, 2010a.

Hrsg. Deutsch, E., Lippert, H.-D. Kommentar zum Arzneimittelgesetz (AMG). Berlin: Springer, 3. Auflage, 2010b.

Hrsg. Schwarz, J.A., Juhl, G., Koch, A., Sickmüller, B. Leitfaden Klinische Prüfungen von Arzneimitteln und Medizinprodukten. Aulendorf ECV: Editio Cantor Verlag, 4. Auflage, 2011.

Hrsg. Deutsch, E., Spickhoff, A. Medizinrecht. Arztrecht, Arzneimittelrecht, Medizinprodukterecht und Transfusionsrecht. Berlin: Springer, 6. Auflage, 2014a.

Hrsg. Lenk, C., Duttge, G., Fangerau, H. Handbruch Ethik und Recht der Forschung am Menschen. Berlin: Springer, 2014b.

Pramann, O., Albrecht, U.-V. Forschung im Krankenhaus. Gestaltung, Chancen, Finanzierung. Düsseldorf: Deutsche Krankenhaus Verlagsgesellschaft, 2015.

Regulation on the application of good clinical practice in the conduct of clinical trials with medicinal products for human use. http://www.gesetze-im-internet.de/gcp-v/BJNR208100004.html (accessed August 26, 2017).

Statutes of the ethics committee of the Hannover Medical School. https://www.mh-hannover.de/16582.html (accessed August 26, 2017).

4

The Use of Transcriptomics in Clinical Applications

Daniel M. Borràs and Bart Janssen

GenomeScan B.V., Leiden, The Netherlands

4.1 Introduction

The associations between genes and diseases have already been the subject of study for many decades. The clearest cases of associations showed that a particular change in a single gene can be the potential cause for a particular disease. There are over 1500 defined genes that were classified as monogenic disorders with an associated phenotype [1], but these do not cover most of the human diseases that are mainly multifactorial. In their expert opinion, Stylianos Antonarakis and Jacques Beckmann state that monogenic disorders are an unfortunate casualty in the race to find the determinants of complex diseases [1–3]. Not all genetic mutations are detected or diagnosed using the same type of material. Isolated DNA is the most common type of patient material used for diagnostics applied to nucleotides. However, the analysis of other nucleic acids such as messenger RNAs (mRNAs) has great potential to elucidate many genetic disorders. In particular, RNA is used to diagnose complex diseases where multiple genes are implicated such as cardiovascular disease, breast cancer, and type 2 diabetes mellitus [4, 5].

Over the years many diagnostic tests have been developed that took advantage of the latest advances in technology for biomarker discovery. The ultimate goal of molecular diagnostics is to accurately predict the presence or absence of a particular genetic disorder or infectious disease, or even response to a particular drug treatment. Toward this end, the latest advances in high-throughput technologies opened up new opportunities for RNA-based diagnostics, allowing the development of new and more sensitive disease diagnostic tools. This goal is achieved by applying advanced statistical methods that reveal hidden patterns within transcriptomics data that characterize multifactorial diseases. This was accomplished to some extent with the appearance of high-throughput technologies such as microarrays and RNA sequencing (RNA-Seq), both able to analyze thousands of mRNAs in a single run. The global analysis of the mRNAs can be quite rewarding and informative, providing information on many variations with base pair accuracy in the case of some diagnostic tools. Analyzing transcription products in diagnostics allowed to detect many diseases caused by genomic alterations as well as specific transcript modifications such as single nucleotide polymorphisms (SNPs), small variants (SV), translocations, inversions, chimeric genes, breakpoints, posttranscriptional modifications, alternative splicing, and gene expression [6–11]. See Figures 4.1, 4.2, and 4.3 for of RNA-Seq visualization examples analyzing disease-associated genes, such as *MECOM*, *NPM1*, and *IRF1*, for the detection of gene expression, SVs, and splicing variants of acute myeloid leukemia, respectively (primary data provided by Prof. Veelken, H., Head of Department of Hematology, Leiden University Medical Center).

In order to perform a diagnostic test on mRNA, the RNA first needs to be accurately extracted. There are many extraction methods available, and their applicability mainly depends on the sample source and type of material. Each method can provide specific advantages depending on the variety of accessible biological sample types and the downstream analysis to be performed. Some of these advantages can be evaluated in several ways, but yield and integrity are generally accepted as criteria of successful extraction. The material obtained from the sample extraction, regardless of the method used and the success of the extraction, is generally in the form of total RNA (tRNA) [12]. In this state, the extracted material from human samples contains mainly ribosomal RNA (rRNA) transcripts (~80%) and tRNA transcripts (~15%). The presence of these two RNA types greatly reduces the abundance of other transcript species that are of diagnostic interest such as mRNAs. Therefore, an additional enrichment of the preferred mRNA is needed, after the extraction of tRNA, for further analysis.

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

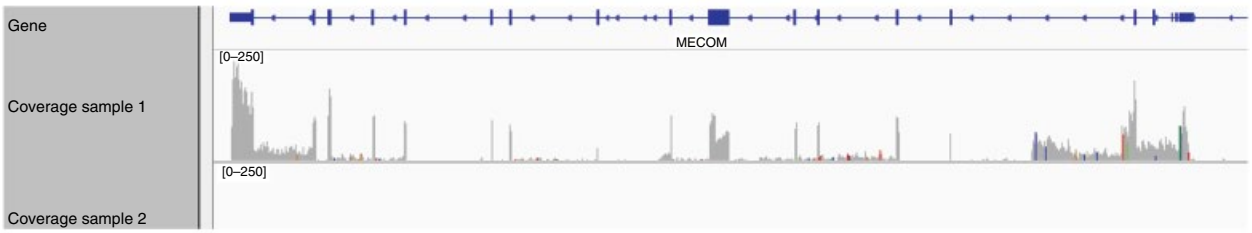


Figure 4.1 Gene coverage for RNA-Seq data. First gene track (blue line) shows the gene structure of MECOM by RefSeq gene annotation. Thick lines show exon locations and arrowed lines intronic locations. The second and third tracks show the coverage of MECOM for samples 1 and 2, respectively. In gray, the depth of coverage of sample 1 indicates that there is expression where the peaks are located. The extent of the expression is proportional to the height of the peaks, and the amplitude represents how much of the gene sequence is expressed, usually coinciding with exon boundaries. *Source:* Data provided by Prof. Veelken, H., Head of Department of Hematology, Leiden University Medical Center.

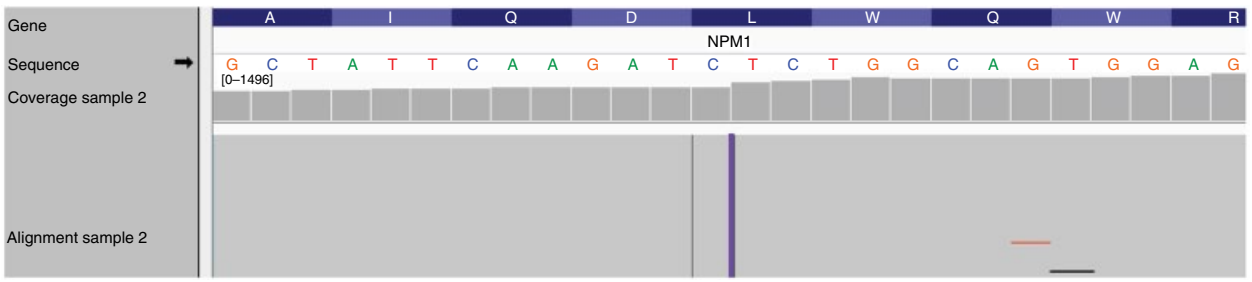


Figure 4.2 Visualization of a SNV detected by RNA-Seq. The 4 bp insertion (TCTG) is represented as a purple line on the alignment track for the *NPM1* gene. The gene track (blue) shows the codons present in this particular location (blue and dim blue) as well as the amino acid it encodes for (letters). The sequence of this particular location is shown, and the black arrow gives information about the strand of the shown nucleotides (forward). The third track shows the coverage for each of the represented bases. Finally, the alignment is shown where horizontal gray lines represent a match with the reference sequence and a colored horizontal line represents a mismatch, which is colored with the same colors as the sequence nucleotides, being T red, A green, C blue, and G orange; black represents a deletion. *Source:* Data provided by Prof. Veelken, H., Head of Department of Hematology, Leiden University Medical Center.

The isolation of tRNA from the biological samples can be currently achieved by several methods including magnetic bead-based extraction, silica columns, and acid phenol/chloroform. The enrichment of mRNA will enhance the reliability and sensitivity of the diagnostic test. This is usually achieved by several steps, which may include degradation steps of nondesired nucleic acids, amplification of DNA, and hybridization and ligation, among others. The degradation of the nondesired RNAs can be performed by an RNase degradation step. Such RNA degradation is, for instance, performed after the reverse transcription (RT) amplification of enriched mRNAs by using poly-A primers. The amplification provides the first strand of the mRNAs as cDNA, which will no longer be degraded by the RNase activity. Similarly, the enrichment can be achieved without RNase degradation by including adapters during the RT steps and/or ligation steps. Then the cDNA is amplified by adding primers that are specific to the adapters introduced. There are other methods that involve physical separation of the mRNA transcripts. In this situation, the use of poly-A ligands is quite usual, also known as poly-A capture enrichment. A common way of physical separation is by pulling down the mRNAs with magnetic beads. The poly-A ligand is attached to the beads, and then the mRNA is hybridized and separated from the sample solution by the use of magnetic forces that pull the magnetic beads down. Another common method involves the use of filtration columns. In this situation, the poly-A ligands are usually attached to the stationary phase of the column. In the majority of the described enrichment cases, commercial kits may be available and used for the enrichment of mRNAs. It is important to note that some kits allow direct isolation of mRNAs without a previous step of tRNA extraction. As general recommendation, the evaluation of the performance of enrichment protocols and available kits is considered as good practice for the success of setting up any diagnostics test. Another method of enrichment is polyacrylamide gel electrophoresis, which is generally used to separate nucleic acid species of desired sequence length [13].

One of the limitations of handling RNA is that it is quite prone to fast degradation. Unless the analysis of the RNA is performed immediately after the sampling

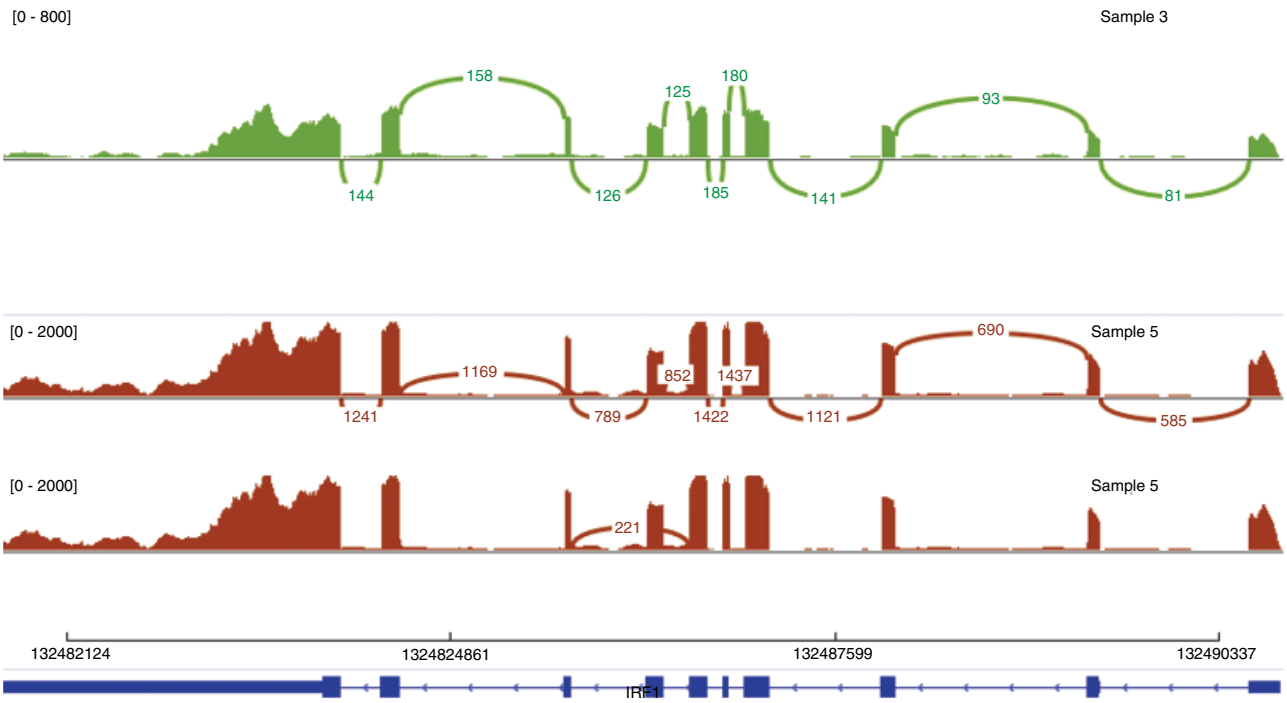


Figure 4.3 Visual representation of splicing detected for two sequenced patients with RNA-Seq. Bottom blue track represents gene structure of *IRF1* gene. Thick lines show exons bases and arrowed line intronic bases. Green track shows the coverage of sample 3 (green blocks) as well as the number of reads that span the distance to the next splice acceptor region (green line). Red tracks show coverage of sample 5 (red blocks) and number of reads that span the distance to the next splice acceptor region (red line) connecting two exons. The last red track shows an alternative splicing event with 221 reads that are connecting exons 3 and 5 but skipping exon 4. This shows that exon 4 is partially skipped by some transcripts in sample 5. *Source:* Data provided by Prof. Veelken, H., Head of Department of Hematology, Leiden University Medical Center.

process, effective precautions against degradation must be taken. Preventing RNA degradation is of particular concern for quantitative analyses and the diagnosis of diseases where high sensitivity is required. This may be crucial in detecting minimal residual disease (MRD), for instance, to quantify *BCR-ABL* mRNA in patients with chronic myeloid leukemia (CML) under imatinib therapy [14]. There exist several procedures and reagents that are used to preserve the integrity of RNA. In the majority of cases, the use of these measures is highly recommended. An extensively used method for preserving the RNA is to snap freeze the samples, which can be done with liquid nitrogen, followed by storage for a longer period at −80°C. Otherwise, during the collection of the samples, cells can be directly embedded in RNA-safeguarding reagents or buffers that can preserve RNA from degradation. In this case, it may be possible to store the samples at a higher temperature depending on the approach and chemicals used during the collection. For all of the aforementioned cases, extra care has to be taken to maintain the samples at the lowest temperature possible during the handling and to work with appropriate RNase-free solutions, material, and working environment at all times [12].

Although it is always preferred to use "fresh" biological material for RNA analysis, the use of fresh material is not always an option. In some cases tissue samples are needed for clinical diagnostics that involve a particularly difficult and painful process for obtaining the material such as biopsies. These, at the same time, may involve complex extraction methods, as in the case of bone samples. Therefore, tissue-specific methods for isolating RNA are both available and required. Continuing with the previous example, the isolation of RNA from bone samples is considered for the understanding of the development of some metabolic disorders, such as type 2 diabetes. Toward that end, work has been done to improve the isolation methods for those difficult samples. As an example, Carter et al. published an improved method for RNA isolation from bone samples in a single tube by using a Bullet Blender homogenizer. This method provided up to eight times more RNA with high quality measured as RNA integrity number (RIN) [15] scores ranging from 6.7 to 9.2. This method could reduce the usually time-consuming steps of bone grinding by nitrogen freezing and hammer smashing the sample, which are commonly a source of contamination, RNA loss, and degradation [16]. The downside of all the tissue-specific approaches

is the increasing variability in terms of efficiency measured in yield and quality that is even more compromised when more generic methods are used with difficult samples. Hence, it is not surprising that the current trend for the development of new diagnostic tools is centered on sample sources that are more easily obtained using noninvasive methods such as blood, saliva, and urine [12].

When RNA is extracted from a biological sample for disease diagnostics, we must also take into consideration that sample composition is another important variability-introducing factor. Tissue samples, for instance, contain different cell types. Each cell type will express its own gene repertoire. A sample mRNA composition may also be influenced by internal and external factors (i.e., nutrition, circadian stage, cellular cycle, stress, exercise, or disease state). Sasagawa et al. showed that single-cell transcriptome analysis using RNA-Seq was able to identify and quantify nongenetic cellular heterogeneity and even differentiate cell types and cell cycle phases of a single cell type [17]. Therefore, it is important to use appropriate methods for targeted cell type enrichment, if possible, such as laser capture microdissection (LCM) for selecting tissue areas from tissue slides, cell sorting for enriching the cell fraction of interest, or centrifugation for separating the desired cell population [12, 17–20].

Many of the aforementioned processes and techniques used for the extraction of RNA are based on the use of a variety of solutions and chemicals. Unfortunately, the procedures are not perfect, which means that during the extraction there is a carryover of chemicals and reminiscent sample components such as DNA, proteins, and salts. Chemicals such as ethanol, chloroform, or phenol, as well as monovalent cation salts such as ammonium acetate, lithium chloride, and sodium acetate, may be present in the sample due to the RNA extraction and precipitation approaches used [21]. Carryover genomic DNA may interfere with amplification or hybridization steps but can be easily removed from the sample by a DNase enzymatic treatment, and remaining proteins can be readily removed by proteases. Sample stabilizers such as citrate, EDTA, and heparin may inhibit RT and should also be removed. Several authors reported approaches for this purpose such as precipitating RNA with lithium chloride to get rid of sample heparin [22] or the use of ultracentrifugation [23]. It is worth mentioning that any single effort toward the removal of potential contaminants may improve the outcome of the diagnostic test applied. These contaminants, in the end, may negatively influence the possible choices for downstream applications and in the worst case have a direct impact on the diagnostic assay and its outcome [12].

While keeping in mind that there are many possible sources of variability, one has to choose from a large series of RNA isolation kits and methods. The choice should be made depending on the sample type to be analyzed. The ideal method is fast and easy, and it allows a reproducible extraction of RNA that is immediately ready to use or store. Many of these kits already include cleaning steps from inhibitors and contaminants and could be automated. Additionally these kits may also directly extract not only tRNA but also targeted subspecies of RNA such as mRNA, miRNA, and even viral RNAs, making the process much easier to set up for diagnostic applications. Several authors reviewed the performance of different RNA isolation kits applied to a variety of tissue samples. A recent example of the assessment of the performance of two purification kits was provided by Akutsu et al. [24]. The authors compared RNeasy Mini Kit (silica column-based kit from Qiagen) with EZ1 RNA Tissue Mini Kit (automatic magnetic bead-based kit from Qiagen). The extraction was applied to two different biofluid samples, saliva and blood, in the form of fresh and dried cotton swab stains. The results of this comparison showed that silica column-based extraction provided a slightly better RNA quality on fresh samples and substantially better on sample stains. Unfortunately, comparisons between multiple kits show that there may be kit-dependent differences in RNA extraction that may impact downstream gene expression measurements [25]. Differences in RNA yield and RNA quality should be taken into account when selecting the best approach for processing clinical samples. Hence, when performing comparisons between studies, authors should be aware of the differences among different isolation procedures used. Further details on the advantages and disadvantages of currently available methods as well as the effect of stabilizers used, such as RNAlater and RNAprotect, are available in the following publications [14, 25–28].

After isolation, there are several ways to investigate the RNA. One can investigate a single gene transcript, several selected transcripts, or a complete transcriptome. The latter is an omics word to define the set of RNA molecules, including mRNA, rRNA, tRNA, and other noncoding RNA, transcribed in one cell or a population of cells. In this chapter, we will restrict ourselves to the analysis of the mRNA part of the transcriptome. Currently, there are two methods to perform transcriptome-wide analysis of mRNAs: next-generation sequencing (NGS) and microarray analysis.

During these past years, the majority of research efforts of NGS applications were focused on the understanding of the etiology of diseases including complex diseases [12]. Few of these research efforts resulted in a diagnostic test that is currently being used in a clinical setting. In some occasions, the diagnostic test for a certain disease could be reduced to a few key genes, making it

Table 4.1 Examples of transcriptomics clinical applications for PCR, microarray, and RNA-seq disease.

	Tissue	Target genes	Type	Method	Application	Source
CML	Peripheral blood, bone marrow	BCR-ABL	Chimeric gene	qPCR	Commercial kit	www.pcrdiagnostics.eu
HIV	Blood plasma	HIV	Retrovirus	qPCR	Commercial kit	www.pcrdiagnostics.eu
Hepatitis	Blood plasma	HCV	Retrovirus	qPCR	Commercial kit	www.pcrdiagnostics.eu
TORCH	Peripheral and umbilical cord blood plasma; saliva; oropharyngeal swabs; amniotic fluid	Rubella virus RNA	Retrovirus	qPCR	Commercial kit	www.pcrdiagnostics.eu
Flu	Nasal and throat swabs or washes; aspirate of trachea; feces; autopsy material	Influenza viruses	Retrovirus	qPCR	Commercial kit	www.pcrdiagnostics.eu
Breast cancer	Tissue biopsy	BRCA1, BRCA2, +19 more	Expression	qPCR	Biomarkers	[29, 30] Oncotype DX
Breast cancer	Tissue biopsy	70-gene signature	Qualitative and quantitative expression	Microarray	Commercial test	Agendia NV
Colon cancer	Tissue biopsy	18-gene signature	Qualitative and quantitative expression	Microarray	Commercial test	Agendia NV [31]
Septic shock	Blood	100-gene signature	Qualitative and quantitative expression	Microarray	Biomarkers	[32]
Colorectal Cancer	Blood	ANXA3, CLEC4D, LMNB1, PRRG4, TNFAIP6, VNN1, IL2RB	Qualitative and quantitative expression	Microarray	Biomarkers	[33]
Colon Cancer	Tissue biopsy	RSPO2, RSPO3, TCF7L2, TET1, TET2, ERBB3, ATM, IGF2, among many others	Fusions, CNV, SNV, gene expression	RNA-Seq	Research	[34]
TNBC	tissue biopsy	PARK2, RBI, PTEN, EGFR, KRAS, IDH1, and ETV6, among many others	SNV, CNV, expression, allelic expression	RNA-Seq	Research	[35]
Transplant rejection	Peripheral blood	IL1R2, CXCR4, HLA-A, WSB1, CD48, SPARC, TYROBP, CD74, HTR1D, and SLC29A1	Expression	RNA-Seq	Research	[36]

much easier to handle with a simple and cost-efficient real-time qPCR (RT-qPCR) test (e.g., *BCR-ABL*, *BRCA1*, *BRCA2*, retroviruses' RNA) (Table 4.1). However, in multifactorial diseases, the number of relevant RNA biomarkers is usually too high to be handled by a single PCR assay. For diagnosing complex diseases, many small dedicated assays are being used in combination. Therefore, the analysis of a complete mRNA transcriptome might be considered as a good alternative. In the succeeding text, we will provide a number of examples showing useful approaches for the analysis of RNA molecules to diagnose inherited or acquired diseases.

4.2 Clinical Applications of Transcriptomics: Cases and Potential Examples

4.2.1 PCR Applications

The diagnostic use of RT-PCR is a very common practice and is frequently applied to detect hundreds of different diseases. The majority of clinical centers and diagnostic facilities offer a wide range catalog of services based on PCR kits to routinely diagnose patients. We will mention some examples of PCR tests that are or could be routinely

applied to diagnostics. It is worth mentioning that there are many tests available that are DNA based and are presented and discussed in another chapter (I3 Genomics). Despite this, PCR-based diagnostics still deserves a place in this chapter for its importance in the field of clinical diagnostics.

Recent advances in PCR technologies have extended the range of possibilities to use PCR for applied diagnostics. These new PCR platforms, considered as the next-generation RT-qPCR, include a wide variety of PCR settings such as microfluidic chips, digital PCR (dPCR), emulsion dPCR also known as digital droplet PCR (ddPCR), microfluidic chip-based dPCR, and steel chips. Many of these take advantage of micro- and nanofluidics combined with miniaturized systems to produce an increased number of reactions with a significant reduction of reagent used, sample volume, and costs. This allowed the increase of throughput of these platforms to thousands of reactions per sample, making these platforms quite suitable for a rapid screening of relatively large series of biomarkers. For a recent review about PCR-based technologies, advances, and advantages, as well as limitations, please refer to Devonshire et al. [12].

Infectious diseases represent some of the clearest examples of PCR use for diagnostic purposes, taking advantage of some PCR tests that are available as commercialized kits. The purpose of these tests is to detect a certain RNA in the analyzed sample such as blood and buccal swabs. These tests have the purpose to be qualitative, providing the answer of the presence or absence of a specific RNA, and in some cases offer quantitative results. Some of these tests for infectious diseases detect the RNA of pathogenic viruses such as the human immunodeficiency virus (HIV) or the hepatitis virus. An example of a commercially available test based on a qPCR kit for detecting HIV may be the AmpliSens HIV-Monitor-FRT kit from PCR Diagnostics. This test uses real-time hybridization fluorescence to detect HIV type 1 RNA in plasma samples. Similar to this test, other commercially available tests exist for detecting infectious diseases targeting RNA in plasma samples as well as other tissue sources. Some examples of available tests are listed in Table 4.1 as well as additional relevant test details and references.

The detection and quantification of mRNAs by PCR tests is also applied to the diagnosis of certain oncogenic diseases such as MRD. Here, the purpose of the test is to sensitively detect cancer recurrence (CR). A recent review, published by *Sherrod* et al., on MRD for multiple myeloma reported the efficiency of several tests to detect CR after stem cell transplantation treatments. The diagnostic tests reviewed included a variety of techniques such as allele-specific oligonucleotide PCR (ASO-PCR) and NGS, among others. ASO-PCR was reported to be very efficient with a sensitivity of up to 10^{-6} malignant plasma cells, which is an order of magnitude more sensitive than commonly used multiparametric flow cytometry (MFC) methods. The ASO-PCR test provided an accurate quantification of the expression of the following immunoglobulins: *IGHV-J*, *IGHD-J*, and *IGKDEL*. However, it also showed some limitations when considering the costs, turnaround time, and availability. Additionally, the development of unique primers for each particular patient reduced the success of this diagnostic test in a clinical practice, invalidating its efficiency when mutations occurred in the clonal antibody gene due to the evolution of the cancer. This effect hampered the success of the ASO-PCR to be applicable in 42–86% of myeloma cases, whereas MFC was successful in >90% of myeloma cases. Results favor MFC for MRD assessment despite ASO-PCR being more sensitive [37].

Another application for PCR tests is the detection and quantification of carcinogenic fusion genes and their expression. For instance, in the particular case of MRD involving CML, the chimeric oncogene *BCR-ABL* is expressed as result of chromosomal abnormalities known as Philadelphia chromosome translocation. For the diagnosis or screening of this chimeric transcript, different methods are commonly used such as fluorescence *in situ* hybridization (FISH) and RT-qPCR. A very recent comparative study assessed these two different diagnostic methods in terms of sensitivity and specificity on a cohort of 78 CML patients. The authors reported that the diagnosis using RT-qPCR methodology was in high concordance with the cytogenetic response category after treatment. Newly diagnosed patients were detected in 100% of the cases, and further analysis follow-up showed that the RT-qPCR was capable to detect and give precise measurements of low-abundant *BCR-ABL* transcript levels as a sensitive indicator of MRD [38].

The analysis of fetal cell-free RNA could also be very helpful to detect, for instance, the gender of the fetus. This was first reported by Poon et al. by using a two-step RT-PCR to detect Y chromosome-specific zinc finger protein (ZFY) mRNA [39]. The presence of multiple nucleic acid sources in maternal blood including fetal cell-free DNA and RNA allowed researchers to develop methodologies to screen for fetal diseases. Many noninvasive prenatal tests (NIPT) are being routinely performed to easily and early screen for the most fatal fetal diseases. These may include a wide variety of affections such as chromosomal abnormalities, sickle cell anemia, hemoglobinopathies, and cystic fibrosis, among other disorders. The detection of fetal nucleic acids in blood of pregnant women provides a wide and open field of study and will lead to the development of many clinical applications. This methodology is not invasive compared with other techniques such as amniotic fluid analysis, and it

has excellent sensitivity/specificity for clinical applications. The majority of these clinical tests are currently targeting cell-free fetal DNA, for instance, to detect Q890X mutation for parental-inherited cystic fibrosis [40] or *HBB* gene mutant alleles for the detection of sickle cell anemia [41, 42]. In a recent publication, Thung et al. reviewed more in depth the NIPT issues and advantages of detecting fetal DNA in maternal blood samples for chromosomal (an)euploidies. The authors provided an extensive insight on practical issues encountered for the implementation of NIPT techniques based on examples of literature data and their own [43]. Examples provided included the detection of chromosomes 13, 18, and 21 and sex chromosome (an)euploidies, as well as other autosomal aneuploidies and genome-wide deletions and duplications. With these examples, Thung et al. proposed an NIPT workflow that includes challenges one may encounter during the implementation, such as sample collection, fetal fraction, sample tracking, automation of DNA isolation and library preparation, required sequencing, scope of testing, data analysis, and discrepant findings. However, even though its use may not be as extensive as is in DNA NIPT, the use of cell-free fetal RNA in maternal blood could develop into an important diagnostic approach for the detection of genetic disorders and fetal defects. In early stages, cell-free maternal RNA could identify the gender of the fetus detecting Y chromosome-specific ZFY mRNA by a two-step RT-PCR [39]. In addition, subsequent work on plasma RNA from pregnant women, reviewed by Wong and Lo., showed that the detection of temporal dynamics from fetal transcriptome is currently a possibility to assess [44]. Apart from other DNA-based NIPT tests, the authors review the advances in the transcriptomics NIPT field including several additional challenges compared with DNA NIPT such as low quantity and quality of mRNA extracted from maternal plasma, as well as varying levels of fetal RNA transcripts due to gene expression differences. To overcome all these challenges, an unbiased RNA-Seq method may still be required for the determination of the fetal transcriptome from maternal plasma. This would open a new field of possibilities for transcriptomics-based NIPT clinical applications, which may include fetal tissue-specific markers for screening of fetal aneuploidies and developmental defects [44].

The analysis of mRNAs for clinical diagnostics usually targets a single transcript of interest. However, some of the latest PCR technologies may allow for multiple parallel PCR tests in a single run and with a single sample. This widens the possibilities for diagnostic applications, based on PCR tests, to screen for several genes/targets at the same time. Hence, disease gene panels are targeted and amplified. There are commercially available tests for specific platforms that take advantage of multiple-gene screening using qPCR-based approaches. As an example, Oncotype DX currently commercializes three diagnostic tests for screening the expression of selective gene panels for breast cancer, colon cancer, and prostate cancer [45]. In all these cases, the objective is to calculate the CR score that may improve personalized cancer treatment. The first test was developed for estrogen receptor-positive (ER+) breast cancer to assess the risk of distant recurrence by targeting 21 breast cancer key transcripts. Oncotype DX for breast cancer is a breast cancer multiple-gene diagnostic assay for individualized treatment planning, such as chemotherapy benefit and distance recurrence. It uses a high-throughput RT-qPCR to analyze the expression of the 21 genes from which 16 are cancer-related genes and provide the majority of information on the recurrence risk in ER+ breast cancer. This 21-cancer panel was developed by screening 250 candidate genes extracted from the literature and tested in a cohort of 447 patients [29, 30]. Then, the 250-gene list was reduced to a panel of 16 cancer-related transcripts (Table 4.1) and 5 reference transcripts (*BACTIN*, *GAPDH*, *GUS*, *RPLPO*, *TFRC*) that were validated. A similar approach was followed to develop another assay for colon CR score assessment, testing a total of 12 key transcripts for the stage II colon CR score [46]. Yet another Oncotype DX assay was designed to calculate the recurrence score using a panel of 17 genes for assessing the risk of recurrence of prostate cancer [29]. Despite all the aforementioned advantages, these approaches still target a limited number of genes and not a full transcriptome. Thus, they cannot be considered as global transcriptomics techniques, even though the latest advances in RT-qPCR throughput are pointing toward that direction. In addition, some of these platforms, for a cost-efficient use, must analyze many samples that should be collected and run at the same time, which may not be the ideal situation for a clinical application where incidences may not be that frequent.

4.2.2 Microarrays

The technology of microarrays is based on the design of multiple DNA probes that are bound on a solid surface such as a glass slide. Microarray technologies have been widely used in research for measuring gene expression changes and elucidating the relationship between genotypes and phenotypes. They are quite cost-effective for profiling gene expression when it comes to model organisms. Microarrays have also been used in clinical diagnostics. Some examples are the detection of copy number variants using SNP arrays such as the Cytogenetics Whole-Genome Array from Affymetrix or the HumanOmni1-Quad BeadChip and HumanCytoSNP-12

DNA Analysis BeadChip from Illumina. In general, microarrays can be used for general screenings, gene expression profiling, genotyping, and many other applications. However, like in PCR-based applications, the use of predefined oligonucleotides (probes) is based on previous knowledge availability. Thus, microarrays are used for quantification of known sequences and not for the discovery of new variants, transcripts, or other unexpected transcriptomics features [47].

In order to fully illustrate the limitations of microarray technology, we should briefly present some basic concepts. Microarray detection is based on hybridization of sample DNA to nucleic acid probes, bound to the surface of a slide. The probes are oligonucleotides with a usual length of 25–120 nucleotides. To further measure the quantity of hybridization to each specific probe, the target sequence (DNA or cDNA) is labeled with fluorescent dyes. Then, after an image is taken and processed, signal intensities can be read and converted to normalized values in order to initiate the data analysis. Due to the nature of microarray probe design, the capabilities of this method are apparently restricted to known sequences and therefore do not allow detection of target sequences beyond the current knowledge. This factor can be a disadvantage for non-model organisms, but diagnostics of well-characterized organisms, such as humans, is feasible, although it relies on the quality of the available bioinformatics data at the moment the microarray was designed. Microarrays can be used for diagnostic transcriptome analysis. If properly designed, they will not only provide information on gene expression and expressed SNPs but also detect exon junctions and fusion genes [48].

Normalization and processing of microarray data can involve quite complex bioinformatics methodologies and statistics. This is a consequence of the nature of the data produced by this technology that may become a limitation for someone not acquainted in the area. However, significant efforts were put into developing standardized procedures for microarray analysis. Some of these procedures as well as suggestions, guidelines, metrics, and thresholds, among other information, are publicly available under the MicroArray Quality Control (MAQC) website, together with the publications that helped to reach consensus on these procedures. Refer to the MAQC project for further details [49].

The human genome has been long studied and annotated, making it easier to use the available information to develop microarray probes for clinical diagnostics. As an example, Agendia N.V. [50] developed clinical tests for complex diseases such as breast and colon cancers based on gene expression profiling microarrays. The Agendia "MammaPrint" assay can be used to classify different types of breast cancer and to calculate the recurrence risk. This assay was tested on a cohort of 6694 early breast cancer patients in a phase III trial (MINDACT) to investigate the utility of the MammaPrint 70-gene signature (Table 4.1) for adjuvant chemotherapy [51]. Similarly, another test for colon cancer could also classify different cancer types as well as calculate recurrence risk factors with an 18-gene signature (Table 4.1) [31]. Please refer to the Agendia N.V. available online resources for further details, publications, and information about the signature assessment, validations, clinical trials, risk assessments, and test efficiency of the aforementioned assays (www.agendia.com).

Genome-wide microarrays can yield a global view of gene expression and are designed without any investigator bias. Compared with custom-designed arrays, genome-wide array analyses provide a better opportunity of resolving complex and heterogeneous clinical syndromes. A review by Wong of several approaches for sepsis and septic shock shows the potential and applications of the genome-wide microarray analysis [52]. One of the outcomes of the studies carried out with septic shock was the characterization of a 100-gene expression signature. The correlation of clinical phenotype of these pediatric patients with array data showed that this expression signature could classify septic shock of three different phenotypes. One of these is a severe phenotype with increased illness severity and higher mortality rate (Table 4.1) [32]. The proper identification of phenotype-correlated marker genes may also lead researchers to find potential therapeutic targets. Hence, proper classification to clinical phenotypes of septic shock would allow for the design of more specific and targeted therapies. With the review of many studies of septic shock showing similar results for genes such as *MMP-8*, highly expressed in patients with septic shock, it was possible to demonstrate that inhibition *MMP-8* resulted in significant improvement of patient survival. Thus, genome-wide microarray approaches can provide insights in the pathology of complex diseases, help to classify patients in groups with specific characteristics, and allow the discovery of novel therapeutic targets [52].

Chao et al. [33] reported that in a microarray blood transcriptome study carried out with 314 colorectal cancer patients, a seven-gene classifier was able to detect left-sided and right-sided lesions (lesions on the left and right side of the colon, respectively) with 78% sensitivity and 66% specificity against control samples with no colonoscopy pathology detected. Treatable cancers were detected with a sensitivity of 76 and 84% for left- and right-sided lesions, respectively. These results supported the plausibility of integrating blood-based screening tests into routine diagnostics for improved colorectal cancer outcome (Table 4.1) [33].

4.2.3 Sequencing

The advances in DNA sequencing, and in particular the advances of NGS, have significantly improved the quantity and quality of genomic information that can be obtained from clinical samples. The reduced cost of NGS as well as the increase in throughput made whole-genome sequencing (WGS), as well as other NGS applications such as whole-exome sequencing (WES) or RNA-Seq, a possible and reliable approach for clinical diagnosis. However, there are still some challenges such as data storage, management, analysis, and interpretation that have to be considered for the proper use of this technology in clinical applications [11]. Following the objectives of this chapter, the tools, applications, approaches, and examples presented here will mainly focus on the use of NGS for the analysis of the transcriptome in clinical applications.

Many different platforms for massive parallel sequencing were developed. The first example, although currently obsolete, is the 454 Genome Sequencer from Roche Applied Sciences. Also outdated is the SOLiD platform from Life Technologies. The current and most widely used technology is the Solexa "Sequencing-by-Synthesis" technology that was acquired by Illumina in 2007. The strength of these technologies relies on a very high throughput at the expense of read accuracy and much shorter read length when compared with the well-known Sanger sequencing. However, the possibilities of use and applications of this technology led to significant scientific discoveries and diagnostic applications [11]. Fortunately, some of the trade-offs are being reduced through continuous platform improvements and developments, which resulted in more advanced sequencer versions such as the Ion Torrent and Ion Proton from Life Technologies and the MiSeq and HiSeq from Illumina. In particular the HiSeq versions have greatly improved in accuracy and read length as well as in significantly higher throughput. Meanwhile, the run time has been decreasing, making it suitable for diagnostic use. Advances and ongoing efforts to improve these platforms even further have made the HiSeq platforms from Illumina the most widely used NGS sequencers.

Depending on the sequencing platform of preference, many options are available for library preparations. The library preparation steps include all transformations the nucleic acids of interest may require prior to being completely ready for sequencing on the platform of choice. In general, NGS library preparations for transcriptomics consist of cDNA synthesis and extension of the cDNA with specific ligated adapters for sequencing. Furthermore, it is quite common that a minimum quantity of RNA is required to ensure a minimal quality. For body fluids and tissues, approximately 10 ng of RNA is often sufficient, while for samples containing degraded RNA, such as FFPE, a minimum of 100 ng is strongly recommended [53]. In addition, many adaptations to library preparation protocols are reported in order to cover different aspects of the complexity of RNA processes and regulations such as posttranscriptional modifications, gene expression, isoforms, regulation, splicing, and degradation [7, 9, 54, 55]. For a better overview of published protocols, please refer to available collections of preparation methods such as the sequencing methods review published from Illumina Technology [56].

The overwhelming quantity of data produced per sample requires advanced bioinformatics analysis to address the wide variety of possible questions. There are many tools and software packages available that can analyze these massive datasets, make inferences from the data, and offer biological interpretations. Despite their differences, there are some data analysis steps that are usually shared among the different approaches. Common steps include quality check of the sequencing data, sequence alignment to a reference genome or de novo assembly in some other cases, and the assessment of the specific experimental results in order to finally provide useful diagnostic information [10, 11, 53]. It is accepted as good practice to perform several quality checks at the different steps in the process of analyzing clinical samples. Several authors reviewed different quality measures and how to use them during the downstream analysis. A recent review by Li et al. exposed many sequencing quality checks specific for RNA-Seq experiments including checks assessing raw sequence quality, nucleotide composition, presence of rRNA or tRNA, and the presence of other contaminant nucleic acids [57]. Another important step is the alignment of the sequenced reads to the reference genome, or transcriptome. The human genome is nowadays quite complete with the latest version 38 released on June 29, 2014, by the Genome Reference Consortium, patch 4 (GRCh38.p4) (www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human). In the alignment, the human genome is used as matching reference for the sequenced reads. RNA-Seq data alignments differ substantially from the DNA-Seq alignments. The nature of read sequences in RNA-Seq provides extra levels of complexity due to the fact that RNA molecules are the product of transcription and posttranscriptional processes such as splicing and RNA editing. The splicing removes part of the transcribed sequences—the introns—leaving the exons present in the sequence. After the library preparation and its fragmentation step, which is an optional step and commonly performed by sonication, some of the shorter reads obtained may come from the region where two exons were joined. In this particular situation, the RNA-Seq aligners have to be flexible enough to be able to map part of the reads to one exon

and the other part to another exon, spanning an exon junction [58]. There are many aligners available that can deal with RNA-Seq data, such as Bowtie2, GSNAP, STAR, and SpliceMap, among many others. Work has been done to review and report available alignment tools to help users through the, sometimes difficult, decision of selecting the best tools for applications in clinical diagnostics [10, 53]. In general, all aligners offer the possibility to modify key parameters in order to adapt their algorithms according to the quality of available data and the question of relevance. Once a decent quality alignment is produced, the proper diagnosis is usually within reach. A common approach is to retrieve transcript abundance, as gene counts, for gene expression profiles or differential expression. However, prior to comparing two RNA-Seq datasets, the raw counts should be normalized to account for some differences introduced by handling during the library preparation steps. Due to this inherent variability, normalization of raw counts is required since these are not directly comparable between or within samples [59]. There are many normalization methods, some correcting for gene length, GC content, and library size, as well as other bias adjustments. For better understanding of the available normalization procedures, Dillies et al. compared several normalization methods in order to clearly present their application in the context of RNA-Seq data. In summary, the available DESeq and TMM normalization methods showed to be able to maintain the power to detect differentially expressed genes while properly controlling the false positive rate [59]. Another way of normalization to deal with extra biases found in cross-platform or inter-laboratory comparisons relies on the inclusion of synthetic spike-in materials. In some cases these external RNA controls developed by the External RNA Controls Consortium (ERCC) became available for the evaluation of cross-platform performance according to GC content, transcript length, and sequencing accuracy [12].

Extended information on RNA-Seq practices as well as some additional recommendations, benchmarking technology comparisons, reproducibility assessments, and evaluations of RNA-Seq for clinical applications was also published by the Sequencing Experiment Quality Control (SEQC) consortium. The SEQC project is the third phase of the MAQC, and it involves 12 countries, 78 organizations, and 180 researchers (http://www.fda.gov/ScienceResearch/BioinformaticsTools).

The wide range of available bioinformatics tools offers the possibility to answer various biological and diagnostic questions. However, bioinformatics analysis may not be able to overcome some limitations that we can still face with NGS data such as highly repetitive sequences, 3′ biases, and biased GC content. In general, the small loss of information due to these limitations is of low impact compared with the significant insights that NGS provides. Repetitive sequences in the human genome are well characterized, making it easier to handle problems related to polymorphic copy number variation in these regions. During the alignment steps, reads that map to many locations of the genome (not uniquely mapped) with equal quality are usually filtered. The enrichment of 3′ end sequences of genes, also known as 3′ bias, is a side effect of the fast degradation of mRNAs from the 5′ end of the transcript, which may be even more prominent when using poly-A enrichment methods during the library preparation. This effect can be widely avoided by using higher-quality RNA, which should be possible in a properly designed diagnostic setting. Additionally, 3′ biases may not affect the outcome of some analysis, such as gene expression measurement, since it is considered that all transcripts exhibit similar degradation and the same library preparation was performed within a particular well-controlled experiment. The last limitation, regarding some difficulties of sequencing high GC regions, is a problem that usually results from several causes. First, it is known that some polymerases may have increased difficulties to transcribe high GC content sequences. This, coupled with the inherent high repetitive nature of GC or AT enriched regions, makes these regions somehow tricky to analyze with higher levels of confidence. However, not all high GC are affected at the same level due to differences in GC percentages and other nucleic acid composition [57]. Hansen et al. worked on an alternative normalization method to acquaint for the GC content as well as gene length of a particular gene using a conditional quartile normalization [60]. However, their method did not outperform other less sophisticated normalization methods [59].

Cancer is commonly regarded as an accumulation of genetic alterations such as single nucleotide variants (SNVs), altered DNA methylation patterns, and chromosomal abnormalities. As a consequence of DNA modifications, there may be dysfunctional genes leading to over- or underactivity and chimeric transcripts or gene fusions. These alterations may disrupt the proper function of the gene, which may become an oncogene, a malfunctioning tumor suppressor, or an incorrect DNA repair gene. The occurrence of one or more of these genetic alterations may affect cellular growth and lead to tumor development. Since the landscape of cancer transcriptome is complex, RNA-Seq can be very useful for clinical diagnostic applications, offering a wider range of screening possibilities to check for the whole diversity of cancer-related alterations in a single run [10]. Many studies have been carried out that contributed in the understanding of molecular determinants of tumor cell types. Cancer characterization is remarkably one of the research fields that has dedicated considerable efforts to

adopt RNA-Seq for research purposes and to assess its potential in clinical applications [6, 10, 11, 53]. Since the accumulation of genetic alterations may be either inherited or somatically acquired, RNA-Seq becomes a strong complementary approach in screening and diagnostic applications.

Among the most common genetic alterations in cancer, we find gene fusions or chimeric genes that result from chromosomal abnormalities such as inversions, deletions, and translocations. Some of these alterations may modify the expression of certain genes, which can be detected with RNA-Seq. Seshagiri et al. [34] could identify some duplications at the DNA level that correlated with the overexpression of a particular gene, *IGF2*, in a subset of colon tumors. In the same study, the analysis of RNA-Seq data identified multiple fusion transcripts. These included recurrent gene fusions from members of the R-spondin family such as *RSPO2* and *RSPO3*. The fusion of these genes occurs in 10% of colon tumors. Additionally, the authors detected an effect of mutually exclusiveness between *RSPO* fusions and the presence of *APC*, indicating a potential role in the activation of tumorigenesis through the activation of the Wnt signaling pathway. This was corroborated by the potentiating effect of *RSPO* fusions detected over the Wnt signaling (Table 4.1) [34].

If somatic mutations occur in a crucial DNA position, this may translate into the development of a tumor. Moreover, uncontrolled cellular divisions combined with defects in DNA repair systems may lead to the evolution of cancers in a clonal expansion manner. This stepwise process and the resulting clonal expansion increase the variability and genetic diversity in a complex pattern. Therapeutic interventions in highly clonal expanded cancers introduce a strong selective pressure that translates in resistant cancer cells. This is regarded to be one of the major causes of therapy failure for some cancer types [61]. However, work has been done to understand clonal expansion and the dynamics of this complex evolution. Shah et al. showed that the detection and quantification of the clonal evolution was possible using RNA-Seq by characterizing the mutational evolution spectrum. In a similar study, the authors analyzed data from SNP arrays, WES, and RNA-Seq for a total of 104 individuals suffering from triple-negative breast cancer (TNBC). The results showed that only 36% of validated SNVs were expressed and detected within the transcriptome. These SNVs as well as the detected splicing variants were mainly accumulated to particular carcinogenic genes such as *TP53*, *PIK3R1*, *AP3B2*, or *TNIP1* (Table 4.1). Additionally, a significant enrichment of somatic mutations was detected in noncoding regulatory regions such as retinoblastoma-associated proteins transcription factor binding (RTFB) sites. Several of these mutations were predicted as damaging the RTFB sites for genes such as *SPATA17*, *KCNE2*, and *SRRM5* (Table 4.1). Additionally, it was possible to differentiate early from late events in the clonal evolution and identify driver genes, potential novel pathways, and oncogenes that were not previously described. Furthermore, authors showed that basal and non-basal TNBC differ in their clonality at the time of diagnosis, basal TNBC having higher clonal frequency than non-basal TNBC [35]. In the end, RNA-Seq was able to detect SNVs and splicing variants as well as to elucidate the clonal evolution of TNBC. These results offer novel insight on cancer biology and will contribute to the revision of the current diagnostic methods for multifactorial diseases such as TNBC.

Apart from complex diseases such as cancer, RNA-Seq showed great potential in other fields of clinical diagnostics such as immunology. Recent tendencies in immunological studies showed that large-scale genomics and transcriptomics approaches are becoming more popular options for immunological studies. These global approaches can help to mitigate, for instance, the limitation of cellular heterogeneity in blood samples, even though the measurement of cell composition in blood samples is not always perfect [6]. Despite this limitation, the use of a global approach, such as transcriptomics analysis of blood samples, has shown to provide very nice advantages toward clinical diagnostics as was seen in early research done on autoimmune disorders, cancer, and infectious diseases [6].

Transplant rejection in heart transplants has been assessed by an endomyocardial biopsy test. But this is an invasive procedure that is characterized by greater risk of morbidity, discomfort for the patient, tissue sampling errors, and late detection of rejection. Due to all these limitations, it was necessary to find an alternative to this type of invasive tests to detect heart transplant rejection and to more accurately and early adapt patient treatment to avoid transplant rejections. With this in mind, Chen et al. [36] used an NGS approach to analyze peripheral blood gene expression profiles, monitor the immune system, and potentially avoid heart transplant rejection by early detection. For this study 12 patients were analyzed from grade 0 (6 quiescent patients) to grade 2R and 3R (6 rejection patients). The results were validated by qPCR of 47 individuals from three different rejection groups. A total of 10 genes (Table 4.1) were identified, which provided a signature of high risk for severe rejection. This 10-gene signature was also tested to be effective in other organ transplants [36].

Other research work in the field of renal diseases also showed the potential of RNA-Seq to identify gene expression profiles, gene pathways, and alternative splicing linked to *TGFB* and *SMAD3* signaling in chronic kidney diseases (CKD) [62]. In summary, research findings using animal models provided insight information about

SMAD3 signaling and its function in renal injury, as well as highlighting potential targets for CKD therapies [62]. The study of *SMAD3*-dependent renal injury was performed using tRNA of kidneys from mice animal models with *SMAD3* wild type and knockouts for immune- and nonimmune-mediated CKD (antiglomerular basement membrane glomerulonephritis and obstructive nephropathy, respectively). Zhou et al. reported nine differentially expressed genes linked to *SMAD3* (*IGHG1*, *IGHG2C*, *IGKV12-41*, *IGHV14-3*, *IGHV5-6*, *IGHG2B*, *UGT2B37*, *SLC22A19*, and *MFSD2A*) and showed that renal injury transcriptomes may mediate pathogenesis of CKD.

4.2.4 Discussion

Recently acquired knowledge, and technology, calls for a wide-scale use of transcriptomics in clinical applications. By using these technologies we may screen for multiple disease determinants in a single run and obtain information on the cause of the disease as well as on the potential response to treatments and many other factors. However, wider screening approaches such as transcriptome profiling are still not commonly preferred over PCR diagnostic tests. The use of PCR for monitoring one or several genes is cheaper than a complete expression profile in the terms of price per sample. However, when considering the price per screened gene, transcriptomics approaches offer a significant advantage. For instance, a 1 tier Fluidigm (multiple-reaction RT-qPCR-based platform) assessing 90 transcripts would cost 22 euro (25$) per sample or 0.24 euro (0.27$) per gene (adapted from Ref. [6]), while an mRNA-Seq would currently cost around 400 euro per sample but much less 0.01 euro per gene, considering that RNA-Seq potentially yields thousands of genes (Table 4.2). This shows that reducing the complexity of disease diagnostics toward a few key genes may reduce the immediate costs of the clinical assay. On the other hand, the cost of NGS will decrease considering the advances in sequencing chemistry.

Some omics approaches in clinical diagnostics make use of quite costly and/or experimentally challenging global approaches such as proteomics and transcriptomics. They replace tests for individual biomarkers, which would be required in increased numbers if global tests would not be available. Transcriptomics assays have the potential to be widely applied in the clinical setting since they are reliable and reproducible [6]. Given the mass of available knowledge about the human genome and disease biomarkers, RNA-Seq approaches or even NGS and genomic-wide microarrays should be considered as options for the future development of clinical diagnostic assays. There are examples of development of clinical applications based on NGS approaches that show their potential toward personalized medicine, despite these being DNA-based applications. Within this context, Renkema et al. reviewed several applications for CKD, highlighting that NGS-based DNA genetic testing can reduce the costs and turnaround time in diagnostics of steroid-resistant nephrotic syndrome and autosomal dominant polycystic kidney disease (PKD). The authors emphasize on the need of WGS approaches to identify nephronophthisis causative genetic variants due to oligogenic inheritance for genes such as *NPHP2*, *NPH3*, and *AHI1*. In addition NGS approaches allowed researchers to achieve a greater understanding of diseases and the pathogenesis of genetic disorders. Renkema et al. remarked that systems biology approaches that integrate data from various sources could be used to screen for potential drug targets. Additionally, the authors provided the example of PKD animal models used to determine efficacy of vasopressin receptor V2 antagonist [63]. Overall, given the latest tendency of clinical applications toward personalized medicine approaches, we may predict an increase of research to develop transcriptomics-based diagnostic applications.

Table 4.2 Technical comparison between PCR, microarray, and RNA-Seq methodologies.

	PCR	Microarray	RNA-Seq
Principle	Hybridization	Hybridization	Sequencing
Resolution	25–100 bp	25–120 bp	Single base
Dependence on available knowledge	High	High	Low
Background noise	High	High	Low
Identification of isoforms	Limited	Limited	High
Differentiate between allelic expressions	Limited	Limited	High
Maximum number of samples per run	384	2	$(96^{+}) \times 8$
Maximum number of target genes per sample	1728	30 455[a]	Everything transcribed

Source: Adapted from Mimura et al. [8] and complemented with Devonshire et al. [12] and personal information.
[a] As of number of CCDS genes from RefSeq genes annotated and the number of hybridization probes designed.

Several studies were recently reviewed by Chaussabel [6] providing a detailed vision of the perspectives of blood transcriptomics in the field of clinical diagnostics. The various applications reviewed included neurological disorders such as autism and Alzheimer; assessing rejection risk signatures for organ transplants such as the liver, heart, kidney, or bone marrow; and a wide variety of different affections such as transcript signatures for exposure of environmental factors, respiratory diseases, allergy, stroke, infections, and diabetes, among others. This highlights the possibilities that a wider approach may offer to clinical diagnostics in terms of types of diseases that can be properly resolved, diagnostic tests that may be developed, and the discovery of therapeutic targets.

Hybridization-based methods such as microarrays and PCR are bound to existing (and potentially limited) knowledge about the transcriptome and its association to possible disease phenotypes. Advances in the understanding of the human genome, function of genes, and their link with disease phenotypes will lead to improvements in the design of PCR tests and microarray experiments. However, currently only a few organisms such as *Homo sapiens* have a well-studied transcriptome, and our understanding of its complexity is still far from being complete [64]. Therefore, with respect to molecular testing in humans, few well-characterized disease-causing genes can be used with confidence as diagnostic tools. Tests performed to analyze more complex and heterogenic diseases are more challenging to implement in the clinical setting. The use of whole transcriptome approaches such as RNA-Seq opens up new possibilities. This, coupled with the increasing number of databases for storage and easy access to data, will provide the basis for larger and more complete studies with data that maintain scientific relevance for many years. The switch to global approaches such as whole transcriptome analysis will enable better prediction of disease onset, outcome, severity, and treatment response and in general easier patient management [6].

Current genome-wide gene expression microarrays do show an improved and reasonably accurate probe design. However, the technology still relies on the hybridization of fluorescent DNA to quantify the expression. This hybridization is known to produce some background noise that can interfere with the true signal. There have been several studies and reviews that address this topic while comparing different technologies in terms of throughput, accuracy, cost, and efficiency. For instance, Marioni et al. [64] assessed the technical differences between microarray technology and RNA-Seq. Despite the decent high-throughput setup that microarrays offer, the authors highlight the implicit high background noise of microarrays due to cross-hybridization. The methodology for controlling this noise in addition to the differences in the design of the hybridization probes makes microarray results almost impossible to merge with other experiments [64]. This effect is mitigated in NGS approaches that showed a higher resolution, fewer artifacts, greater coverage, and a wider dynamic range than microarrays [65].These factors as well as many other characteristics of microarray technology and RNA-Seq have been extensively compared and reviewed since the release of RNA-Seq in 2009 [8, 64, 66]. It must be emphasized that the detection of low-abundance transcripts can only be performed by using RNA-Seq technology [67]. This, as well as the ability to detect previously unknown transcripts, makes sequencing approaches more sensitive and complete than microarrays detecting up to 25% more differentially expressed genes [68]. Nevertheless, microarrays are still widely used in clinical diagnostics and will most likely still have a complementary role in clinical transcriptomics applications [69].

System-scale approaches are not applied in the clinical setting. Despite the numerous advantages that RNA-Seq may offer to clinical applications, it is still a fact that wider transcriptome approaches are yet to be implemented. This may be due to the understanding of transcriptome diagnostic tools as a cost–benefit risk when considered from a monetary but not a medical care point of view. In fact, if everything that needs to be measured is indeed measured by a transcriptome assay, there would be no market for dedicated assays, kits, and instruments for different tests and diagnostics [6]. The benefits of this shift would potentially improve disease diagnostics and reduce healthcare burden especially for more complex diseases. In this hypothetic situation, diagnostics for autoimmunity, cancer, cardiovascular diseases, infectious diseases, neurological diseases, nutrition deficiencies, pregnancy tests, and disease severity, onset, outcome, and response to treatment could be monitored from a single centralized laboratory [6, 33, 70–72].

The significant progress in high-throughput technologies such as genomics, transcriptomics, proteomics, and peptidomics, among others, will lead to personalized high precision medicine. In this setting the traditional symptom-oriented diagnosis and treatment would be complemented with individual molecular profiles of the patients, allowing a better treatment [73]. In this situation, transcriptomics as a continuously improving high-throughput precision omics would greatly facilitate the process. The detailed information of the transcriptome profile would reflect potential physiological changes at the sample collection time. Additionally, the integration with other omics would enhance the scientific research process. Eventually this would translate into improved healthcare by monitoring the patient health status and by applying personalized and preventive treatments.

At first, one may think that this would increase the costs of the healthcare system since some of these omics assays are currently quite expensive. However, as has been reviewed in the "RNA-Seq" section, this would especially help the understanding of complex diseases [35, 74, 75] and in the long term reduce the healthcare burden globally [73].

The successful integration of several omics approaches led to a better understanding of complex diseases. In this situation, one omics technology may be compensating some limitations of another omics approach. In the following example genomics and transcriptomics data were combined. In this approach, the two technologies used for improving complex multifactorial pathologies were the analysis of RNA-Seq combined with DNA sequencing such as WGS or WES. In this setup DNA sequencing offers a very accurate detection of potential DNA variants, which is then complemented with the gene expression profile of RNA-Seq. Using this approach Codina-Solà et al. [74] identified rare de novo mutations, inherited mutations including chromosomal abnormalities, and multiple hits in autism spectrum disorders (ASD). In this process, RNA-Seq was crucial to identify low frequency and rare mutations that were initially missed in WES due to usual filtering such as intronic causative mutations. The functional consequences of some of these mutations that were identified due to the combination of WES and RNA-Seq included aberrant transcripts, deregulated expression, allele-specific expression, and nonsense-mediated decay [74]. In conclusion, integrative whole sequencing approaches such as WES and RNA-Seq have proven to mutually increase the mutation detection efficiency. They enable to detect rare inherited and acquired mutations including intronic mutations affecting transcript splicing, overall expression, and allelic expression. This strategy contributed in the assessment of risk factors in a complex and highly heterogeneous etiology such as ASD.

Sample availability is usually limited in the case of human tissue specimens for research purposes or for clinical applications. This makes the process of sample selection a key step for the correct understanding of tissue-associated biological processes. Blood samples are among other biofluids such as saliva, one of the easiest tissue samples to obtain [12]. Therefore, many studies have been performed with blood samples to detect biomarkers and molecular determinants. However, the differentially expressed transcripts detected in blood may be underrepresented or only detected in a later state [73]. Despite this limitation, there are some successful studies characterizing blood samples by transcriptomics in immunological diseases [6]. Other types of biofluids may also be easily obtainable, making them an interesting target source for many researchers that want to develop new diagnostic tools using these noninvasive tissue sources such as saliva or urine [12]. Other tissue sample types are either not that easily accessible such as the heart, liver, and other internal organ samples or completely unavailable such as brain tissue. If sample availability is scarce, it makes the switch toward whole transcriptome approaches even more necessary since all genes expressed in such samples would be measured at once. This would serve two purposes: first, perform the required diagnostics for the patient at the moment of the test, and second, provide a wider overview of the patient transcriptome that could be stored in a database for future research or develop new diagnostic applications. Using this approach, samples would be available to be further included in larger clinical studies that provide sufficient power with the collected samples [73]. These changes in diagnostic application workflows would certainly make a big difference for those diseases where sample numbers are a limitation for the design of a clinical trial, as well as provide enough time to collect samples that would more accurately represent population frequencies.

Nevertheless, if transcriptomics approaches are to be implemented in clinical applications, the quantity of available data would increase the need of data handling measures. High-throughput applications can produce a tremendous amount of data, which is challenging to process, handle, and properly annotate with the purpose of further clinical interpretation. However, advances in computer technologies and databases as well as bioinformatics and available knowledge continuously improve. In the end, the objective of storing all these data and clinical results is to make it easier for clinicians to mine the databases with appropriate algorithms and make more accurate medical decisions. For this, comprehensive databases are required, which would store health records, variant calls, expression profiles, and all other patient-related molecular information [73].

There are currently many databases that can provide comprehensive functional annotations such as the Catalogue of Somatic Mutations in Cancer (COSMIC) [76], which is a comprehensive collection of somatic mutations for human cancer, or the Leiden Open (source) Variation Database (LOVD) [77], which provides a tool to collect and display DNA variants. These databases may potentially facilitate the complex process of annotation of high-throughput sequencing data. It is through the small effort of global sharing of particular findings that these databases are greatly improving over the years in quantity and quality of annotations. Other efforts are thrown into database collections of gene expression datasets such as the Expression Atlas [78]. In this case gene expression profiles are publicly available and accessible, which provides information about different organisms, expression patterns, and biological conditions, among others.

The Expression Atlas includes RNA-Seq experiment data as well as microarray experiment data that can be reanalyzed through their web portal [79]. There are other similar initiatives with a more focused character such as Nephroseq and Renal Gene Expression Database [80], which both center in collecting expression profiles related to human nephrology diseases. Please refer to the available online resources for further details [81, 82]. Additionally, there is also the possibility of using publicly available RNA-Seq and microarray datasets in combination with clinical data. This is the case of the Gene Expression Omnibus (GEO) from NCBI, which offers an international repository of microarray and NGS datasets submitted by the research community [83]. With this type of initiatives, the research community may benefit of the work of other researchers that may increase the statistical power of analysis and certainly increase the accuracy of clinical diagnostics applications.

References

1 Brinkman, R.R., Dubé, M.-P., Rouleau, G.A., Orr, A.C., Samuels, M.E., 2006. Human monogenic disorders—a source of novel drug targets. *Nat Rev Genet* 7, 249–260.

2 Antonarakis, S.E., Beckmann, J.S., 2006. Mendelian disorders deserve more attention. *Nat Rev Genet* 7, 277–282.

3 O'Connor, T.P., Crystal, R.G., 2006. Genetic medicines: treatment strategies for hereditary disorders. *Nat Rev Genet* 7, 261–276.

4 Herder, C., Karakas, M., Koenig, W., 2011. Biomarkers for the prediction of type 2 diabetes and cardiovascular disease. *Clin Pharmacol Ther* 90, 52–66.

5 Ross, J.S., Hatzis, C., Symmans, W.F., Pusztai, L., Hortobagyi, G.N., 2008. Commercialized multigene predictors of clinical outcome for breast cancer. *Oncologist* 13, 477–493.

6 Chaussabel, D., 2015. Assessment of immune status using blood transcriptomics and potential implications for global health. *Semin Immunol* 27, 58–66.

7 Dominissini, D., Moshitch-Moshkovitz, S., Amariglio, N., Rechavi, G., 2011. Adenosine-to-inosine RNA editing meets cancer. *Carcinogenesis* 32, 1569–1577.

8 Mimura, I., Kanki, Y., Kodama, T., Nangaku, M., 2014. Revolution of nephrology research by deep sequencing: ChIP-seq and RNA-seq. *Kidney Int* 85, 31–38.

9 Sánchez-Pla, A., Reverter, F., Ruíz de Villa, M.C., Comabella, M., 2012. Transcriptomics: mRNA and alternative splicing. *J Neuroimmunol* 248, 23–31.

10 Shyr, D., Liu, Q., 2013. Next generation sequencing in cancer research and clinical application. *Biol Proced Online* 15(1), 4.

11 Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., Shi, L., 2011. Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev Mol Diagn* 11(3), 333–343.

12 Devonshire, A.S., Sanders, R., Wilkes, T.M., Taylor, M.S., Foy, C.A., Huggett, J.F., 2013. Application of next generation qPCR and sequencing platforms to mRNA biomarker analysis. *Methods* 59, 89–100.

13 Petrov, A., Wu, T., Puglisi, E.V., Puglisi, J.D., 2013. RNA purification by preparative polyacrylamide gel electrophoresis, in: *Methods in Enzymology*. Elsevier, London, pp. 315–330.

14 Thorn, I., Olsson-Stromberg, U., Ohlsen, C., Jonsson, A.-M., Klangby, U., Simonsson, B., Barbany, G., 2005. The impact of RNA stabilization on minimal residual disease assessment in chronic myeloid leukemia. *Haematologica* 90, 1471–1476.

15 Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., Ragg, T., 2006. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol* 7, 3.

16 Carter, L.E., Kilroy, G., Gimble, J.M., Floyd, Z.E., 2012. An improved method for isolation of RNA from bone. *BMC Biotechnol* 12, 5.

17 Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., Ueda, H.R., 2013. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. *Genome Biol* 14, R31.

18 Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S.B., Buil, A., Yurovsky, A., Bryois, J., et al., 2015. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet* 11, e1004958.

19 Taussig, D.C., Vargaftig, J., Miraki-Moud, F., Griessinger, E., Sharrock, K., Luke, T., Lillington, D., et al., 2010. Leukemia-initiating cells from some acute myeloid leukemia patients with mutated nucleophosmin reside in the CD34-fraction. *Blood* 115, 1976–1984.

20 Todd, R., Kuo, M.W.L.W.P., 2002. Gene expression profiling using laser capture microdissection. *Expert Rev Mol Diagn* 2, 497–507.

21 Walker, S.E., Lorsch, J., 2013. RNA purification—precipitation methods, in: *Methods in Enzymology*. Elsevier, Amsterdam, pp. 337–343.

22 Del Prete, M.J., Vernal, R., Dolznig, H., Mullner, E.W., Garcia-Sanz, J.A., 2007. Isolation of polysome-bound

mRNA from solid tissues amenable for RT-PCR and profiling experiments. *RNA* 13, 414–421.
23 Ding, M., Bullotta, A., Caruso, L., Gupta, P., Rinaldo, C.R., Chen, Y., 2011. An optimized sensitive method for quantitation of DNA/RNA viruses in heparinized and cyropreserved plasma. *J Virol Methods* 176, 1–8.
24 Akutsu, T., Kitayama, T., Watanabe, K., Sakurada, K., 2015. Comparison of automated and manual purification of total RNA for mRNA-based identification of body fluids. *Forensic Sci Int Genet* 14, 11–17.
25 Jeffries, M.K.S., Kiss, A.J., Smith, A.W., Oris, J.T., 2014. A comparison of commercially-available automated and manual extraction kits for the isolation of total RNA from small tissue samples. *BMC Biotechnol* 14, 94.
26 Kim, J.-H., Jin, H.-O., Park, J.-A., Chang, Y.H., Hong, Y.J., Lee, J.K., 2014. Comparison of three different kits for extraction of high-quality RNA from frozen blood. *SpringerPlus* 3, 76.
27 Schagat, T., Kiak, L., Mandrekar, M., 2008. *RNA Purification Kit Comparison: Yield Quality and Real-Time RT-PCR Performance*. Promega Corporation, Madison.
28 Tavares, L., Alves, P.M., Ferreira, R.B., Santos, C.N., 2011. Comparison of different methods for DNA-free RNA isolation from SK-N-MC neuroblastoma. *BMC ResNotes* 4, 3.
29 Grenader, T., Yerushalmi, R., Tokar, M., Fried, G., Kaufman, B., Peretz, T., Geffen, D.B., 2014. The 21-gene recurrence score assay (Oncotype DX^{TM}) in estrogen receptor-positive male breast cancer: experience in an Israeli cohort. *Oncology* 87, 1–6.
30 Paik, S., Shak, S., Tang, G., Kim, C., Baker, J., Cronin, M., Baehner, F.L., et al., 2004. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *New Engl J Med* 351 (27), 2817–2826.
31 Salazar, R., P. Roepman, G. Capella, V. Moreno, I. Simon, C. Dreezen, A. Lopez-Doriga, et al., 2011. Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. *J Clin Oncol* 29(1), 17–24.
32 Wong, H.R., Cvijanovich, N., Lin, R., Allen, G.L., Thomas, N.J., Willson, D.F., Freishtat, R.J., et al., 2009. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. *BMC Med* 7, 34.
33 Chao, S., Ying, J., Liew, G., Marshall, W., Liew, C.-C., Burakoff, R., 2013. Blood RNA biomarker panel detects both left-and right-sided colorectal neoplasms: a case-control study. *J Exp Clin Cancer Res* 32, 44.
34 Seshagiri, S., Stawiski, E.W., Durinck, S., Modrusan, Z., Storm, E.E., Conboy, C.B., Chaudhuri, S., et al., 2012. Recurrent R-spondin fusions in colon cancer. *Nature* 488, 660–664.
35 Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., et al., 2012. The clonal and mutational evolution spectrum of primary triple negative breast cancers. *Nature* 486, 395–399.
36 Chen, Y., Zhang, H., Xiao, X., Jia, Y., Wu, W., Liu, L., Jiang, J., Zhu, B., Meng, X., Chen, W., 2013. Peripheral blood transcriptome sequencing reveals rejection-relevant genes in long-term heart transplantation. *Int J Cardiol* 168, 2726–2733.
37 Sherrod, A.M., Hari, P., Mosse, C.A., Walker, R.C., Cornell, R.F., 2015. Minimal residual disease testing after stem cell transplantation for multiple myeloma. *Bone Marrow Transplant* doi:10.1038/bmt.2015.164
38 Sag, S.O., Yakut, T., Gorukmez, O., Gorukmez, O., Ture, M., Karkucak, M., Gulten, T., Ali, R., 2015. Qualitative and quantitative evaluation of the BCR-ABL fusion gene in chronic myelogenous leukemia by flourescence in situ hybridization and molecular genetic methods. *Genet Test Mol Biomark* 19(10), 584–588.
39 Poon, L.L.M., Leung, T.N., Lau, T.K., Lo, Y.M.D., 2000. Presence of fetal RNA in maternal plasma. *Clin Chem* 46, 1832–1834.
40 González-González, M.C., García-Hoyos, M., Trujillo, M.J., Rodríguez de Alba, M., Lorda-Sánchez, I., Díaz-Recasens, J., Gallardo, E., Ayuso, C., Ramos, C., 2002. Prenatal detection of a cystic fibrosis mutation in fetal DNA from maternal plasma. *Prenat Diagn* 22, 946–948.
41 Barrett, A.N., McDonnell, T.C.R., Chan, K.C.A., Chitty, L.S., 2012. Digital PCR analysis of maternal plasma for noninvasive detection of sickle cell anemia. *Clin Chem* 58, 1026–1032.
42 Gahan, P., 2013. Circulating nucleic acids in plasma and serum: applications in diagnostic techniques for noninvasive prenatal diagnosis. *Int J Womens Health* 177. doi:10.2147/IJWH.S34442
43 Thung, D.T., Beulen, L., Hehir-Kwa, J. Faas, B.H., January 2, 2015. Implementation of whole genome massively parallel sequencing for noninvasive prenatal testing in laboratories. *Expert Rev Mol Diagn* 15(1), 111–124.
44 Wong, A.I.C. Dennis Lo, Y.M., February 2015 Noninvasive fetal genomic, methylomic, and transcriptomic analyses using maternal plasma and clinical implications. *Trends Mol Med* 21(2), 98–108.
45 Oncotype DX. www.oncotypedx.com (accessed August 25, 2017).
46 Cartwright, T., Chao, C., Lee, M., Lopatin, M., Bentley, T., Broder, M., Chang, E., February 2014. Effect of the 12-gene colon cancer assay results on adjuvant treatment recommendations in patients with stage II colon cancer. *Curr Med Res Opin* 30(2), 321–328.

47 Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., Wold, B., 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5, 621–628.

48 Stefano, G.B., 2014. Comparing bioinformatic gene expression profiling methods: micro-array and RNA-Seq. *Med Sci Monit Basic Res* 20, 138–142.

49 Micro-array Quality Control (MAQC). www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject (accessed September 7, 2015).

50 Agendia. www.agendia.com (accessed September 7, 2015).

51 Viale, G., Slaets, L., de Snoo, F.A., Bogaerts, J., Russo, L., van't Veer, L., Rutgers, E.J.T., et al., February 2016. Discordant assessment of tumor biomarkers by histopathological and molecular assays in the EORTC randomized controlled 10041/BIG 03-04 MINDACT trial breast cancer: intratumoral heterogeneity and DCIS or normal tissue components are unlikely to be the cause of discordance. *Breast Cancer Res Treatment* 155(3), 463–469.

52 Wong, H.R., 2012. Clinical review: sepsis and septic shock–the potential of gene arrays. *Crit Care* 16, 204.

53 Wan, M., Wang, J., Gao, X., Sklar, J., 2014. RNA sequencing and its applications in cancer diagnosis and targeted therapy. *North Am J Med Sci* 7(4), 156–162.

54 Feng, H., Qin, Z., Zhang, X., 2013. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer Lett* 340, 179–191.

55 McGettigan, P.A., 2013. Transcriptomics in the RNA-seq era. *Curr Opin Chem Biol* 17, 4–11.

56 Sequencing Methods Review. www.illumina.com/content/dam/illumina-marketing/documents/products/research_reviews/sequencing-methods-review.pdf (accessed September 7, 2015).

57 Li, X., Nair, A., Wang, S., Wang, L., 2015. Quality control of RNA-seq experiments. *Methods Mol Biol Clifton NJ* 1269, 137–146.

58 Au, K.F., 2015. Accurate mapping of RNA-Seq data, in: Picardi, E. (Ed.), *RNA Bioinformatics*. Springer, New York, pp. 147–161.

59 Dillies, M.-A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., et al., 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform* 14, 671–683.

60 Hansen, K. D., Irizarry, R. A., Wu, Z., April 1, 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13(2), 204–216.

61 Greaves, M., Maley, C.C., 2012. Clonal evolution in cancer. *Nature* 481, 306–313.

62 Zhou, Q., Xiong, Yuanyan, Huang, Xiao R., Tang, Patrick, Yu, Xueqing, Lan, Hui Y., December 9, 2015. Identification of genes associated with Smad3-dependent renal injury by RNA-Seq-based transcriptome analysis. *Scientific Rep* 5, 17901.

63 Renkema, K. Y., Stokman, M.F., Giles, R.H., Knoers, N. V. A. M., June 10, 2014. Next-generation sequencing for research and diagnostics in kidney disease. *Nat Rev Nephrol* 10(8), 433–444.

64 Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., Gilad, Y., 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18, 1509–1517.

65 Park, P.J., 2009. ChIP-Seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 10, 669–680.

66 Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10, 57–63.

67 Brennan, E.P., Morine, M.J., Walsh, D.W., Roxburgh, S.A., Lindenmeyer, M.T., Brazil, D.P., Gaora, P.Ó., et al., 2012. Next-generation sequencing identifies TGF-β1-associated gene expression profiles in renal epithelial cells reiterated in human diabetic nephropathy. *Biochim Biophys Acta* 1822, 589–599.

68 Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., et al., 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* 321, 956–960.

69 Schumacher, S., Muekusch, S., Seitz, H., 2015. Up-to-date applications of micro-arrays and their way to commercialization. *Micro-arrays* 4, 196–213.

70 Berry, M.P.R., Graham, C.M., McNab, F.W., Xu, Z., Bloch, S.A.A., Oni, T., Wilkinson, K.A., et al., 2010. An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466, 973–977.

71 Fehlbaum-Beurdeley, P., Sol, O., Désiré, L., Touchon, J., Dantoine, T., Vercelletto, M., Gabelle, A., et al., 2012. Validation of AclarusDx™, a blood-based transcriptomic signature for the diagnosis of Alzheimer's disease. *J Alzheimers Dis* 32, 169.

72 Sarwal, M., Sigdel, T., 2013. A common blood gene assay predates clinical and histological rejection in kidney and heart allografts. *Clin Transpl* 2013, 241–247.

73 Chen, R., Snyder, M., 2013. Promise of personalized omics to precision medicine. *Wiley Interdiscip Rev Syst Biol Med* 5, 73–82.

74 Codina-Solà, M., Rodríguez-Santiago, B., Homs, A., Santoyo, J., Rigau, M., Aznar-Laín, G., del Campo, M., et al., 2015. Integrated analysis of whole-exome sequencing and transcriptome profiling in males with autism spectrum disorders. *Mol Autism* 6, 21.

75 Huang, Y., Mucke, L., 2012. Alzheimer mechanisms and therapeutic strategies. *Cell* 148, 1204–1222.

76 Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., et al., 2015. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 43, D805–D811.

77 Fokkema, I.F.A.C., Taschner, P.E.M., Schaafsma, G.C.P., Celli, J., Laros, J.F.J., den Dunnen, J.T., 2011. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* 32, 557–563.

78 Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N.A., Gonzalez-Porta, M., Hastings, E., Huber, W., et al., 2014. Expression Atlas update—a database of gene and transcript expression from micro-array- and sequencing-based functional genomics experiments. *Nucleic Acids Res* 42, D926–D932.

79 Expression Atlas, EMBL-EBI. www.ebi.ac.uk/gxa (accessed September 7, 2015).

80 Zhang, Q., Yang, B., Chen, X., Xu, J., Mei, C., Mao, Z., September 24, 2014. Renal Gene Expression Database (RGED): a relational database of gene expression profiles in kidney disease. *Database 2014*, bau092.

81 Nephroseq. www.nephromine.org (accessed August 25, 2017).

82 Renal Gene Expression Database (RGED). http://rged.wall-eva.net/ (accessed August 25, 2017).

83 Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., et al., 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41, D991–D995.

Further Reading

Ecoli S.R.O.—PCR Diagnostics. www.pcrdiagnostics.eu (accessed August 25, 2015).

Human genome overview—Genome Reference Consortium. www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human (accessed September 7, 2015).

5

miRNA Analysis

Theofilos Papadopoulos[1,2], Julie Klein[1,2], Jean-Loup Bascands[3], and Joost P. Schanstra[1,2]

[1] *Renal Fibrosis Laboratory, Institut National de la Santé et de la Recherche Médicale (INSERM), U1048, Institute of Cardiovascular and Metabolic Disease, Toulouse, France*
[2] *Renal Fibrosis Laboratory, Université Toulouse III Paul-Sabatier, Toulouse, France*
[3] *Institut National de la Santé et de la Recherche Médicale (INSERM), U1188- DéTROI- Université de La Réunion, France*

MicroRNAs are short (18–23 nucleotides in length), noncoding, endogenous, single-stranded RNA molecules involved in posttranscriptional regulation of gene expression. Although discovered only two decades ago, these small RNA molecules are one of the "hottest" trends in biological research.

The discovery of the first miRNA took place in 1993. In that year, the two teams of Ambros and Ruvkun independently published complementary studies on a gene, called lin-4 in *Caenorhabditis elegans* [1, 2]. Both teams identified that the lin-4 gene produces a small transcript that does not encode a protein but negatively regulates the level of the LIN-14 protein involved in the development of *C. elegans* [1, 2]. The team of Ambros determined that the transcript of lin-4 was complementary to a repeated sequence in the 3′ UTR of lin-14 mRNA, proposing an antisense RNA–RNA mechanism of regulation [1]. In parallel, the team of Ruvkun reported that posttranscriptional control of lin-14 mRNA by the formation of duplexes in the 3′ UTR leads to the downregulation of lin-14 gene expression [2]. These two groups had identified the first miRNA!

Seven years passed until the identification of a second miRNA, let-7 [3]. In 2006, the Nobel Prize in Physiology or Medicine was jointly given to Fire and Mello for their research in 1998 on the discovery of the silencing mechanism of mRNA expression by small interfering RNA molecules, which are now known as the action mechanism of miRNAs. Since 2001, microRNAs or miRNAs have been shown to be involved in a variety of physiological functions and diseases and forced the scientific interest and research to evolve accordingly toward detecting and identifying these small molecules [4]. The continuously increasing number of reports on miRNAs, starting from some hundreds per year until 2007 and reaching approximately 7000 in 2013 and 2014 (PubMed search for "microRNAs" and "miRNA"), suggests their importance in physiology and disease. This increase is also the result of the more extensive use of next-generation sequencing (NGS) technologies in combination with simplified extraction, quantification, and expression analysis methods of miRNAs. Improved bioinformatics tools have also contributed to the number of newly identified miRNAs [5]. Today, the number of miRNA entries in miRBase (the first and largest database of miRNAs) in the last update of June 2014 (version 21.0) reached 28 645 miRNA entries when considering all species and 2 588 entries for human miRNAs only [6].

5.1 miRNA Biogenesis, Function, and Annotation

The biogenesis of an miRNA is initiated with the generation of a so-called primary miRNA (pri-miRNA) [7]. This transcript originates from specific noncoding genes or from introns of coding genes in the form of a cistronic or polycistronic transcript [8]. In the canonical pathway, the pri-miRNA exhibits a first cleavage at the flanking transcript sequence by the Drosha/DGCR8 complex and reforms to the precursor miRNA (pre-miRNA) with a characteristic hairpin loop structure (Figure 5.1). Alternatively, in the noncanonical pathway, the pre-miRNA is released from the introns of the host mRNA transcript and bypasses the Drosha cleavage. Next in the non-canonical pathway the mRNA transcript is processed by a complex called spliceosome to produce the "mirtron," and next, the mirtron refolds to the typical pre-miRNA hairpin form [8].

The next step includes the transfer of the pre-miRNA from the nucleus to the cytoplasm via exportin-5 in a Ran-GTP-dependent manner. Subsequently DICER and the HIV-1 transactivating response (TAR) RNA-binding

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

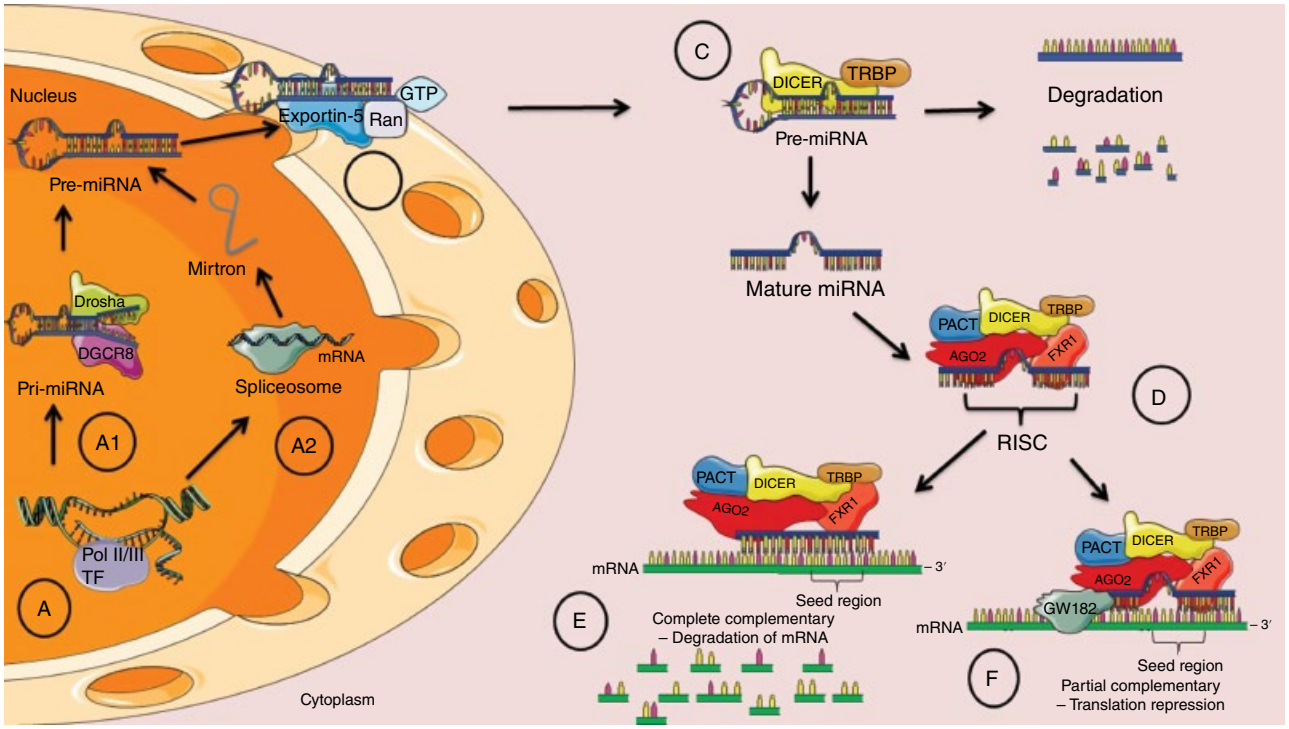


Figure 5.1 mRNA biogenesis and function. (A) In the first step of miRNA biogenesis, miRNA is transcribed similar to mRNA transcription with the help of polymerase II or III and various transcription factors followed by precursor miRNA (pre-miRNA) generation via a canonical and noncanonical pathway. (A1) **Canonical pathway**: miRNA is transcribed in a primary form (pri-miRNA with a hairpin loop) and then spliced by the Drosha/DGCR8 complex to form the pre-miRNA. (A2) **Noncanonical pathway**: A "mirtron" is generated from mRNA after splicing in a complex called spliceosome. Next, the mirtron reforms to the shape of a pre-miRNA. (B) The pre-miRNA is transferred from the nucleus to the cytoplasm with the help of exportin-5 in a Ran-GTP-dependent manner. (C) In the cytoplasm the protein DICER together with TRBP cleaves the pre-miRNA hairpin loop and releases the two strands. One strand will serve as the mature miRNA (18–22 nt length), while the other, in most cases, will be degraded. (D) The mature miRNA then binds to a protein complex including AGO2, forming the so-called RISC. AGO2 then leads the miRNA to the mRNA target. (E) If the miRNA sequence is completely complementary to the 3′ UTR of the mRNA, it binds and AGO2 cleaves the mRNA, completely stopping the translation of the protein. (F) If the miRNA sequence is partially complementary with the mRNA's 3′ UTR, GW182 is recruited in the RISC to aid with the connection and inhibit translation. AGO2, Argonaute 2; DGCR8, DiGeorge syndrome critical region 8; FXR1, fragile X mental retardation-related protein 1; PACT, protein kinase R-activating protein; RISC, RNA-induced silencing complex; TRBP, HIV-1 transactivating response (TAR) RNA-binding protein. The picture was designed using Servier Medical ART (http://smart.servier.fr/servier-medical-art).

protein (TRBP) bind and cleave the double-stranded pre-miRNA, releasing the mature single-stranded miRNA and loading it on the Argonaute2 (AGO2) protein [9]. The miRNA strand with the less stable paired 5′ end is preferentially loaded into the AGO2 protein [10]. From this point, the mature miRNA is active. The other single-stranded miRNA is usually degraded, but on some occasions it is also used in the same way as the mature single-stranded miRNA [11]. Next, the RNA-induced silencing complex (RISC) is formed. The full composition of this RISC is still unknown, but the most important protein is represented by AGO2 [8]. AGO2 belongs to the Argonaute (AGO) family composed of four member proteins. AGO2 binds directly to its small RNA partners and its function is similar to endonuclease-mediated cleavage of RNA [10]. Other proteins identified to be part in the RISC are DICER, TRBP, protein kinase R-activating protein (PACT), and fragile X mental retardation-related protein 1 (FXR1) [12, 13]. The RISC guides the miRNA to the mRNA target based on a 2–8 nucleotide sequence (also known as the seed) at the 3′ UTR of the target mRNA [14]. From this point in time, the miRNA initiates its biological function, which is to block the translation of the mRNA by two possible pathways: (i) in case of full complementarity of the miRNA and the 3′ UTR target, the mRNA is cleaved [9], or (ii) in case of partial complementarity of the miRNA and 3′ UTR target sequence, the GW182 protein is recruited to the RISC interacting with AGO2 to aid in the target identification, leading to translation repression [15]. In any case the result is inhibition of gene expression. Interestingly, one miRNA gene can lead to the generation of more than one mature miRNA, and on the other hand, different miRNA genes can be clustered into so-called miRNA families based on the miRNA sequence [16].

5.2 Annotation of miRNAs

The nomenclature of miRNAs is complex. In 2003 Ambros et al. proposed guidelines for newly identified miRNA annotation [16]. These rules were applied later to miRBase and are followed until today for every new miRNA identified. In brief [17],

- mir and miR: A small (r) represents the pre-miRNA, while the (R) represents the mature miRNA, for example, mir-15 is the pre-miRNA-15 and miR-15 the mature miRNA.
- The three letters in front of the miR represent the origin of the species in which the miRNA was found, for example, hsa (*Homo sapiens*) and mmu (*Mus musculus*).
- The numbers to register the miRNAs are ascending. Different numbers are given when an miRNA has significant sequence differences. The accession number of each miRNA is the only unique identifier. Lin-4 and let-7 (lethal-7) are an exception for historical reasons.
- The same miRNA in different species receives the same number to preserve homology among the database, for example, hsa-miR-16-5p (human) is the ortholog of mmu-miR-16-5p (mouse).
- Sequences with one or two different nucleotides are assigned with the same number but an additional letter to distinguish, for example, hsa-miR-15a and hsa-miR-15b.
- If the same miRNA is found in different loci of a chromosome, the difference is at the pre-miRNA level, and a number is added to distinguish those, for example, hsa-mir-16-1 and hsa-mir-16-2.
- miR and miR*: The mature microRNA found from one arm of the hairpin is usually much more abundant than that found from the other arm [7], in which case an asterisk following the name indicates the mature species found at low levels from the opposite arm of a hairpin. Recent studies however suggested that not in all cases the mature miRNA is released from the same arm [18], and as a result the suffix -5p or -3p is added to define from which arm of the pre-miRNA the mature is originating.
- If miRNAs share the same sequence over a stretch of 2–8 nucleic acids, they are derived from the same precursor and belong to the same cluster generating an miRNA family; for example, the mir-15 family consists of miR-15a and miR-15b sequences, as well as miR-16-1, miR-16-2, miR-195, and miR-497.

5.3 miRNAs: Location, Stability, and Research Methods

5.3.1 miRNA Analysis and Tissue Distribution

Since miRNAs are basically transcribed the same way as mRNAs, miRNAs are expected to be expressed in every tissue. Numerous reports, using a variety of molecular biology methods, provide information about the expression levels of miRNAs in different tissues [19, 20]. In addition, most methods for extraction, quantification, and expression of miRNAs are highly similar to the methods used for DNA and mRNA analysis (real-time qRT-PCR, microRNA arrays, *in situ* hybridization, bead-based profiling, and NGS) [21, 22]. In addition, *in silico* analysis has been shown to be a useful tool for the prediction of the localization and tissue specificity of miRNAs by collecting information from many different tissues [23]. This analysis led to the suggestion that miRNAs distribute unequally in tissue and that some miRNAs can be listed as "tissue specific" and can be used as identifiers for these tissues, for example, miR-122a for the liver, miR-1 and miR-133a for the heart and skeletal muscle, miR-9 for the brain [23], and miR-192 and miR-194 for gastrointestinal organs and kidney [19].

5.3.2 miRNAs in Body Fluids

Since miRNAs are involved in many biological pathways and functions and found in all tissues, it is a reasonable assumption that they can also be found in body fluids and can represent changes in adjacent tissues. Indeed, in 2009 Hanson et al. evaluated for the first time miRNA expression in five dried relevant body fluids including blood, saliva, semen, vaginal secretions, and menstrual blood. They were able to identify a number of miRNAs that could be used to distinguish between different fluid samples by monitoring differential expression levels of these miRNAs in these samples. These miRNAs are miR-451 and miR-16 for blood, miR-135b and miR-10b for semen, miR-658 and miR-205 for saliva, miR-124a and miR-372 for vaginal secretions, and miR-412 (in combination with miR-451) for menstrual blood [24]. Further studies by Weber et al. showed that miRNAs can be found in at least 12 different body fluids including milk, colostrum, saliva, seminal fluid, tears, urine, amniotic fluid, bronchial lavage, cerebrospinal fluid, plasma, pleural fluid, and peritoneal fluid [25]. Other reports followed to confirm these findings and also suggest additional miRNAs as identifiers of these body fluids, such as miR-185 and miR-214 for menstrual blood; miR-20a, miR-126, and miR-486 for blood; miR-943, miR-135a, miR-888, and miR-891a for semen; miR-583, miR-200c, miR-203, miR-205, and miR-138-2 for saliva; and miR-617, miR819a, and miR-124a for vaginal secretions [26–28].

But how do these miRNAs end up in body fluids? MiRNAs are secreted from the cells in which they are produced and packed in microparticles such as microvesicles, exosomes, or apoptotic bodies and are also attached to RNA-binding proteins (RBPs) or lipoprotein complexes

explaining their presence in body fluids [29] (Figure 5.2). Exosomes are the best studied miRNA vehicles. These vacuoles transfer cellular components between cells and promote cell-to-cell communication and interactions [30, 31]. Studies have shown the existence of exosomal transfer of functional mRNAs and miRNAs from one cell to another. One of the strongest arguments for the existence of this phenomenon is the study of Valadi et al. in which exosomes secreted by mouse mast cells transferred into human mast cells produced mouse proteins in the human cells [32]. Except being the means of transport, exosomes, similar to the other miRNA carriers and binding proteins, also act as "bodyguards" of microRNAs by protecting them from RNAse activity (see following text). Studies reported that exosomal miRNAs are intact and protected from degradation rather than cell-free miRNAs in, for example, urine, which are more likely to be degraded by RNAses [33, 34].

As mentioned, miRNAs can be protected and transported by binding to proteins, allowing circulation in body fluids and cell-to-cell communication. In particular, RBP AGO2 (Figure 5.2) was shown to be involved in miRNA transport in the circulation, which potentially leads to the transport of a functional miRNA-induced silencing complex [35]. Another RBP with a similar action as AGO2 is nucleophosmin 1, which was found to carry miRNA in serum in humans and is possibly involved in cell-to-cell communication [36]. High-density lipoprotein (HDL) lipoproteins, surprisingly, were found to have a role in the transport of miRNAs in a similar way as RBPs [37]. Finally, circulating apoptotic cells seem to contribute to miRNA transport as well. MiRNAs are released in the circulation after apoptosis bound to the previously mentioned RBPs and can be absorbed by other cells and perform their function [38, 39].

MiRNAs can be found in urine. These urinary miRNAs are attached to RBPs or contained in exosomes, making the urinary pool of miRNAs potentially suitable for detection and monitoring of both renal and nonrenal diseases [34, 40]. One point that needs to be highlighted is that miRNAs in urine are less abundant than in plasma or serum, and this is most likely due to higher RNAse activity in urine [34]. The urinary miRNAs, as potential biomarkers or pointing to potential therapeutic targets, could be useful in the treatment and the detection of the major and constantly rising worldwide health problem of kidney diseases [41]. The origin of urinary miRNAs is still not fully elucidated. Urinary miRNAs are likely to be shedded from cells all along the urinary pathway [38, 42]. Also, miRNAs could potentially be filtered from the

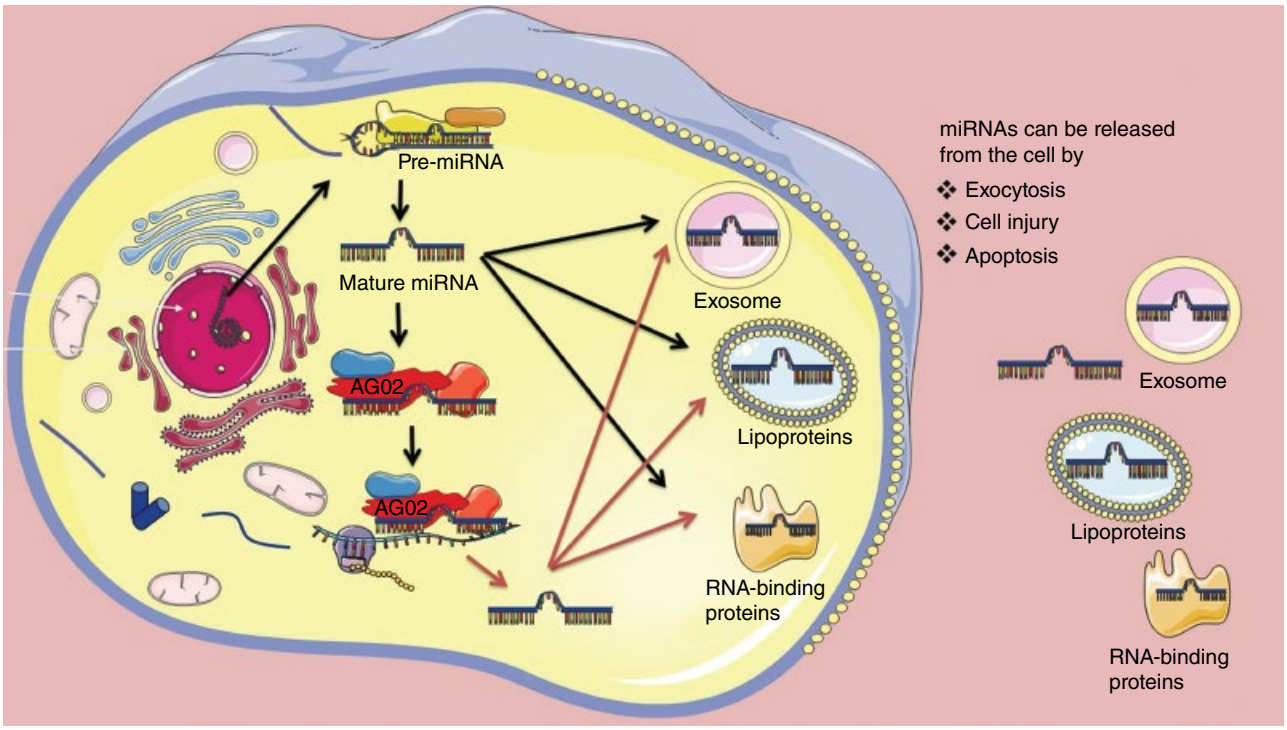


Figure 5.2 miRNA-transporting molecules. During biogenesis or even after fulfilling their biological function, miRNAs can be incorporated into exosomes or bound to lipoproteins (e.g., HDL) or RNA-binding proteins (e.g., AGO2). Release of the encapsulated or bound miRNAs can occur via exocytosis during cell lysis and apoptosis. Encapsulation or binding to these proteins most likely protects the miRNAs from degradation in the hostile biofluid environment. The figures were designed using Servier Medical ART (http://smart.servier.fr/servier-medical-art).

plasma, although the exosome size (30–100 nm) is at the limit value for glomerular filtration (fenestrations' diameter is 60–80 nm) [43, 44]. A recent study has shown that exosomes can cross physiological barriers, such as the brain–blood barrier, despite their large size [45].

The role of urinary miRNAs is still unknown. Either urinary miRNAs can be considered as waste or miRNAs could use urine as a vehicle to move through the urinary tract and function in areas downstream from their site of production. Further research to answer these questions is needed.

5.3.3 Stability of miRNAs

Due to their small size and the protective mechanisms described earlier, miRNAs are quite stable [46, 47]. The stability of the intracellular mature miRNA has been shown to be regulated by the thermostability of the AGO2/miRNA interaction [48–51].

For research purposes, it is possible to extract miRNAs from biological samples stored at least up to one year under laboratory conditions (relatively constant humidity and ambient temperature, no UV exposure, and dust-free). MiRNAs extracted from such samples show no signs of degradation [28]. Plasma and serum miRNAs are stable for 24 h at room temperature and resist to eight cycles of freezing/thawing [52]. Further stability studies on plasma showed that exosomal miRNAs are more stable compared with plasma mRNA when storing the samples at 4, −20, and −80°C for 2 weeks, 2 months, 3 years, and 5 years [53]. Similar observations hold for formalin-fixed paraffin-embedded (FFPE) tissue samples, suggesting that miRNAs are more suitable than mRNAs as potential biomarkers [54]. This stability is an important feature of miRNAs over mRNAs because in clinical practice the availability of sample storage at −20°C is significantly higher compared with −80°C.

Another field that needs more detailed investigation is the stability of urinary miRNAs, either free, protein-bound, or in exosomes. Few studies have examined this topic. One of the first studies that included the investigation of the stability of miRNAs in urine was from Yun et al. They tested whether cell-free miRNAs in urine are able to be used as prognostic and diagnostic biomarkers for bladder cancer. They showed that after seven freeze/thaw cycles or after storing urine at room temperature for 3 days, miRNAs showed only minimal signs of degradation [55]. Other studies demonstrated successfully the stability of urinary miRNAs after, up to 10 freeze/thaw cycles, different storage temperatures between 4 and −80°C for short and long periods of time (between 5 days and 2 years) [56, 57]. This apparent stability of urinary miRNAs opens the window for multiple applications of miRNAs in research and in clinical practice.

Overall, the stability of miRNAs and their presence in nearly all tissues and body fluids have placed miRNAs in the center of attention. Circulating miRNAs have become the study material of choice as possible biomarkers and therapeutic targets of disease.

5.3.4 Methods to Study miRNAs

5.3.4.1 Sampling

High-quality miRNAs can be extracted from a wide range of cells and tissue samples, such as cell lines and fresh or FFPE tissue samples as well as any kind of body fluid (plasma, serum, urine, etc.). Even if it was estimated that the miRNA fraction is about approximately 0.01% of the total RNA mass, the technological progress, together with the aforementioned stability of the miRNAs, yields in general enough miRNA material and of high quality for downstream studies (reverse transcription polymerase chain reaction (RT-qPCR), microarrays, NGS, etc.) [58].

5.3.4.2 Extraction Protocols

miRNA extraction is quite straightforward, with minor protocol modifications for different tissues and body fluids. The progress in technology reduced the amount of starting material needed to perform miRNA expression analysis from 50 ml of urine or plasma to 0.1–0.5 ml. For example, as little as 200 µl of urine can be used for (semi)quantitative RT-PCR analysis [59]. The available commercial miRNA extraction kits are suitable for all biological samples following similar protocols and offer two options for the isolation of the miRNAs from total RNA after the use of a lysis buffer and protein denaturation using guanidinium isothiocyanate: (i) The first option includes the use of affinity columns in which miRNAs bind to the stationary phase, while larger molecules like DNA or RNA and impurities (proteins, lipids) are washed out and purified miRNAs are eluted from the columns. In this category the available commercial kits are "miRNeasy" from Qiagen (one of the most used), "Norgen" from Biotek Corp., "miRCURY" from Exiqon, and "mirVana" of Life Technologies. (ii) The second option is to perform RNA precipitation using organic reagents (ethanol, isopropanol) and in some cases acidic phenol/chloroform, leading to the partitioning of miRNA into aqueous supernatant in order to separate the RNA from the DNA and the other molecules ("TRIzol" from Life Technologies and "Epicentre" from Illumina) [59]. The resulting miRNA samples can be diluted in water with diethyl pyrocarbonate or elution buffer (most usually Tris-EDTA buffer) and stored at −20 or −80°C. The extracted miRNA represents approximately 0.01% of the total RNA [58].

5.3.4.3 miRNA Detection Techniques

The techniques that are being used in miRNA expression analysis are similar to the techniques used in gene expression analysis with some modifications to adjust to the size and quantity of miRNAs.

Initially, miRNA detection and study was carried out using Northern blot analysis [1]. In Northern blotting, miRNAs are separated by electrophoresis, transferred to nitrocellulose membrane, and visualized with ^{32}P-labeled DNA probes complementary to the miRNAs (e.g., representing the mRNA target sequences). This technique enables a quantitative assessment of mature miRNAs—pri-miRNAs and pre-miRNAs—and also the complexes of the miRNAs with Drosha, DICER, and RISC due to their different electrophoretic motilities [60–62]. The drawbacks of this approach include low sensitivity in the nM–pM range [63], low throughput, and the requirement for high miRNA input (around 5–50 μg of total miRNA) in order to get a signal [64]. A major issue in Northern blotting is the use of radioisotopes that require careful handling and generate radioactive waste. Efforts have been made to change the radioisotopes with user and environmentally friendly reagents, for example, digoxigenin (DIG)-labeled oligonucleotide probes containing locked nucleic acids (LNA) [65] or carbodiimide-mediated cross-linking [64], but in any case the sensitivity is lower than with radioisotopes.

The gold standard technique in miRNA research is the quantitative RT-qPCR. RT-qPCR is well known and used for many years in mRNA research and brings the advantages of sensitivity, low cost, precision, reproducibility, and simplicity to the miRNA research [66]. The difficulty RT-qPCR faces in the study of miRNAs is their small size. The mature miRNAs with length 18–22 nt have approximately the size of the primers in an mRNA PCR. To deal with this problem, there are two kinds of RT primers for the generation of cDNA from miRNA: miR-specific primers and universal primers. In the first case, the primers have a stem-loop structure in their 5′ end and an antisense sequence on the 3′ end against the miRNA of interest. This method, except capturing only the miRNA of interest, distinguishes the mature miRNA from pri- and pre-miRNA [58]. The universal primer method uses either polyadenylate polymerase (PAP), which adds a poly-A tail to the extracted miRNA, or T4 ligase, which adds a sequence to the extracted miRNA and then follows the normal RNA procedure: the synthesis of the cDNA together with the universal RT primers [58]. Next, qPCR is as common and straightforward as for cDNA generated from mRNA with the use of TaqMan or SYBR Green probes. In order to detect specifically the mature miRNAs, the use of LNA-modified primers for qPCR is required (like miRCURY LNA qPCR platform from Exiqon) [21]. Several manufacturers, including Applied Biosystems, Exiqon, Fluidigm, and SA Biosystems, offer qPCR kits that can assess hundreds of microRNAs in parallel, and some offer customizable assays [67].

The major issue right now in RT-qPCR analysis of miRNAs is the lack of standardized normalization methods. Most groups use as housekeeping genes noncoding small nuclear (sn)RNA (U6, SNORDs) or miRNAs that are reported to be expressed at a stable level within the tissue or body fluid under study [68]. RT-qPCR will remain for many years to come as the easiest method of detecting and validating miRNAs in samples after large-scale nontargeted screening using microarrays or NGS.

One of the technologies that made a major contribution in miRNA research, and still does, is microarrays. Microarray platforms are designed with fixed probes (e.g., representing the mRNA target sequences) capable to capture labeled miRNAs via hybridization, enabling the detection of miRNAs present in a sample by fluorescence [69]. The relatively low cost, reproducibility, sensitivity, and specificity are some of the advantages of microarrays. In addition, this technology was one of the first to allow the comparison of the relative expression of multiple miRNAs simultaneously in a semiquantitative manner (e.g., healthy vs. disease). However the shortcomings are that the fixed probes do not allow the detection of novel miRNA molecules and that the hybridization between the probes and the labeled molecules is not always efficient due to low affinity and needs an additional validation step to confirm the findings using, for example, RT-qPCR [21]. The introduction of LNA into the microarray probes has partially solved this hybridization problem, setting an equal melting temperature (Tm) among the probes [68]. Microarray protocols require enzymatic or chemical labeling of the samples. In order to remove the dye labeling bias and differences due to hybridization bias and scanning, it is necessary to remove background signals and perform noise correction and signal normalization [58]. Different methods of background correction and normalization have been proposed: global mean/median, linear loess, robust linear loess, quadratic loess, robust quadratic loess, rank invariant, and the most frequently used and stable quantile normalization [70]. Advice from the experts is suggested to choose the most suitable method on a case-by-case basis. Label-free assays like stacking-hybridized universal tag (SHUT) [71] and antibody-like protein-based assay with PAZ domain [72] have been introduced in an attempt to minimize the problems derived from labeling and simplify the data analysis process. Freely available packages in "R-Bioconductor" are available for microarrays analysis based on the "limma" package. Also, another free platform, Gene ARMADA [70], offers a user-(biologist-) friendly environment for microarray data processing for users not versed in programming language. Besides the

drawbacks discussed earlier, microarrays will remain as a reliable technique for analyzing and comparing multiple miRNAs in a fast, simple, and cost-effective manner. It is important to note that this technology is a technology used for discovery of miRNAs associated with a specific condition and the results from microarrays should always be validated before inferring biological conclusions.

Nowadays, NGS is becoming the leading technology in miRNA research. The ability to detect with high sensitivity known and unknown miRNAs, without relying on prefixed probes, and isoforms of the miRNAs (pri- or pre- or mature miRNAs), without the need of labeling, places NGS on the top of the list to use in research [73]. However, NGS demands highly trained personnel for performing the experiment and the downstream data analysis. The data produced are large and complex, but as the technology progresses, novel reliable analytical packages including "DEseq" from "R-Bioconductor" become available. The cost of NGS is also an issue but is getting closer to the microarray costs [21]. RNA-seq (as deep sequencing NGS for RNA is called) bases the detection and the calculation of the abundance of an RNA species on the percentage of total "reads" (or, in some cases, total mappable reads) obtained in a sample. A read is being defined as a data string of A, T, C, and G bases corresponding to a fragment of the sample DNA. The higher the number of reads of a specific DNA fragment, the higher its abundance in the sample. The comparison between different samples is made by examining the overall frequency distribution of miRNA reads between the samples. The difference in the distributions represents the differential expression results that, as in microarray experiments, need to be validated with RT-qPCR [69]. Currently, Illumina's HiSeq 2500 and SOLiD are the leaders in NGS technologies. Illumina technology is based on the "sequencing-by-synthesis" method: a signal is detected every time a nucleotide is attached to the under construction single-stranded cDNA [21, 73]. NGS technologies are the present and the future of miRNA research (and in general for genomics and transcriptomics) and will surely massively expand the available data and our understanding of miRNA biology once they become more accessible in terms of cost and data analysis.

New technologies are being developed for miRNA detection that can potentially complement the "traditional" techniques. Such techniques are called biosensor techniques (BTs). The difference between the use of BTs and the aforementioned technologies is the rapid and highly sensitive results from complex samples such as blood and urine, through the use of a selective molecular probe, avoiding any need for polymerase-based amplification steps [63]. A biosensor is defined as a device for quantitative analytical information via the use of a biorecognition element in direct contact with a transduction element [74]. A DNA probe of complementary sequence to the miRNA target is the selective biorecognition element in miRNA biosensors, and the DNA probe hybridizes with the miRNA target through changes in a measurable output signal [63]. Various miRNA biosensor designs exist based on alternative transduction mechanisms, including electrochemical [75], electromechanical [76], and optical-based detection [77], each of which has achieved femtomolar sensitivity levels to date, with multi-log dynamic range and short time to results. Although different in appearance, they share the defining feature of a biosensor, an integrated molecular recognition agent and transduction element [63].

One final technology worth mentioning is NanoString nCounter system [78]. This method was firstly designed to capture mRNA transcripts, but now miRNA sequences can also act as a guide for adjacent hybridization of a sequence-specific capture probe. This probe is labeled with biotin and a reporter probe labeled with a unique four-color, seven-position barcode. The hybridized constructs are purified and bound to a streptavidin-coated slide. Afterward, a voltage is applied to elongate the molecules, which allows for the digital imaging and counting of the uniquely barcoded miRNA targets.

All technologies have their advantages and disadvantages, and a user must take under consideration all parameters before choosing the most suitable one. Sensitivity, cost, time to results, and data processing are some of the most important parameters to consider. Even though NGS is winning the race of the most robust method for detection and discovery of new miRNAs, validation of findings with other technologies is currently mandatory. Table 5.1 summarizes the features of the most important and popular detection techniques.

5.3.4.4 Data Processing and Molecular Integration

The difficulty in correlating miRNA expression with mRNA targets for clinical applications is the fact that miRNAs in physiological and pathological conditions control more than one mRNA (tens, even hundreds) and that one mRNA can be controlled by more than one miRNA [79]. The major issue in miRNA studies is to identify the targets of the miRNAs and narrow the options to the most important and relevant to the specific case.

Because the *in vitro* and *in vivo* discovery of the real-life targets of miRNAs is a challenging and time-consuming procedure, web tools have been designed to tackle this problem and produce lists of the possible targets. The first bioinformatic approach on this subject was developed by Eric Lai in 2002 [80]. Lai compared the sequence of a subset of miRNAs with the K box and Brd box motifs that were previously shown to mediate negative posttranscriptional regulation. He determined that a

Table 5.1 miRNA detection methods.

Method	Principle	Advantage	Disadvantage	Variants
Northern blot	Visualize of the miRNA by electrophoresis with ^{32}P-labeled DNA probes	• Quantitative detection of mature miRNA, primary miRNA and precursor miRNA and miRNA complexes with Drosha, DICER and RISC	• Low sensitivity • Low throughput • Need of high RNA input	(DIG)-labeled oligonucleotide probes with LNA EDC-mediated cross-linking LED DSLE
RT-qPCR	The extracted miRNAs are converted to the complementary DNAs generating cDNA followed by the PCR protocol with specific primers of interest	• Sensitivity • Specificity • Reproducibility • Low cost • Simplicity • Fast	• Lack of standardized normalization molecules • Depends on purity and quality of RNA input • Demanding design of the primers to deal with the small miRNA size	• miRNA specific or universal primers • TaqMan probe or Sybr Green
Microarrays	Hybridization of the target miRNAs to the complementary immobilized probes and detection via fluorescence	• Parallel analysis of hundred miRNAs in a single sample • Easy, straightforward and automated analysis, • Short turn-around time • Suitable for comparison between two conditions	• Prefixed probes • Not possible to detect new molecules • Potential hybridization bias • Semi-quantitative method • Limited dynamic range	• -5′ hairpin • SHUT • Label-free PAZ-dsRBD method • LASH
Next generation sequencing (NGS)	Massively parallel sequencing of millions of fragments of DNA from a single sample	• Suitable for detection of new molecules and miRNA heterogeneity, • Increased dynamic range and sensitivity	• Size of data files • Complex data analysis • Long time • Costly	• Semiconductor sequencing (Life Technologies-Ion Torrent) • Pyrosequencing (Roche—454) • Sequencing by ligation (Life Technologies—SOLiD) • Reversible terminator—sequence by synthesis (Illumina -Solexa) • Single-molecule real-time DNA sequencing by synthesis (Pacific biosciences—PacBio)
Biosensor techniques	A biorecognition element (a DNA probe complementary to the miRNA of interest) is in direct contact with a transduction element	• High sensitivity • Very fast • Label-free protocols	• Hybridization bias • Not useful for multiplexing • Mass transfer challenge • Reliability of measurements	• Electrochemical • Electromechanical • Optical-based
NanoString nCounter	Molecular “barcodes” and microscopic imaging are used to detect and count up to several hundred unique transcripts in one hybridization reaction	• Multiplex detection • Direct digital detection • No need for amplification • High specificity	• Semi-quantitative • Not suitable for detection of new molecules • limited dissemination of the instrument • Need for increased open-source software tools for data analysis	None

DIG, digoxigenin; EDC, 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide; DSLE, DIG-labeled, splinted-ligation and EDC cross-linking method; LASH, ligase-assisted sandwich hybridization; LED, LNA modified probes, EDC crosslinking and DIG-labeled); LNA, locked nucleic acids; PAZ, Piwi/Argonaute/Zwille; RBD, RNA binding protein; RT-qPCR, reverse transcription—quantitative polymerase chain reaction; SHUT, stacking-hybridized universal tag.

series of eight nucleotides in the beginning of an miRNA had perfect complementarity to the motifs, concluding that this sequence is responsible for the posttranscriptional regulation mediated by miRNAs. This eight-nucleotide sequence is now known as the seed of an miRNA and is the main area responsible for miRNA function. Nowadays the algorithms in the available web tools that predict possible miRNA targets focus their search on four basic parameters:

1) Seed matching: As mentioned earlier, the seed matching refers to the complementarity of the first 2–8 nucleotides of an miRNA in the 5′ end to the 3′ UTR of the mRNA target. It is the basic parameter that most (if not all) tools take under consideration. But perfect complementarity is not always observed. The main types of seed matching include 6 mer (complementary for 6 nucleotides), 7 mer (complementary for 7 nucleotides), and 8 mer (perfect complementary) [81].
2) Conservation of a sequence across different species: The maintenance of the seed region of an miRNA among species enforces the proof of the existence of the seed region-miRNA and reduces the number of false positive predictions based on the seed sequence [82].
3) Thermodynamic stability of the miRNA–RNA duplex: The free energy (Gibbs energy, ΔG) is a measure of the stability of a system. Highly negative ΔG values characterize very stable RNA complexes. By predicting the ΔG between an miRNA and its candidate target, it is possible to predict the stability of the duplex and conclude if this complex can form [81]. Also, the energy needed for the unfolding of the secondary structure of the mRNA allowing the accessibility to the miRNA [83] can be an additional parameter to consider [81].
4) Multiple target sites in the 3′ UTR of the mRNA and possible sites in the coding regions: Studies have shown that in the 3′ UTR of an mRNA, there are multiple target sites for one miRNA [84]. The different algorithms allow analysis of mRNAs with multiple predicted sites for the same miRNA, leading to more reliable results.

Many web tools are available, each with a different algorithm for calculating the predicted targets of miRNAs, recently reviewed in Ref. [81]. One of the first web tools to identify miRNA targets is miRANDA [85]. Even if this tool lacks updates (according to the website), it is still very useful. Using a machine learning approach (mirSVR) [86], miRANDA suggests the targets but also scores them according to the effect an miRNA may have on the target. Another frequently used web tool (and one of the first to predict human targets) is DIANA-microT-CDS [87]. DIANA provides information on the predicted target location (chromosome, location of the binding point on the transcript), binding type according to the number of nucleotides matching (8 mers, 7 mers, etc.), a correlation score, degree of conservation among species, and links to databases such as Ensembl, miRBase, and PubMed. Finally, a third popular targeting web tool is TargetScan [88–91], which is easy to use and actively maintained.

Except for web tools that predict miRNA targets, the existence of three databases that contain information on experimentally validated miRNA targets is worth mentioning. These databases contain information extracted from citations where a well-documented interaction of an miRNA–mRNA target using methods including luciferase assays, NGS, microarrays, and qPCR Western and Northern blot was used **26286669**. MirRecords, last updated in 2013, hosts 2705 records of interactions between 644 miRNAs and 1901 target genes in 9 different species. Among these records, 2028 were curated from "low-throughput" experiments [92]. miRTarBase last updated in September 2015 includes approximately 5000 articles that describe around 3700 miRNAs and 366.000 miRNA interactions validated with reporter assays, Western blots, NGS, microarrays, and other technologies (the 348.000 interactions are the result of NGS technologies, like CLIP-seq, and are considered as "less strong evidence"). The search for an miRNA and its interaction in MiRTarBase offer multiple options and information on the miRNA and the target gene, how the validation was performed, and the corresponding publications as well as information on the expression, sequences, and molecular networks [93]. StarBase is a database for decoding Pan-Cancer and Interaction Networks of long noncoding RNAs (lncRNAs), miRNAs, competing endogenous RNAs (ceRNAs), RBPs, and mRNAs from large-scale Cross-linking immunoprecipitation (CLIP-Seq) data originating from HITS-CLIP, PAR-CLIP, iCLIP, and CLASH on more than 6000 samples and 14 cancer types [94, 95]. Finally, TarBase is part of the DIANA Tools project. TarBase v7.0 provides more than half a million miRNA–gene interactions curated from published experiments on 356 different cell types from 24 species with detailed metadata. DIANA-TarBase v7.0 shows information about positive or negative experimental results, the utilized experimental methodology, and experimental conditions including cell/tissue type and treatment and also information on the binding site location, as identified experimentally as well as *in silico*, and the primer sequences used for cloning experiments [96]. The combination of the prediction tools with the experimentally validated targets databases offers in general a good starting point in evaluating new datasets. Still, the experimental validation of the possible miRNA targets remains the most reliable proof of direct connection between miRNA and mRNA. Tables 5.2 and 5.3

Table 5.2 Available web-tools for miRNA target prediction.

Web-tool name	Description	Category	Organisms	Calculation features
miRanda, http://www.microrna.org/	Detects the maximum complementary between the 3′-UTR and the miRNA, together with the binding energy and of the mRNA-miRNA duplex and the evolutionary conservation	Seed-based	All	Seed match, conservation, and free energy
miRanda-mirSVR, http://www.microrna.org/	Provides a score to represent the effect of a miRNA on the expression of the target gene	Seed-based	Humans, rats, mice, flies, and worms	Seed match, conservation, free energy, site accessibility
TargetScan, http://www.targetscan.org	Identifies the 8 mer, 7 mer and 6 mer sites plus mismatches in conserved pairing at 3′-UTR and centered sites. Scores inform about various features of the miRNA-mRNA binding	Seed-based	Mammals, flies, and worms	Seed match and conservation
DIANA-microT-CDS, http://www.microrna.gr/microT-CDS	Recognises miRNA target in both 3′-UTR and coding sequences (CDS). Provides a score of prediction confidence, together with multiple information on the target site and binding type	Seed-based/ machine learning	Humans, mice, flies, and worms	Seed match, conservation, free energy, site accessibility, target-site abundance
MirTarget2 or miRDB, http://mirdb.org	Predicts miRNA targets by learning the miRNA-mRNA binding features from high-throughput sequencing experiments	Machine learning	Humans, mice, rats, dogs, and chickens	Seed match, conservation, free energy, site accessibility
RNA22-GUI, https://cm.jefferson.edu/rna22v1.0/	Uses patterns to identify target islands on the on the mRNA target based on the conserved miRNA sequence and free energy of the predicted bond	Pattern-based	Humans, mice, flies, and worms	Seed match and free energy
TargetMiner, http://www.isical.ac.in/~bioinfo_miu/targetminer20.htm	The prediction is based on a learning process via positive and negative miRNA—mRNA datasets	Machine learning	Any	Seed match, conservation, free energy, site accessibility, target-site abundance
PITA, http://genie.weizmann.ac.il/pubs/mir07/	The prediction is based on the accessibility of the target site	Target structure	Humans, mice, flies, and worms	Seed match, conservation, free energy, site accessibility and target-site abundance
RNAhybrid, http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	Based on favourable hybridization sites avoiding intramolecular duplexes	Thermodynamics	Any	Seed match, free energy, target-site abundance
miRGate, http://mirgate.bioinfo.cnio.es/miRGate/	A database of miRNAs and mRNA target sites which calculates the prediction by the combination of other target prediction tools (miRanda, PITA, RNAHybrid, Microtar, TargetScan)	Multiplex approach	Human, mice, rats	Seed match, conservation, free energy, site accessibility, target-site abundance
miRWalk 2.0, http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/index.html	Hosts miRNA—target interactions based on seed matching and offers a comparison between 13 other available prediction web-tools	Seed-based	Human, mice, rats and all transcripts and mitochondrial genomes	Seed match and conservation

Table 5.3 Available web-tools for miRNA target validation.

Web-tool name	Species included	Description
mirRecords, http://c1.accurascience.com/miRecords/	9 species including human	Contains manually curated experimental evidence for 2705 records of interactions between 644 miRNAs and 1901 target genes
StarBase, http://starbase.sysu.edu.cn/	14 cancer types (human, mouse, *C. elegans*)	Large-scale CLIP-Seq data originating from HITS-CLIP, PAR-CLIP, iCLIP and CLASH on more than 6000 samples
DIANA—TarBase, http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index	24 species including human	High quality manually curated experimentally validated miRNA : gene interactions, enhanced with detailed meta-data.
miRTarBase, http://mirtarbase.mbc.nctu.edu.tw/index.php	18 species including human	Data for experimental miRNA—target interactions collected from literature from reporter assays, western blot, northern blot, RT-qPCR, microarrays, SILAC, NGS
miRWalk 2.0, http://www.umm.uni-heidelberg.de/apps/zmf/mirwalk/index.html	15 species including human	Experimentally verified miRNA interaction information associated with genes, pathways, organs, diseases, cell lines, OMIM disorders and literature on miRNAs. Around 668.000 interactions
PhenomiR 2.0, http://mips.helmholtz-muenchen.de/phenomir/	Human and mice	Manually curated database with information about differentially regulated miRNA expression in diseases and other biological processes
miR2Disease Base, http://www.mir2disease.org/	Human	A manually curated database that provides a comprehensive resource of miRNA deregulation in various human diseases. Hosts 349 miRNAs related to 163 diseases

display a list of prediction and validated target web tools with their most prominent features.

5.3.4.5 *In Vitro* Target Validation

Currently, *in vitro* studies are the gold standard for validating miRNA targets and monitoring any changes of the target gene's expression due to the action of the miRNA. The principle behind these cellular assays is that loss or gain of function of specific miRNAs is linked to a corresponding change in the expression of a potential target. This is achieved mainly through artificial manipulation of the miRNA concentration with the use of either double-stranded miRNA mimics or miRNA inhibitors and afterward by investigating the effects on the expression of possible targets of these miRNA sequences. MiRNA mimics are artificial, chemically modified miRNA-like small RNAs with the ability to mimic the function of an miRNA guide strand while bypassing the maturation steps of Drosha and DICER of endogenous miRNAs and causing a rapid decrease in the expression of the potential target mRNA [97]. On the other hand, miRNA inhibitors act by binding to mature miRNAs and block their activity. Figure 5.3a and b summarizes the actions of antagomirs and miRNA mimics. The most used and promising miRNA inhibition approaches are anti-miRs and miRNA sponges [98, 99]. Both bind to the miRNA target, preventing it from connecting to the mRNA in the RISC and thus enabling the translation of the mRNA. Cross-linking immunoprecipitation (CLIP)-based approaches study the interaction of miRNA–mRNA–AGO protein and allow to verify the effect of over- or underexpression of the miRNAs on their targets [100].

One of the first methods described was based on co-immunoprecipitation of tagged AGO protein (the main protein of the RISC that connects to the miRNA–mRNA duplex) with miRNAs and mRNAs [101–103]. Quantitative analysis of miRNAs and mRNAs with microarrays or NGS after treating cells with miRNA mimics/inhibitors can provide a panel of the affected mRNAs and thus possible direct targets. A novel approach is high-throughput sequencing of RNA isolated by cross-linking and immunoprecipitation (HITS-CLIP). HITS-CLIP uses UV light to cross-link or freeze RISC including miRNA–mRNA–AGO2, followed by immunoprecipitation and sequencing of isolated miRNA and mRNA [104]. A modification of this technique is photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP). PAR-CLIP displays more efficient cross-linking and RNA recovery. However, this approach uses live cells that need to be supplemented with the nucleoside analog 4-thiouridine (4SU) that is incorporated into nascent mRNAs; thus it cannot be applied to tissue samples [105]. Both techniques have helped significantly in the identification of the genuine miRNA targets [106].

Alternatively high-throughput approaches have contributed to the identification of potential miRNA targets.

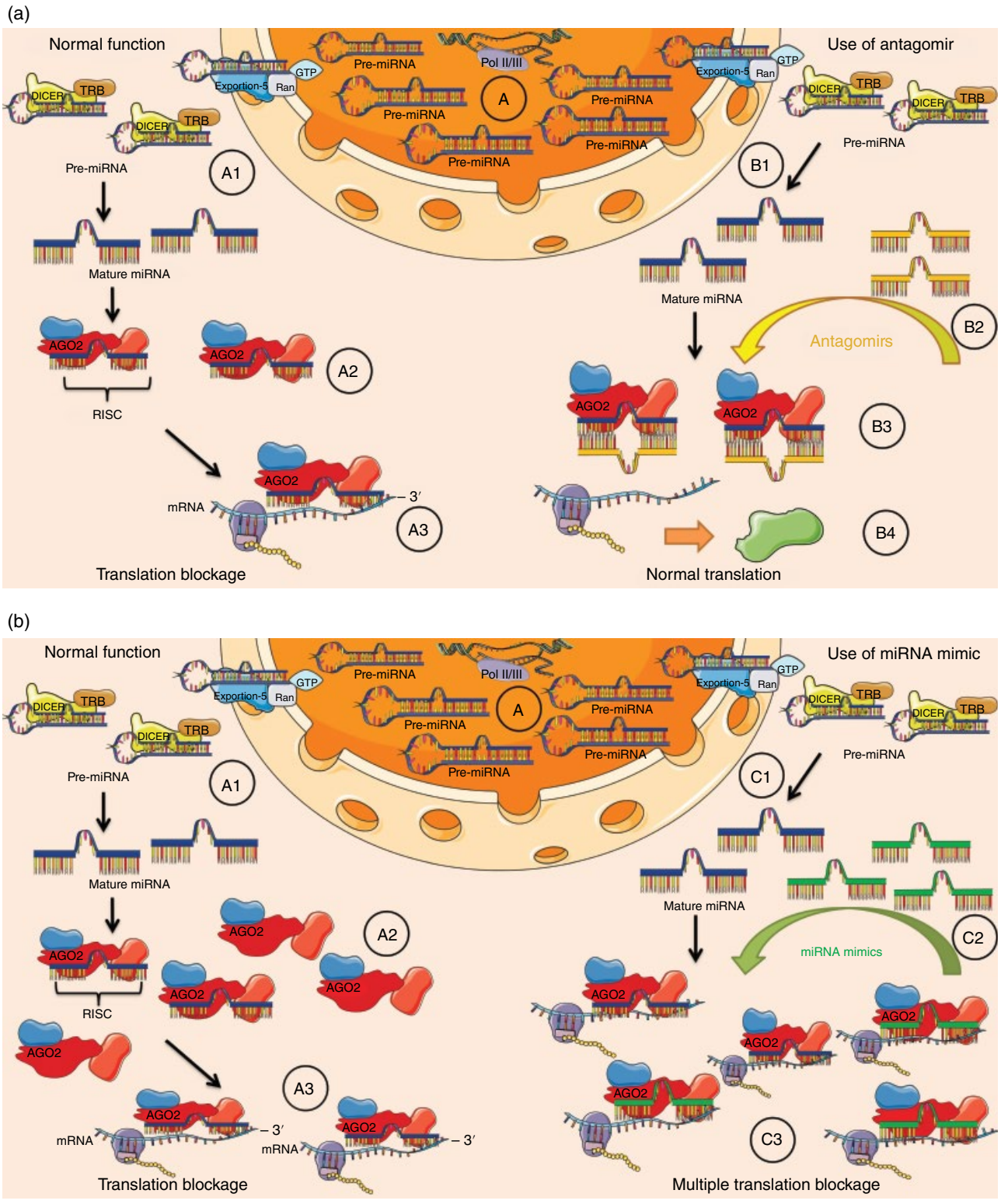


Figure 5.3 (a) Antagomir function. The normal pathway of the miRNA biogenesis and action is shown in summary in A, A1, B1, A2, and A3 with the final result to be the suppression of gene expression. With the addition of the antagomirs (B2), the action of the miRNAs is inhibited. Antagomirs bind in a complementary manner to the miRNAs, leading to translation of the targeted mRNA (B4). (b) miRNA mimics function. The normal pathway of the miRNA biogenesis and action is shown in summary in A, A1, and C1. AGO2 molecules are available in excess to any mature miRNA and complete its function (A2 and A3). With the addition of the miRNA mimics (C2), the free AGO2 proteins are capable to bind to more miRNAs than normal, leading to additional inhibition of translation of the mRNA target molecules.

Microarray profiling of mRNAs or RNA-seq allow studying changes in mRNA expression in cells treated with a specific antagomir or mimic miRNA [107]. Quantitative proteomic analysis as a high-throughput method has also been used including stable isotope labeling with amino acids in culture (SILAC) [108] and two-dimensional difference in-gel electrophoresis (2D-DIGE) of samples with modified miRNA expression [109]. Targeted mass spectrometry-based strategies are also starting to claim their place as quantitative method for observing targeted proteins (selected reaction monitoring/multiple reaction monitoring (SRM/MRM)) [110] and can be used to monitor the expression of particular miRNA targets. The disadvantage of these proteomics approaches is that they cannot determine whether the observed changes in the abundance of a protein are due to direct or indirect miRNA targeting. Bioinformatics can help in detecting direct miRNA targets by comparing the differentially expressed proteins with the seed region of the miRNA under study if there is significant complementarity and the target gene expression is validated by RT-qPCR, Western blotting, and immunohistochemical analysis, which increases the validity of the existence of an miRNA–mRNA pair [110].

Another target validation method is based on the luciferase assay. In brief, the potential miRNA target site is cloned and added to the open reading frame of a reporter gene, for example, luciferase of *Renilla* or firefly. Next the recombinant plasmid is transfected into mammalian cells together with a mimic miRNA, the cells are incubated for 24–48 h, and then the luciferase activity or fluorescence intensity is measured. If the target site of the mRNA is a valid target for the miRNA under study, a reduced signal intensity compared with control plasmid will be observed [82].

An interesting alternative method on target detection was proposed by Vatolin et al. [111]. Instead of trying to find the mRNA expression changes of the mRNA seeds complementary to the miRNA of interest, Vatolin et al. used the sequence of the miRNA to generate primers via reverse transcription matching of the mRNA targets. These primers were subsequently used to generate cDNA from the mRNA of freshly prepared cytoplasmic extracts, creating a 3′-cDNA–miRNA-5′ hybrid molecule. With this hybrid as cDNA primer, it was possible to initiate the synthesis of detectable cDNA from a RNA sample extract and amplify the possible RNA target of the miRNA of choice. Sequencing of the cDNA allows identification of the specific miRNA target.

Biotin-tagged miRNA is another biochemical method for enrichment of miRNA targets. It was first introduced by Orom and Lund [112]. In this approach, biotinylated synthetic miRNA is transfected into cells followed by purification of miRNA–mRNA complexes connected to the seed sequence with streptavidin-agarose beads. A similar method is based on DIG-labeled synthetic miRNAs in combination with anti-DIG agarose beads [113].

Finally, methods designed to capture the miRNA-mediated cleavage products have been developed based on RNA ligase-mediated 5′ rapid amplification of cDNA ends (RLM-RACE) [114]. 5′ RLM-RACE is a PCR-based technique, whereby an RNA adapter is ligated to the free 5′ phosphate of an uncapped mRNA produced from, among other nucleolytic activities, Argonaute2-directed mRNA cleavage. The ligation product can be reverse transcribed using a forward primer directed against the linker and a gene specific reverse primer that is subsequently PCR amplified, cloned, and identified by sequencing [115]. Next-generation RLM-RACE uses the parallel analysis of RNA ends (PARE) that offers a genome-wide identification of the miRNA-induced cleavage products. The major modification, compared with the original method, is the addition of a MmeI restriction site at the 5′ RNA adapter, which after reverse transcription and cDNA synthesis is digested with MmeI, leaving as result 20–21 nt tag sequences attached to the adapter. The next step is to ligate a DNA adapter to the 3′ end of the tag and amplify the target sequence using PCR and 5′ and 3′ adapter-specific primers. Finally, the tags are analyzed with high-throughput sequencing to reveal the corresponding target genes and infer regulatory miRNAs [116, 117].

The previous summary of the key assays to determine miRNA targets shows that only a combination of methods will yield high confidence miRNA targets. Screening methods, including microarrays or NGS, are reliable to provide a panel of possible targets, but specific target assays, such as luciferase or PCR based, are needed to validate the candidates. A list with additional target assays is available in Table 5.4.

5.4 Use of miRNA *In Vivo*

The results from the *in vitro* experiments are transferred to *in vivo* research with the hope to reproduce a similar outcome. The basic approach is similar *in vitro* and *in vivo*: deliver a synthetic miRNA mimic or inhibitor and then detect the changes at the protein level with emphasis on the predicted and *in vitro* validated targets. A further step is the observation of any side effects, alterations in neighboring molecular pathways, and pathologies and effects on the phenotype. The ultimate goal is to identify a possible therapeutic effect of an miRNA that can pass the clinical trials and enter the clinical practice.

Until now the main focus has been on improving the stability of the molecules interfering with miRNA, since unprotected oligonucleotides can easily be degraded by

Table 5.4 miRNAs and target validation methods.

Method/molecule	Full name	Principle	System check
HITS-CLIP	High-throughput sequencing of RNA isolated by crosslinking and immunoprecipitation	UV light to cross-link–freeze—the RISC complex including miRNA-mRNA-AGO2, followed by immunoprecipitation, and sequencing of isolated miRNA and mRNA	*In vitro*—miRNAs and their target mRNAs
PAR-CLIP	Photoactivable-ribonucleoside-enhanced cross linking and immunoprecipitation	Cells are first supplemented with the nucleoside analog 4-Thiouridine (4SU) that is incorporated into nascent mRNAs. Next steps are the same as HITS-CLIP	*In vitro*—miRNAs and their target mRNAs
SILAC	Stable-isotope labeling with amino acids in culture	Relative protein abundance is measured by mass spectrometry of samples labeled with different isotopes	*In vitro*—the level of the protein originating from the mRNA target
2D-DIGE	Two dimensional gel electrophoresis	Electrophoresis on a single gel of two samples labeled with different fluorescent dyes, separating the proteins by iso-electric focusing and SDS–PAGE and then identification by mass spectrometry	*In vitro*—the level of the protein originating from the mRNA target
RT-qPCR	Reverse transcription quantitative polymerase chain reaction	Investigate the level of the miRNA or mRNA of interest with TaqMan or Sybr green probes	*In vitro*—the level of the miRNAs and mRNAs
Western blot		Electrophoresis to separate proteins by molecular mass which are subsequently detected with antibodies specific to the target protein.	*In vitro*—the level of the proteins predicted to be affected by miRNA deregulation
IHC	Immunohistochemistry	*In situ* hybridization with LNA probes complementary against the miRNA of interest and/or immunohistochemical detection of specific proteins	In tissue samples—detect the level of miRNAs and proteins of interest
Microarrays		Hybridization of a panel of miRNAs or genes to the complementary immobilized probes and detection via fluorescence	In all samples examine the expression profile of miRNAs and/or mRNAs
RNA sequencing		Massive parallel sequencing of millions of fragments of DNA from a single sample	In all samples examine the expression profile of miRNAs and/or mRNAs
Luciferase assay		A luciferase reporter contains the 3′-UTR target sites of a mRNA target which the miRNA mimics of the miRNA of interest should target and reduce the fluorescence signal	*In vitro*—validation of the miRNA target
GFP reporter	Green fluorescent protein	A GFP reporter is under the control of multiple 3′-UTR target sites of a specific miRNA. The presence of the miRNA will lower the GFP signal	*In vitro* and *in vivo*—validation of miRNA target
Biotin-tagged miRNA		Biotinylated synthetic miRNA is transfected into cells followed by purification of miRNA:mRNA complexes connected to the seed sequence with streptavidin-agarose beads	*In vitro*—capture of miRNA targets
PARE	Parallel analysis of RNA ends	Genome-wide identification of the miRNA-induced cleavage products	*In vitro* and *in vivo*—capture the miRNA-mediated cleavage products
RLM-RACE	RNA ligase mediated-5′ rapid amplification of cDNA ends	The 5′ phosphate of the cleaved, uncapped poly-A RNAs an RNA adapter is ligated followed by reverse transcription. Then the cDNAs are amplified with the adapter and the gene-specific primers to be cloned and sequenced	*In vitro* and *in vivo*—capture the miRNA-mediated cleavage products

LAMP	Labeled microRNA (miRNA) pull-down assay system	The pre-miRNA is labeled with digoxigenin (DIG), mixed with cell extracts, and immunoprecipitated by anti-DIG antiserum	*In vitro*—identification of the target gene of known miRNAs
Reverse transcription of targets		Use of the sequence of the miRNA to generate primers via reverse transcription matching the mRNA targets followed by PCR	*In vitro*—capture miRNA target genes
Polysome profiling		Cyclohexamide is used to trap elongating ribosomes followed by pull-down of the enclosed mRNA	*In vitro*—detection of the effect of deregulation of a miRNA to the transcriptome
Ribosome profiling		Cyclohexamide is used to trap elongating ribosomes, cells are lysed and the mRNA enclosed to the monosomes is sequenced	*In vitro*—detection of the effect of deregulation of a miRNA to the transcriptome
Direct injection of ant-miRs or miRNA mimics		Direct injection of antagomirs and miRNA mimics into animal model	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology
miRNA sponges or decoys		Single piece of RNA containing multiple seed regions complementary to the miRNA family of interest. By delivering this system into cells, the miRNAs bind strongly to the RNA sequence resulting in silencing the miRNA activity	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology
Modified viruses		Lentivirus, adenovirus and adeno-associated virus are modified accordingly to transfer and deliver miRNA regulators in a model system	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology
Exosomes		Microvesicles that naturally transfer miRNAs and can be used to deliver modified miRNA regulators	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology
Liposomes (DOTMA, SLNs, MaxSuppressor In Vivo RNA-LANCEr II, LPH, scFv, iNOPs)		Lipid bilayers (cationic or neutral) with an internal aqueous phase to load and carry the desired molecules	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology
PLGA	poly(lactide-*co*-glycolic acid)	PLGA polymers are transformed into microparticles or nanoparticles enclosing biological molecules that can release via endolysosomal escape after entering the cell	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology
PEI	polyethylenimine	Cargo release follows the so-called "proton sponge effect"; once the polymer interacts with the surface of the cell, endocytosis takes place. The PEI causes endosome swelling by influx of protons and water (proton sponge effect) which leads to endosome destabilization and the release of the miRNA	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology
Inorganic nanoparticles (AUNP-S-PEG, GD2)	Gold nanoparticles, disialoganglioside antibody	Inorganic nanoparticles are bound and deliver miRNA regulators	*In vitro—in vivo* antagomir or miRNA mimic system and detection/ observation of the expression level of the possible targets and/or physiology

For the validation phase, miRNA antagomirs or mimics are delivered to an *in vitro* experimental system or *in vivo* to reduce or increase the levels of the miRNAs of interest respectively, following by measuring the levels of the predicted mRNA targets or the protein level originating by the mRNA target.

RNAses in the blood. As a consequence the design of carriers has become essential in order to transfer and deliver the miRNA mimic or antagomir to the desired target while minimizing degradation. These carriers need to have particular features. They need to be large enough to avoid renal and hepatic filtration but also small enough to allow passage through the tissue and cellular barriers and release efficiently their load. If the miRNA is not delivered intact, gene silencing will not be achieved. Moreover carriers must not activate the immune system and should not be toxic for the organism. Finally, and most importantly, all possible side effects must be detected and examined. Since one miRNA has multiple targets, strategies must be designed to tackle any off-target effects [118, 119]. These different strategies are described in the following text.

5.4.1 Chemically Modified miRNAs

Many chemical modifications in order to enhance the stability and efficiency of miRNAs have been proposed. The first attempt of systemic delivery involved chemically modified miRNA molecules. In this category are included modifications at the 2′OH group in the ribose ring (most vulnerable to nucleases) with a 2′-*O*-methyl (OMe) or with 2′-*O*-methoxyethyl (MOE). MOE modifications bring another advantage since they also show higher affinity and specificity toward target RNAs [120]. These modifications have been coupled to linkage to cholesterol in order to improve cell entry [121]. LNA have also been frequently used and are basically a category of antagomirs in which a 2′,4′-methylene bridge is introduced in the ribose part of the miRNAs and creates a bicyclic nucleotide, providing higher stability and affinity [122]. Finally, modifications in the mimic strand have been proposed, especially at the passenger strand so that the guide strand remains untouched to execute its biological role, in order to increase stability and resistance against degradation [123, 124].

5.4.2 miRNA Sponges or Decoys

Similar to the antagomirs' action, miRNA sponges or decoys are molecules that can trap miRNAs. Ebert et al. introduced this method, in which basic idea is to create a single piece of RNA containing multiple seed regions complementary to the miRNA family of interest. By delivering this system into cells, miRNAs bind strongly to these RNA seeds, resulting in silencing of the miRNA activity [98]. An *in vivo* example of the use of miRNA sponges showed that miRNA-133 is an indirect regulator of GLUT4 by directly targeting KLF15 in cardiac disease, illustrating a complex of miR-133–KLF15–GLUT4 [125].

5.4.3 Modified Viruses

Modified viruses have also been used as vehicles to deliver miRNA regulators. In particular, lentivirus, adenovirus, and adeno-associated virus (AAV) are widely used as carriers of antagomirs or mimics of miRNAs. Modifications of the structure of viruses, mostly with bifunctional polyethylene glycol (PEG), prevent immune responses. Genetic modifications aid to the specific cell targeting [121]. Non-integrating adenoviruses and AAVs seem to be preferred over lentiviruses because the latter integrate their own reversed transcribed DNAs in the host cells. This integration can lead to serious side effects including mutations or activation of gene expression [118]. To avoid these effects, nonviral delivery systems were developed.

5.4.4 Microvesicles

As mentioned earlier, exosomes protect miRNAs from RNAses present in all body fluids [30, 31]. Zhang et al. were able to monitor the upregulation of miR-150 after delivering exosomes from THP-1 cells to mice intravenously, showing that exosomes can be used as a delivery system of miRNAs *in vivo* [126]. It is also possible to include exosomes in a viral system, either by including viruses with miRNAs in the exosomes or by the use of virosomes (exosomes modified as viruses) in order to increase stability and targeted delivery. However, this approach is far from reaching clinical practice [127, 128].

Liposomes are widely used to transfer reagents *in vitro* and start to find their application as *in vivo* carriers. The first example of use of liposomes *in vivo* is transport chemotherapy drugs and nucleic acids to tumors [118]. Liposomes are lipid bilayers with an internal aqueous phase used to load and carry the desired molecules [118]. Since MiRNAs are negatively charged hydrophilic molecules, they bind to cationic lipids and are kept stable and protected from degradation in the liposome interior. The positive charge on the surface of liposomes forms ionic interactions with the negatively charged cell membrane and allows delivery of their cargo [118]. Several different liposome designs exist. A cationic lipid nanoparticle composed of 2-dioleyloxy-*N*,*N*-dimethyl-3-aminopropane (DOTMA) and cholesterol was used to deliver systemically the pre-miR-133b in mice, resulting in a 30% accumulation in lung tissue and 52-fold increase of the miR-133b levels in lung compared with untreated mice [129], indicating that DOTMA is a promising delivery system. Similarly, a miR-122 mimic was delivered in hepatocellular carcinoma using the DOTMA system (with some modifications), resulting in approximately 50% growth suppression of HCC xenographs by repressing a number of miR-122 targets [130]. Another cationic

system is the solid lipid nanoparticles (SLNs) that were used to deliver a miR-34a mimic to cancer stem cells, leading to tumor growth suppression and better survival rate for the treated mice [131]. The most significant drawback of the cationic liposomes as carriers is that their positive charge activates interferons and is toxic for the liver [118]. To overcome these drawbacks, neutral liposome systems were developed. The commercially available and patented MaxSuppressor™ In Vivo RNA-LANCEr II is a proprietary formulation composed of neutral lipids, a nonionic detergent, and small molecules providing high efficiency transfer of miRNA *in vivo* and has been used in a number of studies, mostly involving cancer, with promising results [132–134]. In the same category, liposome–polycation–hyaluronic acid (LPH) aims to enrich the loading capacity and transfer efficiency. A modification with GC4 single-chain variable fragment (scFv) efficiently transferred miR-34a, leading to suppression of lung tumor [135]. Finally, interfering nanoparticles (iNOPs) are dendrimers prepared from lipids and have shown some significant results. Anti-miR-122 was delivered using iNOPs and efficiently reduced the expression level of miR-122 in the liver, avoiding any immune response or toxicity [136]. In general, the lipid-based delivery systems may prove valuable for clinical use due to their low toxicity and side effects.

5.4.5 The Polymers

Polymers have also been used to deliver miRNAs and their inhibitors. The polymer-based miRNA delivery system is represented by two main systems. The first is the poly(lactide-*co*-glycolic acid) (PLGA)-based system. PLGA has been used since 1970s, is FDA approved, and has shown high safety, biocompatibility, stability, and production efficiency [118, 119, 137]. These polymers can be easily transformed into microparticles or nanoparticles enclosing biological molecules that can be released via endolysosomal escape after entering the cell [138, 139]. Their high and multiple loading capacity and the fact that multiple modifications on the PLGA surface can be introduced have placed PLGA on top of the list for gene and drug delivery [119]. Especially the modification of the surface with PEG leads to increased half-life in the circulation of the PLGA-based nanoparticles [140, 141].

The second polymer system is based on polyethylenimine (PEI). It is one of the most widely used systems for gene delivery because with its positive charge it binds very strong to nucleic acids and can easily transfer and release its load [119]. Cargo release follows the so-called proton sponge effect; once the polymer interacts with the surface of the cell, endocytosis takes place. The PEI causes endosome swelling by influx of protons and water (proton sponge effect), which leads to endosome destabilization and the release of the miRNA [142]. Successful transfer of miR-145—which directly targets the oncogenes Oct4 and Sox2—in lung adenocarcinoma and cancer stem cells with polyurethane short branch polyethylenimine (PU-PEI) in combination with radiotherapy resulted in greatly reduced tumor growth *in vitro* and *in vivo* and higher survival rate [143, 144]. Synthetic polymers are promising miRNA nanocarriers since (i) their cationic profile enables to strongly bind to miRNA and thus protect it from degradation and (ii) they easily release miRNAs in the cells. Additional studies are necessary to evaluate the toxicity of these polymers before clinical use can be envisaged [118].

5.4.6 Inorganic Nanoparticles

Inorganic nanoparticles are another category of nonviral delivery systems. The mostly used elements are gold, silica, and carbon, and by taking advantage of their chemical properties, it is easy to manipulate and construct a vehicle with suitable size to transfer miRNAs without activating the immune system [118, 119]. Gold nanoparticles or AuNP-S-PEG have shown to deliver efficiently miR-31 and miR-1323 and silence gene expression up to 70% of E2F2, STK40, and CEBRA (targets of miR-31) and over 85% of CASP8AP2, DDX4, and AAK1 (targets of miR-1323) and also inhibited proliferation of cancer cells with a low toxicity score [145]. In another study, silica-based nanoparticles modified with disialoganglioside GD2 antibody that binds to the GD2 antigen delivered specifically miR-34a into neuroblastoma, causing down-regulation of MYCN, inhibition of the tumor growth, and increased apoptosis of cancer cells [146]. Unfortunately, extensive *in vivo* studies are not available for evaluation of inorganic nanoparticle toxicity, loading capacity, and delivery efficiency.

5.5 miRNAs as Potential Therapeutic Agents and Biomarkers: Lessons Learned So Far

5.5.1 miRNAs as Potential Therapeutic Agents

The most extensively studied miRNA to be potentially introduced in clinical practice is miRNA-122, an abundant liver-specific miRNA with a key role in liver function (fatty acid and cholesterol metabolism) and active role in hepatitis C (HCV) progression [147]. Elmen et al. showed that reduction of miRNA-122 with LNA-ant-miR122 led to reduction of cholesterol levels in plasma of mice and nonhuman primates [99, 148]. Lanford et al. showed that suppression of HCV viremia by an antagomir for miR-122 in chimpanzees led to improvement HCV liver pathology [149]. In these three studies lasting

from 3 weeks to 3 months, no short-term toxicity was observed, opening the way for use of miRNAs as drugs in liver disease. Indeed, these preclinical results led to the development of miravirsen, the first antagomir drug candidate against HCV by Santaris Pharma A/S [150]. Miravirsen binds to the stem-loop structure of pri- and pre-miR-122 with nanomolar affinity and inhibits both Dicer- and Drosha-mediated processing of miR-122 precursors and has entered phase II trials, after displaying no side effects on healthy volunteers [151].

Another therapeutic miRNA candidate that is now in phase I trials is the liposome-formulated mimic miRNA-34, or MRX34, as it is named by the developing company Mirna Therapeutics. The miRNA-34 family is found to be decreased in many cancer types. It was observed that the miRNA-34 family can regulate the p53 pathway by direct inhibition of many oncogenes including Myc, c-Met, BCL-2, CDK4, and CDK6. Treatment with miR-34 mimics reduced tumor size or growth in many cancer model systems [152–155]. MRX34 could become one of the first mimic miRNA system to restore the level of an miRNA and show significant antitumor activity (http://www.mirnatherapeutics.com/pipeline/mirna-pipeline.html).

Miragen Therapeutics has initiated several preclinical studies for a number of miRNAs. MiR-155 is studied in hematological malignancies and amyotrophic lateral sclerosis and MiR-29 in cutaneous and pulmonary fibrosis, which have passed the preclinical studies stage and now entering clinical trials. MiR-92 is a candidate for peripheral artery disease, now in preclinical studies. In collaboration with Servier, Miragen Therapeutics is performing preclinical studies in heart failure targeting miR-208 and miR-15 (http://miragentherapeutics.com/pipeline/).

Finally, Regulus Therapeutics has initiated their anti-miR programs for miR-122 in HCV and miR-10b in glioblastoma. In collaboration with Sanofi, they are developing anti-miR-221 against hepatocellular carcinoma and miR-21 for kidney fibrosis, the leading complication in kidney disease, and with AstraZeneca miR-103/107 for the treatment of nonalcoholic steatohepatitis (NASH) in patients with type 2 diabetes/prediabetes (http://www.regulusrx.com/therapeutic-areas/).

5.5.2 miRNAs as Potential Biomarkers

Another application is the use of miRNAs as potential diagnostic and prognostic biomarkers. A simple search in PubMed with the words "miRNA" AND "Biomarker" returns more than 5000 publications, the majority related to cancer.

5.5.2.1 Cancer

In 2002, the first miRNAs characterized as being downregulated in patients with B-cell chronic lymphocytic leukemia were miR-15a and miR-16-1 [156], a year after the discovery of the second miRNA, let-7. Later on the miR-17-92 cluster was identified to be upregulated in many kinds of cancers (mostly in lymphoma and leukemia) and associated with upregulation of the oncogene c-Myc, providing a first well-characterized cancer-associated miRNA cluster and introducing a new class of miRNAs, the oncomiRs [157–159]. Additional miRNAs including miR-21 [160], miR-34 family [152, 161], and let-7 [162, 163] are by now accepted as cancer-associated miRNAs. In 2008 Markou et al. published for the first time evidence for the prognostic value of the miRNAs in a clinical study on lung cancer. The authors characterized the overexpression of miR-21 in patients with lung cancer as negative prognostic factor [164]. Other reports are regularly published on miRNAs and their involvement in molecular pathways in nearly every kind of cancer and on their use as potential biomarkers and/or therapeutic agents including miR-21, miR-155, miR-10b, miR-214, and miR-105 [165] in breast cancer; miR-141, miR-125b, and miR-145 in prostate cancer; miR-486, miR-30d, miR-1, miR-499, miR-19b, and miR-29b in lung cancer [166, 167]; and miR-1246 and miR-155 in leukemia [168].

5.5.2.2 Metabolic and Cardiovascular Diseases

The use of a LNA-antagomir for miR-122 reduced the expression of miR-122 in hepatocytes in a high-fat mouse model, resulting in reduced serum cholesterol and triglyceride levels [169, 170]. These reports may point to another therapeutic ability of antagomir for miR-122 against metabolic and cardiovascular diseases. Using the same mouse model, miR-33 has been linked with the high triglyceride and HDL [171, 172], and let-7 overexpression in the pancreas led to impaired glucose tolerance and reduced glucose-induced pancreatic insulin secretion [173, 174].

5.5.2.3 Miscellaneous Diseases

Many other circulating miRNAs have been found to be associated with other diseases including a 9-miRNA plasma/serum panel for Alzheimer's disease [175], a 16-miRNA plasma/serum panel for epilepsy [176], and 7 plasma/serum miRNAs in preeclampsia [177] and cardiovascular diseases [178].

5.6 Conclusion

miRNAs have a key role in maintaining cellular homeostasis by regulating almost all biological function, and this justifies the observations of deregulation of miRNAs in many pathological conditions. The biological

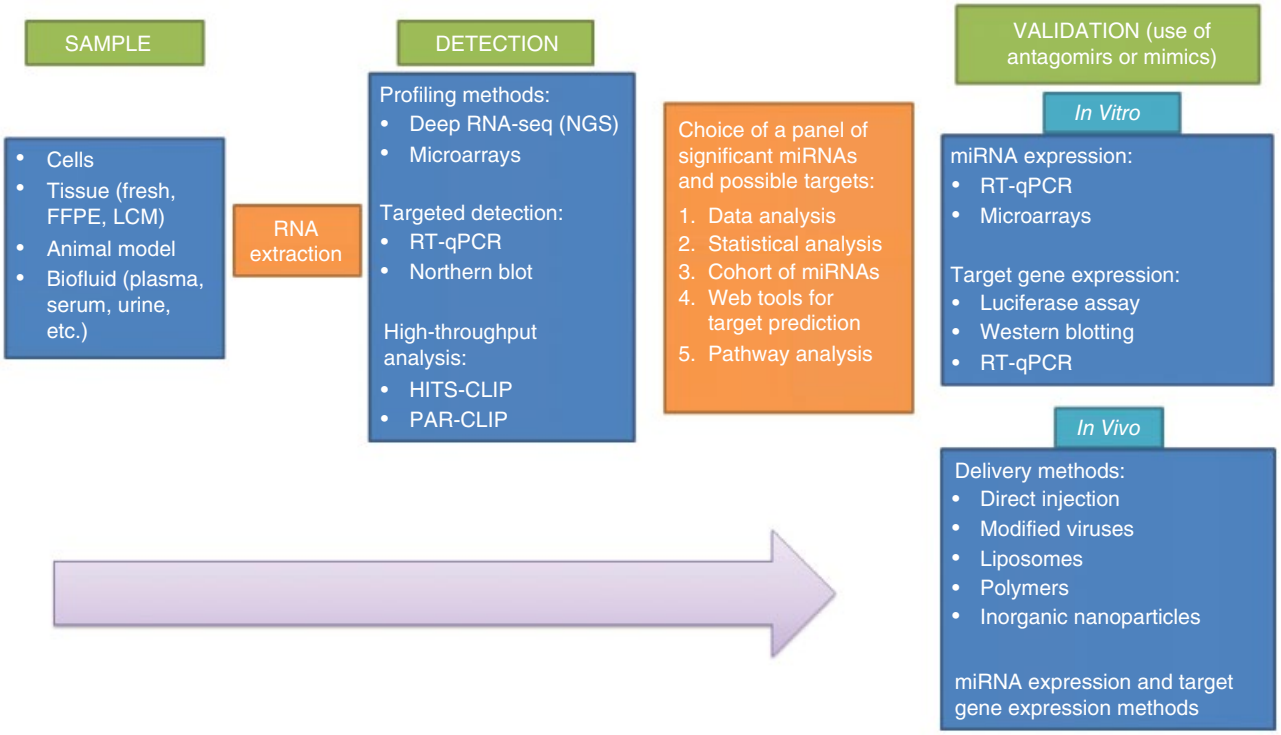


Figure 5.4 A possible workflow of how research on miRNAs could lead to clinically useful miRNAs.

complexity of their actions, however, blocks a fast evolution in transforming them as tools for clinical practice as miRNAs often target multiple genes different functional pathways. However miRNA-based therapies could offer a distinct advantage over other approaches.

The properties of miRNAs set them as potential effective diagnostic and prognostic biomarkers. The stability of miRNAs in FFPE tissues and body fluids is the main advantage for the use of miRNAs as biomarkers of disease. In addition, miRNAs can be extracted from small biopsy specimens and body fluid samples. Finally, miRNAs have the potential to become therapeutic agents for personalized management of disease [79].

The early output of miRNA research is encouraging. However, the current potential utilization of miRNAs in the clinical practice is primarily limited to expression profiling for diagnostic or prognostic purposes. This is due to the fact that little is known about the phenotypical consequences of miRNA targeting when the results are transferred from the *in vitro* to the *in vivo* setting. In addition, the multiple targets of miRNAs and the regulation of a single mRNA from a variety of miRNAs render the targeting therapy difficult to monitor, for example, the off-target side effects [107]. Moreover, the lack of *in vivo* validation can be due to the lack of robust and efficient miR target prediction tools. Development of effective bioinformatics analysis methods will yield confident miRNA–mRNA interaction results. Therefore a more integrated collaboration between clinical and molecular methods with a strong biostatistical and bioinformatics foundation to bridge the gap between research and clinical applications and increase the possibilities for validating miRNAs as important biomarkers or drug targets is needed [21]. A proposed research workflow on miRNAs, from detection to validation, is given in Figure 5.4. As technology is evolving, new and exciting information will be added to the miRNA biology research; pathway regulations will be revealed and allowed to be manipulated, potentially allowing entry of miRNAs in personalized medicine.

References

1 Lee, R. C., Feinbaum, R. L. & Ambros, V. 1993. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75, 843–854.

2 Wightman, B., Ha, I. & Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans*. *Cell*, 75, 855–862.

3 Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R. & Ruvkun, G. 2000. The 21-nucleotide let-7 RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403, 901–906.

4 Almeida, M. I., Reis, R. M. & Calin, G. A. 2011. MicroRNA history: discovery, recent applications, and next frontiers. *Mutat Res*, 717, 1–8.

5 Friedlander, M. R., Lizano, E., Houben, A. J., Bezdan, D., Banez-Coronel, M., Kudla, G., Mateu-Huertas, E., Kagerbauer, B., Gonzalez, J., Chen, K. C., Leproust, E. M., Marti, E. & Estivill, X. 2014. Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol*, 15, R57.

6 Kozomara, A. & Griffiths-Jones, S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42, D68–D73.

7 Bartel, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116, 281–297.

8 Hammond, S. M. 2015. An overview of microRNAs. *Adv Drug Deliv Rev*, 87, 3–14.

9 Winter, J., Jung, S., Keller, S., Gregory, R. I. & Diederichs, S. 2009. Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol*, 11, 228–234.

10 Meister, G. 2013. Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet*, 14, 447–459.

11 Yang, J. S., Phillips, M. D., Betel, D., Mu, P., Ventura, A., Siepel, A. C., Chen, K. C. & Lai, E. C. 2011. Widespread regulatory activity of vertebrate microRNA* species. *RNA*, 17, 312–326.

12 Lewkowicz, P., Cwiklinska, H., Mycko, M. P., Cichalewska, M., Domowicz, M., Lewkowicz, N., Jurewicz, A. & Selmaj, K. W. 2015. Dysregulated RNA-Induced Silencing Complex (RISC) assembly within CNS corresponds with abnormal miRNA expression during autoimmune demyelination. *J Neurosci*, 35, 7521–7537.

13 Macfarlane, L. A. & Murphy, P. R. 2010. MicroRNA: biogenesis, function and role in cancer. *Curr Genomics*, 11, 537–561.

14 Ha, M. & Kim, V. N. 2014. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol*, 15, 509–524.

15 Eulalio, A., Tritschler, F. & Izaurralde, E. 2009. The GW182 protein family in animal cells: new insights into domains required for miRNA-mediated gene silencing. *RNA*, 15, 1433–1442.

16 Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G. & Tuschl, T. 2003. A uniform system for microRNA annotation. *RNA*, 9, 277–279.

17 Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res*, 32, D109–D111.

18 Guo, L., Yu, J., Yu, H., Zhao, Y., Chen, S., Xu, C. & Chen, F. 2015. Evolutionary and expression analysis of miR-#-5p and miR-#-3p at the miRNAs/isomiRs levels. *Biomed Res Int*, 2015, 168358.

19 Liang, Y., Ridzon, D., Wong, L. & Chen, C. 2007. Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics*, 8, 166.

20 Tang, Y., Liu, D., Zhang, L., Ingvarsson, S. & Chen, H. 2011. Quantitative analysis of miRNA expression in seven human foetal and adult organs. *PLoS One*, 6, e28730.

21 Hunt, E. A., Broyles, D., Head, T. & Deo, S. K. 2015. MicroRNA detection: current technology and research strategies. *Annu Rev Anal Chem (Palo Alto Calif)*, 8, 217–237.

22 Kong, W., Zhao, J. J., He, L. & Cheng, J. Q. 2009. Strategies for profiling microRNA expression. *J Cell Physiol*, 218, 22–25.

23 Sood, P., Krek, A., Zavolan, M., Macino, G. & Rajewsky, N. 2006. Cell-type-specific signatures of microRNAs on target mRNA expression. *Proc Natl Acad Sci U S A*, 103, 2746–2751.

24 Hanson, E. K., Lubenow, H. & Ballantyne, J. 2009. Identification of forensically relevant body fluids using a panel of differentially expressed microRNAs. *Anal Biochem*, 387, 303–314.

25 Weber, J. A., Baxter, D. H., Zhang, S., Huang, D. Y., Huang, K. H., Lee, M. J., Galas, D. J. & Wang, K. 2010. The microRNA spectrum in 12 body fluids. *Clin Chem*, 56, 1733–1741.

26 Courts, C. & Madea, B. 2011. Specific micro-RNA signatures for the detection of saliva and blood in forensic body-fluid identification. *J Forensic Sci*, 56, 1464–1470.

27 Wang, Z., Zhang, J., Luo, H., Ye, Y., Yan, J. & Hou, Y. 2013. Screening and confirmation of microRNA markers for forensic body fluid identification. *Forensic Sci Int Genet*, 7, 116–123.

28 Zubakov, D., Boersma, A. W., Choi, Y., Van Kuijk, P. F., Wiemer, E. A. & Kayser, M. 2010. MicroRNA markers for forensic body fluid identification obtained from microarray screening and quantitative RT-PCR confirmation. *Int J Legal Med*, 124, 217–226.

29 Beltrami, C., Clayton, A., Phillips, A. O., Fraser, D. J. & Bowen, T. 2012. Analysis of urinary microRNAs in chronic kidney disease. *Biochem Soc Trans*, 40, 875–879.

30 Camussi, G., Deregibus, M. C., Bruno, S., Cantaluppi, V. & Biancone, L. 2010. Exosomes/microvesicles as a mechanism of cell-to-cell communication. *Kidney Int*, 78, 838–848.

31 Ramachandran, S. & Palanisamy, V. 2012. Horizontal transfer of RNAs: exosomes as mediators of intercellular communication. *Wiley Interdiscip Rev RNA*, 3, 286–293.

32 Valadi, H., Ekstrom, K., Bossios, A., Sjostrand, M., Lee, J. J. & Lotvall, J. O. 2007. Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. *Nat Cell Biol*, 9, 654–659.

33 Cheng, L., Sharples, R. A., Scicluna, B. J. & Hill, A. F. 2014. Exosomes provide a protective and enriched source of miRNA for biomarker profiling compared to intracellular and cell-free blood. *J Extracell Vesicles*, 3, 1–14.

34 Cheng, L., Sun, X., Scicluna, B. J., Coleman, B. M. & Hill, A. F. 2014. Characterization and deep sequencing analysis of exosomal and non-exosomal miRNA in human urine. *Kidney Int*, 86, 433–444.

35 Arroyo, J. D., Chevillet, J. R., Kroh, E. M., Ruf, I. K., Pritchard, C. C., Gibson, D. F., Mitchell, P. S., Bennett, C. F., Pogosova-Agadjanyan, E. L., Stirewalt, D. L., Tait, J. F. & Tewari, M. 2011. Argonaute2 complexes carry a population of circulating microRNAs independent of vesicles in human plasma. *Proc Natl Acad Sci U S A*, 108, 5003–5008.

36 Wang, K., Zhang, S., Weber, J., Baxter, D. & Galas, D. J. 2010. Export of microRNAs and microRNA-protective protein by mammalian cells. *Nucleic Acids Res*, 38, 7248–7259.

37 Vickers, K. C., Palmisano, B. T., Shoucri, B. M., Shamburek, R. D. & Remaley, A. T. 2011. MicroRNAs are transported in plasma and delivered to recipient cells by high-density lipoproteins. *Nat Cell Biol*, 13, 423–433.

38 Mlcochova, H., Hezova, R., Stanik, M. & Slaby, O. 2014. Urine microRNAs as potential noninvasive biomarkers in urologic cancers. *Urol Oncol*, 32, 41, e1–e9.

39 Turchinovich, A., Weiz, L., Langheinz, A. & Burwinkel, B. 2011. Characterization of extracellular circulating microRNA. *Nucleic Acids Res*, 39, 7223–7233.

40 Lorenzen, J. M. & Thum, T. 2012. Circulating and urinary microRNAs in kidney disease. *Clin J Am Soc Nephrol*, 7, 1528–1533.

41 Haase, M. & Mertens, P. R. 2015. Biomarkers: more than just markers! *Nephrol Dial Transplant*, 30, 33–38.

42 Fang, D. Y., King, H. W., Li, J. Y. & Gleadle, J. M. 2013. Exosomes and the kidney: blaming the messenger. *Nephrology (Carlton)*, 18, 1–10.

43 Haraldsson, B., Nystrom, J. & Deen, W. M. 2008. Properties of the glomerular barrier and mechanisms of proteinuria. *Physiol Rev*, 88, 451–487.

44 Satchell, S. 2013. The role of the glomerular endothelium in albumin handling. *Nat Rev Nephrol*, 9, 717–725.

45 Alvarez-Erviti, L., Seow, Y., Yin, H., Betts, C., Lakhal, S. & Wood, M. J. 2011. Delivery of siRNA to the mouse brain by systemic injection of targeted exosomes. *Nat Biotechnol*, 29, 341–345.

46 Gupta, S. K., Bang, C. & Thum, T. 2010. Circulating microRNAs as biomarkers and potential paracrine mediators of cardiovascular disease. *Circ Cardiovasc Genet*, 3, 484–488.

47 Koberle, V., Pleli, T., Schmithals, C., Augusto Alonso, E., Haupenthal, J., Bonig, H., Peveling-Oberhag, J., Biondi, R. M., Zeuzem, S., Kronenberger, B., Waidmann, O. & Piiper, A. 2013. Differential stability of cell-free circulating microRNAs: implications for their utilization as biomarkers. *PLoS One*, 8, e75184.

48 Bail, S., Swerdel, M., Liu, H., Jiao, X., Goff, L. A., Hart, R. P. & Kiledjian, M. 2010. Differential regulation of microRNA stability. *RNA*, 16, 1032–1039.

49 Hibio, N., Hino, K., Shimizu, E., Nagata, Y. & Ui-Tei, K. 2012. Stability of miRNA 5′ terminal and seed regions is correlated with experimentally observed miRNA-mediated silencing efficacy. *Sci Rep*, 2, 996.

50 Rissland, O. S., Hong, S. J. & Bartel, D. P. 2011. MicroRNA destabilization enables dynamic regulation of the miR-16 family in response to cell-cycle changes. *Mol Cell*, 43, 993–1004.

51 Winter, J. & Diederichs, S. 2011. Argonaute proteins regulate microRNA stability: increased microRNA abundance by Argonaute proteins is due to microRNA stabilization. *RNA Biol*, 8, 1149–1157.

52 Mitchell, P. S., Parkin, R. K., Kroh, E. M., Fritz, B. R., Wyman, S. K., Pogosova-Agadjanyan, E. L., Peterson, A., Noteboom, J., O'briant, K. C., Allen, A., Lin, D. W., Urban, N., Drescher, C. W., Knudsen, B. S., Stirewalt, D. L., Gentleman, R., Vessella, R. L., Nelson, P. S., Martin, D. B. & Tewari, M. 2008. Circulating microRNAs as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A*, 105, 10513–10518.

53 Ge, Q., Zhou, Y., Lu, J., Bai, Y., Xie, X. & Lu, Z. 2014. miRNA in plasma exosome is stable under different storage conditions. *Molecules*, 19, 1568–1575.

54 Jung, M., Schaefer, A., Steiner, I., Kempkensteffen, C., Stephan, C., Erbersdobler, A. & Jung, K. 2010. Robust microRNA stability in degraded RNA preparations from human tissue and cell samples. *Clin Chem*, 56, 998–1006.

55 Yun, S. J., Jeong, P., Kim, W. T., Kim, T. H., Lee, Y. S., Song, P. H., Choi, Y. H., Kim, I. Y., Moon, S. K. & Kim, W. J. 2012. Cell-free microRNAs in urine as diagnostic and prognostic biomarkers of bladder cancer. *Int J Oncol*, 41, 1871–1878.

56 Lv, L. L., Cao, Y., Liu, D., Xu, M., Liu, H., Tang, R. N., Ma, K. L. & Liu, B. C. 2013. Isolation and quantification of microRNAs from urinary exosomes/microvesicles for biomarker discovery. *Int J Biol Sci*, 9, 1021–1031.

57 Mall, C., Rocke, D. M., Durbin-Johnson, B. & Weiss, R. H. 2013. Stability of miRNA in human urine supports its biomarker potential. *Biomark Med*, 7, 623–631.

58 Dong, H., Lei, J., Ding, L., Wen, Y., Ju, H. & Zhang, X. 2013. MicroRNA: function, detection, and bioanalysis. *Chem Rev*, 113, 6207–6233.

59 Papadopoulos, T., Belliere, J., Bascands, J. L., Neau, E., Klein, J. & Schanstra, J. P. 2015. miRNAs in urine: a mirror image of kidney disease? *Expert Rev Mol Diagn*, 15, 361–374.

60 Ebhardt, H. A., Fedynak, A. & Fahlman, R. P. 2010. Naturally occurring variations in sequence length creates microRNA isoforms that differ in argonaute effector complex specificity. *Silence*, 1, 12.

61 Koscianska, E., Starega-Roslan, J., Sznajder, L. J., Olejniczak, M., Galka-Marciniak, P. & Krzyzosiak, W. J. 2011. Northern blotting analysis of microRNAs, their precursors and RNA interference triggers. *BMC Mol Biol*, 12, 14.

62 Starega-Roslan, J., Krol, J., Koscianska, E., Kozlowski, P., Szlachcic, W. J., Sobczak, K. & Krzyzosiak, W. J. 2011. Structural basis of microRNA length variety. *Nucleic Acids Res*, 39, 257–268.

63 Johnson, B. N. & Mutharasan, R. 2014. Biosensor-based microRNA detection: techniques, design, performance, and challenges. *Analyst*, 139, 1576–1588.

64 Pall, G. S., Codony-Servat, C., Byrne, J., Ritchie, L. & Hamilton, A. 2007. Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. *Nucleic Acids Res*, 35, e60.

65 Kim, S. W., Li, Z., Moore, P. S., Monaghan, A. P., Chang, Y., Nichols, M. & John, B. 2010. A sensitive non-radioactive northern blot method to detect small RNAs. *Nucleic Acids Res*, 38, e98.

66 Mirco, C., Paul, C., Tania, N. & Vladimir, B. 2013. *Expression Profiling of MicroRNAs by Quantitative Real-Time PCR*. PCR Technology. CRC Press.

67 Baker, M. 2010. MicroRNA profiling: separating signal from noise. *Nat Methods*, 7, 687–692.

68 Chugh, P. & Dittmer, D. P. 2012. Potential pitfalls in microRNA profiling. *Wiley Interdiscip Rev RNA*, 3, 601–616.

69 Pritchard, C. C., Cheng, H. H. & Tewari, M. 2012. MicroRNA profiling: approaches and considerations. *Nat Rev Genet*, 13, 358–369.

70 Chatziioannou, A., Moulos, P. & Kolisis, F. N. 2009. Gene ARMADA: an integrated multi-analysis platform for microarray data implemented in MATLAB. *BMC Bioinformatics*, 10, 354.

71 Duan, D., Zheng, K. X., Shen, Y., Cao, R., Jiang, L., Lu, Z., Yan, X. & Li, J. 2011. Label-free high-throughput microRNA expression profiling from total RNA. *Nucleic Acids Res*, 39, e154.

72 Lee, J. M., Cho, H. & Jung, Y. 2010. Fabrication of a structure-specific RNA binder for array detection of label-free microRNA. *Angew Chem Int Ed Engl*, 49, 8662–8665.

73 Metzker, M. L. 2010. Sequencing technologies—the next generation. *Nat Rev Genet*, 11, 31–46.

74 Thevenot, D. R., Toth, K., Durst, R. A. & Wilson, G. S. 2001. Electrochemical biosensors: recommended definitions and classification. *Biosens Bioelectron*, 16, 121–131.

75 Hamidi-Asl, E., Palchetti, I., Hasheminejad, E. & Mascini, M. 2013. A review on the electrochemical biosensors for determination of microRNAs. *Talanta*, 115, 74–83.

76 Johnson, B. N. & Mutharasan, R. 2012. Sample preparation-free, real-time detection of microRNA in human serum using piezoelectric cantilever biosensors at attomole level. *Anal Chem*, 84, 10426–10436.

77 Zhang, J. & Cui, D. 2013. Nanoparticle-based optical detection of MicroRNA. *Nano Biomed Eng*, 5(1), 1–10.

78 Geiss, G. K., Bumgarner, R. E., Birditt, B., Dahl, T., Dowidar, N., Dunaway, D. L., Fell, H. P., Ferree, S., George, R. D., Grogan, T., James, J. J., Maysuria, M., Mitton, J. D., Oliveri, P., Osborn, J. L., Peng, T., Ratcliffe, A. L., Webster, P. J., Davidson, E. H., Hood, L. & Dimitrov, K. 2008. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol*, 26, 317–325.

79 Rossi, S. & Calin, G. A. 2013. Bioinformatics, non-coding RNAs and its possible application in personalized medicine. *Adv Exp Med Biol*, 774, 21–37.

80 Lai, E. C. 2002. Micro RNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30, 363–364.

81 Peterson, S. M., Thompson, J. A., Ufkin, M. L., Sathyanarayana, P., Liaw, L. & Congdon, C. B. 2014. Common features of microRNA target prediction tools. *Front Genet*, 5, 23.

82 Ritchie, W., Rasko, J. E. & Flamant, S. 2013. MicroRNA target prediction and validation. *Adv Exp Med Biol*, 774, 39–53.

83 Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V. & Ding, Y. 2007. Potent effect of target structure on microRNA function. *Nat Struct Mol Biol*, 14, 287–294.

84 Ritchie, W., Flamant, S. & Rasko, J. E. 2009. Predicting microRNA targets and functions: traps for the unwary. *Nat Methods*, 6, 397–398.

85 Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. & Marks, D. S. 2003. MicroRNA targets in drosophila. *Genome Biol*, 5, R1.

86 Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. 2010. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol*, 11, R90.

87 Paraskevopoulou, M. D., Georgakilas, G., Kostoulas, N., Vlachos, I. S., Vergoulis, T., Reczko, M., Filippidis, C., Dalamagas, T. & Hatzigeorgiou, A. G. 2013. DIANA-microT web server v5.0: service integration into

miRNA functional analysis workflows. *Nucleic Acids Res*, 41, W169–W173.

88 Friedman, R. C., Farh, K. K., Burge, C. B. & Bartel, D. P. 2009. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19, 92–105.

89 Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A. & Bartel, D. P. 2011. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol*, 18, 1139–1146.

90 Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P. & Bartel, D. P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27, 91–105.

91 Lewis, B. P., Burge, C. B. & Bartel, D. P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120, 15–20.

92 Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. & Li, T. 2009. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res*, 37, D105–D110.

93 Hsu, S. D., Tseng, Y. T., Shrestha, S., Lin, Y. L., Khaleel, A., Chou, C. H., Chu, C. F., Huang, H. Y., Lin, C. M., Ho, S. Y., Jian, T. Y., Lin, F. M., Chang, T. H., Weng, S. L., Liao, K. W., Liao, I. E., Liu, C. C. & Huang, H. D. 2014. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res*, 42, D78–D85.

94 Li, J. H., Liu, S., Zhou, H., Qu, L. H. & Yang, J. H. 2014. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*, 42, D92–D97.

95 Yang, J. H., Li, J. H., Shao, P., Zhou, H., Chen, Y. Q. & Qu, L. H. 2011. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res*, 39, D202–D209.

96 Vlachos, I. S., Paraskevopoulou, M. D., Karagkouni, D., Georgakilas, G., Vergoulis, T., Kanellos, I., Anastasopoulos, I. L., Maniou, S., Karathanou, K., Kalfakakou, D., Fevgas, A., Dalamagas, T. & Hatzigeorgiou, A. G. 2015. DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res*, 43, D153–D159.

97 Sokilde, R., Newie, I., Persson, H., Borg, A. & Rovira, C. 2015. Passenger strand loading in overexpression experiments using microRNA mimics. *RNA Biol*, 12, 787–791.

98 Ebert, M. S., Neilson, J. R. & Sharp, P. A. 2007. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat Methods*, 4, 721–726.

99 Elmen, J., Lindow, M., Schutz, S., Lawrence, M., Petri, A., Obad, S., Lindholm, M., Hedtjarn, M., Hansen, H. F., Berger, U., Gullans, S., Kearney, P., Sarnow, P., Straarup, E. M. & Kauppinen, S. 2008. LNA-mediated microRNA silencing in non-human primates. *Nature*, 452, 896–899.

100 Wen, J., Parker, B. J., Jacobsen, A. & Krogh, A. 2011. MicroRNA transfection and AGO-bound CLIP-seq data sets reveal distinct determinants of miRNA action. *RNA*, 17, 820–834.

101 Easow, G., Teleman, A. A. & Cohen, S. M. 2007. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13, 1198–1204.

102 Hendrickson, D. G., Hogan, D. J., Herschlag, D., Ferrell, J. E. & Brown, P. O. 2008. Systematic identification of mRNAs recruited to argonaute 2 by specific microRNAs and corresponding changes in transcript abundance. *PLoS One*, 3, e2126.

103 Karginov, F. V., Conaco, C., Xuan, Z., Schmidt, B. H., Parker, J. S., Mandel, G. & Hannon, G. J. 2007. A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci U S A*, 104, 19291–19296.

104 Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., Darnell, J. C. & Darnell, R. B. 2008. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456, 464–469.

105 Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr., Jungkamp, A. C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M. & Tuschl, T. 2010. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141, 129–141.

106 Haecker, I. & Renne, R. 2014. HITS-CLIP and PAR-CLIP advance viral miRNA targetome analysis. *Crit Rev Eukaryot Gene Exp*, 24, 101–116.

107 Tarang, S. & Weston, M. D. 2014. Macros in microRNA target identification: a comparative analysis of in silico, in vitro, and in vivo approaches to microRNA target identification. *RNA Biol*, 11, 324–333.

108 Ong, S. E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. & Mann, M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol Cell Proteomics*, 1, 376–386.

109 Zhu, S., Si, M. L., Wu, H. & Mo, Y. Y. 2007. MicroRNA-21 targets the tumor suppressor gene tropomyosin 1 (TPM1). *J Biol Chem*, 282, 14328–14336.

110 Mermelekas, G., Vlahou, A. & Zoidakis, J. 2015. SRM/MRM targeted proteomics as a tool for biomarker validation and absolute quantification in human urine. *Expert Rev Mol Diagn*, 15, 1441–1454.

111 Vatolin, S., Navaratne, K. & Weil, R. J. 2006. A novel method to detect functional microRNA targets. *J Mol Biol*, 358, 983–996.

112 Orom, U. A. & Lund, A. H. 2007. Isolation of microRNA targets using biotinylated synthetic microRNAs. *Methods*, 43, 162–165.

113 Hsu, R. J., Yang, H. J. & Tsai, H. J. 2009. Labeled microRNA pull-down assay system: an experimental approach for high-throughput identification of microRNA-target mRNAs. *Nucleic Acids Res*, 37, e77.

114 Llave, C., Xie, Z., Kasschau, K. D. & Carrington, J. C. 2002. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science*, 297, 2053–2056.

115 Thomson, D. W., Bracken, C. P. & Goodall, G. J. 2011. Experimental strategies for microRNA target identification. *Nucleic Acids Res*, 39, 6845–6853.

116 Addo-Quaye, C., Eshoo, T. W., Bartel, D. P. & Axtell, M. J. 2008. Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol*, 18, 758–762.

117 German, M. A., Pillay, M., Jeong, D. H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., De Paoli, E., Lu, C., Schroth, G., Meyers, B. C. & Green, P. J. 2008. Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol*, 26, 941–946.

118 Chen, Y., Gao, D. Y. & Huang, L. 2015. In vivo delivery of miRNAs for cancer therapy: challenges and strategies. *Adv Drug Deliv Rev*, 81, 128–141.

119 Zhang, Y., Wang, Z. & Gemeinhart, R. A. 2013. Progress in microRNA delivery. *J Control Release*, 172, 962–974.

120 Broderick, J. A. & Zamore, P. D. 2011. MicroRNA therapeutics. *Gene Ther*, 18, 1104–1110.

121 Chistiakov, D. A., Sobenin, I. A. & Orekhov, A. N. 2012. Strategies to deliver microRNAs as potential therapeutics in the treatment of cardiovascular pathology. *Drug Deliv*, 19, 392–405.

122 Grunweller, A. & Hartmann, R. K. 2007. Locked nucleic acid oligonucleotides: the next generation of antisense agents? *BioDrugs*, 21, 235–243.

123 Henry, J. C., Azevedo-Pouly, A. C. & Schmittgen, T. D. 2011. MicroRNA replacement therapy for cancer. *Pharm Res*, 28, 3030–3042.

124 Peacock, H., Fucini, R. V., Jayalath, P., Ibarra-Soza, J. M., Haringsma, H. J., Flanagan, W. M., Willingham, A. & Beal, P. A. 2011. Nucleobase and ribose modifications control immunostimulation by a microRNA-122-mimetic RNA. *J Am Chem Soc*, 133, 9200–9203.

125 Horie, T., Ono, K., Nishi, H., Iwanaga, Y., Nagao, K., Kinoshita, M., Kuwabara, Y., Takanabe, R., Hasegawa, K., Kita, T. & Kimura, T. 2009. MicroRNA-133 regulates the expression of GLUT4 by targeting KLF15 and is involved in metabolic control in cardiac myocytes. *Biochem Biophys Res Commun*, 389, 315–320.

126 Zhang, Y., Liu, D., Chen, X., Li, J., Li, L., Bian, Z., Sun, F., Lu, J., Yin, Y., Cai, X., Sun, Q., Wang, K., Ba, Y., Wang, Q., Wang, D., Yang, J., Liu, P., Xu, T., Yan, Q., Zhang, J., Zen, K. & Zhang, C. Y. 2010. Secreted monocytic miR-150 enhances targeted endothelial cell migration. *Mol Cell*, 39, 133–144.

127 Koppers-Lalic, D., Hogenboom, M. M., Middeldorp, J. M. & Pegtel, D. M. 2013. Virus-modified exosomes for targeted RNA delivery; a new approach in nanomedicine. *Adv Drug Deliv Rev*, 65, 348–356.

128 Pegtel, D. M., Cosmopoulos, K., Thorley-Lawson, D. A., Van Eijndhoven, M. A., Hopmans, E. S., Lindenberg, J. L., De Gruijl, T. D., Wurdinger, T. & Middeldorp, J. M. 2010. Functional delivery of viral miRNAs via exosomes. *Proc Natl Acad Sci U S A*, 107, 6328–6333.

129 Wu, Y., Crawford, M., Yu, B., Mao, Y., Nana-Sinkam, S. P. & Lee, L. J. 2011. MicroRNA delivery by cationic lipoplexes for lung cancer therapy. *Mol Pharm*, 8, 1381–1389.

130 Hsu, S. H., Yu, B., Wang, X., Lu, Y., Schmidt, C. R., Lee, R. J., Lee, L. J., Jacob, S. T. & Ghoshal, K. 2013. Cationic lipid nanoparticles for therapeutic delivery of siRNA and miRNA to murine liver tumor. *Nanomedicine*, 9, 1169–1180.

131 Shi, S., Han, L., Gong, T., Zhang, Z. & Sun, X. 2013. Systemic delivery of microRNA-34a for cancer stem cell therapy. *Angew Chem Int Ed Engl*, 52, 3901–3905.

132 Liu, X., Feng, J., Tang, L., Liao, L., Xu, Q. & Zhu, S. 2015. The regulation and function of miR-21-FOXO3a-miR-34b/c signaling in breast cancer. *Int J Mol Sci*, 16, 3148–3162.

133 Trang, P., Wiggins, J. F., Daige, C. L., Cho, C., Omotola, M., Brown, D., Weidhaas, J. B., Bader, A. G. & Slack, F. J. 2011. Systemic delivery of tumor suppressor microRNA mimics using a neutral lipid emulsion inhibits lung tumors in mice. *Mol Ther*, 19, 1116–1122.

134 Wiggins, J. F., Ruffino, L., Kelnar, K., Omotola, M., Patrawala, L., Brown, D. & Bader, A. G. 2010. Development of a lung cancer therapeutic based on the tumor suppressor microRNA-34. *Cancer Res*, 70, 5923–5930.

135 Chen, Y., Zhu, X., Zhang, X., Liu, B. & Huang, L. 2010. Nanoparticles modified with tumor-targeting scFv deliver siRNA and miRNA for cancer therapy. *Mol Ther*, 18, 1650–1656.

136 Su, J., Baigude, H., Mccarroll, J. & Rana, T. M. 2011. Silencing microRNA by interfering nanoparticles in mice. *Nucleic Acids Res*, 39, e38.

137 Kulkarni, R. K., Moore, E. G., Hegyeli, A. F. & Leonard, F. 1971. Biodegradable poly(lactic acid) polymers. *J Biomed Mater Res*, 5, 169–181.

138 Blum, J. S. & Saltzman, W. M. 2008. High loading efficiency and tunable release of plasmid DNA encapsulated in submicron particles fabricated from PLGA conjugated with poly-L-lysine. *J Control Release*, 129, 66–72.

139 Panyam, J. & Labhasetwar, V. 2003. Biodegradable nanoparticles for drug and gene delivery to cells and tissue. *Adv Drug Deliv Rev*, 55, 329–347.

140 Paulmurugan, R., Sekar, N. M. & Sekar, T. V. 2012. Biodegradable polymer nanocarriers for therapeutic antisense microRNA delivery in living animals, *Proc SPIE*, 8232, 823208.

141 Van Vlerken, L. E., Vyas, T. K. & Amiji, M. M. 2007. Poly(ethylene glycol)-modified nanocarriers for tumor-targeted and intracellular delivery. *Pharm Res*, 24, 1405–1414.

142 Boussif, O., Lezoualc'h, F., Zanta, M. A., Mergny, M. D., Scherman, D., Demeneix, B. & Behr, J. P. 1995. A versatile vector for gene and oligonucleotide transfer into cells in culture and in vivo: polyethylenimine. *Proc Natl Acad Sci U S A*, 92, 7297–7301.

143 Chiou, G. Y., Cherng, J. Y., Hsu, H. S., Wang, M. L., Tsai, C. M., Lu, K. H., Chien, Y., Hung, S. C., Chen, Y. W., Wong, C. I., Tseng, L. M., Huang, P. I., Yu, C. C., Hsu, W. H. & Chiou, S. H. 2012. Cationic polyurethanes-short branch PEI-mediated delivery of Mir145 inhibited epithelial-mesenchymal transdifferentiation and cancer stem-like properties and in lung adenocarcinoma. *J Control Release*, 159, 240–250.

144 Yang, Y. P., Chien, Y., Chiou, G. Y., Cherng, J. Y., Wang, M. L., Lo, W. L., Chang, Y. L., Huang, P. I., Chen, Y. W., Shih, Y. H., Chen, M. T. & Chiou, S. H. 2012. Inhibition of cancer stem cell-like properties and reduced chemoradioresistance of glioblastoma using microRNA145 with cationic polyurethane-short branch PEI. *Biomaterials*, 33, 1462–1476.

145 Ghosh, R., Singh, L. C., Shohet, J. M. & Gunaratne, P. H. 2013. A gold nanoparticle platform for the delivery of functional microRNAs into cancer cells. *Biomaterials*, 34, 807–816.

146 Tivnan, A., Orr, W. S., Gubala, V., Nooney, R., Williams, D. E., Mcdonagh, C., Prenter, S., Harvey, H., Domingo-Fernandez, R., Bray, I. M., Piskareva, O., Ng, C. Y., Lode, H. N., Davidoff, A. M. & Stallings, R. L. 2012. Inhibition of neuroblastoma tumor growth by targeted delivery of microRNA-34a using anti-disialoganglioside GD2 coated nanoparticles. *PLoS One*, 7, e38129.

147 Takata, A., Otsuka, M., Yoshikawa, T., Kishikawa, T., Ohno, M. & Koike, K. 2013. MicroRNAs and liver function. *Minerva Gastroenterol Dietol*, 59, 187–203.

148 Elmen, J., Lindow, M., Silahtaroglu, A., Bak, M., Christensen, M., Lind-Thomsen, A., Hedtjarn, M., Hansen, J. B., Hansen, H. F., Straarup, E. M., Mccullagh, K., Kearney, P. & Kauppinen, S. 2008. Antagonism of microRNA-122 in mice by systemically administered LNA-antimiR leads to up-regulation of a large set of predicted target mRNAs in the liver. *Nucleic Acids Res*, 36, 1153–1162.

149 Lanford, R. E., Hildebrandt-Eriksen, E. S., Petri, A., Persson, R., Lindow, M., Munk, M. E., Kauppinen, S. & Orum, H. 2010. Therapeutic silencing of microRNA-122 in primates with chronic hepatitis C virus infection. *Science*, 327, 198–201.

150 Janssen, H. L., Reesink, H. W., Lawitz, E. J., Zeuzem, S., Rodriguez-Torres, M., Patel, K., Van Der Meer, A. J., Patick, A. K., Chen, A., Zhou, Y., Persson, R., King, B. D., Kauppinen, S., Levin, A. A. & Hodges, M. R. 2013. Treatment of HCV infection by targeting microRNA. *N Engl J Med*, 368, 1685–1694.

151 Gebert, L. F., Rebhan, M. A., Crivelli, S. E., Denzler, R., Stoffel, M. & Hall, J. 2014. Miravirsen (SPC3649) can inhibit the biogenesis of miR-122. *Nucleic Acids Res*, 42, 609–621.

152 He, L., He, X., Lim, L. P., De Stanchina, E., Xuan, Z., Liang, Y., Xue, W., Zender, L., Magnus, J., Ridzon, D., Jackson, A. L., Linsley, P. S., Chen, C., Lowe, S. W., Cleary, M. A. & Hannon, G. J. 2007. A microRNA component of the p53 tumour suppressor network. *Nature*, 447, 1130–1134.

153 Li, J., Lam, M. & Reproducibility Project: Cancer, B. 2015. Registered report: the microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. *Elife*, 4, e06434.

154 Liu, C., Kelnar, K., Liu, B., Chen, X., Calhoun-Davis, T., Li, H., Patrawala, L., Yan, H., Jeter, C., Honorio, S., Wiggins, J. F., Bader, A. G., Fagin, R., Brown, D. & Tang, D. G. 2011. The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. *Nat Med*, 17, 211–215.

155 Roy, S., Levi, E., Majumdar, A. P. & Sarkar, F. H. 2012. Expression of miR-34 is lost in colon cancer which can be re-expressed by a novel agent CDF. *J Hematol Oncol*, 5, 58.

156 Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F. & Croce, C. M. 2002. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci U S A*, 99, 15524–15529.

157 Dews, M., Homayouni, A., Yu, D., Murphy, D., Sevignani, C., Wentzel, E., Furth, E. E., Lee, W. M., Enders, G. H., Mendell, J. T. & Thomas-Tikhonenko, A. 2006. Augmentation of tumor angiogenesis by a Myc-activated microRNA cluster. *Nat Genet*, 38, 1060–1065.

158 He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J. & Hammond, S. M. 2005. A microRNA polycistron as a potential human oncogene. *Nature*, 435, 828–833.

159 Lu, Y., Thomson, J. M., Wong, H. Y., Hammond, S. M. & Hogan, B. L. 2007. Transgenic over-expression of the microRNA miR-17-92 cluster promotes proliferation and inhibits differentiation of lung epithelial progenitor cells. *Dev Biol*, 310, 442–453.

160 Medina, P. P., Nolde, M. & Slack, F. J. 2010. OncomiR addiction in an in vivo model of microRNA-21-induced pre-B-cell lymphoma. *Nature*, 467, 86–90.

161 Bommer, G. T., Gerin, I., Feng, Y., Kaczorowski, A. J., Kuick, R., Love, R. E., Zhai, Y., Giordano, T. J., Qin, Z. S., Moore, B. B., Macdougald, O. A., Cho, K. R. & Fearon, E. R. 2007. p53-mediated activation of miRNA34 candidate tumor-suppressor genes. *Curr Biol*, 17, 1298–1307.

162 Esquela-Kerscher, A., Trang, P., Wiggins, J. F., Patrawala, L., Cheng, A., Ford, L., Weidhaas, J. B., Brown, D., Bader, A. G. & Slack, F. J. 2008. The let-7 microRNA reduces tumor growth in mouse models of lung cancer. *Cell Cycle*, 7, 759–764.

163 Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K. L., Brown, D. & Slack, F. J. 2005. RAS is regulated by the let-7 microRNA family. *Cell*, 120, 635–647.

164 Markou, A., Tsaroucha, E. G., Kaklamanis, L., Fotinou, M., Georgoulias, V. & Lianidou, E. S. 2008. Prognostic value of mature microRNA-21 and microRNA-205 overexpression in non-small cell lung cancer by quantitative real-time RT-PCR. *Clin Chem*, 54, 1696–1704.

165 Zhou, W., Fong, M. Y., Min, Y., Somlo, G., Liu, L., Palomares, M. R., Yu, Y., Chow, A., O'Connor, S. T., Chin, A. R., Yen, Y., Wang, Y., Marcusson, E. G., Chu, P., Wu, J., Wu, X., Li, A. X., Li, Z., Gao, H., Ren, X., Boldin, M. P., Lin, P. C. & Wang, S. E. 2014. Cancer-secreted miR-105 destroys vascular endothelial barriers to promote metastasis. *Cancer Cell*, 25, 501–515.

166 Fernandez-Mercado, M., Manterola, L., Larrea, E., Goicoechea, I., Arestin, M., Armesto, M., Otaegui, D. & Lawrie, C. H. 2015. The circulating transcriptome as a source of non-invasive cancer biomarkers: concepts and controversies of non-coding and coding RNA in body fluids. *J Cell Mol Med*, 19, 2307–2323.

167 Ma, J., Lin, Y., Zhan, M., Mann, D. L., Stass, S. A. & Jiang, F. 2015. Differential miRNA expressions in peripheral blood mononuclear cells for diagnosis of lung cancer. *Lab Invest*, 95, 1197–1206.

168 Hornick, N. I., Huan, J., Doron, B., Goloviznina, N. A., Lapidus, J., Chang, B. H. & Kurre, P. 2015. Serum exosome MicroRNA as a minimally-invasive early biomarker of AML. *Sci Rep*, 5, 11295.

169 Esau, C., Davis, S., Murray, S. F., Yu, X. X., Pandey, S. K., Pear, M., Watts, L., Booten, S. L., Graham, M., Mckay, R., Subramaniam, A., Propp, S., Lollo, B. A., Freier, S., Bennett, C. F., Bhanot, S. & Monia, B. P. 2006. miR-122 regulation of lipid metabolism revealed by in vivo antisense targeting. *Cell Metab*, 3, 87–98.

170 Krutzfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M. & Stoffel, M. 2005. Silencing of microRNAs in vivo with "antagomirs". *Nature*, 438, 685–689.

171 Najafi-Shoushtari, S. H., Kristo, F., Li, Y., Shioda, T., Cohen, D. E., Gerszten, R. E. & Naar, A. M. 2010. MicroRNA-33 and the SREBP host genes cooperate to control cholesterol homeostasis. *Science*, 328, 1566–1569.

172 Rayner, K. J., Suarez, Y., Davalos, A., Parathath, S., Fitzgerald, M. L., Tamehiro, N., Fisher, E. A., Moore, K. J. & Fernandez-Hernando, C. 2010. MiR-33 contributes to the regulation of cholesterol homeostasis. *Science*, 328, 1570–1573.

173 Frost, R. J. & Olson, E. N. 2011. Control of glucose homeostasis and insulin sensitivity by the Let-7 family of microRNAs. *Proc Natl Acad Sci U S A*, 108, 21075–21080.

174 Zhu, H., Shyh-Chang, N., Segre, A. V., Shinoda, G., Shah, S. P., Einhorn, W. S., Takeuchi, A., Engreitz, J. M., Hagan, J. P., Kharas, M. G., Urbach, A., Thornton, J. E., Triboulet, R., Gregory, R. I., Consortium, D., Investigators, M., Altshuler, D. & Daley, G. Q. 2011. The Lin28/let-7 axis regulates glucose metabolism. *Cell*, 147, 81–94.

175 Cheng, L., Doecke, J. D., Sharples, R. A., Villemagne, V. L., Fowler, C. J., Rembach, A., Martins, R. N., Rowe, C. C., Macaulay, S. L., Masters, C. L., Hill, A. F., Australian Imaging, B. & Lifestyle Research, G. 2015. Prognostic serum miRNA biomarkers associated with Alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment. *Mol Psychiatry*, 20, 1188–1196.

176 Wang, J., Tan, L., Tan, L., Tian, Y., Ma, J., Tan, C. C., Wang, H. F., Liu, Y., Tan, M. S., Jiang, T. & Yu, J. T. 2015. Circulating microRNAs are promising novel biomarkers for drug-resistant epilepsy. *Sci Rep*, 5, 10201.

177 Luque, A., Farwati, A., Crovetto, F., Crispi, F., Figueras, F., Gratacos, E. & Aran, J. M. 2014. Usefulness of circulating microRNAs for the prediction of early preeclampsia at first-trimester of pregnancy. *Sci Rep*, 4, 4882.

178 Goretti, E., Wagner, D. R. & Devaux, Y. 2014. miRNAs as biomarkers of myocardial infarction: a step forward towards personalized medicine? *Trends Mol Med*, 20, 716–725.

6

Proteomics of Body Fluids

Szymon Filip and Jerome Zoidakis

Proteomics Laboratory, Biomedical Research Foundation, Academy of Athens, Athens, Greece

6.1 Introduction

Proteomics is a term used to describe large-scale protein studies. Since proteins control the key biological processes in the organism, one of the main aims of proteomics has become the identification of disease biomarkers [1]. There are two main sources for the identification of disease biomarkers: tissue samples and body fluids. The advantages of the latter compared with the former are lower invasiveness, lower cost, easier sample collection and storage, and less demanding sample processing [2]. However, the applicability of each body fluid to study a specific disease should be evaluated with caution, and body fluids that have a direct contact with a tissue of interest should be generally taken into consideration. For example, urine is an ideal source to study kidney diseases; however to identify biomarkers for Alzheimer's disease, cerebrospinal fluid (CSF) is more applicable [3, 4]. Additionally, body fluids can be very complex, and protein concentrations in such samples span over several orders of magnitude. Considering that the potential biomarker can be present at low concentrations, its identification can become challenging, since highly abundant proteins can mask the presence of low-abundance molecules. Therefore, to overcome the aforementioned obstacle, sample fractionation is commonly required prior to MS analysis [1, 2, 5]. The general comparison of advantages and disadvantages of tissue and body fluid proteomics is presented in Table 6.1.

Proteomics analysis of various body fluids that ensures obtaining high-quality results is similar in several steps, regardless of the type of analyzed sample. Each step should be performed according to well-established and generally accepted protocols and guidelines to enhance the reproducibility of the results [1].

There are many possible sources that can affect the reproducibility, for example, sample collection, storage, or preparation procedures. These include patient sample information, collection procedure of the body fluid, sample shipment conditions, storage temperature, protein fractionation and digestion, and so on. Additionally, instrument specifications (e.g., collision energy or isolation width) as well as data analysis procedures (e.g., data normalization, statistical tests) should be performed appropriately [6, 7].

The application of proteomics to study human diseases and translate the findings into clinical practice is called clinical proteomics. These findings can be used for early detection of the disease or evaluation of the prognosis. Additionally, proteomics findings can elucidate the mechanism of the disease and, thus, contribute to the identification of novel therapeutic targets [8, 9].

This chapter will concentrate on various aspects of body fluids proteomics, including sample collection and storage procedures, protein fractionation techniques, sample preparation for MS/MS analysis, commonly used analytical instruments, and bioinformatics aspects (differential expression analysis and statistical testing). Afterward, the focus will be set on the validation of findings. Finally, selected examples of the application of body fluids proteomics for the discovery of biomarkers will be presented. It is worth noting that most of the aspects of proteomics analysis do not vary between the analysis of body fluids and other samples (e.g., tissues or cell cultures).

6.2 General Workflow for Obtaining High-Quality Proteomics Results

Appropriate sample and data processing according to well-established guidelines are not the only aspects that guarantee obtaining high-quality proteomics results. Good quality proteomics analysis begins already on the level of study design, followed by proper and reproducible sample and data analysis and application of appropriate

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

Table 6.1 Comparison of tissue and body fluids proteomics in human.

	Body fluid	Tissue
Collection	Depends on a body fluid, but in most cases collected easily and with noninvasiveness or low invasiveness	Invasive procedure hinders the collection process
Availability	Typically available in large quantities and high numbers	Commonly available in low numbers, which can compromise the statistical power of performed analysis
Representation of a proteome	Typically represent the proteome of tissues of which the body fluid has contact with. However, blood represents the proteome of the whole organism	Proteome specific for a collected tissue
Protein concentration	Variable. Blood has very high protein content. However, proteins in other body fluids are present in lower concentration	Relatively high
Number of protein identifications	Without applying fractionation methods, generally low due to two reasons: (i) low protein concentration causes low-abundance molecules to be present below the limit of detection, and (ii) high-abundance proteins can mask the presence of low-abundance molecules	Generally high
Type of protein identifications	Without fractionation, mostly highly abundant proteins	Both high- and low-abundant proteins can be easily detected
Suitability for biomarker identification in the sample	Large number of available samples increases the statistical power. Low number of protein identifications can hinder the identification of putative biomarkers (especially those present at low concentrations)	Low sample availability reduces the statistical power. Large number of protein identifications facilitates the identification of putative biomarkers (especially those present at low concentrations)
Translation of biomarkers	Useful in clinical practice for diagnosis or prognosis of the disease	Useful for elucidation of the disease mechanism and identification of drug targets

Due to large number and variety of available body fluids, only general comparison is demonstrated, and thus, presented comparison cannot be translated to all body fluids used in proteomics.

statistical tests, finishing with validation of findings [10]. For biomarker discovery studies, obtaining high-quality proteomics results does not ensure that obtained findings will be translated into clinical practice. The path for translation of findings is laborious and time consuming with no guarantee of success. The main obstacles on this path include lacking funding, absence of platforms for validation of findings, cost-effectiveness of a biomarker, and inadequate knowledge of researches about steps required to translate their findings [11]. The general workflow for proteomics analysis is depicted in Figure 6.1.

Aspects that need to be taken into consideration in proteomics studies can be summarized as follows [10]:

1) Clinical questions and outcomes should be clearly defined:
   - Define the purpose of the biomarker.
   - Use standardized and well-accepted criteria to document the outcomes.
2) Study subjects should be properly chosen:
   - Define study groups.
   - Healthy controls are not always adequate controls for defining a biomarker specific for a disease and subjects with related or similar disease and should be taken into consideration.
   - Avoid unequal distribution of patients that can lead to potential bias during study design (e.g., unequal distribution of age or gender in the study group).
   - During the discovery phase, patients with unclear diagnosis can be omitted.
   - Use clinically accepted parameters as endpoints.
   - If possible, ensure that the same criteria were used for the classification of patients; if not possible, report this information.
   - Ensure the collection of sufficient demographic and clinical data on the subjects.
3) Each step of the experiment should be conducted appropriately according to available guidelines:
   - Ensure that the sample is collected properly and detailed description of this procedure is available.
   - Store the sample in appropriate conditions (e.g., at −20 or −80°C, unless stated otherwise).
   - Perform the experiments according to the guidelines and/or established protocols.
   - Use appropriate software for data processing.
4) Proper statistical tests for the identification of biomarker candidates should be applied:
   - Have sufficient samples size, according to power calculations.

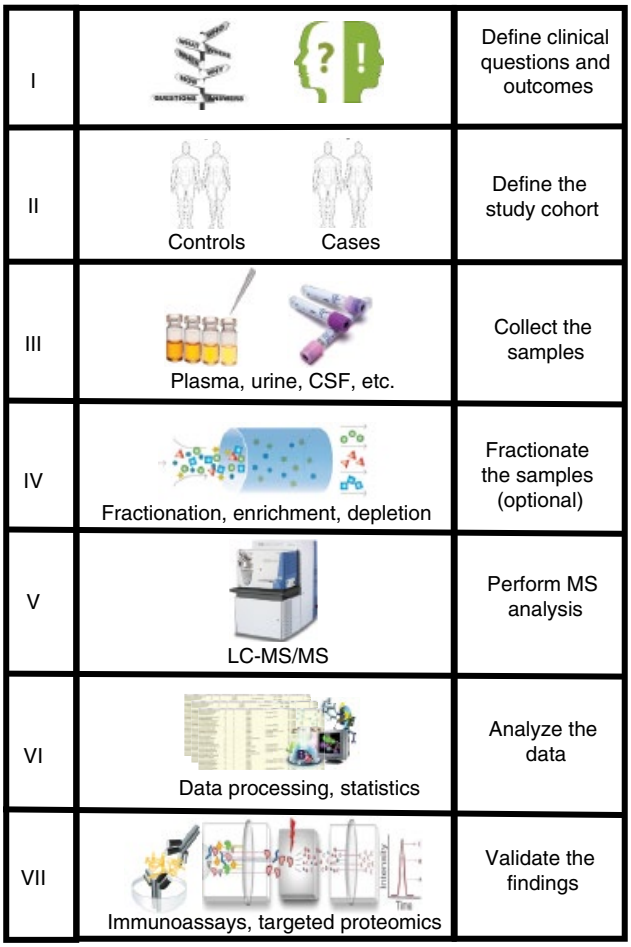


Figure 6.1 Graphical depiction of a general workflow of proteomics analysis.

- Try to avoid sample pooling.
- Use correct statistical tests and adjust for multiple testing.

5) The findings should be validated:
   - Use different sample cohort to confirm your findings.
   - If possible, use different methodology to validate the biomarker candidates.

6.3 Body Fluids

A number of body fluids can be analyzed by proteomics techniques. Blood (plasma and serum) and urine are the most common choices for the analysis; however these are not the only ones used. In this chapter, four body fluids, namely, blood, urine, CSF and saliva, will be discussed. The advantages and disadvantages of these fluids are summarized in Table 6.2.

6.3.1 Blood

Blood is one of the most complex sources of the human proteome [12]. This is mostly due to the enormous concentration differences in the proteins present, which span from millimolar to femtomolar. These differences are responsible for the masking effect and reduce the effectiveness of LC-MS/MS analysis, hindering the detection of low-abundance proteins. Therefore, if no fractionation is applied, the identification of blood proteins is usually restricted to highly abundant molecules. Additionally, blood proteins often undergo additional processing events and modifications, making the analysis even more complicated [13]. Two blood components are typically applied for the proteomics studies: plasma and serum.

Due to the easy accessibility and ability to "mirror" the organism status, blood is commonly applied for the identification of disease biomarkers, including various cancers (e.g., prostate, liver, breast, pancreatic), autoimmune diseases (e.g., systemic lupus erythematosus, multiple sclerosis), infectious diseases, diabetes, nephropathies, and cardiovascular disease [12, 14, 15].

6.3.1.1 Plasma

Plasma is the liquid portion of blood [12]. This straw-colored body fluid is equivalent of approximately 55% of total blood volume. Plasma consists mostly of water and is required for transportation of various substances within the body (e.g., hormones or glucose), immunological response, and withholding the pH and pressure of blood [3].

Table 6.2 Characteristics of four body fluids.

	Blood (plasma or serum)	Urine	Cerebrospinal fluid	Saliva
Applicability to study a disease	Any disease	Several diseases, mostly kidney or urogenital tract related	Some diseases, mostly neurological	Several diseases, but commonly for dental
Availability	+++	+++	+	++
Noninvasive collection	++	+++	–	+++
Protein concentration	+++	+[a]	+	+
Translation into clinical practice	+++	+++	++	+++

–, +, ++, +++ indicate rate of feasibility or concentration in the increasing order.
[a] Variable depends on studied disease.

The great advantage of plasma (and serum as described in the following) is its ubiquitous contact with every tissue in the body, making it a perfect source of information about the organism [5]. However, proteins derived from the tissues become substantially diluted in plasma, commonly being below the limit of detection by currently available MS/MS instruments [16]. Therefore, protein fractionation or enrichment strategies are commonly applied [2, 17]. Current advancements in mass spectrometry (MS) allow for more in-depth analysis of this body fluid. However large protein concentration range and complexity impair the proteomics biomarker studies.

6.3.1.2 Serum

Serum is derived from blood after removal of fibrinogen and clotting factors [12]. It is a liquid of a yellow color, obtained by centrifugation of previously coagulated blood. This process reduces the protein concentration, due to the removal of clotting factors. Still, these factors are not completely removed from the sample, and due to the putative interactions of proteins with the clotting factors and additional coagulation processes, the final levels of peptides and proteins in the sample are affected [3].

Similarly to plasma, serum is an easily accessible body fluid, which, due to its nature, represents the overall status of the organism. However, the same challenges for proteomics analysis characteristic of serum remain—enormous differences in individual protein concentrations and vast complexity [14].

6.3.2 Urine

Urine is a liquid secreted by the kidneys to the urinary bladder and ultimately excreted through the urethra. Urine from healthy individuals contains mostly water (95%). The rest of the components are waste substances produced by filtering of plasma by the kidneys. Due to these reasons, the protein concentration in urine (usually below 0.08 μg/μl) is approximately 1000-fold lower than in plasma (average protein concentration for plasma: 60–80 μg/μl) [3], unless the patient is proteinuric and demonstrates higher and variable protein concentrations in urine than healthy subjects [18]. The fact that 70% of proteins present in urine derive from kidney makes this body fluid a preferable source for studying various renal diseases [19].

Urine has several advantages: it can be obtained easily and repeatedly in large quantities, while proteins and peptides are already solubilized in the fluid. Importantly, urine demonstrates higher stability compared with plasma, since it is detained in the bladder for hours before collection, and thus, proteolytic processes have already finished before the collection. However, daily intake of fluids affects the concentration of peptides and proteins in urine. Additionally, due to several aspects of daily routine (e.g., diet, exercises, and hormonal changes), variation in the proteome can be observed during the day. The first issue can be tackled by normalizing the data based on creatinine values. In the case of the latter, these changes are mostly related to a small portion of the proteome and do not affect the majority of urinary proteins [20].

As mentioned urine is commonly applied to study a variety of renal diseases [21]. However, it was also applied to study irritable bowel syndrome, coronary artery disease [22], sepsis, and lung cancer [1].

6.3.3 Cerebrospinal Fluid (CSF)

CSF is a colorless fluid that surrounds the brain and the spinal cord. This fluid is in contact with central nervous system and removes waste substances, transports nutrients, and acts as a mechanical support for the brain. These qualities make it a perfect source for studying neurodegenerative disorders. CSF is produced by the choroid plexus. It is generally an ultrafiltrate of plasma; however, the interstitial fluid of the nervous tissue also impacts the molecular content of CSF [3, 4].

Most of the protein content (80%) comes from blood and the rest from central nervous system. Therefore, highly abundant plasma proteins are also abundant in CSF. Still, the protein concentration in CSF (0.2–0.8 μg/μl) is substantially lower than in plasma (60–80 μg/μl) [4].

Even though CSF seems to be an ideal source to study neurological diseases, it has some disadvantages. Firstly, its collection is invasive and requires a lumbar puncture. Secondly, alcohol consumption or smoking can influence the composition of the fluid [4]. Thirdly, identification of disease biomarkers in CSF can be hindered, due to the fact that most of the proteins derive from plasma [2]. Lastly, blood contamination of the CSF is relatively common, and 14–20% of all lumbar puncture procedures end with the sample contamination. Considering very low protein concentration of CSF compared with plasma, even small leakage of the latter can significantly change protein concentrations in the sample, skewing the analysis and leading to the identification of nonspecific biomarkers [4].

6.3.4 Saliva

Saliva is fluid that contains a secreted mixture of several salivary glands. Its function is not only restricted to food digestion, tasting, and swallowing, but it also plays a role in lubricating the oral tissues, ensuring integrity of teeth and protection against bacteria and viruses [23].

Saliva contains mainly water (99%). The remaining 1% consists of inorganic ions, organic secretion substances, products of food degradation, lipids, and proteins. The most abundant proteins can be classified into two classes: gland derived (e.g., α-amylase, cystatins, histatins) and plasma derived (albumin, serotransferrin, sIgA) [24]. Average protein concentration in saliva is in the range of 0.5–2 mg/ml [25].

Saliva could be considered as an alternative source to plasma for proteomics studies, since it can be obtained easily, noninvasively and in large quantities, while its collection and storage do not require a highly trained personnel. Due to very low sample collection cost, saliva seems to be a preferable diagnostic body fluid in Third World countries. However, sample collection procedures for this body fluid have not been well established yet. Additionally, protein concentration in saliva is very low, while the range of the concentration of individual molecules is large (e.g., in mg/ml for α-amylase in pg/ml for cytokines). Therefore, identification of biomarkers in this body fluid requires sensitive MS instruments [23–28].

Salivary proteomics has been implemented in the study of a variety of diseases including periodontal disease, squamous cell carcinoma, head and neck cancer, breast cancer, and diabetes [23, 24, 27].

6.4 Sample Collection and Storage

Sample collection procedures are very well described in the following chapter [3]. Additionally, such protocols are commonly available on several websites and/or publications. Therefore, instead of describing the guidelines, they will be summarized in Table 6.3.

6.5 Sample Preparation for MS/MS Analysis

Sample preparation for MS/MS analysis depends on study requirements and should be adjusted to each study individually. Factors that affect the preparation procedures are type of sample (plasma, urine, etc.), sample complexity, protein concentration, type of proteins of interest (all proteins, glycoproteins, phosphoproteins, etc.), aim of the study (protein profiling, biomarker discovery, etc.), analysis method, and instrument running conditions. Protocols or guidelines for most of the sample processing steps have already been established. Still, many variations exist, while some steps are to some degree established empirically in each lab (e.g., LC-MS/MS running conditions). Since all of these aspects are interconnected, there are no standard guidelines available that would cover all of them.

This section will focus on three main parts: fractionation of intact proteins, sample preparation for MS/MS runs (tryptic digestion methods), and peptide fractionation techniques.

6.5.1 Protein Separation

A number of separation techniques are available for proteomics analysis. The samples are commonly processed by one or more fractionation techniques. Still, the application of such strategies is optional, and samples can be prepared without applying protein separation prior to the MS analysis.

In the literature the most common categories of separation techniques include gel-based and gel-free approaches [31–34] and electrophoresis-based and liquid chromatography approaches [1, 35]. Both categories

Table 6.3 Sample collection guidelines references.

Sample type	Guideline source	Link	References
Plasma	National Institutes of Health	http://edrn.nci.nih.gov/resources/standard-operating-procedures/standard-operating-procedures/plasma-sop.pdf	[3]
Serum	National Institutes of Health	http://edrn.nci.nih.gov/resources/standard-operating-procedures/standard-operating-procedures/serum-sop.pdf	[3]
Urine	Urine and Kidney Proteomics COST Action or Human Kidney and Urine Proteome Project	http://eurokup.org/sites/default/files/StandardProtocolforUrineCollection.pdf or http://www.hkupp.org/Urine%20collectiion%20Documents.htm	[3, 29]
Cerebrospinal fluid	—	—	[3, 30]
Saliva	World Health Organization, International Agency for Research on Cancer	http://www.iarc.fr/en/publications/pdfs-online/wrk/wrk2/Standards_ProtocolsBRC.pdf	[3]

share many similarities: most of the electrophoretic methods (excluding free-flow electrophoresis (FFE)) are gel based, while chromatographic approaches are included in gel-free techniques.

In this chapter, the presented methods will be divided into electrophoresis-based and liquid chromatography (LC) techniques, while the focus will be set mostly on the separation of intact proteins. The advantages and disadvantages of applying each technique are summarized in Table 6.4. The application of fractionation methods in individual studies is well described in the following reviews [1, 31–33].

6.5.1.1 Electrophoresis-Based Methods

Electrophoresis-based methods include one-dimensional electrophoresis (1DE or more commonly SDS-PAGE), two-dimensional electrophoresis (2DE), isoelectric focusing (IEF), and two-dimensional fluorescence difference gel electrophoresis (2D DIGE) [31]. These techniques rely on the migration of charged proteins in the medium [1]. The most common method is SDS-PAGE, where proteins are denaturated and coated with sodium dodecyl sulfate (SDS) detergent. In such conditions, the separation is based mostly on the molecular weight of each protein [34, 35].

IEF, on the other hand, uses the differences in the isoelectric points (p*I*) of proteins. The proper resolving power is achieved by applying pH gradient strips. Proteins migrate in the electric field until they reach their pI, where the charge is equal to zero [34, 36].

2DE can be considered as a combination of both gel-based 1DE and IEF. Proteins are firstly separated based on the pI as in the case of IEF (first dimension) and afterward based on their molecular weight, as for SDS-PAGE (second dimension). 2DE allows to visualize up to thousands of gel spots and differentiates between various isoforms of the same protein [35]. Since the main issue of 2DE is the comparison of multiple gels, a modification of the method was introduced: 2D DIGE [1]. This method overcomes the obstacle related to comparative analysis of different samples, as proteins from different samples are labeled with distinct fluorescent dyes, mixed together and loaded onto the same gel. This way, multiple samples can be compared with each other in one experiment [35].

The presented methods are gel based. However, some liquid-based modifications exist. For example, liquid-phase IEF is characterized by higher loading capacity and increased protein recovery compared with its gel-based counterpart [33]. FFE technology was also introduced, which, in contrast to most of the electrophoretic methods, does not require any solid matrix to separate proteins. In this method, the sample is injected continuously and is moving in a laminar flow in a chamber composed of two glass plates. The electric field is applied vertically to the laminar flow direction, and the samples can be divided based on their charge, p*I*, or mobility. This method, in comparison with traditional electrophoresis, improves the possibility of detecting low-abundance proteins [33].

6.5.1.2 Liquid Chromatography Methods

The separation principle of LC is based on the interactions of the analytes with the stationary phase. Since each of the components in the sample interacts differentially with the sorbent, the sample can be fractionated. These methods can be divided into two categories: physiochemical strategies and affinity based. Due to the fact that physiochemical strategies are well known and established, this section will provide only short description of such methods. More focus will be put on affinity-based methods that are becoming more and more popular, as they are applicable to achieve rapid and selective purification of the proteins. However, in some cases (e.g., abundant protein depletion), the available data contain contradictory results, questioning the added value of such strategies in proteomics studies.

6.5.1.2.1 Physiochemical-Based Separation

The most commonly applied chromatography techniques using physiochemical properties for the separation of intact proteins are reverse-phase liquid chromatography (RP-LC), hydrophobic interaction liquid chromatography (HILIC), and ion-exchange [34] or size-exclusion chromatography (SEC) [1].

In RP-LC the analytes are separated based on their differences in hydrophobicity, where nonpolar stationary phase and a polar mobile phase are applied. Hydrophilic proteins are eluted first. As the polarity of the mobile phase is gradually reduced, other (more hydrophobic) proteins can elute [37].

Similarly to RP-LC, HILIC separates proteins on the basis of their hydrophobicity and uses comparable types of eluents. However, in contrast to the former, it applies hydrophilic stationary phase. Therefore, hydrophobic proteins are eluted first, and as the polarity of the mobile phase increases, other (more hydrophilic) proteins can elute [38].

Ion-exchange chromatography separation principle is based on the attraction of oppositely charged molecules. There are two types of stationary phases used in ion-exchange chromatography: negatively charged cation exchangers and positively charged anion exchangers. Proteins, based on their pI and the pH of the environment, carry different surface charges. Molecules are eluted by increasing salt concentration of the elution buffer, hence its ionic strength. Proteins with the lowest charge elute first, followed by other molecules that bind stronger with the resin [39].

Table 6.4 Advantages and disadvantages of various separation techniques.

Separation type	Separation technique	Separation basis	Advantages	Disadvantages
Electrophoretic approaches	SDS-PAGE	Molecular weight	Simple; cheap	Moderately time consuming; not automated; low resolving power
	IEF	Isoelectric point	Relatively simple with moderately good separation power	Resolving strips need to be purchased from a vendor; strips covering whole pH range are commonly not available
	2DE	Molecular weight and isoelectric point	Allows for identification of protein isoforms	Low reproducibility if more than one gel is used; low number of replicates is typically analyzed; laborious
Liquid chromatography: physiochemical separation	RP	Hydrophobicity	Moderate resolving power; desalting capability; adjustable gradient; compatible with mass spectrometry	Carryover effect; not completely reproducible
	HILIC	Hydrophobicity	Good resolving power	Not compatible with RP in an online mode
	IEX	Protein charge	Good method to be used in combination with other separation techniques; very adjustable, due to the presence of various ion exchangers	Not compatible directly with MS
	SEC	Protein size	Simple; wide separation mass range	Protein binding to the stationary phase cannot be entirely eliminated; low resolving power; requires high elution volumes; requires high sample loads; requires appropriate peptides or proteins for calibration
Liquid chromatography: affinity methods	Protein depletion	Affinity binding of protein targets	Simple; reproducible; selective	Added value for proteomics analyses is questionable; presence of a co-depletion mechanism
	Hexapeptide ligand library beads	Protein capturing	Effective; allow for sample enrichment in low-abundance proteins	Requires large starting protein amounts; proteins that saturate the beads cannot be quantified
	IMAC	Binding of proteins to metal ions	Specific for phosphoproteins and peptides as well as proteins exhibiting acidic amino acids	Low adsorption of some phosphoproteins and peptides
	Glycoprotein enrichment methods	Affinity binding of glycoproteins to lectins or through chemical modifications to a resin	Adjustable specificity due to the availability of many lectins	Analysis of O-glycosylated molecules is still challenging
	Phosphoprotein enrichment methods	Affinity binding of phosphorylated proteins to, for example, IMAC, TiO_2, or ZnO_2 resins	High specificity and recovery	Enrichment typically performed on the level of peptides, causing loss of information regarding phosphorylation interdependency (i.e., if a phosphorylation of one side in a protein is dependent on the phosphorylation of another)

2DE, two-dimensional electrophoresis; HILIC, hydrophilic interaction chromatography; IEF, isoelectric focusing; IEX, ion-exchange chromatography; IMAC, immobilized metal affinity chromatography; RP, reverse-phase chromatography; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; SEC, size-exclusion chromatography.

In SEC the separation is based on the differences in molecular size. The principle assumes that no interactions between the molecules and the stationary or mobile phase exist. Therefore, the separation is an entropy-controlled process where proteins are transported with a mobile phase through a stationary phase composed of carefully chosen porous particles. Large proteins do not pass through the pores and elute first. The smaller is the protein, the more pores it has to pass through, increasing the time of elution [40].

6.5.1.2.2 Affinity Chromatography

Affinity chromatography methods are becoming commonly used in proteomics, due to their broad range of applications and high specificity. The main goal of affinity chromatography is significant reduction of sample complexity, allowing for identification of low-abundance proteins. These methods can be used for the identification posttranslational modifications (PTMs), protein–protein interactions, and interaction networks as well as for the separation of specific type of proteins from the sample (e.g., phospho- or glycoproteins). In general, two main categories of affinity techniques are established: (i) depletion-based strategies, aiming at removal of highly abundant molecules from the sample, thus unmasking low-abundance molecules, and (ii) enrichment-based strategies, where protein of interest is separated from the total proteome [41]. An interesting type of affinity technique that combines both depletion and enrichment strategies is based on combinatorial peptide ligand libraries and will also be described in this section.

Depletion strategies allow for rapid and efficient removal of targeted proteins, reducing the complexity of the sample, hence, in theory, facilitating detection of low-abundance molecules. There are many depletion kits commercially available that vary in the number of depleted proteins. Immuno-based methods are most common, due to their high specificity and efficiency [33, 42]. The typical target for depletion kits is albumin: highly abundant plasma protein with a concentration of 30–50 μg/μl. Its removal reduces the total protein content by approximately 50%. However, these kits often target more proteins, including immunoglobulins, haptoglobin, and fibrinogen [42]. Abundant protein depletion kits are designed to be used for plasma or serum, but their applicability was also tested in urine. Still, the added value of applying these strategies is questionable, since many studies report contradictory results: some describing increased number of protein identifications, while others no added value. Kulloli et al. [43] depleted 14 highly abundant proteins from plasma, which allowed to increase the number of protein identifications from 71 to 130 compared with non-depleted sample. Tu et al. [44] depleted 7 or 14 abundant proteins from plasma, increasing the number of identified proteins by 25% in comparison with unfractionated sample. Still, the applicability of this method for biomarker discovery purposes is questionable, since only 6% of proteins were classified as low abundance. Additionally, they did not find any added value of removing 14 abundant proteins compared with 7, as both of the methods performed similarly. Echan et al. [45] also depleted up to six highly abundant proteins from serum or plasma, and the samples were afterward subjected to 2DE. They found a slight increase in the number of visualized spots. After protein identification, it appeared that these spots represented mostly minor forms of highly abundant proteins (e.g., ceruloplasmin or complement factors). Similar contradictory results are presented in the case of applying protein depletion in urine. Kushnir et al. [46] identified 2.5-fold more proteins in urine after depleting 6 highly abundant proteins. However, Afkarian et al. [47] did not report any increase in the number of identified proteins after albumin and IgG depletion in normo- or macroalbuminuric patients. In our hands [48], the application of four different depletion strategies in urine from healthy controls or CKD patients did not lead to an increase in the number of protein identifications. The analysis was performed using low starting urine volumes (500 μl), mimicking conditions typical for large cohort biomarker discovery studies. Even though the protein depletion was efficient in most cases and the concentration of targeted proteins substantially decreased, no added value was observed after applying such strategies. Altogether, the results imply that depletion of only few highly abundant proteins from body fluids might not be sufficient. In order to detect low-abundance proteins, a higher number of proteins need to be previously depleted. "SuperMix" resin might offer a solution to this problem, since it is able to deplete a large number of highly and moderately abundant proteins. On the downside, the application of such depletion strategies requires high starting protein content, and the depletion targets are not specified by the producer. SuperMix is used in a combination with resins based on avian antibodies, designated for removal of several (12–14) highly abundant plasma proteins. The aforementioned system was successfully applied by Qian et al. [49] and Patel et al. [50] in plasma. Qian et al. [49] compared IgY12 depletion strategy alone with IgY12–SuperMix system using one- or two-dimensional LC-MS/MS. The application of SuperMix system resulted in an increase of 60–80% of proteome coverage compared with IgY12 depletion. Patel et al. [50] compared Qproteome kit, which removes albumin and immunoglobulins, with SepproIgY14–SuperMix column. The latter resulted in the identification of 276 proteins that were not detectable after Qproteome depletion, from which most were classified as low-abundance proteins. However, Bandow [51]

applied three fractionation strategies on plasma samples (i.e., IgY14 depletion, IgY14–SuperMix set, and ProteoMiner) and analyzed them by 2DE. Even though obtained gel protein patterns between applied methods were different in comparison with the unfractionated plasma, analyzed spots represent only highly abundant proteins.

ProteoMiner (a commercialized name of combinatorial peptide ligand libraries) relies on depletion of highly abundant proteins with simultaneous enrichment of low-abundance ones. This is achieved as a result of the application of millions of randomized hexapeptides, which can capture specific proteins. In contrast to depletion strategies though, the method is not specific for one or a group of proteins, but rather affects the whole proteome. Highly abundant proteins rapidly saturate their binding sites and thus, the excess passes through the beads (depletion mechanism). Low-abundance proteins, however, do not saturate their binding sites that quickly, and therefore, the constant flow of the sample will cause gradual increase in the number of bound low-abundance proteins (enrichment mechanism) [52]. It is worth mentioning that if none of the ligands in the library demonstrate an affinity to a given protein, the latter will not be retained and thus detected by MS. Moreover, the strategy substantially affects the relative protein concentration in the sample (especially when the beads are saturated), increasing the complexity of quantitation procedures [53]. Similarly to protein depletion strategies, ProteoMiner was also criticized in some papers, but according to the "defenders" of the strategy, stated complaints are caused by an incorrect application of ProteoMiner rather than flaws of the technique [54]. In more details, Bandow [51] analyzed plasma samples by applying different depletion strategies (including ProteoMiner) followed by 2DE analysis. As already described in the paragraph earlier regarding depletion strategies, analyzed gel spots led to the identification of only highly abundant plasma proteins. In the defense of ProteoMiner, Righetti et al. [54] stated that they used too mild eluant able to desorb less than 30% of bound proteins. For desorption of all the proteins and thus proper application of ProteoMiner, stronger elution protocol should have been applied. Additionally, the capturing of proteins was performed at high ionic strength, reducing the interaction efficiency of ProteoMiner. In the second example, Keidel et al. [55] compared the enrichment efficiency in plasma of ProteoMiner with five different chromatographic beads (Sepabeads) that are based on the concept of hydrophobicity. They found only a weak increase in the concentration of low-abundance proteins, while all of the tested methods gave similar results. Therefore, they suggested that ProteoMiner uses the hydrophobic binding mechanism, while the ligand diversity plays a minor role. Righetti et al. [54] again stated that the capturing of proteins was performed at high ionic strength, decreasing the efficiency of formed interactions. Thus, the statement that ProteoMiner applies hydrophobic mechanism is incorrect and exploits the lack of knowledge of authors regarding the applied binding mechanisms. It is therefore very important to perform the analysis according to available guidelines and taking into consideration all of the recommendations. Such guidelines are very well described in the following chapter to which the readers interested in the technique are referred to Ref. [52].

Other available affinity-based methods do not seem as controversial as abundant protein depletion or ProteoMiner strategies. These methods rely clearly on protein enrichment, rather than depletion and in contrast to aforementioned "controversial strategies," target specific protein groups instead of the whole proteome. First method that is worth mentioning is the immobilized metal affinity chromatography (IMAC), where the enrichment is based on the affinity of proteins to metal ions. Such binding can be achieved, due to the exposure of functional groups by several proteins (i.e., Cys, His, Glu, Asp, Tyr) and their interaction with the metal ions (e.g., Al^{3+} or Ni^{2+}). These moderately specific methods usually enrich the samples in proteins with acidic amino acids [1]. However, the method can be also used for the enrichment of phosphorylated amino acids as further described in this section [56].

Two other techniques, aiming at enrichment of glycoproteins or phosphoproteins, are widely used in the field of proteomics. Since differences in PTMs between healthy and disease subjects can be used as biomarkers, these techniques evolved into two new subfields of proteomics: glycoproteomics and phosphoproteomics.

Two approaches are commonly applied for the enrichment of glycoproteins: lectin affinity and hydrazide chemistry-based enrichment. In the case of the former, specific oligosaccharide epitopes bind to the lectin resin. A large number of lectins with different selectivity are commercially available. For example, concanavalin A recognizes mannosyl and glucosyl epitopes, while wheat germ agglutinin sialic acid and *N*-acetylglucosamine. It is worth noting that the specificity of lectins binding O-glycosylated proteins is lower compared with these binding N-glycosylated. The hydrazide chemistry-based enrichment method relies on the chemical reaction (formation of covalent bonds) between the glycosylated proteins and the medium. The method however can be only used for the enrichment of N-glycosylated proteins due to the lack of an enzyme that would cut O-glycosylated bonds [57]. It is very important to take into consideration that the glycoproteomics workflow may differ from the one presented in this chapter, including application of specific software for bioinformatics analysis.

Therefore, readers interested in glycoproteomics are referred to the following review [58].

In the case of phosphoproteomics, most of the separation methods are used at the peptide level (exception: immunoprecipitation), which ensures high efficiency and high depletion rate of non-phosphorylated molecules. However, protein digestion leads to loss of the information on the phosphorylation interdependence (i.e., if a phosphorylation of one site in a protein is dependent on the phosphorylation of another). One of the strategies for the enrichment of the phosphoproteome (IMAC) was already mentioned in this section. The "typical" IMAC however is not specific for phosphoproteins and also enriches for proteins or peptides containing acidic amino acids. Therefore, in the case of phosphoproteomics, several modifications of the method have been introduced, for example, esterification of the peptides, adjustment of the pH, and use of various chelate ions. Another phosphoproteome enrichment strategy is called metal oxide affinity chromatography, which uses titanium dioxide- or zirconium dioxide-based chromatography methods. This strategy shows high affinity toward phosphorylated proteins and is more tolerant to the presence of salts, denaturation factors, or detergents compared with IMAC, making it a preferable strategy in phosphoproteomics compared with the latter. Other methods used for the enrichment of the phosphoproteome include ion-exchange chromatography, HILIC, chemical modification of phosphopeptides, and calcium or barium precipitation. All of the methods presented are well described in the following review [56].

6.5.2 Sample Preparation for MS/MS (Tryptic Digestion)

Before describing commonly applied tryptic digestion methods, the difference between bottom-up and top-down proteomics is worth mentioning, as only the former requires protein digestion. In the case of bottom-up proteomics, proteins are digested (usually with trypsin) prior to MS analysis, and detected peptides are aligned to associated proteins. However, the method has several limitations: (i) reduced specificity (identified peptide sequences may belong to more than one protein) and (ii) incomplete protein sequence coverage (large protein fragments may not be identified at all by MS, resulting in the loss of information regarding sequence variants or present PTMs). Top-down proteomics, aiming at the analysis of intact (non-digested) proteins, can tackle these problems, allowing to achieve complete protein sequence coverage, identify various proteoforms, and present PTMs. Nevertheless, the method is more applicable for the identification of proteins in simple mixtures. For proteome-wide studies, bottom-up proteomics show higher throughput, proteome coverage, and sensitivity compared with top-down proteomics. Therefore, in this chapter the focus will be set on bottom-up proteomics, while readers interested in top-down proteomics are referred to the following review [34].

There are three commonly applied types of protein digestion methods: in-solution digestion, in-gel digestion, and filter-aided sample preparation (FASP). Advantages and disadvantages of each method can be found in the following review [35].

In-solution digestion is probably the most-known sample preparation method, which, in the case of body fluids, typically involves the following steps: protein precipitation, resolubilization in a urea buffer, reduction, alkylation, and trypsin digestion. Due to the common use of the method, it is challenging to find a publication concentrating on its description per se. However, many protocols can be found online (e.g., Ref. [59]).

In-gel digestion offers the possibility to divide the proteome into multiple bands, reducing the sample complexity and allowing for in-depth analysis of complex samples. Moreover, the procedure allows to remove impurities of low molecular weight (e.g., buffer components or detergents). However, higher amounts of trypsin are required compared with in-solution digestion process, which may result in formation of trypsin autolysis products. Additionally, casting the gels can increase the possibility of contaminating the samples with keratins. Both 1DE and 2DE can be used and applied for this method, and the protocol for in-gel digestion is described in the following paper [60].

In FASP, all of the steps are performed using centrifugal filter units. According to the original protocol, two filter cutoffs are applicable: 10 or 30 kDa. As expected, the application of 30 kDa filter is faster, due to more rapid sample concentration compared with 10 kDa filters. The digestion comprises of four steps: (i) removal of components of low molecular weight in urea buffer, (ii) reduction and alkylation, (iii) protein digestion, and (iv) peptide isolation by filtration. The use of filters ensures that before protein digestion all low molecular weight components will be removed, while, after digestion, the same filter allows to retain high molecular weight components on the filters, allowing for isolation of the peptide mixture. The methodology is described in the original paper from Wiśniewski et al. [61].

6.5.3 Separation of Peptides

In proteomics, the separation on peptide level is usually achieved by coupling the liquid chromatography system with mass spectrometer (LC-MS-based proteomics).

This procedure reduces the number of analytes that enter the mass spectrometer and at the same time (i) decreases the potential of ionization suppression, taking place when molecules cannot be detected, due to the dynamic range limitation of the detector; (ii) reduces the undersampling effect, related to the choice of ions by mass spectrometer; and (iii) increases the concentration of analytes, enhancing the sensitivity of MS detection. In the MS setup the LC typically refers to RP chromatography [62]. However, recently combinatorial approaches of more than one chromatography methods are employed and can be used on- or offline. Such techniques in proteomics are recognized as multidimensional protein identification technology (MudPIT). The most common setup for MudPIT is the use of strong ion-exchange chromatography as a first dimension and RP as the second one. However, other commonly applied systems include high-pH RP coupled with RP or HILIC coupled with RP [62, 63]. The separation of peptides using various one- or multidimensional approaches for MS analysis is well described in the following review [62]. Of note, instead of applying RP coupled to MS, it is possible to use capillary electrophoresis (CE) system. Nevertheless, CE is more commonly applied in peptidomics studies and, thus, will be described in details in the corresponding chapter.

6.6 Analytical Instruments

Available analysis instruments and ionization techniques will not be covered in this chapter, since, even though the instrument setup applied has an enormous effect of the analysis (e.g., number of identified proteins, comprehensiveness of the analysis, sequence coverage), the users are commonly bound to a specific instrument. In other words, this aspect is not commonly modifiable for every analysis performed. Moreover, applied instrument settings (e.g., collision energy, intensity threshold required to trigger MS/MS event or even protein loaded onto the column) commonly vary from lab to lab, as they are adjusted empirically. There are two aspects of LC-MS/MS setup that should be mentioned though: instrument run time and the column length. In general, the longer the run time and length of the chromatography column, the more proteins can be identified, as demonstrated by Hsieh et al. [64] based on *Caenorhabditis elegans* analysis (same rules are applicable for body fluids). However, if the concentration of the peptides is below the detection limit of applied mass spectrometer, increase in the run time will not improve the number of protein identifications. Commonly applied instrument setups in proteomics are well described in the following review [65].

6.7 Data Processing and Bioinformatics Analysis

In this section bioinformatics tools for peptide and protein identification in shotgun proteomics will be briefly presented. Afterward, the focus will be set on protein quantitation methods, followed by description of data normalization and statistical analysis, which ensures obtaining good quality results.

6.7.1 Peptide and Protein Identification

The output from MS/MS analysis comprises of m/z and intensity values. However, it can be present in many different formats, depending on the instrument provider, for example, files ".raw" are produced by Thermo Scientific instruments, ".RAW" from Waters Corporation, while ".wiff" from Applied Biosystems (ABI) [66].

Various programs can be used for the identification of peptides and proteins. These include software based on sequence database searches (e.g., MASCOT, SEQUEST, OMSSA, X!Tandem), de novo sequencing of peptides (e.g., PEAK, MS BLAST), tag searches (e.g., InsPect), or searches in spectral libraries (e.g., X!Hunter). Other programs were created to validate the assignments of MS/MS spectra to corresponding peptides (e.g., PeptideProphet or ProteinProphet) [66]. For shotgun proteomics, the most commonly applied search algorithm is based on sequence database search engines, where obtained fragment ion spectra are matched to corresponding predicted and theoretical spectra from the database [67]. Since the matching is not perfect, the obtained identifications are subjected to quality control procedures by calculating the false discovery rate (FDR). This parameter estimates the number of false identifications (e.g., 1% of FDR means that 1% of the identifications in the dataset may not be correct) [1].

The lack of a common file format produced by different programs is an obstacle for further processing of the data, since it is difficult for a software to cover all of the available data formats. To deal with that issue, two unification formats have been established. Firstly, XML type of formats have become commonly applied by open-source software, including mzXML for processed mass spectra, pepXML for identifications and quantitation of peptides, and protXML for identification and quantitation of proteins. The second common format, mzML, was developed by Human Proteome Organization Proteomics Standards Initiative (HUPO-PSI) [66].

6.7.2 Protein Quantitation

In proteomics there are two commonly applied quantification methods: label-free and label-based methods.

In label-free methods, protein abundance is usually calculated on the basis of the number of MS/MS spectra (spectral counting) or by integrating signal intensities. The principle of spectral counting is based on the assumption that highly abundant proteins are analyzed/detected more often by the mass spectrometer and, thus, generate a higher number of MS/MS spectra. The spectral counting results exhibit good correlation with protein abundance. However, it is limited by the MS/MS saturation or undersampling effects, while the linear concentration range is lower compared with MS1-based (peptide ion intensity-based) quantification techniques [68]. Moreover, spectral counting can be biased toward large proteins, which can produce more peptide digests and, thus, more spectra. Therefore, the obtained data needs to be corrected, taking into consideration the protein length. A spectral count-based quantitation method called absolute protein expression (APEX) corrects the observed peptide count using learned probabilities to identify the peptides. The method is based on the assumption that a protein in the sample is represented by a fraction of injected peptides. The role of APEX is the application of correction factors that makes these fractions proportional to each other. Therefore, protein abundance is calculated as a fraction of the peptide spectra that is correlated with that protein using the correction factors of observing each peptide. These factors include expected contribution of a protein in the pool of identified peptides, total number of putative peptides generated by a protein, probability of peptide being ionized and analyzed by mass spectrometer, and peptide sequence characteristics that can affect the analysis (e.g., effect of PTMs). The method is described in details in the original paper [69]. In another method known as protein abundance index (PAI), the relative protein abundance is calculated based on the number of identified peptides divided by the number of theoretically observable (tryptic) peptides. A derivation of this method, used for absolute quantification, is called exponentially modified protein abundance index (emPAI), which is calculated as 10^(PAI)-1 [70]. Lastly, quantification based on ion intensity depends on the integration of intensity values for each peptide. Two steps are important for this method: detection and alignment of chromatograms and quantification of peptide intensities [66]. Each peptide needs to be detected in two-dimensional LC-MS space. Detection in LC dimension attributes appropriate retention time for each peptide, and since LC runs are not completely reproducible, it ensures that correct regions are evaluated in different runs. The quantification in this method is most commonly performed by calculating the area under the curve for each peak.

Label-based methods rely on labeling the peptides with stable isotopes. The frequently applied label-based approach in body fluid proteomics is called isobaric tags for relative and absolute quantification (iTRAQ). In this technique, N-terminus and lysine side chains are labeled with up to eight different isobaric tags comprising a reporter group, mass balancing group, and a group that reacts with peptides. The m/z of reporter groups varies from 113 to 121, while the m/z of balance groups from 32 to 24. However, the sum of the m/z of a reporter and balancing group for each of the labels remains the same, resulting in the generation of reporter ions with characteristic m/z ratios. During peptide fragmentation, the balancing group is lost as a neutral fragment. The abundance of the peptide is reflected by the intensity of the corresponding reporter ion. Peptide relative abundance is determined by calculating the ratio of reported ion intensity in different samples. Combining individual peptide abundances allows estimation of protein abundance. The method however has some limitations: underevaluation of protein fold changes and presence of cross-labeled impurities that causes interference during the analysis [67].

6.7.3 Data Normalization (Example of Label-Free Proteomics Using Ion Intensities)

Obtained proteomics data needs to be normalized to compensate for differences in the amount of sample loaded onto the column, variability in ionization efficiency, and detector saturation or LC column carryover effect. Correcting for these differences increases the accuracy of the analysis. The bias related to protein amount injected into the column can be firstly reduced by evaluating the total protein content by protein assays (e.g., Bradford protein assay) or by taking into consideration the area under the curve of previous runs based on UV trace signals. The normalization can be also based on the "housekeeping proteins," whose levels should not change between compared conditions or on creatinine values in urine [68]. The number of available normalization methods in proteomics is vast, and there are a number of papers focusing on comparison of normalization methods in proteomics datasets. For example, Callister et al. [71] compared four different normalization methods using LC-FTICR MS instrument: (i) central tendency normalization (global normalization), which centers ratios of peptide abundance over a mean; (ii) linear regression normalization, which relies on the principle that systematic bias depends linearly on the range of peptide abundances; (iii) local regression normalization, which, in contrast to linear regression, assumes that the systematic bias does not linearly depend on the range of peptide abundances; and (iv) quantile normalization, which assumes that distribution of peptides in different samples is similar. They suggest that global and linear

normalization performed the best from the four. In our laboratory, for protein ion intensity normalization, we usually apply a method based on the total ion current, where the ion intensity of a given protein is divided by the total ion intensity of all proteins in a sample.

6.7.4 Statistics in Proteomics Analysis

In order to correct for the variability characterizing biological samples, an appropriate sample size is required. The optimum sample size should be determined based on statistical power calculations. Underpowered analyses are at a very high risk of identifying false positive biomarkers or of failing to discriminate the true disease biomarkers. What is worth noting is that a sample size smaller than 12 does not allow to estimate the mean and variance based on normal distribution, and thus, larger sample sizes are recommended. Sample pooling should also be avoided, as it does not allow the extraction of information related to sample variability and the identification of the outliers. Application of strict and correct statistical tests should be considered for every proteomics study. Moreover, the findings should be corrected for multiple testing in order to reduce the number of false identifications. If a protein is found statistically significant prior to application of multiple testing correction, but not after, the solution is not to avoid the correction, but rather increase the sample size or accept that this protein cannot be identified as a biomarker in this study. There are many tests available for multiple testing correction. Frequentist, false discovery, or Bayesian approaches are applicable for proteomics studies [10].

6.8 Validation of Findings

Techniques typically used for the validation of findings in body fluids are immunoassay-based methods like ELISA and Western blot. These techniques are well established and used for many years for validation of proteomics findings. However, many disadvantages have been associated with the use of antibody-based strategies. Firstly, in order to perform the analysis, a specific antibody has to be produced, which is not always possible. Secondly, nonspecific binding can take place, increasing the variability of the method [72]. Finally, several commercially available kits show poor analytical performance. This issue was demonstrated by Kift et al. [73] where five ELISA assays for neutrophil gelatinase-associated lipocalin were tested on urine samples. Only two out of five evaluated kits showed good performance, the rest of them failed either on testing one (e.g., limit of quantitation or parallelism) or more parameters. Recently, however, another technique is becoming increasingly popular for the validation purposes—targeted MS. This technology has many more applications apart from validation of proteomics findings [74]. Nevertheless, in this chapter, the focus will be set on the application of targeted proteomics for validation purposes.

The technology used for targeted proteomics is called selected reaction monitoring (SRM) or multiple reaction monitoring (MRM). It is usually performed on a triple-quadrupole mass spectrometer, which selects specific precursor and product ion pairs, known as transitions. The first quadrupole is used for filtering a selected precursor peptide, which is fragmented in the second quadrupole by collision-induced dissociation (CID). The third quadrupole is used to filter a specific fragment ion that is finally measured by the detector. The quantitation of peptides is based on the measurements of the intensities of fragment ions. Obtained peptide measurements can be afterward assigned to a corresponding protein [75]. In MRM it is important to choose appropriate peptides as targets. These must be unique for a given protein and easily detectable by the mass spectrometer. Several databases are available to make the choice of proper target feasible, including PeptideAtlas, the Global Proteome Machine Database, Pride, and Tranche. Such databases contain information from many experiments, enhancing statistical significance of findings, providing information regarding the global number of observations for the peptide, and ensuring that the submitted data was obtained according to generally acceptable criteria. If a protein of interest does not have peptides in the database, the user can use bioinformatics tools that can predict the proteotypic peptides for a given protein. These tools include Peptide Sieve, ES Predictor, and Detectability Predictor. Similarly to parent peptides, transitions that show high intensity and low interference levels should be selected. These ion fragments are commonly chosen based on empirically obtained data from triple-quadrupole experiments [74]. The verification of findings can be performed using publicly available spectral libraries. A typical procedure in MRM experiments involves the addition of heavy isotope-labeled peptide standards that improve the confidence of findings. These peptides co-elute at the same time as their non-labeled analogs, and their fragment ion intensity ratio should also be the same, improving the confidence of identifications [76]. Execution of this complicated MRM procedure has been substantially facilitated by the development of sophisticated bioinformatics tools. The most commonly used open-access software are MRMer and Skyline, both well suited for the analysis of complex MRM experiments. Description of various software that facilitate MRM analysis can be found in the following review [75].

6.9 Clinical Applications of Body Fluid Proteomics

The following section will describe several studies relative to the application of body fluids proteomics for biomarker discovery. The search was performed with Web of Science using following words: Biomarker *AND* Proteomics *AND* (Urin* OR Plasm* OR Ser* OR Bloo* OR CSF OR Cerebrospin* OR Body Fluid). Articles from the last 5 years were taken into consideration. In total 1989 papers were retrieved. The three top cited articles from each year, concentrating on biomarker discovery studies in human subjects and applying MS, were taken into consideration. Therefore, in total 15 papers will be presented.

Addona et al. [77] analyzed plasma samples for the identification of early biomarkers of cardiac injury. For the discovery phase, blood from the coronary sinus from patients ($n = 3$) with therapeutic planned myocardial infarction to treat hypertrophic cardiomyopathy was collected before, 10 and 60 min after the procedure (total of 9 samples). The samples were depleted from 12 highly abundant proteins, followed by applying strong cation-exchange (SCX) chromatography, yielding 80 fractions. These fractions were subjected to LC-MS/MS, leading to the identification of a total of 1105 proteins, from which 70% were identified in all 9 samples. In order to select the candidates for validation, they applied several criteria, including (i) fivefold change in the precursor ion intensity of minimum two peptides per protein after the injury, (ii) prioritization using AIMS technology (targeted MS approach on Orbitrap or other high-performance instrument), where proteins identified in coronary sinus plasma were also searched in peripheral blood plasma; and (iii) evaluation of expression enrichment trend. Biomarker candidates were validated by ELISA or MRM in peripheral plasma from patients with therapeutic planned myocardial infarction, spontaneous myocardial infarction, or ischemia (cases) and with routine catheterization (controls). A number of novel cardiovascular biomarker candidates were identified (i.e., MYL3, TPM1, and FHL1), while discovery of previously reported was confirmed in this study (e.g., creatinine kinase, MB, or FABP) [77].

Shang et al. [78] studied the proteome of hepatocellular carcinoma patients in order to identify novel disease biomarkers. In the discovery set they analyzed plasma samples from 18 cirrhosis (controls) and 17 hepatocellular carcinoma patients (cases). High-abundance proteins were depleted from plasma samples, and the remaining proteins were separated using SDS-PAGE. Gel pieces were afterward destained, digested, and analyzed by 2D-LC-MS/MS. The levels of osteopontin were found to be upregulated in cases compared with controls. This finding was validated in two independent sample cohorts comprising a total of 312 plasma samples of hepatocellular carcinoma patients (cases), cirrhosis subject (controls), chronic hepatitis C and B individuals (controls), and healthy subjects (controls). Higher osteopontin levels were confirmed for cases in both validation sets compared with controls.

Da Costa et al. [79] searched for diagnostic biomarkers for HBV-related hepatocellular carcinoma. High-abundance proteins were depleted from plasma samples from patients with hepatocellular carcinoma and chronic liver disease and healthy controls. Subsequently, depleted samples were subjected to 2D chromatography, followed by SDS-PAGE and LC-MS/MS analysis. After prioritization of findings, latent transforming growth factor-β binding protein 2 and osteopontin with levels significantly higher in carcinoma patients compared with other groups were chosen for validation. ELISA was performed in a total of 684 plasma samples. A biomarker panel comprising these two proteins showed an area under the curve of 0.85.

Cima et al. [80] applied a genetically guided strategy for the identification of serum biomarkers for diagnosis and prognosis of prostate cancer. In general, they were interested in biomarkers related to an inactivation of phosphatase and tensin homolog (PTEN) tumor suppressor gene, which is commonly observed in prostate cancer patients. In the first step, sera and tissue samples from Pten-null mice were analyzed using glycoprotein enrichment (hydrazide chemistry-based method) followed by LC-MS/MS. This led to the identification of 775 glycoproteins, and after bioinformatics prioritization, 39 biomarker candidates were chosen for identification in human sera of prostate cancer patients ($n = 52$) and healthy controls ($n = 40$). The validation was performed using SRM or ELISA. Thirty-nine proteins were quantified consistently in the validation samples. From these proteins, the best candidates for construction of a predictive model for detection of normal and abnormal PTEN were chosen. Ultimately a biomarker panel comprising four proteins—thrombospondin-1 (THBS1), metalloproteinase inhibitor 1 (TIMP-1), complement factor H (CFH), and prolow-density lipoprotein receptor-related protein 1 (LRP-1)—was able to discriminate between normal and aberrant PTEN with 79.2% sensitivity and 76.7% specificity.

Zeng et al. [81] analyzed serum samples to identify biomarkers of breast cancer. Five samples from breast cancer patients (stage 2) and five from healthy controls were pooled and subjected to abundant protein depletion, followed by multilectin affinity chromatography separation into fractions comprising unretained and retained proteins. Such prepared fractions were separated using IEF technique and finally analyzed by LC-MS/MS.

After bioinformatics prioritization, four proteins were chosen as putative biomarker candidates: thrombospondin 1 and 5, α-1B-glycoprotein, serum amyloid, P-component, and tenascin-X. However, only tenascin-X was validated using ELISA in the same samples (but this time not pooled) of five breast cancer patients and five controls.

Bukhari et al. [82] applied an affinity proteomics strategy for the identification of biomarkers of colorectal cancer. In brief, they utilized immunological mechanisms and constructed an antibody set in rabbits against the secretome of human colon adenocarcinoma HT29. This secretome was used as an affinity reagent to capture possible tumor-related antigens from sera of pooled samples from five colon cancer patients, five rectal cancer, and five healthy controls. The samples were subjected to 2DE and spots exclusively present in tumor samples were excised and analyzed by MALDI-TOF. Two proteins (soluble vimentin and keratin type II cytoskeletal) were identified uniquely in colon cancer sera, while TGF-beta-inhibited protein was found only in rectal cancer sera. Pathway and protein–protein interaction analyses revealed that only vimentin was linked to tumorigenic pathways of colorectal pathways. The levels of vimentin were validated in 43 preoperative patients with colorectal cancer and 20 controls, revealing that the protein levels were approximately 5 times higher in colon cancer patients compared with controls or subjects with rectal cancer.

Cynthia Martin et al. [83] studied the serum proteome in order to identify biomarkers for progression of Duchenne muscular dystrophy. For the discovery set, four pooled sample groups were analyzed: adult controls, child controls, Duchenne muscular dystrophy ambulant patients, and Duchenne muscular dystrophy non-ambulant subjects. Serum pools were processed by the ProteoMiner kit and fractionated with SDS-PAGE. The gels were afterward cut into 10 pieces and subjected to LC-MS/MS testing 2 replicates for each pool. After performing bioinformatics analysis, fibronectin, in which levels were found upregulated in Duchenne muscular dystrophy patients compared with controls, was chosen for further validation. This validation was performed using ELISA, where the levels of fibronectin were compared with 68 Duchenne muscular dystrophy patients with 71 subjects with various muscular dystrophies and 15 controls. Upregulation of fibronectin in Duchenne muscular dystrophy patients compared with subjects with other dystrophies and healthy controls was confirmed by ELISA. Additionally, levels of fibronectin were tested in 22 Duchenne muscular dystrophy patients over a period of time (6 months to 4 years), demonstrating its gradual increase.

Nie et al. [84] analyzed the serum glycoproteome in order to discover putative biomarkers for pancreatic cancer. In this study, proteomics profile of pancreatic cancer patients ($n = 37$) was compared with subjects with conditions that are related to pancreas: diabetes ($n = 30$), cyst ($n = 30$), chronic pancreatitis ($n = 30$), and obstructive jaundice ($n = 22$). From each of the disease groups, four to six samples were chosen randomly for the analysis. Additionally, pooled sample from 30 healthy controls was used as an internal standard. Chosen samples were subjected to abundant protein depletion and glycoprotein enrichment (lectin chromatography), followed by label-free or tandem mass tag (TMT)-labeled LC-MS/MS analysis. This led to the identification of 243 proteins in non-labeled analysis and 354 in TMT-labeled approach. After bioinformatics analysis, significantly different proteins were validated by ELISA and lectin ELISA assays in a total of 179 samples. This led to a development of a biomarker panel comprising four proteins (α-1-antichymotrypsin, thrombospondin-1, haptoglobin, and carbohydrate antigen 19-9) that can distinguish pancreatic cancer from other diseases with an area under the curve values in the range of 0.92 to 0.95.

Linden et al. [85] analyzed urine samples for the identification of putative biomarkers of non-muscle-invasive bladder cancer. Urine samples from bladder cancer patients ($n = 5$) and healthy controls ($n = 4$) were concentrated and depleted from highly abundant proteins. Afterward, pooled or individual samples were analyzed by LC-MS/MS in two replicates each. These led to the identification of 231 and 298 unique protein identifications in pooled or individual urine, respectively. In total, 387 unique proteins were identified, from which, after the analysis, 29 overexpressed in non-muscle-invasive bladder cancer candidate biomarkers were discovered. Higher levels of four proteins were confirmed with Western blot on 17 cases and 26 controls: fibrinogen β-chain precursor, apolipoprotein E, α-1 antitrypsin, and leucine-rich α-2 glycoprotein 1. Dot-blot analysis on individual urine samples from 99 bladder cancer patients and 13 healthy controls further prioritized fibrinogen β-chain precursor and α-1 antitrypsin as the most interesting candidates showing sensitivity and selectivity in the range of 66–85%.

Chen et al. [86] aimed for the identification of potential bladder cancer biomarkers in urine samples. Hernia patients were used as controls ($n = 9$) and bladder cancer subjects as cases ($n = 9$), which were pooled into one control and one case sample. They applied hexapeptide library beads or abundant protein depletion column, followed by iTRAQ labeling. Afterward the samples were subjected to basic RP chromatography, and ultimately analyzed fractionated or unfractionated urine was analyzed by LC-MS/MS. The application of depletion strategies increased the number of identified proteins from approximately 300 (for unfractionated sample) to

500 (after any of applied fractionations) per run. Levels of six apolipoproteins (APOA1, APOA2, APOB, APOC2, APOC3, and APOE) and of serum amyloid A-4 protein were elevated in bladder cancer patients compared with hernia individuals, while pro-epidermal growth factor was found downregulated in bladder cancer subjects. The findings were confirmed by Western blot or Bio-Plex assay in an independent cohort.

Kentsis et al. [87] analyzed urine samples to identify biomarkers for Kawasaki disease. The study cohort for discovery is comprised of six patients with Kawasaki disease, six subjects with febrile illness that mimics Kawasaki disease, and three patients with Kawasaki disease after complete response to treatment. Urine samples were subjected to centrifugation, protein precipitation, SDS-PAGE, and RP-LC, followed by LC-MS/MS analysis. In total 2131 unique protein were identified. Proteome of patients with Kawasaki disease was enriched in biomarkers of cellular injury and immune regulators compared with two other tested groups. Elevated levels of meprin A and filamin C were validated in urine or serum samples in two independent sample cohorts comprising a total of 236 patients. A biomarker panel composed of meprin A and filamin C showed very good diagnostic performance for diagnosis of Kawasaki disease showing an area under the curve value of 0.98.

Sylvester et al. [88] analyzed urine samples from infants to identify diagnostic and prognostic biomarkers of necrotizing enterocolitis. The study was divided into two phases: discovery set, composed of 45 necrotizing enterocolitis patients, 12 sepsis, and 2 healthy controls, and validation set, which included 40 necrotizing enterocolitis subjects, 5 sepsis, and 15 healthy controls. Samples from discovery set were analyzed by LC-MS, which led to identification of 7 putative biomarkers (α-2-macroglobulin-like protein 1, cluster of differentiation protein 14, fibrinogen alpha chain, cystatin 3, pigment epithelium-derived factor, retinol binding protein 4, and vasolin), which levels were confirmed by ELISA in the validation set. The panel allowed to differentiate medical versus surgical NEC, NEC versus sepsis, and NEC versus controls with high sensitivity (89–96%) and specificity (80–90%).

Ringman et al. [89] studied the proteome of patients with familial Alzheimer's disease, caused by mutation in *PSEN1* and *APP* genes. Fourteen mutation carriers were used as cases and five non-carriers as controls. Gathered CSF samples were depleted from highly abundant proteins and analyzed by LC-MS/MS. This led to identification of 600 proteins, from which 46 were upregulated and 10 downregulated in carriers compared with non-carriers. Fourteen of differentially expressed proteins were previously reported in the literature (e.g., APP, transferrin, $\alpha_1\beta$-glycoprotein, or plasminogen). Novel findings included secreted phosphoprotein 1, calsyntenin 3, and CD99 antigen. Unfortunately, the findings were not validated, and considering the relatively small sample size, they should be evaluated with caution.

Kroksveen et al. [90] aimed for the identification of differentially expressed proteins between patients with multiple sclerosis and controls for future evaluation of these proteins as disease biomarkers in large sample cohorts. In the discovery set, the CSF from patients with clinically isolated syndrome ($n = 5$) and relapsing-remitting multiple sclerosis that had clinically isolated syndrome when the lumbar puncture was performed ($n = 5$) and from controls with other inflammatory neurological diseases ($n = 5$) was analyzed. The samples were firstly depleted from 14 highly abundant proteins using MARS column, digested, and labeled with iTRAQ reagent. Such prepared peptides were subjected to SCX chromatography, followed by clean-up from the salts, and analyzed by LC-MS/MS. This led to the identification of a total of 1291 proteins. From the list, 20 proteins were found differentially expressed between 10 patients with clinically isolated syndrome, and the controls were validated by MRM. The validation population consisted of a total of 131 patients divided into following groups: 16 patients with relapsing-remitting multiple sclerosis that had clinically isolated syndrome at the time of lumbar puncture, 15 patients with clinically isolated syndrome, 36 patients with relapsing-remitting multiple sclerosis, 33 patients with other inflammatory neurological diseases, and 32 patients with other neurological diseases. Five proteins (α-1-antichymotrypsin, contactin-1, apolipoprotein D, clusterin, and kallikrein-6) were found differentially expressed between various groups. Nevertheless, none of the validated proteins were able to discriminate between multiple sclerosis patients and controls on its own, and thus, in the future, an establishment of a biomarker panel was recommended.

Liong et al. [91] analyzed the cervicovaginal fluid to predict spontaneous preterm labor in symptomatic pregnant women between 22 and 36 weeks of gestation. The fluid from eight women who spontaneously delivered at term and from four women who spontaneously delivered preterm was pooled, respectively. The samples were subjected to 2D-DIGE, and subsequently, spots of interest were excised and analyzed by LC-MS/MS. This led to identification of 12 differentially expressed proteins between compared groups. Elevated levels of albumin and vitamin D-binding protein in women with preterm delivery were confirmed by ELISA in an independent samples cohort of 129 samples. The biomarker panel comprising two proteins showed 66.7% sensitivity, 100% specificity, 100% positive predictive value, and 96.7% negative predictive value.

6.10 Conclusions

Proteomics analysis of body fluids is a complex and multistep procedure. Every step of sample preparation, starting from collection procedures ending with LC-MS/MS analysis conditions, needs to be optimized. Similarly to sample preparation, data processing can be performed in many different ways. There is no ideal way to conduct the analysis, and many steps should be optimized based on various factors (e.g., sample type analyzed, type of proteins of interest, applied instrument setup or disease studied, and many more). All these aspects severely affect the final outcome of the analysis. It is therefore recommended to optimize the workflow prior to performing the analysis of samples of interest, and this optimization should be adjusted specifically to study requirements.

The analysis of body fluids, compared with tissue, shows a notable advantage: it is easier to apply the identified biomarkers in clinical practice for diagnosis or prognosis of a disease. This is possible simply due to easy, noninvasive, and low-cost sample collection. It is worth mentioning though that the analysis of tissue is more suitable for the elucidation of disease mechanism and, thus, for identifying novel drug targets. Ultimately, analysis of body fluids is required for diagnostic purposes, while analysis of tissue samples is required for drug development. However, it is impossible to treat a disease if it is not diagnosed (i.e., having discovered drug targets, but not being able to diagnose the disease). Along the same lines, even early diagnosis of a disease is not advantageous, if treatment options do not exist (i.e., having diagnostic biomarkers, but not developed the drugs). In conclusions, research on both types of samples (tissue and body fluids) is needed to reduce the health and economic burden related to a disease.

Proteomics analysis of body fluids is for now, to some degree, limited by the analytical characteristics of current instrumentation/techniques (e.g., resolving power, detection limit). In the future, advances in MS analysis and bioinformatics should allow for more in-depth analysis of body fluids proteomes, facilitating the development of disease biomarkers.

References

1 Filip S, Zoidakis J, Vlahou A, Mischak H. Advances in urinary proteome analysis and applications in systems biology. *Bioanalysis*. 2014;6(19):2549–2569.
2 Hu S, Loo JA, Wong DT. Human body fluid proteome analysis. *Proteomics*. 2006;6(23):6326–6353.
3 Lygirou V, Makridakis M, Vlahou A. Biological sample collection for clinical proteomics: existing SOPs. *Methods in molecular biology.* 2015;1243:3–27.
4 Kroksveen AC, Opsahl JA, Aye TT, Ulvik RJ, Berven FS. Proteomics of human cerebrospinal fluid: discovery and verification of biomarker candidates in neurodegenerative diseases using quantitative proteomics. *Journal of proteomics*. 2011;74(4):371–388.
5 Ahn SM, Simpson RJ. Body fluid proteomics: prospects for biomarker discovery. *Proteomics clinical applications*. 2007;1(9):1004–1015.
6 Parker CE, Borchers CH. Mass spectrometry based biomarker discovery, verification, and validation—quality assurance and control of protein biomarker assays. *Molecular oncology*. 2014;8(4):840–858.
7 Gelfand CA, Omenn GS. Preanalytical Variables for Plasma and Serum Proteome Analyses. In: Ivanov AR, Lazarev AV, editors. *Sample Preparation in Biological Mass Spectrometry*: Springer, Dordrecht; 2011. p. 269–289.
8 Savino R, Paduano S, Preiano M, Terracciano R. The proteomics big challenge for biomarkers and new drug-targets discovery. *International journal of molecular sciences*. 2012;13(11):13926–13948.
9 Blonder J, Issaq HJ, Veenstra TD. Proteomic biomarker discovery: it's more than just mass spectrometry. *Electrophoresis*. 2011;32(13):1541–1548.
10 Mischak H, Allmaier G, Apweiler R, Attwood T, Baumann M, Benigni A, et al. Recommendations for biomarker identification and qualification in clinical proteomics. *Science translational medicine*. 2010;2(46):46ps2.
11 Mischak H, Ioannidis JP, Argiles A, Attwood TK, Bongcam-Rudloff E, Broenstrup M, et al. Implementation of proteomic biomarkers: making it work. *European journal of clinical investigation*. 2012;42(9):1027–1036.
12 Ray S, Reddy PJ, Jain R, Gollapalli K, Moiyadi A, Srivastava S. Proteomic technologies for the identification of disease biomarkers in serum: advances and challenges ahead. *Proteomics*. 2011;11(11):2139–2161.
13 Zhu P, Bowden P, Zhang D, Marshall JG. Mass spectrometry of peptides and proteins from human blood. *Mass spectrometry reviews*. 2011;30(5):685–732.
14 Zhang AH, Sun H, Yan GL, Han Y, Wang XJ. Serum proteomics in biomedical research: a systematic review. *Applied biochemistry and biotechnology*. 2013;170(4):774–786.
15 Beck HC, Overgaard M, Rasmussen LM. Plasma proteomics to identify biomarkers—application to cardiovascular diseases. *Translational proteomics.* 2015;7:40–48.

16 Surinova S, Schiess R, Huttenhain R, Cerciello F, Wollscheid B, Aebersold R. On the development of plasma protein biomarkers. *Journal of proteome research*. 2011;10(1):5–16.

17 Zhang Q, Faca V, Hanash S. Mining the plasma proteome for disease applications across seven logs of protein abundance. *Journal of proteome research*. 2011;10(1):46–50.

18 Gorriz JL, Martinez-Castelao A. Proteinuria: detection and role in native renal disease progression. *Transplantation reviews*. 2012;26(1):3–13.

19 Kim MJ, Frankel AH, Tam FW. Urine proteomics and biomarkers in renal disease. *Nephron experimental nephrology*. 2011;119(1):e1–e7.

20 Decramer S, Gonzalez de Peredo A, Breuil B, Mischak H, Monsarrat B, Bascands JL, et al. Urine in clinical proteomics. *Molecular & cellular proteomics*. 2008;7(10):1850–1862.

21 Shao C, Wang Y, Gao Y. Applications of urinary proteomics in biomarker discovery. *Science China life sciences*. 2011;54(5):409–417.

22 Voss J, Goo YA, Cain K, Woods N, Jarrett M, Smith L, et al. Searching for the noninvasive biomarker holy grail: are urine proteomics the answer? *Biological research for nursing*. 2011;13(3):235–242.

23 Spielmann N, Wong DT. Saliva: diagnostics and therapeutic perspectives. *Oral diseases*. 2011;17(4):345–354.

24 Liu J, Duan Y. Saliva: a potential media for disease diagnostics and monitoring. *Oral oncology*. 2012;48(7):569–577.

25 Schulz BL, Cooper-White J, Punyadeera CK. Saliva proteome research: current status and future outlook. *Critical reviews in biotechnology*. 2013;33(3):246–259.

26 Pfaffe T, Cooper-White J, Beyerlein P, Kostner K, Punyadeera C. Diagnostic potential of saliva: current state and future applications. *Clinical chemistry*. 2011;57(5):675–687.

27 Zhang A, Sun H, Wang P, Wang X. Salivary proteomics in biomedical research. *Clinica chimica acta*. 2013;415:261–265.

28 Al Kawas S, Rahim ZH, Ferguson DB. Potential uses of human salivary protein and peptide analysis in the diagnosis of disease. *Archives of oral biology*. 2012;57(1):1–9.

29 Zurbig P, Dihazi H, Metzger J, Thongboonkerd V, Vlahou A. Urine proteomics in kidney and urogenital diseases: moving towards clinical applications. *Proteomics clinical applications*. 2011;5(5–6):256–268.

30 Teunissen CE, Petzold A, Bennett JL, Berven FS, Brundin L, Comabella M, et al. A consensus protocol for the standardization of cerebrospinal fluid collection and biobanking. *Neurology*. 2009;73(22):1914–1922.

31 Puangpila C, Mayadunne E, El Rassi Z. Liquid phase based separation systems for depletion, prefractionation, and enrichment of proteins in biological fluids and matrices for in-depth proteomics analysis-An update covering the period 2011-2014. *Electrophoresis*. 2015;36(1):238–252.

32 Selvaraju S, Rassi ZE. Liquid-phase-based separation systems for depletion, prefractionation and enrichment of proteins in biological fluids and matrices for in-depth proteomics analysis—an update covering the period 2008-2011. *Electrophoresis*. 2012;33(1):74–88.

33 Ly L, Wasinger VC. Protein and peptide fractionation, enrichment and depletion: tools for the complex proteome. *Proteomics*. 2011;11(4):513–534.

34 Catherman AD, Skinner OS, Kelleher NL. Top Down proteomics: facts and perspectives. *Biochemical and biophysical research communications*. 2014;445(4):683–693.

35 Camerini S, Mauri P. The role of protein and peptide separation before mass spectrometry analysis in clinical proteomics. *Journal of chromatography A*. 2015;1381:1–12.

36 Stoyanov A. IEF-based multidimensional applications in proteomics: toward higher resolution. *Electrophoresis*. 2012;33(22):3281–3290.

37 Fekete S, Veuthey JL, Guillarme D. New trends in reversed-phase liquid chromatographic separations of therapeutic peptides and proteins: theory and applications. *Journal of pharmaceutical and biomedical analysis.* 2012;69:9–27.

38 Boersema PJ, Mohammed S, Heck AJ. Hydrophilic interaction liquid chromatography (HILIC) in proteomics. *Analytical and bioanalytical chemistry*. 2008;391(1):151–159.

39 Jungbauer A, Hahn R. Ion-exchange chromatography. *Methods in enzymology.* 2009;463:349–371.

40 Fekete S, Beck A, Veuthey JL, Guillarme D. Theory and practice of size exclusion chromatography for the analysis of protein aggregates. *Journal of pharmaceutical and biomedical analysis.* 2014;101:161–173.

41 Medvedev A, Kopylov A, Buneeva O, Zgoda V, Archakov A. Affinity-based proteomic profiling: problems and achievements. *Proteomics*. 2012;12(4–5):621–637.

42 van den Broek I, Niessen WM, van Dongen WD. Bioanalytical LC-MS/MS of protein-based biopharmaceuticals. *Journal of chromatography B.* 2013;929:161–179.

43 Kullolli M, Warren J, Arampatzidou M, Pitteri SJ. Performance evaluation of affinity ligands for depletion of abundant plasma proteins. *Journal of chromatography B.* 2013;939:10–16.

44 Tu C, Rudnick PA, Martinez MY, Cheek KL, Stein SE, Slebos RJ, et al. Depletion of abundant plasma proteins and limitations of plasma proteomics. *Journal of proteome research*. 2010;9(10):4982–4991.

45 Echan LA, Tang HY, Ali-Khan N, Lee K, Speicher DW. Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *Proteomics*. 2005;5(13):3292–3303.

46 Kushnir MM, Mrozinski P, Rockwood AL, Crockett DK. A depletion strategy for improved detection of human proteins from urine. *Journal of biomolecular techniques*. 2009;20(2):101–108.

47 Afkarian M, Bhasin M, Dillon ST, Guerrero MC, Nelson RG, Knowler WC, et al. Optimizing a proteomics platform for urine biomarker discovery. *Molecular & cellular proteomics*. 2010;9(10):2195–2204.

48 Filip S, Vougas K, Zoidakis J, Latosinska A, Mullen W, Spasovski G, et al. Comparison of depletion strategies for the enrichment of low-abundance proteins in urine. *PloS one*. 2015;10(7):e0133773.

49 Qian WJ, Kaleta DT, Petritis BO, Jiang H, Liu T, Zhang X, et al. Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. *Molecular & cellular proteomics*. 2008;7(10):1963–1973.

50 Patel BB, Barrero CA, Braverman A, Kim PD, Jones KA, Chen DE, et al. Assessment of two immunodepletion methods: off-target effects and variations in immunodepletion efficiency may confound plasma proteomics. *Journal of proteome research*. 2012;11(12):5947–5958.

51 Bandow JE. Comparison of protein enrichment strategies for proteome analysis of plasma. *Proteomics*. 2010;10(7):1416–1425.

52 Righetti PG, Boschetti E. Sample treatment methods involving combinatorial peptide ligand libraries for improved proteomes analyses. *Methods in molecular biology.* 2015;1243:55–82.

53 Guerrier L, Righetti PG, Boschetti E. Reduction of dynamic protein concentration range of biological extracts for the discovery of low-abundance proteins by means of hexapeptide ligand library. *Nature protocols*. 2008;3(5):883–890.

54 Righetti PG, Fasoli E, Boschetti E. Combinatorial peptide ligand libraries: the conquest of the 'hidden proteome' advances at great strides. *Electrophoresis*. 2011;32(9):960–966.

55 Keidel EM, Ribitsch D, Lottspeich F. Equalizer technology—equal rights for disparate beads. *Proteomics*. 2010;10(11):2089–2098.

56 Leitner A, Sturm M, Lindner W. Tools for analyzing the phosphoproteome and other phosphorylated biomolecules: a review. *Analytica chimica acta*. 2011;703(1):19–30.

57 Pan S, Chen R, Aebersold R, Brentnall TA. Mass spectrometry based glycoproteomics—from a proteomics perspective. *Molecular & cellular proteomics*. 2011;10(1):R110 003251.

58 Thaysen-Andersen M, Packer NH. Advances in LC-MS/MS-based glycoproteomics: getting closer to system-wide site-specific mapping of the N- and O-glycoproteome. *Biochimica et biophysica acta*. 2014;1844(9):1437–1452.

59 Andersen RF, Palmfeldt J, Jespersen B, Gregersen N, Rittig S. Plasma and urine proteomic profiles in childhood idiopathic nephrotic syndrome. *Proteomics clinical applications*. 2012;6(7–8):382–393.

60 Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature protocols*. 2006;1(6):2856–2860.

61 Wisniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. *Nature methods*. 2009;6(5):359–362.

62 Xie F, Smith RD, Shen Y. Advanced proteomic liquid chromatography. *Journal of chromatography A*. 2012;1261:78–90.

63 Stoll DR, Li X, Wang X, Carr PW, Porter SE, Rutan SC. Fast, comprehensive two-dimensional liquid chromatography. *Journal of chromatography A*. 2007;1168(1–2):3–43; discussion 2.

64 Hsieh EJ, Bereman MS, Durand S, Valaskovic GA, MacCoss MJ. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *Journal of the American society for mass spectrometry*. 2013;24(1):148–153.

65 Holcapek M, Jirasko R, Lisa M. Recent developments in liquid chromatography-mass spectrometry and related techniques. *Journal of chromatography A*. 2012;1259:3–15.

66 Gonzalez-Galarza FF, Lawless C, Hubbard SJ, Fan J, Bessant C, Hermjakob H, et al. A critical appraisal of techniques, software packages, and standards for quantitative proteomic analysis. *Omics: a journal of integrative biology*. 2012;16(9):431–442.

67 Drabovich AP, Martinez-Morillo E, Diamandis EP. Toward an integrated pipeline for protein biomarker development. *Biochimica et biophysica acta*. 2015;1854(6):677–686.

68 Christin C, Bischoff R, Horvatovich P. Data processing pipelines for comprehensive profiling of proteomics samples by label-free LC-MS for biomarker discovery. *Talanta*. 2011;83(4):1209–1224.

69 Lu P, Vogel C, Wang R, Yao X, Marcotte EM. Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nature biotechnology*. 2007;25(1):117–124.

70 Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, et al. Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein. *Molecular & cellular proteomics*. 2005;4(9):1265–1272.

71 Callister SJ, Barry RC, Adkins JN, Johnson ET, Qian WJ, Webb-Robertson BJ, et al. Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *Journal of proteome research*. 2006;5(2):277–286.

72 Gan SD, Patel KR. Enzyme immunoassay and enzyme-linked immunosorbent assay. *The journal of investigative dermatology*. 2013;133(9):e12.

73 Kift RL, Messenger MP, Wind TC, Hepburn S, Wilson M, Thompson D, et al. A comparison of the analytical performance of five commercially available assays for neutrophil gelatinase-associated lipocalin using urine. *Annals of clinical biochemistry*. 2013;50(Pt 3):236–244.

74 Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. *Nature methods*. 2012;9(6):555–566.

75 Boja ES, Rodriguez H. Mass spectrometry-based targeted quantitative proteomics: achieving sensitive and reproducible detection of proteins. *Proteomics*. 2012;12(8):1093–1110.

76 Rauh M. LC-MS/MS for protein and peptide quantification in clinical chemistry. *Journal of chromatography B*. 2012;883–884:59–67.

77 Addona TA, Shi X, Keshishian H, Mani DR, Burgess M, Gillette MA, et al. A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. *Nature biotechnology*. 2011;29(7):635–643.

78 Shang S, Plymoth A, Ge S, Feng Z, Rosen HR, Sangrajrang S, et al. Identification of osteopontin as a novel marker for early hepatocellular carcinoma. *Hepatology*. 2012;55(2):483–490.

79 Da Costa AN, Plymoth A, Santos-Silva D, Ortiz-Cuaran S, Camey S, Guilloreau P, et al. Osteopontin and latent-TGF beta binding-protein 2 as potential diagnostic markers for HBV-related hepatocellular carcinoma. *International journal of cancer*. 2015;136(1):172–181.

80 Cima I, Schiess R, Wild P, Kaelin M, Schuffler P, Lange V, et al. Cancer genetics-guided discovery of serum biomarker signatures for diagnosis and prognosis of prostate cancer. *Proceedings of the national academy of sciences of the United States of America*. 2011;108(8):3342–3347.

81 Zeng Z, Hincapie M, Pitteri SJ, Hanash S, Schalkwijk J, Hogan JM, et al. A proteomics platform combining depletion, multi-lectin affinity chromatography (M-LAC), and isoelectric focusing to study the breast cancer proteome. *Analytical chemistry*. 2011;83(12):4845–4854.

82 Bukhari S, Mokhdomi TA, Chikan NA, Amin A, Qazi H, Wani SH, et al. Affinity proteomics led identification of vimentin as a potential biomarker in colon cancers: insights from serological screening and computational modelling. *Molecular bioSystems*. 2015;11(1):159–169.

83 Cynthia Martin F, Hiller M, Spitali P, Oonk S, Dalebout H, Palmblad M, et al. Fibronectin is a serum biomarker for Duchenne muscular dystrophy. *Proteomics clinical applications*. 2014;8(3-4):269–278.

84 Nie S, Lo A, Wu J, Zhu J, Tan Z, Simeone DM, et al. Glycoprotein biomarker panel for pancreatic cancer discovered by quantitative proteomics analysis. *Journal of proteome research*. 2014;13(4):1873–1884.

85 Linden M, Lind SB, Mayrhofer C, Segersten U, Wester K, Lyutvinskiy Y, et al. Proteomic analysis of urinary biomarker candidates for nonmuscle invasive bladder cancer. *Proteomics*. 2012;12(1):135–144.

86 Chen CL, Lin TS, Tsai CH, Wu CC, Chung T, Chien KY, et al. Identification of potential bladder cancer markers in urine by abundant-protein depletion coupled with quantitative proteomics. *Journal of proteomics.* 2013;85:28–43.

87 Kentsis A, Shulman A, Ahmed S, Brennan E, Monuteaux MC, Lee YH, et al. Urine proteomics for discovery of improved diagnostic markers of Kawasaki disease. *EMBO molecular medicine*. 2013;5(2):210–220.

88 Sylvester KG, Ling XB, Liu GY, Kastenberg ZJ, Ji J, Hu Z, et al. Urine protein biomarkers for the diagnosis and prognosis of necrotizing enterocolitis in infants. *The journal of pediatrics*. 2014;164(3):607–12 e1-7.

89 Ringman JM, Schulman H, Becker C, Jones T, Bai Y, Immermann F, et al. Proteomic changes in cerebrospinal fluid of presymptomatic and affected persons carrying familial Alzheimer disease mutations. *Archives of neurology*. 2012;69(1):96–104.

90 Kroksveen AC, Aasebo E, Vethe H, Van Pesch V, Franciotta D, Teunissen CE, et al. Discovery and initial verification of differentially abundant proteins between multiple sclerosis patients and controls using iTRAQ and SID-SRM. *Journal of proteomics.* 2013;78:312–325.

91 Liong S, Di Quinzio MK, Fleming G, Permezel M, Rice GE, Georgiou HM. New biomarkers for the prediction of spontaneous preterm labour in symptomatic pregnant women: a comparison with fetal fibronectin. *BJOG: an international journal of obstetrics and gynaecology*. 2015;122(3):370–379.

7

Peptidomics of Body Fluids

Prathibha Reddy[1,], Claudia Pontillo[2,*], Joachim Jankowski[1], and Harald Mischak[2]*

[1] *Institute for Molecular Cardiovascular Research, University Hospital RWTH Aachen, Aachen, Germany*
[2] *Mosaiques Diagnostics GmbH, Hannover, Germany*

7.1 Introduction

Peptides consist of a chain of 2 to approximately 50 amino acids linked by an amide bond [1, 2]. There is no clear demarcation between peptides and proteins based on molecular weight or the number of amino acids [2]. In the post-genomic era, the term "peptidomics" was firstly defined as the study of low molecular weight proteins, ranging from 0.5 to 15 kDa [3].

In other words, the peptidome concept could be related to biological peptides such as hormones, neuropeptides, cytokines, growth factors, and also inhibitors, activators, or substrates of a pathway [4, 5], which are generated from larger precursors. These molecules are involved in signal exchange between cells, and the transport of these messengers is most often performed through body fluids that enable communication even between cells that are too remote to interact directly or by migration [6]. Another type of peptides is fragments derived by the enzymatic cleavage of proteins *in vivo*. Particularly, those proteolytic events could reflect specific biological states of individuals, which is very interesting in clinical chemistry and modern medicine. Moreover, changes in excretion and modification of peptides may be associated with pathological events.

In proteomics, proteins are digested to shorter fragments by using enzymes to enable analysis by mass spectrometry (MS), but no such cleavage pattern is employed in peptidomics. The native or endogenous peptides are formed during protein processing or degradation processes by the action of proteases; therefore, the peptides are no longer part of the precursor protein. As these peptides are in general found in nanomolar or picomolar range, they are masked by the predominant proteins, hence not detectable by the standard proteomics approaches.

7.2 Clinical Application of Peptidomics

Peptidomics studies aim at the identification of peptides present in a biological sample that could offer reliable information on biological processes. The high concentration of the analytes in a compartment can alter their diffusion or secretion to other compartments. The determination of the concentration of peptides in body fluids, tissues, and cells could contribute to the understanding of human pathology. In particular, peptides could be used as biomarkers when their concentration varies significantly in association with a pathological condition.

Analysis of biological fluids could reveal the health status of an individual, provide informative biomarkers across a wide spectrum of diseases, and elucidate the cause of a disease [7, 8]. MS-based approaches are used for the discovery of novel biomarkers; the results of these experiments could improve clinical diagnosis and prognosis and eventually lead to lifesaving medical treatments [9].

7.3 Different Types of Body Fluids Used in Biomarker Research

7.3.1 Blood

Blood is a routinely used clinical specimen in disease diagnosis and therapeutic monitoring. It is composed of cells and extracellular fluid that circulates in the whole body; thus it reflects the health status of an individual due to the leakage of components from different organs/cells in addition to the classical plasma proteins (albumin, immunoglobulins, transferrin, etc.) [10, 11]. Plasma is the liquid obtained after removal of cells from whole blood,

*Equal contribution.

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

while serum is the liquid collected after blood clotting. For peptidomic profiling, blood plasma is preferred over serum because clotting involves the activation of numerous proteases that cleave highly abundant proteins, thus releasing large amounts of peptides that can prevent the identification of low-abundance native plasma peptides [12]. Moreover, during coagulation process, the low-abundance peptides might be retained in the clot, thus impeding the identification of potential biomarkers. The blood plasma proteome is the most complex and heterogeneous human-derived proteome, and the concentration of total protein is typically in the range of 60–80 mg/ml [13]. The plasma proteome database developed by the Human Proteome Organization (HUPO) contains more than 10 500 proteins. These proteins have various posttranslational modifications (PTMs), so the actual number of isoforms present in plasma is significantly higher [14, 15].

7.3.2 Urine

Urine can be considered as the "fluid biopsy" of the kidney and urogenital tract [16]. Theoretically, urinary proteins can originate from glomerular filtration of plasma proteins, secretion of proteins from renal tubular epithelial cells, and shedding of whole cells along the urinary tract: shedding of apical membranes of renal tubular epithelial cell and exosome secretion. Under normal conditions, low molecular weight proteins and only a small fraction of proteins with middle molecular weight pass freely through the glomerular barriers and reach renal tubules. Because of the high efficacy of the reabsorption process by proximal tubular epithelial cells, all proteins in the tubular lumen are excreted in small amounts under physiological conditions into urine [17].

A protocol for urine collection has been reported by the Human Kidney and Urine Proteome Project (HKUPP) and the European Urine and Kidney Proteomics (EuroKUP) initiatives (see http://www.hkupp.org and http://www.eurokup.org for detailed information).

Urine collection is not invasive and a volume sufficient for peptidomics analysis can be easily obtained. The urinary proteome is quite stable because urine stays in the bladder for a considerable time before collection. This provides sufficient time for total proteolytic processing at 37°C by endogenous proteases. Urine is less complex than blood, which allows simpler sample preparation procedures. Samples can be stored for several years at –20°C without significant alteration of the urinary proteome [18, 19].

Due to these advantages, urine has been often used as a convenient source for biomarker discovery. In order to incorporate the huge amount of data generated from urinary studies, to collect those data, and to allow researchers to discover new relationships between diseases and proteins, several databases were created [20–24]. A large dataset including over 20 000 data of human naturally occurring urinary peptides under pathophysiological conditions was developed by Mischak et al. (http://mosaiques-diagnostics.de/diapatpcms/mosaiquescms/front_content.php?idcat=257).

A list of urinary databases is reported in Table 7.1.

Besides blood and urine, researchers are investigating other biological fluids such as cerebrospinal fluid (CSF), saliva, interstitial fluid, amniotic fluid, and follicular fluid for diagnostic biomarker discovery. The usefulness of these alternative body fluids for biomarker discovery has not been yet clearly established as that of plasma and urine specimens. The main aim of the clinical proteomics is to acquire reliable data possible for diagnostic and therapeutic purposes. Therefore, in this chapter, the use of plasma and urine in clinical peptidomics via MS is presented.

Table 7.1 Proteomics and peptidomics urinary databases.

Database	Link	Topic	Reference
HKUPP Database	http://www.hkupp.org	Proteome of normal kidney and urine	
EuroKUP	http://www.eurokup.org	Human and kidney proteomics	
MAPU Urine Dataset	http://www.mapuproteome.com	Human urinary proteins	[20]
Clinical Urine Proteomics Database	http://alexkentsis.net/urineproteomics/	Proteomics database	[21]
Urinary Exosome Protein Database	https://hpcwebapps.cit.nih.gov/ESBL/Database/Exosome/	Proteins identified from exosomes in normal human urine	[22]
Urinary Peptide Biomarker Database	http://mosaiques-diagnostics.de/mosaiques-diagnostics/human-urinary-proteom-database	CE-MS results of naturally occurring human urinary peptides under pathophysiological conditions	[23]
Urinary Protein Biomarker Database	http://122.70.220.102/biomarker	Manually curated human and animal urine protein biomarker database	[24]

7.4 Sample Preparation and Separation Methods for Mass Spectrometric Analysis

7.4.1 Depletion Strategies

Analysis of plasma by mass spectrometric techniques either quantitatively or qualitatively offers the possibility for identification of biomarkers for a particular disease. However, 90% of the total protein content corresponds to highly abundant proteins like albumin and immunoglobulins [25]. Moreover, the identification of low-abundance proteins/peptides is hindered by the presence of high-abundance proteins/peptides, and thus the discovery of novel biomarkers is challenging. For example, angiotensin II levels were found to be 18 pM in healthy subjects and gradually increase as the stage of chronic kidney disease (CKD) progresses [26]. The prostate-specific antigen (PSA) and other biomarkers are also present at low levels, that is, in pg/ml [27]. Consequently, the quantification of the whole proteome in a single assay is quite challenging and constitutes a major bottleneck for new biomarker discovery. In addition, potential biomarkers are often found in concentrations below the limit of detection of the available MS [28]. Therefore, improvements in both sample preparation and peptide analysis are required in order to increase the analytical performance of peptidomics assays.

Currently, different fractionation methods are employed in order to fractionate the peptides and reduce the sample complexity prior to mass spectrometric analysis, such as chromatography (reversed phase, affinity, size exclusion, ion exchange, etc.), electrophoresis (1D-PAGE, 2D-PAGE, capillary, etc.), ultrafiltration, and precipitation [29, 30]. These techniques differ in their principles and instrumentation and achieve fractionation by exploiting the differences in physicochemical properties of peptides. Nowadays, a combination of different techniques is applied in a single step by packing different matrices in a single column interfaced with a mass spectrometer enabling fast sample processing. One such example is Multidimensional Protein Identification Technology (MuDPIT) where a strong cation-exchange resin and a reverse-phase resin are packed in a single column that is interfaced with electrospray ionization (ESI) tandem mass spectrometry (MS/MS) [31]. Affinity-based depletion of highly abundant proteins is also applied in peptidomics studies prior to mass spectrometric analysis [32].

In any multidimensional technique, the primary step in the workflow involves the depletion of predominant proteins that substantially increases the sensitivity for identifying low-abundant proteins/peptides. However, other proteins are present in the high-abundance protein fraction. So it is important to analyze all the fractions in order to ensure that no other vital proteins are omitted by chance. In the past, albumin was depleted using the Cibacron blue F3GA column, a hydrophobic chlorotriazine dye conjugated to sepharose, which has more affinity toward albumin [33, 34]. This is a simple and inexpensive method for removing albumin from plasma. Other synthetic dyes are developed, which mimic the Cibacron blue dye and have an even greater affinity toward albumin. However, these dyes bind nonspecifically to other proteins too [35, 36]. Immunoglobulins are removed using protein G/A by affinity chromatography that can be combined with the dye ligand for removal of albumin [37–40]. Nowadays, antibody/immunoaffinity ligands for albumin and immunoglobulins are used more widely than other less specific methods (dyes, protein G/A). Multiple Affinity Removal Column or immunoaffinity capturing columns allow the partial elimination of the most abundant plasma proteins, namely, albumin, IgG, antitrypsin, IgA, transferrin, haptoglobin, and apolipoproteins [41, 42]. Chicken IgY antibodies have been used successfully for depletion of high-abundance plasma proteins, allowing the enrichment for low-abundance peptides [43–49]. However, these approaches cannot totally remove a high-abundance protein; hence a more appropriate term is immunoaffinity-based protein subtraction chromatography (IASC).

Besides the antibody-based methods, Bio-Rad developed a protein enrichment technology under the trade name of ProteoMiner™ where a combinatorial hexapeptide library is utilized to achieve enrichment of low-abundance proteins based on hydrophobic and charge-to-charge interactions between proteins and immobilized hexapeptides [50, 51]. Moreover, the isolated peptides or proteins are in their native state, hence retaining their biological activity. Another approach aiming to enrich for phosphopeptides is based on titanium dioxide chromatography. The hydrophilic phosphate of phosphopeptides interacts with porous titanium dioxide surface by forming coordinate covalent bonds [52]. A recent report presents data on high-abundance protein depletion with the use of anionic hydrogel particles of poly(*N*-isopropylacrylamide-*co*-acrylic acid) [53].

7.4.1.1 Ultrafiltration

Ultrafiltration is a simple, fast, and affordable sample preparation method that uses defined molecular weight cutoff membrane filters to separate proteins into low molecular weight and high molecular weight fractions by centrifugation [54]. High molecular weight molecules are retained on the membrane, whereas low molecular weight molecules pass through the membrane. Amicon Ultra Centrifugal filters are generally used for protein purification and concentration and for desalting [55]. Zougman et al. separated the low molecular weight

peptides (neuropeptides) from the CSF by ultrafiltration, which led to the identification of 563 peptides by nano-LC-MS/MS [54]. Prior to ultrafiltration, acetonitrile (ACN) is added to the samples to disrupt the interaction between proteins and peptides in order to increase the recovery of low molecular weight peptides [56].

7.4.1.2 Precipitation

By addition of organic solvents, acids, or salts, high molecular weight proteins are precipitated, and low molecular weight peptides remain in solution. Normally, in organic precipitation, proteins are efficiently precipitated using ice-cold acetone or ACN [29, 55]. Chertov et al. extracted low molecular weight proteins by ACN with 0.1% TFA as an ion pairing reagent that can disturb the interaction of peptides or low molecular weight proteins with predominant proteins [57]. Different percentages of ammonium sulfate can be used for precipitation of proteins; however this salt interferes with the MS analysis. Kawashima et al. developed a new peptide extraction method called differential solubilization based on plasma dilution with a denaturating solution and addition of ice-cold acetone followed by centrifugation. Later, the precipitate was dissolved in 70% ACN containing 12 mM HCl followed by centrifugation. The supernatant contained low molecular proteins/peptides that were lyophilized and then fractionated by RP-HPLC [58]. This method is more effective and reproducible than ACN precipitation and ultrafiltration. Recently, Pena et al. used 70% perchloric acid for the precipitation of proteins followed by neutralization with potassium hydroxide. After centrifugation the peptides contained in the supernatant were fractionated by RP-HPLC [8].

7.4.1.3 Liquid Chromatography

Size-exclusion chromatography (SEC) is used to separate the molecules based on the size or the hydrodynamic volume of a molecule. This technique is also applicable to separate low molecular weight proteins/peptides from abundant proteins of higher molecular weight. The beads utilized in this technique are available in different pore sizes [59]. However, this technique has certain disadvantages: it requires a large sample volume, the eluted peptides are diluted, and the resolving power is limited [60]. Ueda et al. used SEC for the enrichment of peptidome of serum samples from patients suffering from lung adenocarcinoma in order to identify biomarkers by using nano-LC-MS/MS [61].

Reverse-phase chromatography is routinely used in proteomics prior to MS. It separates the peptides according to their hydrophobicity and can be directly coupled to ESI. In general, the stationary phase is made of C18, C8, and C4 carbon chains that provide a hydrophobic surface for peptide binding. In peptidomics, C18 columns are commonly used, and the peptides are eluted by using an increasing concentration of ACN. Factors that influence the resolution include the type of hydrophobic stationary phase, particle size, sample volume, column length, and the pH of the mobile phase. The efficiency can be improved by increasing the column length and reducing the inner diameter. However, a small inner diameter increases the backpressure and thus requires higher pressure [62, 63].

7.4.1.4 Capillary Electrophoresis

Numerous protocols for extraction, purification, and analysis of proteins are well established. However, few of these methods are applicable in the case of peptides. For example, a standard separation technique for proteins is 2D electrophoresis combined with MS, which cannot be applied for smaller proteins and native peptides (<10 kDa). The concentration of peptides in biological samples is usually low, and the presence of highly abundant proteins prevents peptide detection by the standard proteomics approaches. Moreover, gel-based methods cannot efficiently fractionate peptides. Trypsin digestion, commonly used for large proteins, is not useful for peptides, resulting in a limited number of fragments. Capillary electrophoresis is a separation technique based on the differential electrophoretic velocities of ions in a high-voltage electric field, in which molecules migrate depending upon the charge of the molecule. A broad number of separation modes are available: capillary zone electrophoresis (CZE), capillary gel electrophoresis (CGE), micellar electrokinetic capillary chromatography (MEKC), capillary electrochromatography (CEC), capillary isoelectric focusing (CIEF), and capillary isotachophoresis (CITP). CZE is widely used among the CE techniques due to the fast and highly efficient separation coupling [64].

In CZE (generally referred to as CE), the separation of the peptides depends on differences in electrophoretic mobility applying high voltage to a capillary filled with electrolyte solution. The migration of peptides is directly proportional to the charge of the molecule and inversely proportional to its size [9].

CE is usually coupled to MS via ESI or matrix-assisted laser desorption/ionization (MALDI). ESI is the most common type of ionization coupled online to the CE due to the high ionization efficiency and the soft nature of the ionization process [65].

The coupling of CE to MALDI can only be achieved offline and requires fractionation (spotting) on a target plate, leading to a separation step that is not physically coupled to the ionization step [66].

CE-ESI-MS has demonstrated robustness, relatively short run times (about 60 min per run), and high reproducibility, which is suitable for clinical application and biomarker discovery [65] (Figure 7.1).

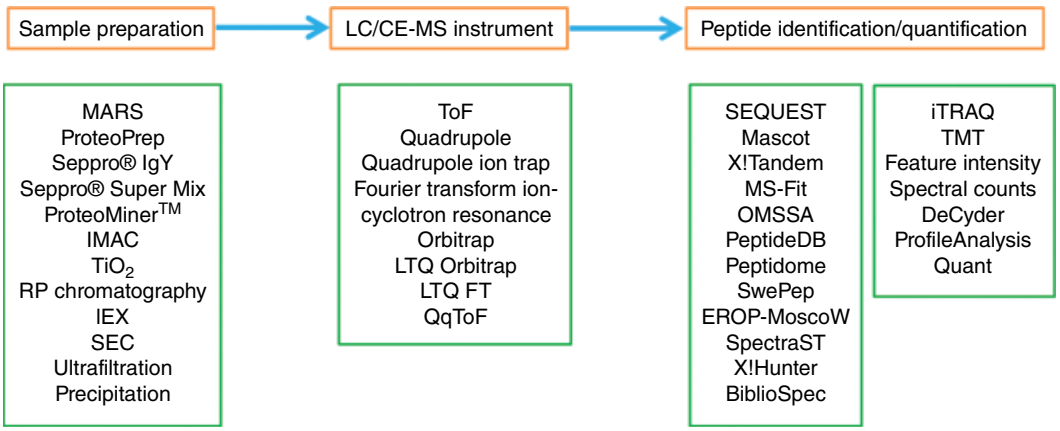


Figure 7.1 Illustration of the workflow for peptidomic analysis.

7.4.1.5 Instrumentation

MS is a technique used to determine the molecular weight, structural information, and elemental composition (isotope pattern) of a wide range of compounds/analytes. The key components of MS are the ion source, the mass analyzer (ToF, quadrupole, quadrupole ion trap, Fourier transform ion cyclotron resonance, Orbitrap), and a detector that records the ion abundance and the mass-to-charge (m/z) ratio. The main requirement for MS analysis is that peptides must be converted to ions in gaseous state. Since peptides are thermally labile and nonvolatile, soft ionization techniques are employed, namely, ESI and MALDI. A wide range of sensitive MS instruments are available. Modern instruments have hybrid mass analyzers that further increase the sensitivity of the instrument [8, 9, 66, 67].

7.5 Identification of Peptides and Their Posttranslational Modifications

In order to identify peptides by MS, the precursor ions must be fragmented. Fragmentation is carried out by collision-induced dissociation (CID), electron transfer dissociation (ETD), or high energy collision dissociation (HCD), either independently or in combination. Thus, a high-resolution fragmentation spectrum is generated and evaluated with different algorithms [68, 69]. The most extensively used algorithms for peptide identifications are SEQUEST [70], Mascot [71], X! Tandem [72], MS-Fit [73], and OMSSA [74]. Many software tools have been developed to deconvolute the spectra for more confident peptide identifications (e.g., Thrash algorithm [75], MS-Decconv [76], Xtract [77], DeconMSn [78]). Yufeng et al. [79] compared different collision methods for plasma peptidome identifications using conventional software tools (SEQUEST, Mascot) and the counts of peptide backbone cleavages (CBC) [80]. Analysis of very large raw spectral that are generated through MS-based peptidomics is quite challenging because naturally occurring peptides have variable termini compared with the peptides that are generated by cleaving proteins with specific enzymes (trypsin). Moreover, amino acids can be modified, and these modifications impede unambiguous peptide identification. In addition, the available databases sometimes yield contrasting results; hence a definitive identification must be supported with additional experimental data [5]. Reisinger et al. [81] introduced an online custom tool, "Database on Demand," that can produce a wide variety of customized sequence databases based on UniProtKB/Swiss-Prot, UniProtKB/TrEMBL, and the International Protein Index (IPI).

Within a single MS experiment, a limited number of spectra are confidently designated to specific peptide sequences, and the rest are ignored due to many reasons. Some of them are (i) flaws in the scoring systems that are employed in the database search tools, (ii) peptide sequence variations due to single nucleotide polymorphisms, (iii) variations from the genomic sequences (splice variants, processed proteins), and (iv) modifications of peptides. Nesvizhskii et al. developed a method ("Spectrum Quality Score") to extract high-quality spectra from tandem spectra that are not identified by the conventional databases and also assign a quality score to each spectrum. The score is assigned based on the spectrum features, sequence tag features, and complementary fragment ions or neutral losses due to loss of ammonia, carbon monoxide, and water [82]. The conventional databases partially utilize the advantages of multiple complementary fragmentation ion spectra generated from Fourier transform instrument. Therefore, Savitski et al. [83] developed a new scoring algorithm (S-Score, ModifiComb) that can filter out the poor data and false positive identifications based on the maximum length of the peptide tag prior to database search. By employing Fisher's linear discriminant analysis, Wu et al. [84] constructed a classifier based on 12 features that evaluate the quality of CID spectra generated

by ion trap instruments in order to distinguish the assigned and unassigned spectra. The spectra with high-quality score that were not matched to a peptide sequence in the database might lead to the identification of novel peptides or PTMs. Analysis of PTMs in peptidomics is important as these PTMs are associated with the stability and activity of the peptides [85].

Spectral clustering is the best approach for identification of novel PTMs from tandem spectra [86]. For identification of PTMs by the search engines, each modification has to be specified in advance, and as result unexpected or novel PTMs cannot be detected. Ma et al. [87] developed a tier-wise scoring algorithm that can identify the unexpected modifications based on the mass shift peaks that are statistically significant. QuickMod, a tool developed to identify the modified peptides from the spectrum libraries without prior specification, utilizes a support vector machine to score the final spectrum based on the spectrum–spectrum match [86]. OpenSea [88] and MS-Alignment [89] are algorithms that improve the identification of PTMs.

Spectral searching became an alternative to sequence database searching. Currently, PeptideDB [90], Peptidome [91], SwePep [92], EROP-Moscow [10, 93], and NeuroPep [94] are used specifically to identify peptides by spectral matching. SpectraST [95], X!Hunter [96], and BiblioSpec [97] are proteomics-based peptide spectral libraries that could be used to construct a particular peptidomics library in order to facilitate the identification process. Finally, to evaluate the error rates of peptide identifications, Kim et al. [98] developed a validation methodology called "generating functions" where spectral energy and probability features are used to identify the peptides. This approach offers an alternative to decoy database search.

In order to apply peptidomics approaches to the clinical setting, peptide identification must be complemented by accurate quantification. Quantification of peptide abundance in clinical samples is an arduous task, especially in the case of biological fluids where peptide concentrations have a significant dynamic range. Variability can be observed due to sample collection, processing, and experimental conditions. Moreover, peptides have highly variable ionization efficiencies based on their chemical properties. Two main quantification methods are applied in peptidomics, namely, label based and label-free. In label-based quantification, the peptides are labeled with different isotopes. Isotope-coded affinity tag (ICAT) is a widely used method [99]. Isobaric tags for relative and absolute quantification (iTRAQ) and tandem mass tags (TMT) are general labeling techniques applied for biofluids. iTRAQ is a single-step chemical labeling procedure where the peptides are identified and quantified simultaneously and has multiplexing capacity [100]. It involves the labeling of primary amine (N-terminal amino and epsilon amino group of lysine) of peptides by *N*-hydroxysuccinimide chemistry [101]. It consists of reporter, balance, and peptide reactive group that provides the quantitative information only in the fragmentation spectra at lower collision energy. The TMT have an extra specific linker group to ensure the fragmentation of reporter ion and relies on the same principle as the iTRAQ method. Vaudel et al. [102] described these quantification methods in detail.

Due to shortcomings in the labeling techniques in untargeted peptidomics approach, label-free quantification methods have gained an immense importance. As the definition itself indicates, labels are not used, and each sample is analyzed separately through LC-MS, and the acquired spectra are compared. Quantification is achieved by extracting feature intensities or by spectral counting. In feature-intensity-based methods, all signal peaks are considered corresponding to one specific charged state of the peptide in the MS/MS spectrum, whereas in spectral counting the peptides are quantified based on the number of fragmentation spectra associated with their identification. Nahnsen et al. [103] presented the available tools for label-free quantification. Label-free quantification methods (i) do not require special sample preparation protocols, (ii) can be applied to small sample volumes, and (iii) are compatible with many different analytical platforms. For differential expression studies, the peptide data are analyzed using different software packages like DeCyder MS Differential Analysis Software (GE Healthcare), ProfileAnalysis (Bruker Daltonics), and Quant.

7.6 Urinary Peptidomics for Clinical Application

7.6.1 Kidney Disease

As thousands of different low molecular weight peptides are naturally occurring in urine and among them a significant proportion is associated with the status of the kidney, urine is typically used as a source of biomarkers for renal diseases. The majority of urinary peptidomics studies for CKD used capillary electrophoresis coupled to mass spectrometry (CE-MS) as an analytical platform. However, this approach does not usually allow the determination of peptide sequence. For peptide sequencing LC-MS/MS is instead superior [104]. MALDI-MS is also reported for biomarker discovery [105]; however the relative abundance of peptides cannot be assessed with confidence based on MALDI-TOF-MS [106]. In this section, we aim to report the main clinical peptidomics studies relative to kidney diseases and urogenital cancers.

CKD is defined by a progressive loss of kidney function and a reduction in the glomerular filtration rate (GFR) or by an increase in urinary albumin excretion.

Good et al. [107] used CE-MS to develop a classifier based on 273 urinary peptides (CKD273 classifier) for diagnosis of CKD. The study included 379 healthy subjects and 230 patients with various kidney diseases. Most of the urinary biomarker peptides were fragments of collagen type (I) chain, downregulated in CKD patients. In contrast, CKD patients displayed increased urinary excretion of fragments of highly abundant plasma proteins (serum albumin and fibrinogen), which may reflect chronic renal damage of the glomerular filtration barrier. The validity of the CKD273 classifier was confirmed in several studies for diagnosis and prognosis of CKD. In a population of diabetic patients, Andersen et al. [108] studied therapeutic effects of irbesartan in microalbuminuric type 2 diabetes patients in urine. They used the CKD273 classifier to evaluate the peptides that showed significant changes upon irbesartan treatment. Eighteen of these CKD markers showed significant differences in urine of patients before and after a 2-year treatment with irbesartan. Zurbig et al. [109] showed that the CKD273 classifier was able to diagnose early the onset of diabetic nephropathy (DN).

Roscioni et al. [110] showed that the CKD273 classifier was independently associated with transition to microalbuminuria or macroalbuminuria and predicted the development and progression of CKD. A recent study by Schanstra et al. [111] demonstrated that the CKD273 classifier performed significantly better in detecting and predicting progression of CKD than the current clinical standard biomarkers (urinary albumin and estimated GFR). The classifier was also more sensitive for identifying patients with rapidly progressing CKD. In this study, novel urinary biomarkers associated with the progression of CKD were identified. These biomarkers mostly included peptides derived from proteins related to inflammation and tissue repair [111].

The application of CE-MS was also explored for autosomal dominant polycystic kidney disease (ADPKD), which is a hereditary kidney disease causing progressive kidney dysfunction and leading to end-stage renal disease (ESRD). The main cause of this disease is mutations on *PKD1* (85% of cases) or *PKD2* gene (15% of cases). Kistler et al. [112] identified 38 urinary biomarkers that efficiently discriminate ADPKD patients from patients with other renal diseases, resulting in a sensitivity of 87.5% and specificity of 97.5% in the validation set. A similar approach was performed also for acute kidney injury (AKI), which is characterized by a sudden increase of serum creatinine levels [113, 114]. Metzger et al. [115] defined 20 potential biomarkers for AKI with a sensitivity of 89% and a specificity of 82% in the validation test. The majority of the identified peptides were derived from collagen type I fragments and α1-antitrypsin.

7.6.2 Urogenital Cancers

Urine has been demonstrated to be a source of biomarkers for urogenital cancers, such as renal cell carcinoma (RCC), bladder cancer (BCa), and prostate cancer (PCa). RCC affects 210 000 patients each year worldwide [116]. It is characterized by the presence of histological necrosis and malignant tumors [117]. Recently, Frantzi et al. [118] identified 86 urinary peptides that could be specifically associated to RCC. A diagnostic classifier was developed and evaluated in an independent set of 76 samples, resulting in 80% sensitivity and 87% specificity for diagnosis of RCC. The peptide biomarkers reported in this study were fibrinogen chains, immunoglobulin Fc regions, hemoglobin subunits, and proteins most likely expressed in the kidney, such as Na/K-transporting ATPase subunit γ, retinitis pigmentosa GTPase regulator, VPS10 domain-containing receptor SorCS3, and the endothelial adhesion molecule CD99 antigen-like protein 2.

BCa is the sixth most common of all malignancies in men [119]. In the context of urinary peptidomics, two studies were performed for this disease. Theodorescu et al. [120] developed and validated a panel of 22 peptides achieving high sensitivity (100%) and specificity (100%) in a study involving 31 patients with urothelial carcinoma, 11 healthy controls, and 138 patients with nonmalignant genitourinary diseases. Schiffer et al. [120] identified polypeptides associated with muscle-invasive BCa. In this study, the majority of identified peptides were derived from four proteins: collagen α-1 (I), collagen α-1 (III), membrane-associated progesterone receptor component 1(PGRMC1), and uromodulin. The results were validated in a blinded cohort of 130 samples from patients with urothelial BCa. In another study, Frantzi et al. [121] used an LC-MS/MS approach for the discovery of native urinary peptides potentially associated with invasive BCa, noninvasive BCa, and benign urogenital diseases [121]. A total of 1845 peptides were identified, corresponding to a total of 638 precursor proteins. Specific enrichment for proteins involved in nucleosome assembly and for zinc finger transcription factors was observed. The differential expression of two candidate biomarkers, histone H2B and NIF-1 (zinc finger 335), in BCa was verified in independent sets of urine samples by ELISA and by immunohistochemical analysis of BCa tissue. The results indicated changes in the expression of both of these proteins with tumor progression, suggesting their potential role as biomarkers for discriminating BCa stages. In addition, the data indicated a possible involvement of NIF-1 in BC progression, likely as a suppressor and through interactions with transcription factors Sox9 and HoxA1 [121].

PCa is the second most common cancer worldwide in men. In the context of urinary peptidomics, some studies have proposed putative biomarkers for PCa diagnosis.

Theodorescu et al. [122] investigated urine samples from patients with PCa and healthy controls and identified a panel of 12 biomarkers for PCa using CE-MS [122].

Schiffer et al. tested the validity of a peptide-based classifier in comparison with PSA, a biomarker used for PCa diagnosis. In 184 participants, PCa was detected in 49 cases. The peptide classifier identified 42 out of 49 tumor patients, showing a sensitivity of 86%. Of 135 PCa-negative patients, 79 had a negative urinary proteome analysis for PCa test (specificity 59%). Negative and positive predictive values were 92 and 43%, respectively. A significant improvement ($P < 0.0005$) in terms of diagnostic accuracy was observed in comparison with serum PSA [123].

M'Koma et al. [124] employed MALDI-TOF to identify potential biomarkers for PCa. Urine samples were fractionated using reverse-phase chromatography, and subsequently peptides were analyzed using MALDI-TOF. 130 signals with a mass range between 1000 and 5000 m/z resulted in a urinary peptide classifier with 71.2% specificity and 67.4% sensitivity in discriminating PCa from benign prostate hyperplasia. However, the sequence of the peptides was not determined.

7.6.3 Blood Peptides as Source of Biomarkers

Blood is in contact with all human tissues, so it is suitable for biomarker discovery. However, it is challenging to discover blood biomarkers for a specific disease.

Serum and plasma have been employed to investigate changes in the peptidome in certain diseases. However, most clinical studies focused on proteome analysis [125, 126]. Serum peptides are susceptible to degradation due to the high proteolytic activity associated with clot formation [127]; thus plasma is preferable for peptidomics studies.

Luczak et al. [128] investigated CKD related to atherosclerosis using plasma as source of possible biomarkers. They used three different sample fractions: high-abundance, low-abundance, and low molecular weight proteins/peptides. Low molecular weight proteins/peptides were directly injected in the LC-MS/MS. The analysis revealed the presence of 36 differentially expressed proteins/peptides; unfortunately only 4 peptides could be identified. Pena et al. [8] studied patients with hypertension ($n = 125$) and type 2 diabetes ($n = 82$) to predict the development of micro- or macroalbuminuria in hypertension or type 2 diabetes. They developed a plasma peptide classifier based on improved risk prediction for transition in albuminuria stage on top of the reference model (C-index from 0.69 to 0.78; $p < 0.01$). Hansen et al. [129] investigated plasma peptidomics in patients with type 1 diabetes. In this study, they identified three candidate biomarkers for DN.

Blood-derived peptides have been investigated also in the context of cancer diagnosis. CE-MS was employed to investigate serum alpha-1-acid glycoprotein one of eight patients with BCa and eight individuals without bladder [130]. Using an enzymatic digestion to deglycosilate glycoproteins, they found higher levels of tri-antennary and tetra-antennary fucosylated oligosaccharides in patients with BCa. Schwamborn et al. [131] investigated the serum peptidome of 41 patients with BCa and 39 healthy individuals. In particular, they established two mathematical models based upon serum peptidome profiles generated by MALDI-MS. Two models were generated using five and six peptides, respectively, ranging from 3.5 to 5.9 kDa. After independent validation in the test set of 64 patients with BCa and 59 healthy individuals, the sensitivity of the two classifiers for the detection of BCa was reported to be 96.4%, while the specificity was 86.5%. However, this study included only a small number of patients and thus requires further validation. Villanueva et al. [132] analyzed by MALDI-MS the serum peptidome of 73 patients with advanced prostate ($n = 32$), breast ($n = 21$), and bladder ($n = 20$) cancer, as well as serum samples from 33 healthy volunteers. The resulting signatures for the three cancer types consisted of 26 (prostate), 50 (bladder), and 25 (breast) naturally occurring peptides, several of which occur in 2 or all 3 cancer groups. One peptide from complement C4a protein and two from the inter-α-trypsin inhibitor heavy chain H4 cluster had consistently higher ion intensities in all cancers than in healthy controls. Instead 3 fibrinopeptide A fragments were lower in all cancers.

7.6.4 Proteases and Their Role in Renal Diseases and Cancer

Numerous studies have been published on the role of more than 550 human proteases in human diseases. Proteases catalyze peptide bond cleavage and contribute to the formation of naturally occurring peptides. They are involved in many physiological processes, and their regulation affects the pathogenesis of various diseases. They are usually present in nature as zymogens that need to be activated. Activators and inhibitors of proteases establish proteolytic networks that regulate many biological pathways. In order to elucidate this complex process, systems biology approaches and conventional biochemical methods are applied. *In vitro* experiments and combinatorial chemistry are used to generate potential substrate libraries for mapping protease specificity [133, 134]. Several bioinformatics tools such as WebLogo [135] and iceLogo [136] helped to identify consensus substrate sequences for each protease. However, these *in vitro* methods cannot identify all the possible substrates cleaved by proteases.

MS-based data coupled to bioinformatics made a substantial contribution to identify *in vivo* substrate cleavage sites. Mapping N- and C-terminal residues generated by *in vivo* digestion of proteins [137–139] define protease substrate sequences with higher accuracy in comparison with the *in vitro* experiments. In order to achieve this goal, it is necessary to measure accurately the

amount of peptides in biological samples. Quantitative methods for MS such as ICAT [140], iTRAQ [141], and SILAC [142] have been used to identify substrates of several proteases. In addition, label-free approaches such as multiple reaction monitoring (MRM), performed on triple-quadrupole instruments, have provided reliable data for peptide concentration. The application of this method provides accurate peptide quantification [143].

In order to assist scientists in disseminating the large amount of data generated by MS-based experiments, there are several websites dedicated to proteases, such as the Proteolysis Map (PMAP [144]), the TopFIND knowledgebase (http://clipserve.clip.ubc.ca/topfind/) [145], MEROPS (http://merops.sanger.ac.uk) [146], and the Mammalian Degradome Database [147]. Recently, a new tool to predict proteases responsible for generating naturally occurring peptides was developed. Proteasix (http://proteasix.org/) allows *in silico* prediction of proteases based on peptide cleavage sites. This tool was already used to test proteases involved in CKD and cardiovascular diseases [104, 148].

Extracellular matrix (ECM) is a noncellular structure present in all the tissues that modulates interaction with epithelial cells, regulating migration, proliferation, cell adhesion, and apoptosis events. Cleavage and remodeling of ECM is promoted by several proteases. The most important class of proteases involved in remodeling of ECM is represented by matrix metalloproteinases (MMPs). Endogenous inhibitors such as specific tissue inhibitors of metalloproteinases (TIMPs) and α-2-macroglobulin regulate metalloproteinase activity [149]. Other important enzymes involved in ECM remodeling are adamalysins (ADAMs) and meprins [150]. Alteration of protease regulation is often associated with pathological conditions, for example, in cancer and renal diseases [151–154]. MMP2 was found to be upregulated in the urine of patients with type 2 diabetes [155], but it was downregulated in the serum proteome of patients with IgAN and lupus nephritis [156]. Moreover, MMP2 was upregulated in the serum of patients with CKD at stages 3 and 4 [157]. Important findings were reported by Caseiro et al. [158] for the regulation of MMP9 in urine, saliva, and serum of type 1 diabetes patients suffering from retinopathy and nephropathy. The study showed the upregulation of MMP9 in saliva and urine; on the contrary, in the serum, MMP9 was downregulated [158].

Interesting results were also reported in BCa research. On the basis of three different studies, upregulation of MMP2 was found in serum [159], plasma [160], and urine [161] of patients with BCa.

In this section we report an overview regarding the proteases involved in pathogenesis of renal disease and BCa (see Table 7.2).

Table 7.2 Protease regulation in human diseases.

Protease	Disease	Regulation in established disease	Type of fluid	Study
MMP2	Diabetic type 2	Up	Urine	[155]
MMP2	IgAN, LN	Down	Serum	[156]
MMP2	CKD stage 3–4	Up	Serum	[157]
MMP2	CKD	Up	Serum	[161]
MMP2	CKD	Up	Plasma	[162]
MMP2	Bladder cancer recurrence	Up	Serum	[159]
MMP2	Bladder cancer	Up	Plasma	[160]
MMP2	Bladder cancer	Up	Urine	[161]
MMP3	Diabetic type 2	Up	Serum	[163]
MMP7	Diabetic type 2	Up	Serum	[163]
MMP7	Bladder cancer, lymph node metastasis	Up	Serum	[164]
MMP7	Bladder cancer	Up	Plasma	[165]
MMP9	Diabetic type 1 with retinopathy and nephropathy	Up	Saliva and Urine	[158]
MMP9	Diabetic type 1 with retinopathy and nephropathy	Down	Serum	[158]
MMP9	Lupus nephritis	Up	Serum	[166]
MMP9	Bladder cancer	Up	Serum	[167]
MMP9	Bladder cancer	Up	Urine	[168]
MMP9	Bladder cancer	Up	Serum	[169]
ADAMTS13	CKD	Down	Plasma	[170]
ADAM10	Glomerular kidney disease	UP	Urine	[171]
ADAM28	Bladder cancer	UP	Urine	[172]

7.7 Concluding Remarks

Peptidomics studies offer novel insights on human diseases by the discovery of clinical biomarkers and biological processes (mainly proteolysis). The main challenge for the future is to increase the efficiency of peptide identification and quantification. Confident determination of known and novel PTMs in peptides will contribute significantly to our understanding of biological processes. The field of peptidomics is rapidly developing and the prospects are bright.

References

1 Rlinghaus R, Shaefer J, Schweet R. Mechanism of peptide bond formation in polypeptide synthesis. *Proc Natl Acad Sci U S A* 1964 Jun;51:1291–1299.

2 Finoulst I, Pinkse M, Van DW, Verhaert P. Sample preparation techniques for the untargeted LC-MS-based discovery of peptides in complex biological matrices. *J Biomed Biotechnol* 2011;2011:245291.

3 Schulz-Knappe P, Zucht HD, Heine G, Jurgens M, Hess R, Schrader M. Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. *Comb Chem High Throughput Screen* 2001 Apr;4(2):207–217.

4 Jankowski J, Schanstra JP, Mischak H. Body fluid peptide and protein signatures in diabetic kidney diseases. *Nephrol Dial Transplant* 2015 Aug;30(Suppl 4):iv43–iv53.

5 Salem S, Jankowski V, Asare Y, Liehn E, Welker P, Raya-Bermudez A, et al. Identification of the vasoconstriction-inhibiting factor (VIF), a potent endogenous cofactor of angiotensin II acting on the angiotensin II type 2 receptor. *Circulation* 2015 Apr 21;131(16):1426–1434.

6 Schrader M, Schulz-Knappe P. Peptidomics technologies for human body fluids. *Trends Biotechnol* 2001 Oct;19(10 Suppl):S55–S60.

7 McDonald WH, Yates JR, III. Shotgun proteomics and biomarker discovery. *Dis Markers* 2002;18(2):99–105.

8 Pena MJ, Jankowski J, Heinze G, Kohl M, Heinzel A, Bakker SJ, et al. Plasma proteomics classifiers improve risk prediction for renal disease in patients with hypertension or type 2 diabetes. *J Hypertens* 2015 Oct;33(10):2123–2132.

9 Pontillo C, Filip S, Borràs DM, Mullen W, Vlahou A, Mischak H. CE-MS-based proteomics in biomarker discovery and clinical application. *Proteomics Clin Appl*. 2015 Apr;9(3–4):322–334.

10 Bandow JE. Comparison of protein enrichment strategies for proteome analysis of plasma. *Proteomics* 2010 Apr;10(7):1416–1425.

11 Ray S, Reddy PJ, Jain R, Gollapalli K, Moiyadi A, Srivastava S. Proteomic technologies for the identification of disease biomarkers in serum: advances and challenges ahead. *Proteomics* 2011 Jun;11(11):2139–2161.

12 Tammen H, Schulte I, Hess R, Menzel C, Kellmann M, Mohring T, et al. Peptidomic analysis of human blood specimens: comparison between plasma specimens and serum by differential peptide display. *Proteomics* 2005 Aug;5(13):3414–3422.

13 Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002 Nov;1(11):845–867.

14 Nanjappa V, Thomas JK, Marimuthu A, Muthusamy B, Radhakrishnan A, Sharma R, et al. Plasma Proteome Database as a resource for proteomics research: 2014 update. *Nucleic Acids Res* 2014 Jan;42(Database issue): D959–D965.

15 Mann M, Jensen ON. Proteomic analysis of post-translational modifications. *Nat Biotechnol* 2003 Mar;21(3):255–261.

16 Mischak H. Pro: urine proteomics as a liquid kidney biopsy: no more kidney punctures! *Nephrol Dial Transplant* 2015 Apr;30(4):532–537.

17 Thongboonkerd V. Recent progress in urinary proteomics. *Proteomics Clin Appl* 2007 Aug;1(8):780–791.

18 Mischak H, Delles C, Klein J, Schanstra JP. Urinary proteomics based on capillary electrophoresis-coupled mass spectrometry in kidney disease: discovery and validation of biomarkers, and clinical application. *Adv Chronic Kidney Dis* 2010 Nov;17(6):493–506.

19 Filip S, Pontillo C, Peter SJ, Vlahou A, Mischak H, Klein J. Urinary proteomics and molecular determinants of chronic kidney disease: possible link to proteases. *Expert Rev Proteomics* 2014 Oct;11(5):535–548.

20 Adachi J, Kumar C, Zhang Y, Olsen JV, Mann M. The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biol* 2006;7(9):R80.

21 Kentsis A, Monigatti F, Dorff K, Campagne F, Bachur R, Steen H. Urine proteomics for profiling of human disease using high accuracy mass spectrometry. *Proteomics Clin Appl* 2009 Sep 1;3(9):1052–1061.

22 Pisitkun T, Shen RF, Knepper MA. Identification and proteomic profiling of exosomes in human urine. *Proc Natl Acad Sci U S A* 2004 Sep 7;101(36):13368–13373.

23 Siwy J, Mullen W, Golovko I, Franke J, Zurbig P. Human urinary peptide database for multiple disease

biomarker discovery. *Proteomics Clin Appl* 2011 Jun;5(5–6):367–374.
24 Shao C, Li M, Li X, Wei L, Zhu L, Yang F, et al. A tool for biomarker discovery in the urinary proteome: a manually curated human and animal urine protein biomarker database. *Mol Cell Proteomics* 2011 Nov;10(11):M111.
25 Tu C, Rudnick PA, Martinez MY, Cheek KL, Stein SE, Slebos RJ, et al. Depletion of abundant plasma proteins and limitations of plasma proteomics. *J Proteome Res* 2010 Oct 1;9(10):4982–4991.
26 Schulz A, Jankowski J, Zidek W, Jankowski V. Absolute quantification of endogenous angiotensin II levels in human plasma using ESI-LC-MS/MS. *Clin Proteomics* 2014;11(1):37.
27 Fortin T, Salvador A, Charrier JP, Lenz C, Lacoux X, Morla A, et al. Clinical quantitation of prostate-specific antigen biomarker in the low nanogram/milliliter range by conventional bore liquid chromatography-tandem mass spectrometry (multiple reaction monitoring) coupling and correlation with ELISA tests. *Mol Cell Proteomics* 2009 May;8(5):1006–1015.
28 Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003 Mar 13;422(6928):198–207.
29 Polson C, Sarkar P, Incledon B, Raguvaran V, Grant R. Optimization of protein precipitation based upon effectiveness of protein removal and ionization effect in liquid chromatography-tandem mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci* 2003 Mar 5;785(2):263–275.
30 Such-Sanmartin G, Bache N, Callesen AK, Rogowska-Wrzesinska A, Jensen ON. Targeted mass spectrometry analysis of the proteins IGF1, IGF2, IBP2, IBP3 and A2GL by blood protein precipitation. *J Proteomics* 2015 Jan 15;113:29–37.
31 Wolters DA, Washburn MP, Yates JR, III. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 2001 Dec 1;73(23):5683–5690.
32 Polaskova V, Kapur A, Khan A, Molloy MP, Baker MS. High-abundance protein depletion: comparison of methods for human plasma biomarker discovery. *Electrophoresis* 2010 Jan;31(3):471–482.
33 Travis J, Bowen J, Tewksbury D, Johnson D, Pannell R. Isolation of albumin from whole human plasma and fractionation of albumin-depleted plasma. *Biochem J* 1976 Aug 1;157(2):301–306.
34 Gianazza E, Arnaud P. A general method for fractionation of plasma proteins. Dye-ligand affinity chromatography on immobilized Cibacron blue F3-GA. *Biochem J* 1982 Jan 1;201(1):129–136.
35 Zolotarjova N, Martosella J, Nicol G, Bailey J, Boyes BE, Barrett WC. Differences among techniques for high-abundant protein depletion. *Proteomics* 2005 Aug;5(13):3304–3313.
36 Gallant SR, Koppaka V, Zecherle N. Dye ligand chromatography. *Methods Mol Biol* 2008;421:61–69.
37 Bjorck L, Kronvall G. Analysis of bacterial cell wall proteins and human serum proteins bound to bacterial cell surfaces. *Acta Pathol Microbiol Scand B* 1981 Feb;89(1):1–6.
38 Greenough C, Jenkins RE, Kitteringham NR, Pirmohamed M, Park BK, Pennington SR. A method for the rapid depletion of albumin and immunoglobulin from human plasma. *Proteomics* 2004 Oct;4(10):3107–3111.
39 Govorukhina NI, Keizer-Gunnink A, van der Zee AG, de JS, de Bruijn HW, Bischoff R. Sample preparation of human serum for the analysis of tumor markers. Comparison of different approaches for albumin and gamma-globulin depletion. *J Chromatogr A* 2003 Aug 15;1009(1–2):171–178.
40 Wang YY, Cheng P, Chan DW. A simple affinity spin tube filter method for removing high-abundant common proteins or enriching low-abundant biomarkers for serum proteomic analysis. *Proteomics* 2003 Mar;3(3):243–248.
41 Bjorhall K, Miliotis T, Davidsson P. Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* 2005 Jan;5(1):307–317.
42 Shah PM. Integrated maternal and child health and family planning. *Indian Pediatr* 1991 Dec;28(12):1445–1451.
43 Hinerfeld D, Innamorati D, Pirro J, Tam SW. Serum/Plasma depletion with chicken immunoglobulin Y antibodies for proteomic analysis from multiple Mammalian species. *J Biomol Tech* 2004 Sep;15(3):184–190.
44 Echan LA, Tang HY, li-Khan N, Lee K, Speicher DW. Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma. *Proteomics* 2005 Aug;5(13):3292–3303.
45 Herosimczyk A, Dejeans N, Sayd T, Ozgo M, Skrzypczak WF, Mazur A. Plasma proteome analysis: 2D gels and chips. *J Physiol Pharmacol* 2006 Nov;57(Suppl 7):81–93.
46 Liu T, Qian WJ, Mottaz HM, Gritsenko MA, Norbeck AD, Moore RJ, et al. Evaluation of multiprotein immunoaffinity subtraction for plasma proteomics and candidate biomarker discovery using mass spectrometry. *Mol Cell Proteomics* 2006 Nov;5(11):2167–2174.
47 Qian WJ, Kaleta DT, Petritis BO, Jiang H, Liu T, Zhang X, et al. Enhanced detection of low abundance human plasma proteins using a tandem IgY12-SuperMix immunoaffinity separation strategy. *Mol Cell Proteomics* 2008 Oct;7(10):1963–1973.
48 Shi T, Zhou JY, Gritsenko MA, Hossain M, Camp DG, Smith RD, et al. IgY14 and SuperMix immunoaffinity

separations coupled with liquid chromatography-mass spectrometry for human plasma proteomics biomarker discovery. *Methods* 2012 Feb;56(2):246–253.

49 Patel BB, Barrero CA, Braverman A, Kim PD, Jones KA, Chen DE, et al. Assessment of two immunodepletion methods: off-target effects and variations in immunodepletion efficiency may confound plasma proteomics. *J Proteome Res* 2012 Dec 7;11(12):5947–5958.

50 Boschetti E, Righetti PG. The ProteoMiner in the proteomic arena: a non-depleting tool for discovering low-abundance species. *J Proteomics* 2008 Aug 21;71(3):255–264.

51 Hartwig S, Czibere A, Kotzka J, Passlack W, Haas R, Eckel J, et al. Combinatorial hexapeptide ligand libraries (ProteoMiner): an innovative fractionation tool for differential quantitative clinical proteomics. *Arch Physiol Biochem* 2009 Jul;115(3):155–160.

52 Klemm C, Otto S, Wolf C, Haseloff RF, Beyermann M, Krause E. Evaluation of the titanium dioxide approach for MS analysis of phosphopeptides. *J Mass Spectrom* 2006 Dec;41(12):1623–1632.

53 Such-Sanmartin G, Ventura-Espejo E, Jensen ON. Depletion of abundant plasma proteins by poly(N-isopropylacrylamide-acrylic acid) hydrogel particles. *Anal Chem* 2014 Feb 4;86(3):1543–1550.

54 Zougman A, Pilch B, Podtelejnikov A, Kiehntopf M, Schnabel C, Kumar C, et al. Integrated analysis of the cerebrospinal fluid peptidome and proteome. *J Proteome Res* 2008 Jan;7(1):386–399.

55 Tucholska M, Scozzaro S, Williams D, Ackloo S, Lock C, Siu KW, et al. Endogenous peptides from biophysical and biochemical fractionation of serum analyzed by matrix-assisted laser desorption/ionization and electrospray ionization hybrid quadrupole time-of-flight. *Anal Biochem* 2007 Nov 15;370(2):228–245.

56 Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* 2003 Oct;2(10):1096–1103.

57 Chertov O, Biragyn A, Kwak LW, Simpson JT, Boronina T, Hoang VM, et al. Organic solvent extraction of proteins and peptides from serum as an effective sample preparation for detection and identification of biomarkers by mass spectrometry. *Proteomics* 2004 Apr;4(4):1195–1203.

58 Kawashima Y, Fukutomi T, Tomonaga T, Takahashi H, Nomura F, Maeda T, et al. High-yield peptide-extraction method for the discovery of subnanomolar biomarkers from small serum samples. *J Proteome Res* 2010 Apr 5;9(4):1694–1705.

59 Barth HG, Boyes BE, Jackson C. Size exclusion chromatography. *Anal Chem* 1994 Jun 15;66(12):595R–620R.

60 Friedman DB, Hoving S, Westermeier R. Isoelectric focusing and two-dimensional gel electrophoresis. *Methods Enzymol* 2009;463:515–540.

61 Ueda K, Saichi N, Takami S, Kang D, Toyama A, Daigo Y, et al. A comprehensive peptidome profiling technology for the identification of early detection biomarkers for lung adenocarcinoma. *PLoS One* 2011;6(4):e18567.

62 MacNair JE, Lewis KC, Jorgenson JW. Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. *Anal Chem* 1997 Mar 15;69(6):983–989.

63 Xu P, Duong DM, Peng J. Systematical optimization of reverse-phase chromatography for shotgun proteomics. *J Proteome Res* 2009 Aug;8(8):3944–3950.

64 Ahmed FE. The role of capillary electrophoresis-mass spectrometry to proteome analysis and biomarker discovery. *J Chromatogr B Analyt Technol Biomed Life Sci* 2009 Jul 15;877(22):1963–1981.

65 Mischak H, Coon JJ, Novak J, Weissinger EM, Schanstra JP, Dominiczak AF. Capillary electrophoresis-mass spectrometry as a powerful tool in biomarker discovery and clinical diagnosis: an update of recent developments. *Mass Spectrom Rev* 2009 Sep;28(5):703–724.

66 Pejchinovski M, Hrnjez D, Ramirez-Torres A, Bitsika V, Mermelekas G, Vlahou A, et al. Capillary zone electrophoresis on-line coupled to mass spectrometry: a perspective application for clinical proteomics. *Proteomics Clin Appl* 2015 Jun;9(5–6):453–468.

67 Glorieux G, Mullen W, Duranton F, Filip S, Gayrard N, Husi H, et al. New insights in molecular mechanisms involved in chronic kidney disease using high-resolution plasma proteome analysis. *Nephrol Dial Transplant* 2015 Jul 9;30(11):1842–1852.

68 Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat Methods* 2007 Oct;4(10):787–797.

69 Shen Y, Tolic N, Xie F, Zhao R, Purvine SO, Schepmoes AA, et al. Effectiveness of CID, HCD, and ETD with FT MS/MS for degradomic-peptidomic analysis: comparison of peptide identification methods. *J Proteome Res* 2011 Sep 2;10(9):3929–3943.

70 Eng JK, McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 1994 Nov;5(11):976–989.

71 Perkins DN, Pappin DJ, Creasy DM, Cottrell JS. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999 Dec;20(18):3551–3567.

72 Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004 Jun 12;20(9):1466–1467.

73 Jimenez CR, Huang L, Qiu Y, Burlingame AL. Searching sequence databases over the internet: protein identification using MS-Fit. *Curr Protoc Protein Sci* 2001 May; Chapter 16:Unit.

74 Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, et al. Open mass spectrometry search algorithm. *J Proteome Res* 2004 Sep;3(5):958–964.

75 Horn DM, Zubarev RA, McLafferty FW. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom* 2000 Apr;11(4):320–332.

76 Liu X, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, et al. Deconvolution and database search of complex tandem mass spectra of intact proteins: a combinatorial approach. *Mol Cell Proteomics* 2010 Dec;9(12):2772–2782.

77 Zabrouskov V, Senko MW, Du Y, Leduc RD, Kelleher NL. New and automated MSn approaches for top-down identification of modified proteins. *J Am Soc Mass Spectrom* 2005 Dec;16(12):2027–2038.

78 Mayampurath AM, Jaitly N, Purvine SO, Monroe ME, Auberry KJ, Adkins JN, et al. DeconMSn: a software tool for accurate parent ion monoisotopic mass determination for tandem mass spectra. *Bioinformatics* 2008 Apr 1;24(7):1021–1023.

79 Shen Y, Tolic N, Purvine SO, Smith RD. Improving collision induced dissociation (CID), high energy collision dissociation (HCD), and electron transfer dissociation (ETD) Fourier transform MS/MS degradome-peptidome identifications using high accuracy mass information. *J Proteome Res* 2012 Feb 3;11(2):668–677.

80 Bythell BJ, Barofsky DF, Pingitore F, Polce MJ, Wang P, Wesdemiotis C, et al. Backbone cleavages and sequential loss of carbon monoxide and ammonia from protonated AGG: a combined tandem mass spectrometry, isotope labeling, and theoretical study. *J Am Soc Mass Spectrom* 2007 Jul;18(7):1291–1303.

81 Reisinger F, Martens L. Database on demand—an online tool for the custom generation of FASTA-formatted sequence databases. *Proteomics* 2009 Sep;9(18):4421–4424.

82 Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, Eddes JS, Gruissem W, et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* 2006 Apr;5(4):652–670.

83 Savitski MM, Nielsen ML, Zubarev RA. New data base-independent, sequence tag-based scoring of peptide MS/MS data validates Mowse scores, recovers below threshold data, singles out modified peptides, and assesses the quality of MS/MS techniques. *Mol Cell Proteomics* 2005 Aug;4(8):1180–1188.

84 Wu FX, Gagne P, Droit A, Poirier GG. Quality assessment of peptide tandem mass spectra. *BMC Bioinf* 2008;9(Suppl 6):S13.

85 Menschaert G, Vandekerckhove TT, Landuyt B, Hayakawa E, Schoofs L, Luyten W, et al. Spectral clustering in peptidomics studies helps to unravel modification profile of biologically active peptides and enhances peptide identification rate. *Proteomics* 2009 Sep;9(18):4381–4388.

86 Falkner JA, Falkner JW, Yocum AK, Andrews PC. A spectral clustering approach to MS/MS identification of post-translational modifications. *J Proteome Res* 2008 Nov;7(11):4614–4622.

87 Ma CW, Lam H. Hunting for unexpected post-translational modifications by spectral library searching with tier-wise scoring. *J Proteome Res* 2014 May 2;13(5):2262–2271.

88 Searle BC, Dasari S, Wilmarth PA, Turner M, Reddy AP, David LL, et al. Identification of protein modifications using MS/MS de novo sequencing and the OpenSea alignment algorithm. *J Proteome Res* 2005 Mar;4(2):546–554.

89 Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol* 2005 Dec;23(12):1562–1567.

90 Liu F, Baggerman G, Schoofs L, Wets G. The construction of a bioactive peptide database in Metazoa. *J Proteome Res* 2008 Sep;7(9):4119–4131.

91 Slotta DJ, Barrett T, Edgar R. NCBI Peptidome: a new public repository for mass spectrometry peptide identifications. *Nat Biotechnol* 2009 Jul;27(7):600–601.

92 Falth M, Skold K, Norrman M, Svensson M, Fenyo D, Andren PE. SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol Cell Proteomics* 2006 Jun;5(6):998–1005.

93 Zamyatnin AA, Borchikov AS, Vladimirov MG, Voronina OL. The EROP-Moscow oligopeptide database. *Nucleic Acids Res* 2006 Jan 1;34(Database issue):D261–D266.

94 Wang Y, Wang M, Yin S, Jang R, Wang J, Xue Z, et al. NeuroPep: a comprehensive resource of neuropeptides. *Database (Oxford)* 2015;2015:bav038.

95 Lam H, Deutsch EW, Eddes JS, Eng JK, Stein SE, Aebersold R. Building consensus spectral libraries for peptide identification in proteomics. *Nat Methods* 2008 Oct;5(10):873–875.

96 Yen CY, Meyer-Arendt K, Eichelberger B, Sun S, Houel S, Old WM, et al. A simulated MS/MS library for spectrum-to-spectrum searching in large scale identification of proteins. *Mol Cell Proteomics* 2009 Apr;8(4):857–869.

97 Frewen B, MacCoss MJ. Using BiblioSpec for creating and searching tandem MS peptide libraries. *Curr Protoc Bioinformatics* 2007 Dec; Chapter 13:Unit.

98 Kim S, Gupta N, Pevzner PA. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J Proteome Res* 2008 Aug;7(8):3354–3363.

99 Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 1999 Oct;17(10):994–999.

100 Evans C, Noirel J, Ow SY, Salim M, Pereira-Medrano AG, Couto N, et al. An insight into iTRAQ: where do we stand now? *Anal Bioanal Chem* 2012 Sep;404(4):1011–1027.

101 Wiese S, Reidegeld KA, Meyer HE, Warscheid B. Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research. *Proteomics* 2007 Feb;7(3):340–350.

102 Vaudel M, Sickmann A, Martens L. Peptide and protein quantification: a map of the minefield. *Proteomics* 2010 Feb;10(4):650–670.

103 Nahnsen S, Bielow C, Reinert K, Kohlbacher O. Tools for label-free peptide quantification. *Mol Cell Proteomics* 2013 Mar;12(3):549–556.

104 Klein J, Papadopoulos T, Mischak H, Mullen W. Comparison of CE-MS/MS and LC-MS/MS sequencing demonstrates significant complementarity in natural peptide identification in human urine. *Electrophoresis* 2014 Apr;35(7):1060–1064.

105 Molin L, Seraglia R, Lapolla A, Ragazzi E, Gonzalez J, Vlahou A, et al. A comparison between MALDI-MS and CE-MS data for biomarker assessment in chronic kidney diseases. *J Proteomics* 2012 Oct 22;75(18):5888–5897.

106 Albalat A, Stalmach A, Bitsika V, Siwy J, Schanstra JP, Petropoulos AD, et al. Improving peptide relative quantification in MALDI-TOF MS for biomarker assessment. *Proteomics* 2013 Oct;13(20):2967–2975.

107 Good DM, Zurbig P, Argiles A, Bauer HW, Behrens G, Coon JJ, et al. Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. *Mol Cell Proteomics* 2010 Nov;9(11):2424–2437.

108 Andersen S, Mischak H, Zurbig P, Parving HH, Rossing P. Urinary proteome analysis enables assessment of renoprotective treatment in type 2 diabetic patients with microalbuminuria. *BMC Nephrol* 2010;11:29.

109 Zurbig P, Jerums G, Hovind P, Macisaac RJ, Mischak H, Nielsen SE, et al. Urinary proteomics for early diagnosis in diabetic nephropathy. *Diabetes* 2012 Dec;61(12):3304–3313.

110 Roscioni SS, de ZD, Hellemons ME, Mischak H, Zurbig P, Bakker SJ, et al. A urinary peptide biomarker set predicts worsening of albuminuria in type 2 diabetes mellitus. *Diabetologia* 2013 Feb;56(2):259–267.

111 Schanstra JP, Zurbig P, Alkhalaf A, Argiles A, Bakker SJ, Beige J, et al. Diagnosis and prediction of CKD progression by assessment of urinary peptides. *J Am Soc Nephrol* 2015 Aug;26(8):1999–2010.

112 Kistler AD, Mischak H, Poster D, Dakna M, Wuthrich RP, Serra AL. Identification of a unique urinary biomarker profile in patients with autosomal dominant polycystic kidney disease. *Kidney Int* 2009 Jul;76(1):89–96.

113 Yohannes S, Chawla LS. Evolving practices in the management of acute kidney injury in the ICU (Intensive Care Unit). *Clin Nephrol* 2009 Jun;71(6):602–607.

114 Ricci Z, Ronco C. Today's approach to the critically ill patient with acute kidney injury. *Blood Purif* 2009;27(1):127–134.

115 Metzger J, Kirsch T, Schiffer E, Ulger P, Mentes E, Brand K, et al. Urinary excretion of twenty peptides forms an early and accurate diagnostic pattern of acute kidney injury. *Kidney Int* 2010 Dec;78(12):1252–1262.

116 Siegel R, Naishadham D, Jemal A. Cancer statistics, 2013. *CA Cancer J Clin* 2013 Jan;63(1):11–30.

117 Fisher R, Gore M, Larkin J. Current and future systemic treatments for renal cell carcinoma. *Semin Cancer Biol* 2013 Feb;23(1):38–45.

118 Frantzi M, Metzger J, Banks RE, Husi H, Klein J, Dakna M, et al. Discovery and validation of urinary biomarkers for detection of renal cell carcinoma. *J Proteomics* 2014 Feb 26;98:44–58.

119 Frantzi M, Latosinska A, Fluhe L, Hupe MC, Critselis E, Kramer MW, et al. Developing proteomic biomarkers for bladder cancer: towards clinical application. *Nat Rev Urol* 2015 Jun;12(6):317–330.

120 Schiffer E, Vlahou A, Petrolekas A, Stravodimos K, Tauber R, Geschwend JE, et al. Prediction of muscle-invasive bladder cancer using urinary proteomics. *Clin Cancer Res* 2009 Aug 1;15(15):4935–4943.

121 Frantzi M, Zoidakis J, Papadopoulos T, Zurbig P, Katafigiotis I, Stravodimos K, et al. IMAC fractionation in combination with LC-MS reveals H2B and NIF-1 peptides as potential bladder cancer biomarkers. *J Proteome Res* 2013 Sep 6;12(9):3969–3979.

122 Theodorescu D, Schiffer E, Bauer HW, Douwes F, Eichhorn F, Polley R, et al. Discovery and validation of urinary biomarkers for prostate cancer. *Proteomics Clin Appl* 2008 Mar 7;2(4):556–570.

123 Schiffer E, Bick C, Grizelj B, Pietzker S, Schofer W. Urinary proteome analysis for prostate cancer diagnosis: cost-effective application in routine clinical practice in Germany. *Int J Urol* 2012 Feb;19(2):118–125.

124 M'Koma AE, Blum DL, Norris JL, Koyama T, Billheimer D, Motley S, et al. Detection of pre-neoplastic and neoplastic prostate disease by MALDI profiling of urine. *Biochem Biophys Res Commun* 2007 Feb 16;353(3):829–834.

125 Kim HJ, Cho EH, Yoo JH, Kim PK, Shin JS, Kim MR, et al. Proteome analysis of serum from type 2 diabetics with nephropathy. *J Proteome Res* 2007 Feb;6(2):735–743.

126 Cho EH, Kim MR, Kim HJ, Lee DY, Kim PK, Choi KM, et al. The discovery of biomarkers for type 2 diabetic nephropathy by serum proteome analysis. *Proteomics Clin Appl* 2007 Apr;1(4):352–361.

127 Kolch W, Neususs C, Pelzing M, Mischak H. Capillary electrophoresis-mass spectrometry as a powerful tool in clinical diagnosis and biomarker discovery. *Mass Spectrom Rev* 2005 Nov;24(6):959–977.

128 Luczak M, Formanowicz D, Marczak L, Pawliczak E, Wanic-Kossowska M, Figlerowicz M, et al. Deeper insight into chronic kidney disease-related atherosclerosis: comparative proteomic studies of blood plasma using 2DE and mass spectrometry. *J Transl Med* 2015;13:20.

129 Hansen HG, Overgaard J, Lajer M, Hubalek F, Hojrup P, Pedersen L, et al. Finding diabetic nephropathy biomarkers in the plasma peptidome by high-throughput magnetic bead processing and MALDI-TOF-MS analysis. *Proteomics Clin Appl* 2010 Sep;4(8–9):697–705.

130 Ongay S, Martin-Alvarez PJ, Neususs C, de FM. Statistical evaluation of CZE-UV and CZE-ESI-MS data of intact alpha-1-acid glycoprotein isoforms for their use as potential biomarkers in bladder cancer. *Electrophoresis* 2010 Oct;31(19):3314–3325.

131 Schwamborn K, Krieg RC, Grosse J, Reulen N, Weiskirchen R, Knuechel R, et al. Serum proteomic profiling in patients with bladder cancer. *Eur Urol* 2009 Dec;56(6):989–996.

132 Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, Olshen AB, et al. Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J Clin Invest* 2006 Jan;116(1):271–284.

133 Diamond SL. Methods for mapping protease specificity. *Curr Opin Chem Biol* 2007 Feb;11(1):46–51.

134 Harris JL, Backes BJ, Leonetti F, Mahrus S, Ellman JA, Craik CS. Rapid and general profiling of protease specificity by using combinatorial fluorogenic substrate libraries. *Proc Natl Acad Sci U S A* 2000 Jul 5;97(14):7754–7759.

135 Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. *Genome Res* 2004 Jun;14(6):1188–1190.

136 Colaert N, Helsens K, Martens L, Vandekerckhove J, Gevaert K. Improved visualization of protein consensus sequences by iceLogo. *Nat Methods* 2009 Nov;6(11):786–787.

137 Urbanska E, Ikonomidou C, Sieklucka M, Turski WA. Aminooxyacetic acid produces excitotoxic lesions in the rat striatum. *Synapse* 1991 Oct;9(2):129–135.

138 McDonald L, Robertson DH, Hurst JL, Beynon RJ. Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides. *Nat Methods* 2005 Dec;2(12):955–957.

139 Agard NJ, Wells JA. Methods for the proteomic identification of protease substrates. *Curr Opin Chem Biol* 2009 Dec;13(5–6):503–509.

140 Dean RA, Butler GS, Hamma-Kourbali Y, Delbe J, Brigstock DR, Courty J, et al. Identification of candidate angiogenic inhibitors processed by matrix metalloproteinase 2 (MMP-2) in cell-based proteomic screens: disruption of vascular endothelial growth factor (VEGF)/heparin affin regulatory peptide (pleiotrophin) and VEGF/Connective tissue growth factor angiogenic inhibitory complexes by MMP-2 proteolysis. *Mol Cell Biol* 2007 Dec;27(24):8454–8465.

141 Prudova A, auf dem KU, Butler GS, Overall CM. Multiplex N-terminome analysis of MMP-2 and MMP-9 substrate degradomes by iTRAQ-TAILS quantitative proteomics. *Mol Cell Proteomics* 2010 May;9(5):894–911.

142 Gioia M, Foster LJ, Overall CM. Cell-based identification of natural substrates and cleavage sites for extracellular proteases by SILAC proteomics. *Methods Mol Biol* 2009;539:131–153.

143 Domon B, Aebersold R. Options and considerations when selecting a quantitative proteomics strategy. *Nat Biotechnol* 2010 Jul;28(7):710–721.

144 Igarashi Y, Heureux E, Doctor KS, Talwar P, Gramatikova S, Gramatikoff K, et al. PMAP: databases for analyzing proteolytic events and pathways. *Nucleic Acids Res* 2009 Jan;37(Database issue):D611–D618.

145 Lange PF, Overall CM. TopFIND, a knowledgebase linking protein termini with function. *Nat Methods* 2011 Sep;8(9):703–704.

146 Rawlings ND, Barrett AJ, Bateman A. MEROPS: the peptidase database. *Nucleic Acids Res* 2010 Jan;38(Database issue):D227–D233.

147 Quesada V, Ordonez GR, Sanchez LM, Puente XS, Lopez-Otin C. The degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Res* 2009 Jan;37(Database issue):D239–D243.

148 Siwy J, Zoja C, Klein J, Benigni A, Mullen W, Mayer B, et al. Evaluation of the Zucker diabetic fatty (ZDF) rat as a model for human disease based on urinary peptidomic profiles. *PLoS One* 2012;7(12):e51334.

149 Thrailkill KM, Clay BR, Fowlkes JL. Matrix metalloproteinases: their potential role in the pathogenesis of diabetic nephropathy. *Endocrine* 2009 Feb;35(1):1–10.

150 Bonnans C, Chou J, Werb Z. Remodelling the extracellular matrix in development and disease. *Nat Rev Mol Cell Biol* 2014 Dec;15(12):786–801.

151 Boyd NF, Martin LJ, Yaffe MJ, Minkin S. Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Res* 2011;13(6):223.

152 Poola I, DeWitty RL, Marshalleck JJ, Bhatnagar R, Abraham J, Leffall LD. Identification of MMP-1 as a putative breast cancer predictive marker by global gene expression analysis. *Nat Med* 2005 May;11(5):481–483.

153 Gonzalez-Avila G, Vadillo-Ortega F, Perez-Tamayo R. Experimental diffuse interstitial renal fibrosis. A biochemical approach. *Lab Invest* 1988 Aug;59(2):245–252.

154 Turck J, Pollock AS, Lee LK, Marti HP, Lovett DH. Matrix metalloproteinase 2 (gelatinase A) regulates glomerular mesangial cell proliferation and differentiation. *J Biol Chem* 1996 Jun 21;271(25):15074–15083.

155 van der Zijl NJ, Hanemaaijer R, Tushuizen ME, Schindhelm RK, Boerop J, Rustemeijer C, et al. Urinary matrix metalloproteinase-8 and -9 activities in type 2 diabetic subjects: a marker of incipient diabetic nephropathy? *Clin Biochem* 2010 May;43(7–8):635–639.

156 Zakiyanov O, Kalousova M, Kratochvilova M, Kriha V, Zima T, Tesar V. Changes in levels of matrix metalloproteinase-2 and -9, pregnancy-associated plasma protein-A in patients with various nephropathies. *J Nephrol* 2013 May;26(3):502–509.

157 Smith ER, Tomlinson LA, Ford ML, McMahon LP, Rajkumar C, Holt SG. Elastin degradation is associated with progressive aortic stiffening and all-cause mortality in predialysis chronic kidney disease. *Hypertension* 2012 May;59(5):973–978.

158 Caseiro A, Ferreira R, Quintaneiro C, Pereira A, Marinheiro R, Vitorino R, et al. Protease profiling of different biofluids in type 1 diabetes mellitus. *Clin Biochem* 2012 Dec;45(18):1613–1619.

159 Gohji K, Fujimoto N, Komiyama T, Fujii A, Ohkawa J, Kamidono S, et al. Elevation of serum levels of matrix metalloproteinase-2 and -3 as new predictors of recurrence in patients with urothelial carcinoma. *Cancer* 1996 Dec 1;78(11):2379–2387.

160 Staack A, Badendieck S, Schnorr D, Loening SA, Jung K. Combined determination of plasma MMP2, MMP9, and TIMP1 improves the non-invasive detection of transitional cell carcinoma of the bladder. *BMC Urol* 2006;6:19.

161 Gerhards S, Jung K, Koenig F, Daniltchenko D, Hauptmann S, Schnorr D, et al. Correspondence re: C. F. M. Sier et al., Enhanced urinary gelatinase activities (matrix metalloproteinases 2 and 9) are associated with early-stage bladder carcinoma: a comparison with clinically used tumor markers. Clin Cancer Res., 6: 2333-2340, 2000. *Clin Cancer Res* 2001 Feb;7(2):445–447.

162 Hsu TW, Kuo KL, Hung SC, Huang PH, Chen JW, Tarng DC. Progression of kidney disease in non-diabetic patients with coronary artery disease: predictive role of circulating matrix metalloproteinase-2, -3, and -9. *PLoS One* 2013;8(7):e70132.

163 Peiskerova M, Kalousova M, Kratochvilova M, Dusilova-Sulkova S, Uhrova J, Bandur S, et al. Fibroblast growth factor 23 and matrix-metalloproteinases in patients with chronic kidney disease: are they associated with cardiovascular disease? *Kidney Blood Press Res* 2009;32(4):276–283.

164 Ban CR, Twigg SM, Franjic B, Brooks BA, Celermajer D, Yue DK, et al. Serum MMP-7 is increased in diabetic renal disease and diabetic diastolic dysfunction. *Diabetes Res Clin Pract* 2010 Mar;87(3):335–341.

165 Szarvas T, Becker M, vom DF, Gethmann C, Totsch M, Bankfalvi A, et al. Matrix metalloproteinase-7 as a marker of metastasis and predictor of poor survival in bladder cancer. *Cancer Sci* 2010 May;101(5):1300–1308.

166 Svatek RS, Shah JB, Xing J, Chang D, Lin J, McConkey DJ, et al. A multiplexed, particle-based flow cytometric assay identified plasma matrix metalloproteinase-7 to be associated with cancer-related death among patients with bladder cancer. *Cancer* 2010 Oct 1;116(19):4513–4519.

167 Jiang Z, Sui T, Wang B. Relationships between MMP-2, MMP-9, TIMP-1 and TIMP-2 levels and their pathogenesis in patients with lupus nephritis. *Rheumatol Int* 2010 Jul;30(9):1219–1226.

168 Gerhards S, Jung K, Koenig F, Daniltchenko D, Hauptmann S, Schnorr D, et al. Excretion of matrix metalloproteinases 2 and 9 in urine is associated with a high stage and grade of bladder carcinoma. *Urology* 2001 Apr;57(4):675–679.

169 Guan KP, Ye HY, Yan Z, Wang Y, Hou SK. Serum levels of endostatin and matrix metalloproteinase-9 associated with high stage and grade primary transitional cell carcinoma of the bladder. *Urology* 2003 Apr;61(4):719–723.

170 Shen L, Lu G, Dong N, Jiang L, Ma Z, Ruan C. Von Willebrand factor, ADAMTS13 activity, TNF-alpha and their relationships in patients with chronic kidney disease. *Exp Ther Med* 2012 Mar;3(3):530–534.

171 Gutwein P, Schramme A, Bdel-Bakky MS, Doberstein K, Hauser IA, Ludwig A, et al. ADAM10 is expressed in human podocytes and found in urinary vesicles of patients with glomerular kidney diseases. *J Biomed Sci* 2010;17:3.

172 Tyan YC, Yang MH, Chen SC, Jong SB, Chen WC, Yang YH, et al. Urinary protein profiling by liquid chromatography/tandem mass spectrometry: ADAM28 is overexpressed in bladder transitional cell carcinoma. *Rapid Commun Mass Spectrom* 2011 Oct 15;25(19):2851–2862.

8

Tissue Proteomics

Agnieszka Latosinska[1,2], Antonia Vlahou[1], and Manousos Makridakis[1]

[1] *Biotechnology Division, Biomedical Research Foundation, Academy of Athens, Athens, Greece*
[2] *Mosaiques Diagnostics GmbH, Hannover, Germany*

8.1 Introduction

Analysis of tissue specimens is well integrated into clinical research and clinical practice, since tissue is a site of different stages of the disease from initiation to progression. The function of the tissue is determined by the set of proteins made by the cells that comprise the tissue. However, analysis at the level of individual proteins does not fully reflect the alterations underlying disease pathophysiology. To meet this challenge, proteomics approaches have been introduced, aiming at comprehensive analysis of all proteins in a tissue. Global analysis of tissue proteome has greatly contributed to putative biomarker discovery as well as to improvement of our knowledge on disease-associated mechanisms, while analysis of individual proteins is broadly applied for validation of biomarker candidates as well as assessment of their prognostic and/or predictive value. An overview of the potential application of tissue proteomics is presented on Figure 8.1.

Advancement in mass spectrometry (MS)-based approaches enabled to move toward more comprehensive characterization of the tissue proteome. Within the past years, several initiatives have been undertaken to describe in depth the human proteome including the investigation of the protein content from specific tissues. As an output, three independent and at the same time complementing human proteome drafts have been developed analyzing more than 60 tissue types (Table 8.1) [1–3]. Moreover, these efforts led to identification of numerous novel proteins, which resulted in improvement on proteome coverage and enhanced further proteomics research. Collected data can be also considered as invaluable reference to support the validity of analytical workflow applied for the collection of new proteomics data. Considering the possibility of false positive identifications in shotgun experiments [4], even accounting for the false discovery rate (FDR), identified proteins may require further verification. The reliability of protein identity can be partially assessed by searching against different proteomics repositories, evaluating, if possible, identifications obtained in the analysis of specific type of tissue. Following the same principle, by using the human proteome maps, the expression of proteins derived from animal disease models can be assessed in human specimens, particularly when the findings discovered in animal models are planned to be investigated in human tissue. Considering also the validation of the proteomics findings in human tissue, the database developed by Uhlen et al. is a powerful tool to help with the selection of antibodies, as the antibodies used in the context of this initiative have undergone a strict validation.

Despite significant advances, application of tissue proteomics is still accompanied with several challenges. First of all, tissue proteome is characterized by high complexity and broad dynamic range of protein concentrations. In an effort to overcome this challenge, the experimental design has to be adjusted, particularly by the introduction of fractionation step(s) prior to MS analysis. This can be achieved at (i) tissue level (subcellular fractionation or laser capture microdissection (LCM)), (ii) protein level (protein fractionation), (iii) peptide level (peptide fractionation), and (iv) MS level (gas-phase fractionation). Moreover, tissue proteomics analysis is also affected by high cellular heterogeneity of the tissue sample. As an example, during resection of the tumor, apart from the tumor cells, the tissue contains also several other non-tumoral elements such as blood vessels, supporting stromal cells, infiltrating lymphocytes, and so on. The homogeneity of studied population of cells can be improved by applying LCM. However, when the size of available tissue is small, the application of LCM might be very difficult. Last but not least, due to the invasive procedure of tissue sampling, availability of

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

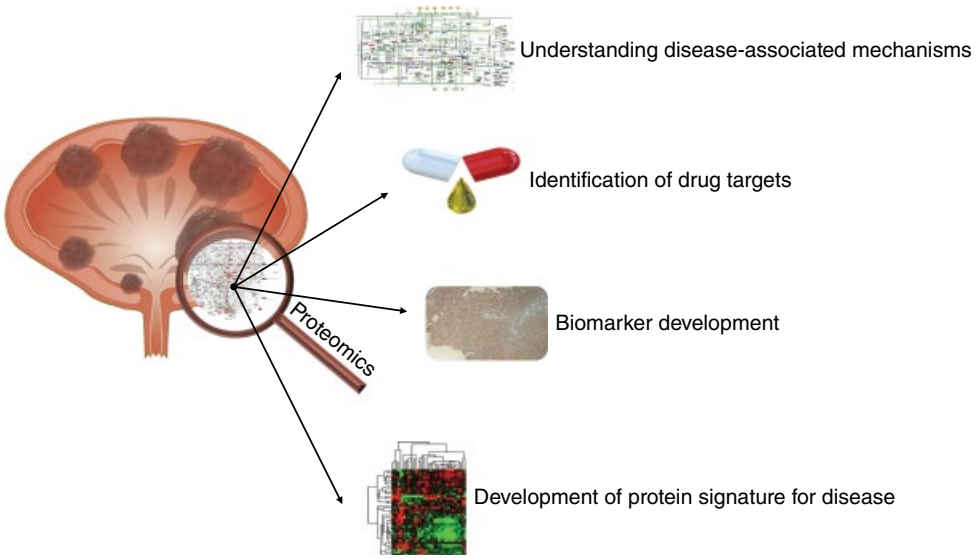


Figure 8.1 Application and significance of tissue proteomics in clinical research. Through a global proteomic profiling of tissue specimens, disease mechanisms can be highlighted at the molecular level, leading to the discovery of potential biomarkers and drug targets.

fresh-frozen tissue sample is often limited. Even though formalin-fixed paraffin-embedded (FFPE) tissue specimens are readily available, formalin fixation hampers their utility in proteomics experiments. Recently, several protocols have been established for processing FFPE samples prior to proteomics analysis [5–7], resulting in more frequent application of this type of specimen. Considering the notable contribution of MS-based techniques in the field of tissue proteomics, we present an overview of sample collection/storage, sample preparation, analytical platforms, and data evaluation.

8.2 Tissue Proteomics Workflow

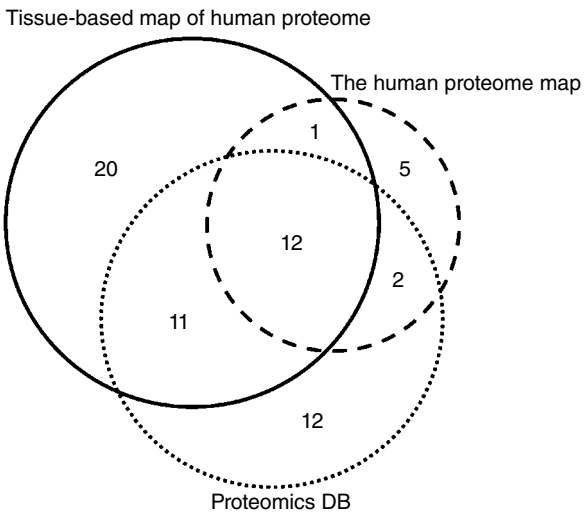
Proteomics analysis of tissue specimens is a complex and time-consuming process, with the main steps illustrated in Figure 8.2.

Tissue proteomics study begins with defining a study design and collection of tissue specimens. In the case of a biomarker discovery study, disease and control groups have to be carefully selected in order to place the biomarker in the context of existing clinical needs. In addition, an added value of the putative biomarker over the current practice has to be shown along with the potential therapeutic consequence [8]. These requirements might be particularly challenging in the tissue-based investigations. Tissue samples are not easily accessible, particularly from control or healthy subjects as well as for monitoring purposes, mainly because of the invasive way of collection. Due to the limited availability of tissue samples, tissue proteomics studies usually suffer in statistical power, and the detected biomarkers might reflect more the intraindividual heterogeneity rather than the association with pathophysiological events. Therefore, the verification of the findings in an independent cohort is required. However, findings from global analysis of tissue proteome can be verified in archival tissues by antibody-based approaches. Since these paraffin-embedded tissue blocks can be preserved and stored in biobanks for years, a high number of samples (of up to thousands) with valuable clinical follow-up data can be available. The main problem is that antibody-based assays have certain limitations, mostly related to antibody availability and high cross-reactivity. On the contrary, when studying physiological events underlying diseased condition, the biological relevance of the findings has to be investigated by using ideally multiple *in vitro* and/or *in vivo* models.

To ensure good quality of the tissue proteomics data, each step has to be adjusted and optimized for a specific study. More importantly, several guidelines and recommendations on study design and reporting of proteomics findings have been introduced [8, 9], particularly in the area of biomarker research. In the frame of the STROBE-ME project [10], guidelines for the epidemiological studies have been developed, while PROBE and REMARK standards have been suggested in the context of predictive [11] and prognostic [12] biomarkers, respectively. However, the detailed description of the guidelines on reporting of proteomics biomarkers is not within the scope of this chapter.

Table 8.1 An overview of the studied tissue proteomes.

Proteome resources/characteristics	Tissues/organs
Tissue-based map of human proteome • Available online at http://www.proteinatlas.org/humanproteome [1] • Methodology: tissue microarray-based IHC • Number of studied tissues/organs: 44	Cerebral cortex, hippocampus, nasopharynx, salivary gland, soft tissue, bronchus, lung, lymph node, liver, adrenal gland, gallbladder, duodenum, small intestine, colon, appendix, smooth muscle, rectum, seminal vesicles, prostate gland, testis, epididymis, skeletal muscle, lateral ventricle, cerebellum, oral mucosa, tonsil, thyroid gland, parathyroid gland, esophagus, heart muscle, breast, stomach, spleen, kidney, pancreas, placenta, fallopian tube, ovary, endometrium, uterine cervix, vagina, urinary bladder, bone marrow, skin
The Human Proteome Map • Available online at http://humanproteomemap.org/ [2] • Methodology: mass spectrometry • Number of studied tissues/organs: 20	Spinal cord, frontal cortex, retina, esophagus, pancreas, colon, rectum, ovary, testis, prostate gland, urinary bladder, kidney, adrenal gland, gallbladder, liver, heart, lung, gut, heart, brain, placenta
ProteomicsDB • Available online at https://www.proteomicsdb.org [3] • Methodology: mass spectrometry • Number of studied tissues/organs: 37	Gallbladder, kidney, seminal vesicle, oral epithelium, esophagus, tonsil, liver, lung, pancreas, colon, rectum, salivary gland, skin, prostate gland, myometrium, uterus, ovary, testis, cardia, stomach, tube, placenta, adrenal gland, thyroid gland, uterine cervix, lymph node, spleen, adipocyte, ascites, cerebral cortex, nasopharynx, vulva, heart, bone, breast, hair follicle, blood platelet
Comparison	
Overlapping investigated proteomes:	
lung, liver, adrenal gland, colon, rectum, prostate gland, testis, esophagus, kidney, pancreas, placenta, ovary	



Proteomics analysis can be performed either on intact proteins (top-down analysis) or on peptides obtained from protein digestion with proteases (bottom-up analysis, also called shotgun proteomics) [13]. These two approaches have been extensively described in the literature (for bottom-up proteomics see, e.g., Refs. [14–16]; for top-down proteomics see, e.g., Refs. [17–19]). Briefly, the peptide-centric analysis is well established and readily applicable for the high-throughput analysis of complex samples like tissue or body fluids, whereas, due to the poor dynamic range of methods employed to resolve intact proteins from complex protein mixtures, the top-down proteomics approach has been rather limited to the analysis of single proteins or simple protein systems [13].

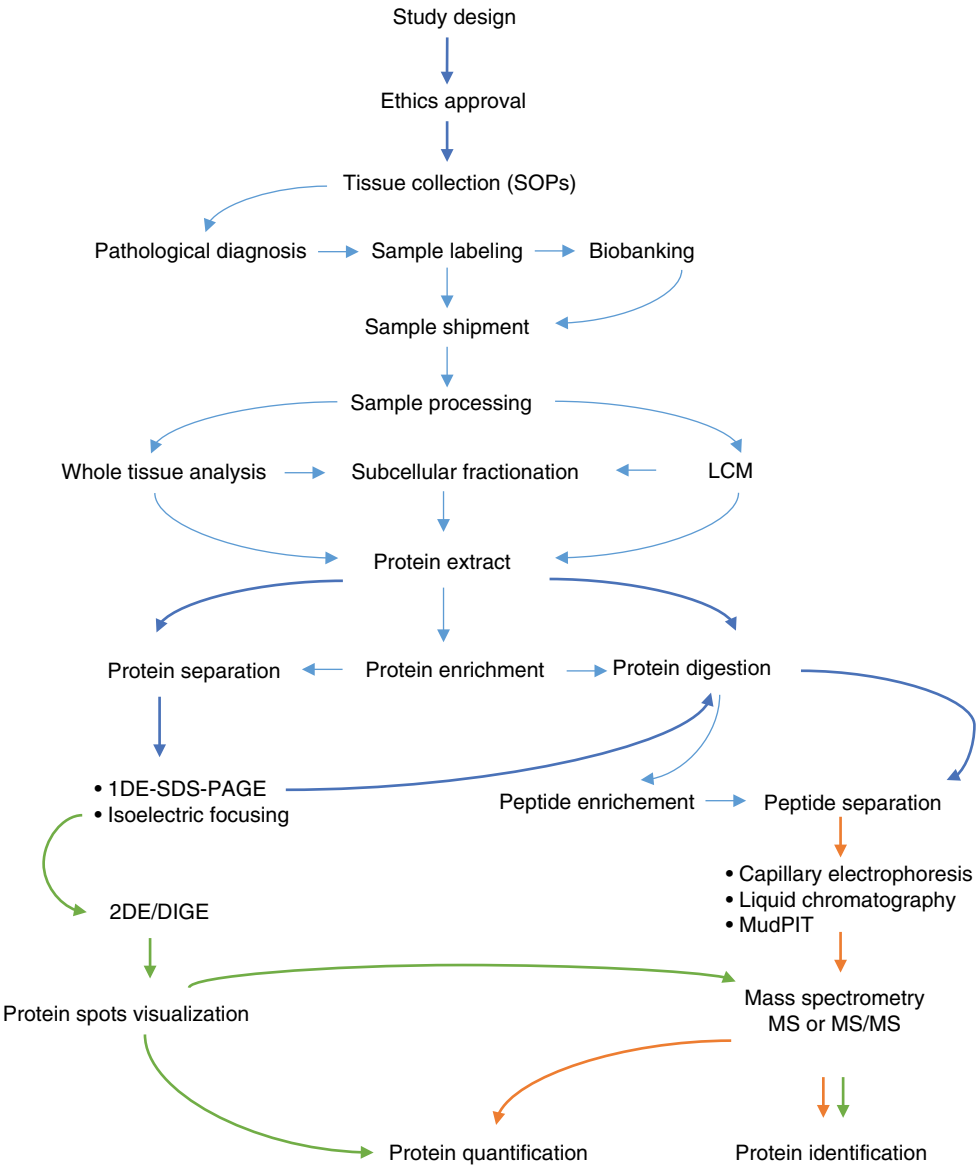


Figure 8.2 Schematic representation of tissue proteomics workflow. The main steps during the tissue proteomics studies are marked in bold.

On the other hand, the top-down approach allows to preserve information on posttranslational modifications (PTMs) present *in vivo* [20], polymorphisms [21], and protein isoforms [22] and provides higher sequence coverage of targeted protein. More intriguingly, recent advancements in fractionation techniques, instrumentation, and software tools allow to observe even thousands of proteins in top-down experiments (reviewed in Ref. [23]). However, when the global characterization of a complex proteome is the main purpose of the study, the bottom-up approach has become the most popular method of choice, and further information on this approach will be more thoroughly described in the next sections.

8.3 Tissue Sample Collection and Storage

Application of an optimal protocol for sample collection and storage has a substantial impact on quality of the proteomics output. Multiple factors including tissue

type, time of collection, materials utilized (container type, additives or preservatives), and procurement (condition and duration) may have a notable impact on quality of the collected specimens and thus stability of assessed proteome [24]. Therefore, samples have to be handled and stored according to well-established protocols that are ideally verified in the initial phase of the study. This is of paramount importance, particularly in the context of the translational multicenter studies, as it allows for the collection of the samples in an unbiased way, and thus, the results of the studies are comparable across the different centers. In order to address these issues and assure for the high quality of collected tissue specimens, standard operating procedures (SOPs) have to be established. Depending on the context of use of tissue specimens, different SOPs have been established. These include guidelines of sample collection for molecular epidemiological studies [24], EORTC-related translational research [25], the European human frozen tumor tissue bank (TuBaFrost) [26], clinical trials (i.e., breast cancer clinical trial [27]), and many other projects (Endometriosis Phenome and Biobanking Harmonisation Project [28]; prospective study of inflammation and the host response to thermal injury [29]). Although the aforementioned SOPs are not specific for the collection of the tissue specimens for proteomics analysis, the reported recommendation/guidelines present basic rules of proper tissue collection for various molecular studies. Some SOPs have been established for tissue sample collection, freezing, and storage for SELDI, MALDI, or 2DE applications [30] or MS-based proteomics analysis of laser microdissected FFPE tissue [31].

As highlighted in the aforementioned SOPs, sample collection involves numerous steps and requires direct collaboration between patients, researchers, clinicians, and hospital staff. The overview on the critical issues concerning the collection of tissue samples is presented in Table 8.2.

Independent of the context of use of tissue specimens, several challenges have been associated with tissue collection. Prior to collection of tissue samples (or any other kind of human biological material), the study has to be approved by an ethics committee and informed consent has to be provided [24]. The main critical point of tissue collection is the time interval between tissue excision and snap freezing. During the excision of the tissue specimen, tissue loses vascular supply (ischemia), leading to increased protease activity, protein degradation, and tissue autolysis. The effect of tissue ischemia time on gene [32–34] or protein expression [34] in the excised tissue has been demonstrated in several investigations. In a study of Spruessel et al. [34], initial changes in gene and protein expression profiles of healthy and tumor colon tissues have been observed after 5–10 min upon tissue excision, while 30 min after surgery, 20% of detectable genes and proteins differed significantly from the baseline values [34]. In general, it has been reported that the time frame between completion of surgery and sample freezing should be around 30 min to assure for the good quality of the material for most of the proteomics techniques [30]. Therefore, to preserve the tissue proteome, the specimens have to be frozen as soon as possible after excision and subsequently stored at −80°C freezer. Moreover, tissue procurement should be arranged without any delay under the proper condition (usually dry ice shipment is a method of choice), as de-freezing of the sample during the shipment may affect to large extent the quality of tissue material. In order to protect the sample from the protease activity, the formalin fixation is often applied. However, application of formalin introduces irreversible modification in proteins (cross-links) and may affect the quality of tissue extracts. Therefore, analysis of the "fresh" tissue is preferable.

8.4 Sample Preparation

Sample preparation has to be performed under controlled conditions and according to standardized procedure. This is a critical point in the context of translational research, as the information based on the established biomarkers or disease models may be used for clinical decision making. Up to now, several sample preparation methods have been established and depend on the type of tissue specimens employed including (i) analysis of fresh-frozen tissue, (ii) analysis of FFPE tissue, or (iii) analysis of specific type of cells isolated using LCM.

8.4.1 Homogenization of Fresh-Frozen Tissue

Homogenization (i.e., sample disruption) is an initial step of sample preparation workflow aiming at disintegration of tissue structure and extraction of tissue-associated molecules such as proteins, nucleic acids, and so on. In general, three homogenization strategies can be distinguish including (i) mechanical homogenization (based on shearing cells using liquid flow, explosion of the cell by pressure differences inside/outside cell, collision forces induced by beads, or combination of different methods), (ii) enzymatic homogenization (lysis by using hydrolytic enzymes), and (iii) chemical homogenization (by using lysis buffers with detergents, chaotropic agents, or other additives) [35]. Very often, a combination of different kind of homogenization strategies is the method of choice. Particularly in the context of tissue analysis,

Table 8.2 An overview on the workflow for collection of fresh-frozen tissue specimens and associated critical factors.

Step	Critical factors
Ethical concerns	Prior beginning of the tissue collection, the study needs to be approved by the ethics committee, and participants of the study have to give informed consent. The ethical concerns should be addressed based on the law applicable in the collecting country
Tissue excision	Storage:
	Upon excision of the tissue, sample should be stored on ice in labeled sterile pot/bag; keeping the tissue in low temperature may delay possible degradation processes
	Timing:
	Excised tissue (fresh and unfixed) has to be transferred for the pathological examination as soon as possible
Pathological examination	Tissue dissection:
	Representative parts of the tissue should be fixed and embedded for routine diagnosis; remaining material (if in sufficient amount) can be dedicated for specific research and/or storage in biobank
	Tissue size:
	The recommended minimal tissue size for freezing should be around 0.5 cm^3
Labeling	Labeling:
	Labeling system depends on local practice; vial containing tissue specimens can be labeled by using unique code, that is, barcode/sequential code/institutional code, etc. Of note, labeling has to be made by using a waterproof pen, suitable for long-term storage at low temperatures. Personal information of the patient should not be included in the labeling process
	Recording:
	All collected samples should be recorded in the inventory book, and upon completion of procedure, the collected data should be transferred into computerized database system
Freezing	Temperature:
	An optimal freezing point for the tissue is −160°C
	Timing:
	Ideally, tissue should be snap frozen within 30 min of excision from patient. However, the lag time between excision of specimen and freezing should be as short as possible. Establishment of proper organizational structure may help to avoid unnecessary delay, which may affect the quality of the specimen
	If the time interval between tissue excision and snap freezing is up to 2 h, the delay has to be reported
Storage	Storage options:
	Tissue samples have to be stored at −80°C freezer or liquid nitrogen storage facility
	Storage details:
	Storage details should be recorded; ideally duplicate samples should be kept in separate freezers, if available
Procurement	Sample should stay frozen during the shipment. Only dry ice shipment is acceptable. The distance and required time of shipment has to be evaluated in advance to assure that the sample stays frozen until reaching the destination place. The package should be also tracked to control the time of the shipment

Source: Adapted from Mager et al. [25] and Morente et al. [26].
The summary presented herein has been prepared based on the information collected from previously published SOPs for tissue collection and biobanking [25, 26].

the mechanical method and its combination with chemical disruption are frequently employed. The enzymatic disruption might not be preferable for the purpose of tissue proteomics; although easy to use, it is associated with the low reproducibility and requires additional steps to remove the utilized enzymes.

There is no universal method applied for homogenization of tissue specimens, and several factors have to be considered prior to deciding on optimal strategy. The latter includes the type of the tissue, sample volume, experimental setup (analysis of total cell lysate; sample enriched in specific organelles; molecules of interest, i.e., proteins, RNA, DNA, etc.), and intended characteristics of final homogenate [36]. Additionally, practical aspects of applicability of respective methods have to be considered in the context of a specific experimental setup, as some of the procedures might not be optimal when a high number of samples have to be processed. In any case, the sample homogenization procedure requires optimization, and the performance of the method should be evaluated in the context of efficiency of protein recovery and reproducibility. Additionally, when reporting the experimental procedure, it is crucial to describe in detail the sample processing workflow as it may have a substantial impact on the outcome. Currently, there are not many guidelines or SOPs on how tissue homogenization (e.g., tumor tissue) should be conducted. To address this issue, EORTC Pathobiology Group developed the SOP protocols for the preparation of tumor tissue extracts suitable for quantitative biomarker analysis [35]. The established method relies on the disruption of the tissue in the deep frozen state by using Mikro-Dismembrator S machine (bead mills technology), resulting in generation of a frozen tissue powder. A detailed protocol is included in the manuscript of Schmitt et al. [35].

An overview on the mechanical methods of tissue homogenization as well as selection of the lysis buffer is summarized in the following sub-sections. Data presented later were assembled from Schmitt et al. [35], Hopkins et al. [37], Goldberg et al. [38], and Burden et al. [36].

8.4.1.1 Mechanical Methods of Tissue Homogenization

8.4.1.1.1 Bead-Based Homogenizers

Bead-based homogenizers rely on collision of the beads with tissue. Beads can be accelerated by vortexing (bead mill homogenizers), shaking (shaking-type bead mills), or spinning (rotor-type bead mills). Depending on the type and volume of the tissue and bead type (density, diameter, material, and quantity), speed of agitation and procedure duration have to be adjusted. However, the temperature during the homogenization has to be controlled, as application of excessive forces may cause sample heating. Application of this method allows for effective homogenization of tissues that are difficult to disrupt. Multiple bead-based homogenizers are currently on the market including Bullet Blender® (Next Advance), Mixer Mill (Retsch), Mini-BeadBeater (BioSpec), and others.

8.4.1.1.2 Blade Homogenizers (Called "Blenders")

Blade homogenizers disrupt tissue by shearing using high speed rotating steel cutting blades. This homogenization method is easy to use and also fast in processing. It allows for an efficient extraction of molecules from both small and large tissue pieces. However, homogenization using blade homogenizers might be associated with aeration and foaming. It may also require cooling of the sample during homogenization or between homogenization steps to avoid overheating. Some examples of blade homogenizers include the following: Bio-Gen PRO200 Hand-Held Homogenizer (PRO Scientific Inc.) or GLH Blade-type Homogenizer (Omni International).

8.4.1.1.3 Rotor-Stator Homogenizers (Also Called Willems Homogenizers)

Rotor-stator homogenizers have a rapidly spinning paddle placed into an open-ended, static tube (stator) with slots close to working end. This allows for quite fast sample processing, although the processing time is associated with toughness and size of the tissue. Rotor-stator homogenizers have been successfully applied to a broad range of sample size/volume and various sample types. However, depending on the application, the geometry of the rotor/stator has to be adjusted, and for small rotors, tissue samples may need to be chopped prior to homogenization. In comparison with blade homogenizer, foaming, aeration, and sample heating are minimized. On the other hand, this type of homogenizer is not easy to clean and it is expensive. Some examples of this type of homogenizer include the following: Polytron (Glen Mills) or Tissue Tearor (Glen Mills).

8.4.1.1.4 Tissue Grinders

Tissue homogenization is based on friction of the sample between two surfaces, causing tearing and ripping of samples. Several tools utilizing grinding as a homogenization strategy were developed including mortar and pestle, tube and pestle, glass homogenizers, CryoGrinder, and others. This homogenization strategy is gentle and has been frequently applied for sample enrichment in subcellular organelles. Tissue grinders are suitable for homogenization of small tissue pieces, while in case of processing of larger tissue specimens, tissue should be

cut into smaller parts. Additionally, sample processing is simple and does not require expensive equipment. However, homogenization is performed manually, although some of the tools can be driven electrically. Therefore, application of grinders can be laborious, particularly for high number of samples, and may result in poor homogenization efficiency, particularly for fibrous and membranous components that remain relatively intact. Some examples of tissue grinders include the following: Potter-Elvehjem grinder (Wheaton), Dounce tissue grinder (Sigma Aldrich), mortar and pestle (Sigma Aldrich), or CryoGrinder™ (CoreCommerce).

8.4.1.1.5 Ultrasonic Homogenizers

Disruption of the tissue occurs through microbubbles that are generated by sonic pressure waves in liquid medium. Depending on the type and volume of the tissue, amplitude has to be adjusted. Ultrasonic homogenizers are suitable for disintegration of "tough" tissue after initial tissue maceration. However, samples can be easily overheated; thus cooling on ice between sonication steps is required. Some examples of ultrasonic homogenizers include the following: Bioruptor® Sonicator (Diagenode) or Sonicator® (Qsonica Sonicators).

8.4.1.2 Chemical Methods of Tissue Homogenization

Chemical homogenization usually assists mechanical disruption of tissue. Prior to downstream proteomics analysis, dissociation of protein complexes, protein denaturation, and solubilization of hydrophobic proteins are required for efficient protein digestion. Achieving a good solubilization of proteins extracted from tissue remains a challenge in MS-based experiments. Several chemicals regularly applied for protein denaturation are not compatible with proteomics analysis, imposing application of additional purification steps. This includes application of chaotropic or denaturating agents. Addition of chaotropic agents such as urea may impair digestion of proteins via (i) denaturation of proteolytic enzymes or (ii) modification of peptides/proteins by isocyanic acid produced by urea upon sample heating. Urea can be easily removed using reverse-phase chromatography (RPC). Sodium dodecyl sulfate (SDS), an ionic detergent frequently used for protein solubilization, interferes with chromatographic separation of proteins/peptides and electrospray ionization. Additionally, a high concentration of SDS might hamper protein digestion. It has been shown that reliable MS analysis is possible for peptide solutions containing up to 0.01% SDS [39]. Considering the benefits associated with SDS use, several methods have been established aiming at removal of SDS prior to LC-MS/MS analysis. These include protein precipitation [40], filter-aided sample preparation (FASP) [41], in-gel protein digestion [42], application of ultrafiltration columns [43], and many others. However, comparison of different protocols for SDS removal is presented elsewhere [39, 44].

Alternatively, several MS-compatible detergents have been developed to facilitate direct proteomics analysis, without additional purification steps. These include acid-labile surfactants [e.g., RapiGest SF (sodium 3-[(2-methyl-2-undecyl-1,3-dioxolan-4-yl) methoxy]-1-propanesulfonate), PPS Silent Surfactant (sodium 3-(4-(1,1-bis(hexyloxy)ethyl)pyridinium-1-yl) propane-1-sulfonate), MS-compatible degradable surfactant (MaSDeS) (sodium 3-((((1-(thiophen-3-yl) undecyl)oxy)carbonyl)amino)propane-1-sulfonate)] that are hydrolyzed at low pH prior to MS analysis or surfactants having different elution time than most of the peptides [e.g., Invitrosol (dimethylbenzylammonium propane sulfonate, 3-(1-pyridinio)-1-propanesulfonate)]. Particularly, application of MaSDeS was well demonstrated for the purpose of tissue proteomics [45]. MaSDeS performance in protein solubilization is comparable with SDS. Importantly, improved protein solubilization as well as number of identified proteins was higher in comparison with other commercially available MS-compatible surfactants (RapiGest, PPS Silent Surfactant, ProteaseMAX, octyl β-D-glucopyranoside, and others) [45].

8.4.2 LCM

Tissues consist of heterogeneous cell populations. Therefore, analysis of proteins extracted from the entire tissue sample may affect interpretation of obtained results, as the origin of a given molecule cannot be easily determined. Investigation of specific cell population might be of high relevance in the context of quantitative tissue proteomics, as the percentage of specific type of cells (e.g., tumor/stromal cells) might differ between samples as well as individual cell populations may behave in a different way under pathological condition. To overcome the problem of tissue heterogeneity and to provide an accurate snapshot on protein alterations, analysis of specific cell populations is required. This can be achieved by using different microdissection techniques. Both manual and automated procedures have been developed, although the former are characterized by limited throughput and reproducibility. These shortcomings were addressed by the introduction of laser-assisted methods. Generally, two laser-based technologies have been implemented including (i) LCM and (ii) laser ablation. However, in the context of this chapter, the focus will be placed on LCM.

Table 8.3 Overview on the application of LCM on human tissues in combination with proteomic analysis.

Proteomics platform	Type of tissue	Biological samples
LC-MS/MS, iTRAQ	Fresh frozen	Oral cancer [51], bladder cancer [52, 53], colon carcinoma [54], lung squamous cell cancer [55, 56], oral cavity squamous cell carcinoma [57]
	FFPE	Nasopharyngeal carcinoma [58], diabetic glomerulosclerosis [59], multiple sclerosis lesions [60]
LC-MS/MS, label-free	Fresh frozen	Lung adenocarcinoma [61], breast cancer [62–64], pancreatic ductal adenocarcinoma [65], biopsies of normal skin, chronic wound keratinocytes from a diabetic patient and glomeruli from needle biopsies of patients with diabetic, lupus and genetic kidney diseases [66], endometrial cancer [67], ductal carcinoma [68]
	FFPE	Pancreatic tissue [69], benign prostatic hyperplasia [70], colonic adenomas [71], colorectal cancer [72], colon cancer [73], pancreatic ductal adenocarcinoma [74] (MudPIT)
LC-MS/MS, O-18/O-16 labeling	Fresh frozen	Gastric adenocarcinoma [75]
LC-MS/MS, SILAC/ SUPER-SILAC	Fresh frozen	Breast cancer [64]
	FFPE	Lung cancer [76]
Reverse-phase protein array	Fresh frozen	Colorectal cancer [77]
	FFPE	Prostate gland [78]
2DE (DIGE) + LC-MS/ MS or MALDI-TOF	Fresh frozen	Gastric cancer [79], nasopharyngeal carcinoma [80], colorectal cancer [81, 82], prostate cancer [83], lung adenocarcinoma [84]

LCM method allows for the analysis of tissue areas of interest by using microscope and laser beam [46, 47]. This method is compatible with different types of tissue preservation techniques, for example, fresh frozen or formalin fixed [30]. Detailed description of sample preparation protocols for each of these techniques has been extensively reviewed here [30, 48–50]. However, microdissection of the cells of interest is a time-consuming and laborious procedure; thus it is not easily applicable when a large number of samples have to be analyzed. Usually the amount of material collected by LCM is limited imposing the need for application of high-resolution proteomics platforms. An overview of recent studies based on LCM in combination with proteomics techniques is presented in Table 8.3.

8.4.3 Protein Digestion

As aforementioned, bottom-up/shotgun proteomics is widely utilized for the analysis of complex samples like tissue. In this approach, prior to MS analysis, proteins are digested to peptides, with a molecular weight (MW) in the range of approximately 500–3000 Da [85]. In comparison with intact proteins, peptides can be fractionated more efficiently and have lower mass and fewer charge states. These properties improve the sensitivity of the analysis [86]. Two types of protein digestion procedures can be followed, that is, enzymatic (using proteolytic enzymes) and nonenzymatic (using chemicals). Trypsin digestion is the most common approach because of the high cleavage efficiency, limited enzyme autolysis, high availability, low cost, and high specificity. Trypsin cleaves peptide residues after Lys and Arg (except when they are followed by Pro), which are found frequently in protein sequence, resulting in generation of peptides comprised of approximately 14 amino acids, with at least 2 positive charges. These peptides are suitable for MS analysis, and their fragmentation results in good quality spectra. However, besides trypsin, several other proteolytic enzymes are available, for example, endoproteinase LysC, LysN, chymotrypsin, and others [85], and can be applied either separately or in combination [87–89]. Since each protease has defined digestion specificity, efficiency, and optimum reaction conditions, utilization of multiple enzymes results in increase of the number of identified proteins and protein sequence coverage (as a result of generation of complementary peptides). However, this approach is not routinely applied in proteomics laboratories, as it typically requires a higher amount of initial material and it may also increase MS run time.

In general, three protein digestion strategies are regularly applied: (i) in-solution digestion [90], (ii) in-gel digestion [42], and (iii) FASP [41]. For the latter method, various modifications have been developed including N-glyco-FASP (a method for generation of deglycosylated peptides from tissue samples for MS analysis

Table 8.4 Comparison of the three main enzymatic digestion strategies applied in bottom-up proteomics.

	In-solution digestion	In-gel digestion	FASP
Principle	Digestion of the extracted protein mixture in solution [90]	In-gel digestion is carried out within polyacrylamide matrix after protein electrophoresis [42]	FASP is employed for the on-filter digestion of detergent lysed cells and tissue prior to MS analysis [41]. By using ultrafiltration devices, lysis buffer is exchanged with urea and ammonium bicarbonate buffer in a series of centrifugations
Practical aspects	• In-solution digestion has to be preceded by protein denaturation, reduction, and alkylation • Prior to LC-MS/MS analysis, the peptide solution has to be purified to remove the interfering substances, for example, using C18 ZipTip	• Applicable after SDS-PAGE or 2D SDS-PAGE • Applicable with no or minor adjustments to gels stained with silver Coomassie colloidal blue • Operations should be performed in laminar flow hood; pipets/tips/tubes should be dust-free	• Filter devices: Amicon ultra centrifugal filter devices (0.5 ml, 30 kDa MWCO, Merck Millipore), 0.5 ml Microcons (Sartorius-Stedim), Vivacon 500 units (30 or 50 kDa MWCO); Recommended MWCO: 30 kDa • High performance for starting protein amount in the range of 25–100 µg [93]
Advantages	• All steps are performed in one tube • Possible automatization, reduction of sample handling • New-generation surfactants (e.g., RapiGest SF, PPS Silent Surfactant) have been developed to improve protein solubilization • Easy in handling and fast	• If the proteins are fractionated by the electrophoresis, the analysis of individual bands/spots increases the dynamic range and depth of the analysis • Removal of low molecular impurities • Enables the analysis of wide range of proteins in the lysate	• Digestion protocol compatible with high detergent concentration • Improved solubilization of membrane fraction due to the high detergent concentration • Allows for collection of protein digest free from nucleic acids and other cellular components • Peptide eluate is clean (no further desalting step required)
Disadvantages	• Incompatible with most detergents • Proteins have to be dissolved in a solution compatible with digestion; otherwise removal of interfering substances is mandatory • Usually, additional desalting step is required	• Increases risk of sample contamination with keratins during casting the gels and processing the excised gel slices • Laborious and time consuming	• Requires multiple centrifugation steps • Due to the application of filters, peptide recovery might be reduced

Examples of studies that have implemented these digestion strategies in the context of tissue proteomics are presented.

[91]), FFPE-FASP (protocol for processing of FFPE tissue samples for MS analysis [92]), and MED-FASP (protocol combining multienzyme sample digestion and FASP [87]). An overview of these digestion strategies is presented in Table 8.4. Several manuscripts have been published aiming at comparing the performance of different digestion techniques for both fresh-frozen [94] and FFPE tissues [95, 96]. However, protein extraction and digestion protocols have to be adjusted for each specific experimental setup. The type of analyzed tissue samples, source of clinical material (fresh frozen or FFPE), sample availability, and amount of starting material should be considered when selecting the optimal protocol.

8.5 Overcoming Tissue Complexity and Protein Dynamic Range: Separation Techniques

Tissue samples are characterized by high complexity and high dynamic range of protein concentration. It is estimated that a single eukaryotic cell contains an average of 20 000–50 000 unique proteins, while the total number of proteins found in tissue may exceed 100 000 [97]. Moreover, protein abundance within a single cell varies in the range between approximately 100 and 100 000 000 copies per cell (10^6 orders of magnitude) and might be even higher at the tissue level [97]. On the contrary, the

dynamic range of proteomics platforms spans only a few orders of magnitude, reducing the success rate for identification of low-abundance proteins. Therefore, to overcome challenges related to tissue complexity and protein dynamic range prior to MS analysis, several fractionation methods have been established. These include fractionation applied at the tissue/cell level (subcellular fractionation) and at the peptide/protein level. Separation techniques applicable at the protein and peptide level are categorized into gel based (one-dimensional (1D) or two-dimensional (2D) gel electrophoresis, isoelectric focusing (IEF)) and gel-free (HPLC, multidimensional chromatography, affinity chromatography). A detailed description and comparison of these approaches was covered in the context of several recently published review articles. Gel-based and gel-free approaches have complementary properties allowing for identification of unique proteins by each approach [98]. The benefits derived from the combination of different strategies to analyze complex samples were demonstrated in several studies [99]. It has been reported that adjustment of certain parameters in proteomics experiments associated with the separation at protein/peptide level (the degree of protein separation, initial amount of peptides subjected to chromatographic separation, and the degree of peptide separation) may lead to improvement of comprehensiveness of proteomics analysis [100]. This indicates that successful and comprehensive proteomics analysis requires a series of optimization steps. Besides classical fractionation strategies, novel methods are being developed to address the high tissue complexity and broad dynamic range in protein expression such as heat stabilization of the tissue proteome [101].

8.5.1 Subcellular Fractionation

Reduction of tissue proteome complexity can be attempted even prior to the peptide/protein separation step. This can be achieved using LCM (as described earlier) and/or subcellular fractionation techniques. In general, subcellular fractionation includes a homogenization step to disrupt cellular structure of analyzed material and separate the cellular components. The latter is possible, since different organelle have specific physicochemical properties such as size, charge, density, and others, allowing for their separation [102]. Several methods have been described to enrich tissue sample in specific type of subcellular organelles (e.g., nuclear envelope [103], nuclear matrix [104], mitochondria [105, 106], plasma membrane [107, 108]). Moreover, some protocols have been developed to capture multiple subcellular fractions in a single experiment. For example, application of differential centrifugation in density gradients allows for isolation of nuclear, mitochondrial, microsomal, and cytosolic proteins within approximately 5 h [109]. Even though these fractionation strategies enable isolation of multiple cellular fractions from a single tissue sample, they may not provide as good purity as methods focused on isolation of an individual fraction. However, applying subcellular fractionation strategies prior to proteomics analysis is associated with additional experimental steps, and may introduce an additional variablity [102]. Various subcellular fractionation methods as well as challenges associated with proteomics analysis of isolated organelles have been extensively reviewed elsewhere [102, 110, 111, 112]. Moreover, some commercially available kits designed for subcellular fractionation of tissue samples have been developed (e.g., Subcellular Protein Fractionation Kit for Tissues from Thermo Fisher, Mitochondria Isolation Kit for Tissue from Abcam, FOCUS™ SubCell Kit from G-Bioscience, and others). Performance of some of the available kits for subcellular fractionation of cell pellets was compared by Bünger et al. [113] and Rockstroh et al. [114]. However, to the best of our knowledge, this comparison has not been performed yet for kits available for tissue fractionation.

8.5.2 Gel-Based Approaches

Gel-based approaches relied on protein separation in gels based on their MW, isoelectric point, or sequential combination of both features. These include separation of proteins using gel-based systems such as 1D (SDS-PAGE, native PAGE), 2D electrophoresis or IEF. Gel-based approaches represent a group of simple and easy-to-use fractionation strategies, allowing for removal of interfering substances and assessment of sample amount and sample complexity. On the other hand, separation of proteins characterized by similar MW or isoelectric point might be difficult in a complex mixture. Additionally, identification of the proteins from gel matrix involves excision of individual bands/spots and further in-gel digestion, which requires usually manual sample handing. A comparison of the gel-based techniques [i.e., 1D PAGE, 1D preparative PAGE, 2D PAGE, and IEF-IPG (immobilized pH gradient)] was recently published by Jafari et al. [115], advocating the complementarity of different methods. However, one of the highest number of identified proteins was reported for 1D SDS-PAGE and IEF-IPG. In addition, benefits derived from sequential application of 1DE SPS-PAGE (protein level), IEF-IPG (peptide level) prior to RP-HPLC- LC-MS/MS resulted in notable increase in number of identified proteins in HeLa cell line [116]. An overview on the gel separation strategies, focusing on their applicability in tissue proteomics, is provided below.

1D gel electrophoresis under denaturing conditions (1DE SDS-PAGE) separates proteins based on their MW.

This technique provides information about sample purity and protein degradation level, efficiency of fractionation procedure, reproducibility of sample preparation, and relative quantification of individual proteins [117]. Therefore, it is widely applied during optimization of experimental protocols. More importantly, 1DE SDS-PAGE followed by the identification of the proteins using LC-MS/MS has become one of the most frequently applied gel-based techniques (called GeLC-MS/MS). Upon protein separation, gel is sliced into several bands or one band, which is subjected to enzymatic digestion and tandem mass spectrometry (MS/MS) analysis. Due to the compatibility with detergents, GeLC-MS/MS is a convenient method applied for the analysis of samples difficult to be homogenized or solubilized. Additionally, it can be applied to process a low amount of starting material (10 μg), which is important in the analysis of clinical tissue specimens, as their availability is usually limited. However, GeLC-MS/MS is a laborious and time-consuming procedure; thus more experimental errors could occur such as keratin contaminations.

2D gel electrophoresis is a classical gel-based separation method, usually combined with protein identification using peptide mass fingerprinting (MALDI-TOF MS) or MS/MS. Using 2DE-SDS-PAGE, proteins are first separated based on their isoelectric point followed by subsequent separation based on MW. Application of this method allows for resolving up to 10 000 of proteins, including also protein isoforms. However, limited gel-to-gel reproducibility and resolving capacity of hydrophobic, low abundance, high MW proteins or proteins with an isoelectric point beyond the pH range of IPG strips are still considered as shortcomings of this method. However, problems associated with gel-to-gel reproducibility have been partially eliminated through development of DIGE. Additionally, identification of complete set of separated proteins is time consuming, as single protein spot has to be excised and analyzed separately. This can be automated though application of spot-picking machines. Numerous studies up to now have utilized 2DE to study tissue proteome. Additionally, several 2DE reference maps for human (liver, kidney) as well as mouse tissues (adipose tissue, gastrocnemius muscle, liver, etc.) were established [118]. Up to date only 36 of 2DE reference maps (including yeast, *Escherichia coli*, *Staphylococcus aureus*, *Mus musculus*, *Homo sapiens*, etc.) have been deposited on SWISS-2DPAGE, which might reflect the trend toward application of LC-MS/MS in the proteomics field.

IEF separates molecules on the basis of their charge. Separation is carried on acrylamide gel matrix with a pH gradient IPG. Besides the classical way of separation using IPG strips, several novel systems have been developed; IEF can be also performed using OFFGEL fractionation system (e.g., Agilent 3100 OFFGEL Fractionator) or capillary system, and these systems will be more thoroughly described in the next section. IEF has been employed to separate both proteins and peptides mixture in combination with other gel-based or gel-free approaches (online or offline), leading to multidimensional separation. This includes application of the IPG-IEF to separate the protein mixture prior to 2DE as well as to separate protein/peptide mixture prior to LC-MS/MS analysis [119, 120]. Benefits derived from application of IPG strips are related to highly reproducible pH gradient, allowing for easy assessment of p*I* range of individual fractions. Upon completion of IEF, IPG strips are cut in small and equal pieces followed by peptide extraction from the strip. However, peptides are diffuse in the strip while cutting the strip into fraction, leading to decreased resolution. Additionally, it has been reported that extraction of peptides from the strip is more efficient than in-gel digestion. Moreover, extracted peptides require further purification prior to MS analysis. Peptides identified from the individual IEF fractions can be further filtered on the basis of their pI supporting the identification process [120]. More details about fractionation of peptides using IPG-IEF can be found in the review by Cargile et al. [120]. Comparison of performance of IPG-IEF peptide separation method with other strategies used for shotgun analysis has been recently reported [120, 121].

8.5.3 Gel-Free Approaches

Gel-free separation techniques have become a standard for the purpose of MS-based proteomics. These include capillary isoelectric focusing (CIEF), high-pressure liquid chromatography (HPLC), multidimensional chromatography, and affinity chromatography [13]. Among these, HPLC is the most commonly applied method in MS-based proteomics, allowing for the separation of molecules based on different properties, depending on the chromatographic materials employed. The following chromatographic material has been broadly applied in MS-based experiments: ion exchange (IEX), reverse phase (RP), hydrophilic interaction chromatography (HILIC), affinity, and hybrid materials [13]. For the purpose of tissue proteomics, application of CIEF has also received an increased attention [122, 123]. The successful application of this technique was also demonstrated for the analysis of laser capture microdissected tissue [124]. Combination of CIEF with RPLC allowed for identification of 6866 peptides and 1820 distinct proteins from glioblastoma multiforme tissue.

Due to the high complexity of biological material, which increases even more in cases of tryptic digests, a single-dimensional separation might not have a sufficient resolving power to separate unfractionated peptides. In

an effort to improve the resolving power, multidimensional separations were introduced by combining various fractionation methods, representing either the same or a combination of different experimental approaches (gel based/gel-free). In addition, each of the individual strategies must separate peptides according to different physiochemical properties [13]. Several setups allowing for the multidimensional separation have been developed and can be either implemented in online or offline system [13]. Furthermore, multidimensional separation strategies to analyze tissue-derived glioblastoma multiforme-derived cancer stem cell (i.e., capillary isotachophoresis (CITP)-based multidimensional separations and multidimensional liquid chromatography (LC) system) were conducted by Fang et al. [125]. The most common approach allowing for multidimensional separation is multidimensional protein identification technology (MudPIT) [126], resolving the molecules using 2D LC (i.e., strong cation-exchange (SCX) and RPC chromatography). The first dimension involves the separation of the molecules based on the charge. Peptides are eluted in a series of washes with increasing salt concentration, and collected fractions are subjected to RPC (separation based on hydrophobicity). Since SCX is not compatible with MS, several improvements of MudPiT procedure were developed. More information on designing on MudPiT experiment can be found in Florens et al. [127]. Several studies have applied MudPiT to investigate tissue proteome. Using MudPiT, Kislinger et al. established an analytical and experimental approach for unbiased proteomics analysis of mammalian cells and tissues (proteomic investigation strategy for mammals (PRISM)), which has been initially applied to investigate proteome from mouse lung and liver tissue [128]. PRISM combines analysis of individual cellular fractions (nucleus, cytosol, soluble mitochondria, insoluble mitochondria, microsomes), multidimensional MS analysis, and further bioinformatics analysis of identified proteins using developed bioinformatics software [128]. MudPiT was also applied to perform a proteomics profiling of nuclear proteins extracted from eight human tissues: brain, heart, liver, lung, muscle, pancreas, spleen, and testis [129]. Many other studies applied also MudPiT in the context of tissue proteomics aiming at establishment of novel biomarkers or better understanding of disease pathophysiology using different types of starting material, for example, fresh-frozen [130] or FFPE tissue sections [74].

8.6 Instrumentation

Due to the technical advancements in MS instruments, MS-based proteomics have become the most prominent branch of proteomics research. Typically, a mass spectrometer consists of an ion source, a mass analyzer, and a detector (to record the number of ions at each m/z). MS analysis starts from conversion of analyte molecules into gas-phase ion (ion source, e.g., MALDI, ESI) following by their storage and separation in a basis of their mass-to-charge ratio (m/z) (mass analyzer). In the last phase, the number of ions at each m/z is recorded (detector) [13]. Several different types of ion sources (e.g., ESI, MALDI), mass analyzers (ion trap (IT), Orbitrap, ion cyclotron resonance (ICR), quadrupoles, and time of flight (TOF)), and detectors have been introduced [13]. Particularly, each type of the mass analyzer exhibits different analytical performance using different strategy to characterize ions. As an example, TOF analyzers measure the flight time of the ions to detector, quadrupoles separate the ions based on the stability of their trajectories in the oscillating electric fields, and in case of IT, Orbitrap, and ICR the separation of ions is based on m/z resonance frequency. This topic has been extensively covered in several excellent reviews. Therefore, the trend to develop hybrid instruments combining different mass analyzers have evolved and resulted in development of several instrument configurations.

For the purpose of proteomics research, several types of instrument configurations have been widely employed including IT, triple quadrupoles, triple TOF, and hybrid instruments such as LTQ Orbitrap, Quadrupole-TOF, and others [13]. In addition, based on publically available MS datasets deposited in ProteomeXchange platform from 2012, LTQ Orbitrap, LTQ Orbitrap Velos, Q Exactive, and TripleTOF 5600 were frequently applied to analyze tissue proteomes. An overview on their analytical performance is provided in Table 8.5. The three most commonly applied instrument configurations are described in the following sub-sections.

8.6.1 LTQ Orbitrap

LTQ Orbitrap (Thermo Scientific) is a hybrid mass spectrometer combining linear IT and Orbitrap technology. The broad applicability of LTQ Orbitrap in proteomics field is attributed to properties of both LTQ (the high sensitivity, speed, capability of MS/MS) and Orbitrap mass analyzers (high resolution and high mass accuracy) [13]. This technology has been used for quantitative bottom-up and top-down proteomics experiments as well as in proteomics analysis of PTMs. Additionally, high mass accuracy allows for improved quantification of low-abundance peptides [131] as well as reduction of false positive identifications, which is of paramount importance for identification of new biomarkers as well as understanding of disease-associated mechanisms. Several proteomics studies demonstrated successful application of LTQ Orbitrap to investigate tissue

Table 8.5 Specification of commonly used mass spectrometers to analyze tissue proteome.

Instrument	Dynamic range[a]	Resolving power	Mass accuracy	Mass range (*m/z*)	Dissociation techniques
LTQ Orbitrap Velos (Thermo Scientific)	>5000	Min. 7 500 Max. 100 000 at *m/z* 400	<3 ppm[b] <1 ppm[c]	50–2000, 200–4000	CID, HCD, optionally ETD
LTQ Orbitrap (Thermo Scientific)	>4000	Min. 7 500 Max. 100 000 at *m/z* 400	<3 ppm[b] <1 ppm[c]	50–2000, 200–4000	CID, PQD, HCD, ETD upgradeable
Q Exactive (Thermo Scientific)	>5000	Max. 140 000 at *m/z* 200	<3 ppm[b] <1 ppm[c]	50–6000	HCD
LTQ Orbitrap Elite (Thermo Scientific)	>5000	Min. 15 000 Max > 240 000 at *m/z* 400	<3 ppm[b] <1 ppm[c]	50–2000, 200–4000	CID, HCD, optionally ETD

The instruments were shortlisted based on publicly available ProteomeXchange datasets using the following search criteria: "Title contains Tissue" and "Species contains Homo sapiens." These instruments were also defined as most frequently used in the field of proteomics research (according to ProteomeXchange datasets collected for *Homo sapiens*).
[a] Dynamic range within a single scan.
[b] Using external calibration.
[c] Using internal calibration.

proteomes (e.g., Refs. [132, 133]). Kume et al. using SCX pre-fractionation and LTQ Orbitrap XL (Thermo Fisher Scientific) identified a total of 5566 proteins from membrane fraction collected from colorectal cancer tissue [132]. In another study, the same technology allowed to establish protein signature for triple-negative breast cancer. Proteomics analysis of a total of 126 tissue samples from patients with lymph node-negative and adjuvant therapy-naive triple-negative breast cancer resulted in development of multi-protein panel comprised of 11 proteins [133].

8.6.2 LTQ Orbitrap Velos

LTQ Orbitrap Velos (Thermo Scientific) is a new generation of LTQ Orbitrap. Velos instrument combines Orbitrap and dual-pressure linear IT technology. Ions are captured and fragmented in the first IT at relatively high pressure, while the second IT allows for a fast scan speed at reduced pressure [134]. Further improvements in the LTQ Orbitrap Velos are related to improved efficiency of vacuum systems as well as the presence of C-trap/HCD collision cell. This allows to achieve increased dynamic range and sensitivity at a higher scan speed. The detailed performance of LTQ Orbitrap versus LTQ Orbitrap Velos was compared in a study by Olsen et al. [134]. The improved performance of LTQ Orbitrap Velos was demonstrated in multiple studies. Proteomics analysis of 95 tumor samples (colon and rectal cancer), characterized before by the Cancer Genome Atlas (TGCA), allows for identification of a total of 124 823 distinct peptides representing 7 526 protein groups, with an FDR of 2.6% [135]. Moreover, by applying multidimensional chromatography and high-resolution proteomics profiling, five molecular subtypes were identified for colon and rectal cancer, two of which overlapped with transcriptomics subtypes established in the frame of TCGA [135].

8.6.3 Q Exactive

Q Exactive (Thermo Scientific) is a hybrid instrument combing the high-performance selection of precursor ions in a quadrupole instrument with the high accuracy and high resolution derived from Orbitrap technology [136]. The performance of Q Exactive and new-generation LTQ Orbitrap Velos was evaluated in RAW 264.7 cell lysate with a range of protein amount 1 ng to 1 µg [137]. Under these conditions, a higher number of peptides and proteins were identified using Q Exactive in comparison with LTQ Orbitrap Velos (HCD) [137]. This is likely attributed to the faster scan rate and higher resolution. On the other hand, Sun et al. [137] pointed out advantages of LTQ Orbitrap Velos including availability of multiple dissociation modes, proving to be beneficial in large-scale proteomics experiments, and analysis of PTMs. Q Exactive is broadly applied in proteomics research. In the context of tissue proteomics, using this technology, Welinder et al. performed an analysis of lymph node metastasis tissue ($n = 10$), which resulted in development of protein sequence database for metastatic melanoma. By analyzing unfractionated and fractionated (SCX) peptide mixture, 5326 unique proteins were identified, among which 2641 proteins overlapped between the two approaches [138]. In another study, Noberini et al. established a method to analyze histone PTMs on FFPE tissues [139]. As a proof of concept, this method was applied to analyze breast cancer samples,

indicating notable changes in histone H3 methylation patterns between triple-negative and luminal A-like disease subtypes [139].

8.7 Quantitative Proteomics

One significant advantage of proteomics analysis is the capability to assess protein abundance. Since tissue is a site of disease initiation and progression, comparative analysis of the protein abundance between different physiological states provides a global "snapshot" on disease-associated changes. Information on the absolute (exact amount of protein or concentration) or relative protein abundance (protein expression trend: up- or downregulation in comparison with control group) are subsequently utilized in a systems biology analysis targeting establishment of *in silico* disease models.

The classical quantification strategy applied in proteomics research combines separation of proteins using 2DE and application of dyes or fluorophores, whereas for the purpose of gel-free MS-based proteomics, two main quantification strategies have been distinguished including label-based and label-free approaches [140]. The former method relies on introduction of isotope labels. Depending on the strategy of incorporation of the labels, several type of label-based quantification approaches were developed such as metabolic labeling (stable isotope labeling with amino acids in cell culture (SILAC), ^{15}N), chemical labeling [isobaric tag for relative and absolute quantitation (iTRAQ), isotope-coded protein labeling (ICPL), isotope-coded affinity tag (iCAT), tandem mass tags (TMT)], or proteolytic labeling (^{18}O) [140]. Both chemical and proteolytic labeling have been utilized to quantify tissue proteomes, with the former being most commonly applicable. Even though metabolic labeling is typically limited to the analysis of cell line models, due to recent developments, SILAC can be also applied for the analysis of tumor tissue proteomes (called super-SILAC). Super-SILAC uses as a reference/internal standard a mixture of different cancer cell lines labeled with SILAC reagent, which is added to tissue extracts in a fixed ratio [141]. Successful application of super-SILAC was reported in the context of breast cancer, brain tumors [141], and lung squamous cell adenocarcinoma [142]. Additionally, a protocol combining super-SILAC with FACS sorting or LCM was developed for quantification of protein changes in cancer cell subpopulations derived from liquid and solid tumors, respectively. This method allows for identification of up to 8000 proteins from patient-derived samples using hybrid quadrupole-Orbitrap MS [76]. An overview on recent developments and application of super-SILAC is provided by Shenoy et al. [143].

On the contrary, label-free approach is easier to use, as it does not require additional labeling steps. Additionally, in the label-free approach there is no limitation with regard to the number of analyzed samples in comparison with label-based methods. However, each sample has to be analyzed individually, which may increase MS instrument use and variability. The accuracy and linearity of the label-free quantification can be affected particularly by the presence of other compounds in the samples, causing suppression effect. Irreproducibility in sample preparation is also a major concern. This might be remediated to some extent using labeled internal standards [140]. Two quantification methods in label-free proteomics are spectral counting and intensity-based quantification [140]. The first method relies on counting the number of MS/MS spectra for a specific protein. Therefore, more abundant proteins generate more abundant peptides, increasing the probability of ion selection for MS/MS analysis. However, differences in the physicochemical properties of peptides might affect detection of peptides by MS and thus may have an impact on quantification using spectral counting. These include peptide length, mass, amino acid sequence, solubility, net charge, and others. Therefore, to address this issue Lu et al. developed a novel method called absolute protein expression (APEX) measurements [144]. In this method, considering the physicochemical properties of individual peptides, probability of their detection is assessed by a supervised classification algorithm. In the intensity-based approach, the quantitation is performed at the MS1 level based on the area under the curve (AUC) from the extracted-ion chromatogram.

Independent of the quantification strategies used, in an effort to accurately compare the quantification results between different samples, data normalization is required. By normalizing the data, an effect associated with differences in protein loading, ionization efficiency, carryover effect, and others can be taken into account. Up to now several normalization methods have been developed and are well described in the context of several manuscripts [145–147].

Based on the aforementioned, numerous techniques are currently being applied to quantify the tissue proteome. It has been shown that both quantification methods were successfully applied either for the analysis of total tissue proteomes [148, 149] or tissues subjected to LCM. Moreover, quantitative proteomics was used to analyze fresh-frozen as well as FFPE tissues. Some studies are listed in Table 8.3. The outcome of these studies is strictly associated with both proteome coverage and accuracy and precision of applied quantification strategies. Comparative analysis of label-free and label-based methods has been broadly described in *in vitro* cultured cells [150–154], while performance in highly complex

tissue samples has not been thoroughly studied. Recently, Latosinska et al. compared label-free (intensity-based) and iTRAQ quantification methods using as starting material BC tissue samples [148]. It has been shown that both methods, that is, label-free and pre-fractioned iTRAQ sample, enable achievement of high proteome coverage and apparently valid predictions in terms of protein differential expression [148]. However, higher sequence coverage and higher number of differentially expressed proteins were demonstrated in the case of label-free approach. However, due to the limited number of analyzed samples, the risk for receiving false associations exists, indicating the need for the analysis of higher sample numbers and/or application of adjustment for multiple testing [148].

8.8 Functional Annotation of Proteomics Data

Untargeted proteomics analysis is capable of identifying hundreds to thousands of proteins, which have to be subsequently interpreted within a specific biological context. Although, in most of the comparative studies, emphasis is given on a shortlist of disease-related proteins (i.e., differentially expressed proteins between case and control groups), the number of features remains high. Therefore, numerous resources and tools have been developed in order to elucidate biological meaning of the proteomics data. Specifically, functional analysis of differentially expressed proteins allows better understanding of disease-associated biological processes and could assist in the selection of specific proteins for further *in vitro*/*in vivo* experiments.

Functional analysis is conducted at the level of individual proteins (analysis of biological processes, molecular function, and subcellular localization), protein–protein interactions (PPIs), or pathways [155, 156]. However, due to the diversity of research aims and proteomics data, there is no consensus on how to extract the biological significance from collected data. Briefly, annotation of proteomics data can be performed either by searching protein databases (e.g., UniProt (http://www.uniprot.org/) [157, 158], neXtProt (http://www.nextprot.org/) [159], Human Protein Atlas (www.proteinatlas.org/) [1, 160], etc.) or applying various computational tools (Table 8.6). Independent of the approach followed, protein identifiers have to be converted to be compatible with selected tools. For that purpose, several ID mapping tools have been developed (e.g., Synergizer [179] (http://llama.mshri.on.ca/synergizer/translate/), PICR [180] (http://www.ebi.ac.uk/Tools/picr/), protein-centric ID mapping service [158, 181] (http://www.uniprot.org/mapping/), and others, as manual adjustment of IDs is prone to mistakes. Moreover, converted identifiers should be checked for the presence of duplicates. Only unique identifiers should be used for functional analysis, as the presence of duplicates may biased results.

Currently, the vocabulary system developed by Gene Ontology (GO) Consortium is commonly applied to annotate *-omics* data according to their contribution in biological processes, molecular function and subcellular localization [182]. The annotation system developed by GO has a hierarchical structure with general annotations at high level of hierarchy and more specific annotations at lower level. Of note, association of GO term to specific protein is supported by specific level of evidence. Most of the evidence codes are assigned manually by a curator (except those defined as Inferred from Electronic Annotation) and are divided into four main categories: (i) experimental, (ii) computational analysis, (iii) author statements, and (iv) curatorial statements. Up to now numerous tools have been developed to both map GO terms and perform enrichment analysis. The latter approach allows identification of biological information that is overrepresented in selected datasets. This feature is frequently used to assess the efficiency of sample preparation procedures, particularly in the context of subcellular fractionation analysis.

Analysis of GO annotation provides an overview on the biological relevance of proteomics findings, but it does not show relations/interactions between proteins as well as associated biological outcome. Since proteins usually function as a part of a complex machinery, PPI analysis (including factional and physical interactions) enables studying of functional modules of proteome. One step further, pathway analysis reflects series of interactions, which leads to specific biological outcome. Available network databases and pathway resources are summarized in the context of meta-database called Pathguide (http://pathguide.org) [183]. Specific resources may have different content, focus, and coverage. Therefore, its selection should be based on the scope of analysis. Similarly to the GO-based analysis, relevance of findings revealed by network and pathway analysis can be priorities using enrichment analysis. However, special caution is advised when evaluating networks, as FDR at the level of retrieved interaction is higher than in other types of analysis. As presented in Table 8.6, many of the currently available tools (e.g., DAVID [173, 174], PANTHER [176], Babelomics [175], and other) integrate information about GO, pathways, domains, interactions, and so on. Some further information on bioinformatics analysis of proteomics data were described by Bhat et al. [184].

Table 8.6 Examples of tools/resources applied for functional annotations of proteomics datasets.

Name	Description	Reference
GO analysis		
AmiGO	http://amigo.geneontology.org/amigo	[161]
GoMiner	http://discover.nci.nih.gov/gominer/index.jsp	[162]
WebGestalt	http://bioinfo.vanderbilt.edu/webgestalt/	[163]
GOrilla	http://cbl-gorilla.cs.technion.ac.il/	[164]
Network analysis		
STRING	http://string-db.org/	[165]
IntAct	http://www.ebi.ac.uk/intact/	[166]
BioGRID	http://thebiogrid.org/	[167]
MINT	http://mint.bio.uniroma2.it/	[168]
Pathway analysis		
IMPaLA	http://impala.molgen.mpg.de/	[169]
KEGG	http://www.genome.jp/kegg/pathway.html	[170]
Reactome	http://www.reactome.org/	[171]
Ingenuity Pathway Knowledge Base[a]	http://www.ingenuity.com/products/	—
MetaCore[a]	https://clarivate.com/products/metacore/	—
Meta-tools		
Cytoscape/Cytoscape Plugins	http://www.cytoscape.org/	[172]
DAVID	https://david.ncifcrf.gov/	[173, 174]
Babelomics	http://babelomics.bioinfo.cipf.es/	[175]
PANTHER Classification System	http://www.pantherdb.org/	[176]
Enrichr	http://amp.pharm.mssm.edu/Enrichr/	[177]
Blast2GO	https://www.blast2go.com/	[178]

[a] Commercially available tools.

8.9 Application of MS-Based Tissue Proteomics in Bladder Cancer Research

To exemplify the applicability of tissue proteomics in BC research, relevant manuscripts were retrieved by literature search using the Web of Science (TOPIC: ("tissue" or "laser capture") AND TOPIC: ("bladder ca*" or "urothelial ca*" or "transitional cell*") AND TOPIC: ("protein*" or "proteom*")). Only original articles published within the last 5 years and utilizing MS-based proteomics platforms were selected and are described below.

The BC tissue proteome has not been extensively studied. Within the last 5 years, a couple of reports have been published offering some insights into tumor biology as well as identification of novel biomarker candidates. The overview on the research activities conducted in the field of BC tissue proteomics is provided in Table 8.7. In most of the studies, proteomics profiles were generated by using LC-MS/MS analysis in combination with either label-free or label-based (iTRAQ) quantification. These two quantification methods were compared by Latosinska et al. using as starting material BC tissue samples from non-muscle-invasive and muscle-invasive cases [148]. The reported result indicated better capability of label-free quantification to identify differentially expressed proteins. However, the results have not been interpreted in the context of tumor biology.

In a series of reports, Niu et al. extensively investigated BC tissue proteins by combining LCM with shotgun proteomics analysis [186–188, 191, 192]. Proteomics profiling of cancer cells/adjacent urothelium as well as cancer stromal cells/adjacent normal stromal cells originated from superficial [185, 186] and muscle-invasive BC [187, 188] was conducted. Through application of various bioinformatics approaches to functionally characterize the identified differentially expressed proteins as well as to predict altered pathways, better understanding of tumor biology could be achieved. In a study of superficial BC, a

Table 8.7 Overview on tissue proteomics studies performed in the context of bladder cancer.

Reference	Methodology	Context of the study	Cohort/proteomics analysis	Type of specimens	Prominent findings[a]
Kato et al. [149]	LC-MS/MS iTRAQ	Biomarker discovery	6 bladder cancers and paired normal mucosa	Proteomics: fresh frozen Verification: FFPE	BCAP31, CCT4, DDX39, FKBP4, IDH1, KRT19, MYH9, P4HB, YBX1
Niu et al. [185]	LCM 2D-LC-MS/MS	Understanding of cancer biology (superficial bladder cancer)	4 paired superficial BC and adjacent normal urothelial tissue sample	Fresh frozen	N/A
Niu et al. [186]	LCM 2D-LC-MS/MS	Understanding of cancer biology by studying stromal cells (superficial bladder cancer)	4 paired superficial BC and adjacent normal urothelial tissue samples	Fresh frozen	N/A
Niu et al. [187]	LCM 2D-LC-MS/MS	Understanding of cancer biology by studying stromal cells (muscle-invasive bladder cancer)	4 paired muscle-invasive BC and adjacent normal urothelial tissue samples	Fresh frozen	N/A
Niu et al. [188]	LCM 2D-LC-MS/MS	Understanding of cancer biology by parallel analysis of urothelial and stromal cells (muscle-invasive bladder cancer)	4 paired muscle-invasive BC and adjacent normal urothelial tissue samples	Fresh frozen	N/A
Liu et al. [189]	LCM 2D-LC-MS/MS iTRAQ	Understanding of heterogeneity for muscle-invasive bladder cancer	30 paired muscle-invasive BC and adjacent normal urothelial tissue samples	Fresh frozen	N/A
Liu et al. [53]	LCM 2D-LC-MS/MS iTRAQ	Understanding of heterogeneity for muscle-invasive bladder cancer by analyzing stromal cells	30 paired muscle-invasive BC and adjacent normal urothelial tissue samples	Fresh frozen	N/A
Latosinska et al. [148]	LC-MS/MS iTRAQ/ LFQ	Evaluation of quantification strategies for tissue proteomics research	4 pTa, 4 pT2+	Fresh frozen	N/A
Chen et al. [52]	LCM LC-MS/MS iTRAQ	Biomarker discovery	4 pairs of primary bladder cancer tumor and adjacent non-tumorous tissue	Proteomics: fresh frozen Verification: FFPE, urine	CA2, PGK1, SFN, SLC3A2, STMN1, TAGLN2, TXN
Oezdemir et al. [190]	MALDI-TOF IMS	Grading of urothelial neoplasms	Samples from patients with pTa BC including 27 low grade (G1), 21 high grade (G3), and 31 high grade (G2)	Fresh frozen	Classification SVM-based model of 23 peaks

LCM, laser capture microdissection; LFQ, label-free quantification; MALDI-TOF IMS, matrix-assisted laser desorption/ionization time-of-flight imaging mass spectrometry; N/A, not applicable.
[a] Most promising proteomics findings selected for validation by using alternative techniques.

total of 580 differentially expressed proteins were defined [185]. These include proteins previously associated with carcinogenesis (e.g., fibronectin, MMP9, galectin-1, alpha-enolase, etc.) as well as novel findings (achaetescute like-2 protein). Further interpretation of the findings in the context of biological functions or pathways indicated the role of differentially expressed proteins in oxidative phosphorylation (e.g., ATP5A1, CYC1, NDUFA3, etc.), metabolic processes (glycolysis, gluconeogenesis, e.g., ADH1A, ALDOA, ENO2), focal adhesion (e.g., ACTN1, COL3A1, FLBN), ribosome (e.g., RPL10A, RPL12, RPS8), and others [185]. In a subsequent study, the same approach was applied to characterize the proteomics profile of cancer/normal stromal cells in superficial BC [186]. The analysis resulted in identification of 637 differentially expressed proteins [186]. Evaluation of those findings revealed an alteration in multiple pathways including ECM receptor interaction (e.g., ITGA5, LAMA4, LAMB2, VNT), cell communication (e.g., TNC, DES, DSG1, COL3A1), focal adhesion (e.g., CDC42, FLNA, FLNC, MYLK), metabolism (fatty acid metabolism, e.g., ACAA1, ACOX1, ALDH1B1), oxidative phosphorylation (e.g., COX17, COX5B, NDUFA8), and others [186]. Some of these pathways and/or associated proteins were also identified by proteomics profiling of cancer/normal cells [185]. Following the same principle, parallel analysis of proteomics profiles in both cancer/cancer stromal cells and normal urothelial/normal stromal cells was performed in the context of muscle-invasive BC [188]. The parallel analysis enabled for identification of 1753 differentially expressed proteins between cancer and normal tissue. These proteins were associated with metabolic pathways (e.g., ACAA1, ACO1, ATP5B), spliceosomes (e.g., ACIN2, EIF4A3, RBM8A), endocytosis (e.g., ACAP2, ARP23, CDC42), regulation of actin cytoskeleton (e.g., ARPC1B, ARPC2, CYFIP1, ROCK), and others. In follow-up studies, Liu et al. quantified protein changes in muscle-invasive BC of different metastatic risk groups by using label-based approach (iTRAQ) [53, 189]. 855, 2210, and 633 proteins were considered as differentially expressed (fold change >1.5) in high-/median-/low-risk groups' relative to normal groups [189]. Interestingly, based on the specific set of changes in each metastatic risk group, the most prominent deregulated pathways were predicted (top 10 pathways based on significance). The main difference was observed between low-risk and medium-/high-risk group. In the higher-/medium-risk group, majority pathways were relevant to genetic information processing or metabolism, while in the low-risk group significantly affected pathways were related to focal adhesion, ECM receptor interaction, complement and coagulation cascade, and others [189]. Along the same lines, cancer stroma proteomics expression profiles of muscle-invasive BC in different metastatic risk groups were conducted [53].

Chen et al. evaluated the diagnostic potential of tissue proteomics-derived findings. Similarly to the previous studies, proteins derived from the homogeneous population of cancer and normal urothelial cells were analyzed by LC-MS/MS [52]. In a series of two iTRAQ experiments, a total of 3217 proteins were identified, with an overlap of 1585 proteins. Considering the magnitude of the fold change (≥1.5 for upregulated and ≤0.67 for downregulated in cancer vs. noncancerous tissue) and frequency of identification of the proteins with this specific change (at least in 2 out of 4 paired samples), 131 and 181 proteins were characterized by increased or decreased expression in BC versus noncancerous samples, respectively. Seven potential biomarkers (i.e., CA2, PGK1, SFN, SLC3A2, STMN1, TAGLN2, TXN) were shortlisted accounting for their biological relevance to human cancer and frequency of detection of differential expression in analyzed samples (proteins overexpressed in at least 3 out of 4 paired microdissected tissue specimens) [52]. Differential expression of STMN1, TAGLN2, and SL3A2 was confirmed in an independent sample set by immunohistochemistry (IHC). In parallel, diagnostic performance of those seven candidates was assessed in urine samples from BC and hernia patients (control group) by using Western blot or ELISA. Collected results indicated proteins that are able to discriminate between early- and late-stage BC (CA2, PGK1, STMN1, TAGLN2), low-grade (LG) and high-grade (HG) tumors (CA2, PKG1, TAGLN2), and BC and control individuals (PKG1, STMN1, TAGLN2), with the AUC value in the range of 0.631–0.712. Among those candidates, tissue and urinary TAGLN2 exhibited the most significant increase in expression in BC patients versus controls [52].

In another study by Kato et al. [149], tissue proteomics analysis was also applied to identify biomarkers for BC progression. From the list of 493 identified proteins, only 15 proteins showed increased expression in cancer tissues compared with adjacent normal tissues (fold change above 1.2) and were detected in at least four out of six sets of BC and paired normal mucosal samples. This includes ACTR3B, BCAP31, CCT4, DDX39, EZR, FKBP4, IDH1, KRT19, MYH9, NPM1, P4HB, PTMA, S100A11, S100P, and YBX1. IHC validation was subsequently performed only for those proteins, which have not been previously investigated in BC. However, in the case of ACTR3B, IHC could not be conducted due to the lack of available antibody. For these proteins, the IHC results support the proteomics findings. Additionally, only the expression of DDX39 was found to be correlated with cancer stage and grade [149]. An inverse correlation between expression of DDX39 and cancer stage and grade was reported. Moreover, based on the follow-up

data, low DDX39 expression was associated with more rapid disease progression ($p = 0.0083$) [149].

Another interesting application of tissue proteomics in the area of BC research was presented by Oezdemir et al. [190]. Compared with other studies described earlier, MALDI-TOF imaging MS was employed to establish the proteomics profile able to discriminate between low grade (LG) and high grade (HG) BC tumors. Classification model was developed based on the data from G1/LG ($n = 27$) and G3/HG ($n = 21$) [190]. From a total of 46 significantly differentially expressed peaks, 23 were selected to be incorporated into a support vector machine (SVM)-based model [190].

8.10 Conclusions

The investigation of tissue proteome is an invaluable approach in clinical research enabling comprehensive molecular characterization of protein complement of specific tissue. This opens up new perspectives to better understand disease-associated mechanisms and thus could contribute to establishment of disease molecular models, allowing for development of biology-driven therapeutic targets. Moreover, application of tissue proteomics may also help to assess the validity of prognostic and/or predictive biomarkers of disease outcome and/or treatment response. Considering a clear clinical demand in the field of cancer research to improve on therapeutic treatment options and develop novel biomarkers, application of tissue proteomics appears to be a suitable tool that can help to fulfill these needs. Recent advances in analytical tools applied in the context of tissue proteomics (e.g., LCM, separation techniques, and MS) as well as developments of novel methodologies (e.g., MS analysis of FFPE tissue specimens) have allowed to some extent to overcome the tissue proteomics associated challenges such as high sample heterogeneity, broad dynamic range of proteins expressed in tissue, or limited sample availability. The impact of these improvements can be further exemplified by their successful application in tissue proteomics-based research. Tissue proteomics has greatly contributed to our understanding of different diseases, but for the moment its application in clinical setting is limited. However, efforts should be made to further validate and interpret the collected data in the context of existing literature by employing a systems biology approach as well as explore their biological role in the cell using *in vitro* and *in vivo* models.

References

1 Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, et al. (2015) Proteomics. Tissue-based map of the human proteome. *Science* 347: 1260419.

2 Kim MS, Pinto SM, Getnet D, Nirujogi RS, Manda SS, et al. (2014) A draft map of the human proteome. *Nature* 509: 575–581.

3 Wilhelm M, Schlegl J, Hahne H, Moghaddas Gholami A, Lieberenz M, et al. (2014) Mass-spectrometry-based draft of the human proteome. *Nature* 509: 582–587.

4 Cargile BJ, Bundy JL, Stephenson JL, Jr. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res* 3: 1082–1085.

5 Crockett DK, Lin Z, Vaughn CP, Lim MS, Elenitoba-Johnson KS (2005) Identification of proteins from formalin-fixed paraffin-embedded cells by LC-MS/MS. *Lab Invest* 85: 1405–1415.

6 Maes E, Broeckx V, Mertens I, Sagaert X, Prenen H, et al. (2013) Analysis of the formalin-fixed paraffin-embedded tissue proteome: pitfalls, challenges, and future prospectives. *Amino Acids* 45: 205–218.

7 Wisniewski JR (2013) Proteomic sample preparation from formalin fixed and paraffin embedded tissue. *J Vis Exp* (79): 50589.

8 Mischak H, Vlahou A, Righetti PG, Calvete JJ (2014) Putting value in biomarker research and reporting. *J Proteomics* 96: A1–A3.

9 Tabb DL (2012) Quality assessment for clinical proteomics. *Clin Biochem* 46: 411–420.

10 Gallo V, Egger M, McCormack V, Farmer PB, Ioannidis JP, et al. (2013) STrengthening the reporting of OBservational studies in Epidemiology-Molecular Epidemiology (STROBE-ME): an extension of the STROBE statement. *Eur J Epidemiol* 26: 797–810.

11 Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD (2008) Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. *J Natl Cancer Inst* 100: 1432–1438.

12 Altman DG, McShane LM, Sauerbrei W, Taube SE (2012) Reporting recommendations for tumor marker prognostic studies (REMARK): explanation and elaboration. *BMC Med* 10: 51.

13 Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng* 11: 49–79.

14 Zhang Y, Fonslow BR, Shan B, Baek MC, Yates JR, 3rd (2013) Protein analysis by shotgun/bottom-up proteomics. *Chem Rev* 113: 2343–2394.

15 Gilmore JM, Washburn MP (2010) Advances in shotgun proteomics and the analysis of membrane proteomes. *J Proteomics* 73: 2078–2091.

16 Horvatovich P, Hoekman B, Govorukhina N, Bischoff R (2010) Multidimensional chromatography coupled to

mass spectrometry in analysing complex proteomics samples. *J Sep Sci* 33: 1421–1437.
17 Cui W, Rohrs HW, Gross ML (2011) Top-down mass spectrometry: recent developments, applications and perspectives. *Analyst* 136: 3854–3864.
18 Armirotti A, Damonte G (2010) Achievements and perspectives of top-down proteomics. *Proteomics* 10: 3566–3576.
19 Calligaris D, Villard C, Lafitte D (2011) Advances in top-down proteomics for disease biomarker discovery. *J Proteomics* 74: 920–934.
20 Zhang H, Ge Y (2011) Comprehensive analysis of protein modifications by top-down mass spectrometry. *Circ Cardiovasc Genet* 4: 711.
21 Nagel T, Meyer B (2014) Simultaneous characterization of sequence polymorphisms, glycosylation and phosphorylation of fibrinogen in a direct analysis by LC-MS. *Biochim Biophys Acta* 1844: 2284–2289.
22 Tran JC, Zamdborg L, Ahlf DR, Lee JE, Catherman AD, et al. (2011) Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480: 254–258.
23 Kellie JF, Tran JC, Lee JE, Ahlf DR, Thomas HM, et al. (2010) The emerging process of Top Down mass spectrometry for protein analysis: biomarkers, protein-therapeutics, and achieving high throughput. *Mol Biosyst* 6: 1532–1539.
24 Holland NT, Smith MT, Eskenazi B, Bastaki M (2003) Biological sample collection and processing for molecular epidemiological studies. *Mutat Res* 543: 217–234.
25 Mager SR, Oomen MH, Morente MM, Ratcliffe C, Knox K, et al. (2007) Standard operating procedure for the collection of fresh frozen tissue samples. *Eur J Cancer* 43: 828–834.
26 Morente MM, Mager R, Alonso S, Pezzella F, Spatz A, et al. (2006) TuBaFrost 2: standardising tissue collection and quality control procedures for a European virtual frozen tissue bank network. *Eur J Cancer* 42: 2684–2691.
27 Leyland-Jones BR, Ambrosone CB, Bartlett J, Ellis MJ, Enos RA, et al. (2008) Recommendations for collection and handling of specimens from group breast cancer clinical trials. *J Clin Oncol* 26: 5638–5644.
28 Fassbender A, Rahmioglu N, Vitonis AF, Vigano P, Giudice LC, et al. (2014) World Endometriosis Research Foundation Endometriosis Phenome and Biobanking Harmonisation Project: IV. Tissue collection, processing, and storage in endometriosis research. *Fertil Steril* 102: 1244–1253.
29 Silver GM, Klein MB, Herndon DN, Gamelli RL, Gibran NS, et al. (2007) Standard operating procedures for the clinical management of patients enrolled in a prospective study of Inflammation and the Host Response to Thermal Injury. *J Burn Care Res* 28: 222–230.
30 Diaz JI, Cazares LH, Semmes OJ (2008) Tissue sample collection for proteomics analysis. *Methods Mol Biol* 428: 43–53.
31 Alkhas A, Hood BL, Oliver K, Teng PN, Oliver J, et al. (2011) Standardization of a sample preparation and analytical workflow for proteomics of archival endometrial cancer tissue. *J Proteome Res* 10: 5264–5271.
32 Blackhall FH, Pintilie M, Wigle DA, Jurisica I, Liu N, et al. (2004) Stability and heterogeneity of expression profiles in lung cancer specimens harvested following surgical resection. *Neoplasia* 6: 761–767.
33 Dash A, Maine IP, Varambally S, Shen R, Chinnaiyan AM, et al. (2002) Changes in differential gene expression because of warm ischemia time of radical prostatectomy specimens. *Am J Pathol* 161: 1743–1748.
34 Spruessel A, Steimann G, Jung M, Lee SA, Carr T, et al. (2004) Tissue ischemia time affects gene and protein expression patterns within minutes following surgical tumor excision. *Biotechniques* 36: 1030–1037.
35 Schmitt M, Mengele K, Schueren E, Sweep FC, Foekens JA, et al. (2007) European Organisation for Research and Treatment of Cancer (EORTC) Pathobiology Group standard operating procedure for the preparation of human tumour tissue extracts suited for the quantitative analysis of tissue-associated biomarkers. *Eur J Cancer* 43: 835–844.
36 Burden D (2012) Guide to the Disruption of Biological Samples—2012. *Random Primers* (12), 1–25.
37 Hopkins T (2016) Cell Disrupters: A Review. http://www.biospec.com/ (accessed August 17, 2017).
38 Goldberg S (2008) Mechanical/physical methods of cell disruption and tissue homogenization. *Methods Mol Biol* 424: 3–22.
39 Botelho D, Wall MJ, Vieira DB, Fitzsimmons S, Liu F, et al. (2010) Top-down and bottom-up proteomics of SDS-containing solutions following mass-based separation. *J Proteome Res* 9: 2863–2870.
40 Zhou JY, Dann GP, Shi T, Wang L, Gao X, et al. (2012) Simple sodium dodecyl sulfate-assisted sample preparation method for LC-MS-based proteomics applications. *Anal Chem* 84: 2862–2867.
41 Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. *Nat Methods* 6: 359–362.
42 Shevchenko A, Tomas H, Havlis J, Olsen JV, Mann M (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* 1: 2856–2860.
43 Antharavally BS, Mallia KA, Rosenblatt MM, Salunkhe AM, Rogers JC, et al. (2011) Efficient removal of detergents from proteins and peptides in a spin column format. *Anal Biochem* 416: 39–44.
44 Bereman MS, Egertson JD, MacCoss MJ (2011) Comparison between procedures using SDS for

shotgun proteomic analyses of complex samples. *Proteomics* 11: 2931–2935.

45 Chang YH, Gregorich ZR, Chen AJ, Hwang L, Guner H, et al. (2015) New mass-spectrometry-compatible degradable surfactant for tissue proteomics. *J Proteome Res* 14: 1587–1599.

46 Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, et al. (1996) Laser capture microdissection. *Science* 274: 998–1001.

47 Espina V, Wulfkuhle JD, Calvert VS, VanMeter A, Zhou W, et al. (2006) Laser-capture microdissection. *Nat Protoc* 1: 586–603.

48 Mustafa D, Kros JM, Luider T (2008) Combining laser capture microdissection and proteomics techniques. *Methods Mol Biol* 428: 159–178.

49 Xu BJ (2010) Combining laser capture microdissection and proteomics: methodologies and clinical applications. *Proteomics Clin Appl* 4: 116–123.

50 Johann DJ, Mukherjee S, Prieto DA, Veenstra TD, Blonder J (2011) Profiling solid tumor heterogeneity by LCM and biological MS of fresh-frozen tissue sections. *Methods Mol Biol* 755: 95–106.

51 Xiao H, Langerman A, Zhang Y, Khalid O, Hu S, et al. (2015) Quantitative proteomic analysis of microdissected oral epithelium for cancer biomarker discovery. *Oral Oncol* 51: 1011–1019.

52 Chen C-L, Chung T, Wu C-C, Ng K-F, Yu J-S, et al. (2015) Comparative tissue proteomics of microdissected specimens reveals novel candidate biomarkers of bladder cancer. *Mol Cell Proteomics* 14: 2466–2478.

53 Liu P-F, Wang Y-H, Cao Y-W, Jiang H-P, Yang X-C, et al. (2014) Far from resolved: stromal cell-based iTRAQ research of muscle-invasive bladder cancer regarding heterogeneity. *Oncol Rep* 32: 1489–1496.

54 Mu Y, Chen Y, Zhang G, Zhan X, Li Y, et al. (2013) Identification of stromal differentially expressed proteins in the colon carcinoma by quantitative proteomics. *Electrophoresis* 34: 1679–1692.

55 Xu Y, Cao LQ, Jin LY, Chen ZC, Zeng GQ, et al. (2012) Quantitative proteomic study of human lung squamous carcinoma and normal bronchial epithelial acquired by laser capture microdissection. *J Biomed Biotechnol* 2012: 510418.

56 Zeng GQ, Zhang PF, Deng X, Yu FL, Li C, et al. (2012) Identification of candidate biomarkers for early detection of human lung squamous cell cancer by quantitative proteomics. *Mol Cell Proteomics* 11: M111.013946.

57 Chang KP, Yu JS, Chien KY, Lee CW, Liang Y, et al. (2011) Identification of PRDX4 and P4HA2 as metastasis-associated proteins in oral cavity squamous cell carcinoma by comparative tissue proteomics of microdissected specimens using iTRAQ technology. *J Proteome Res* 10: 4935–4947.

58 Xiao Z, Li G, Chen Y, Li M, Peng F, et al. (2010) Quantitative proteomic analysis of formalin-fixed and paraffin-embedded nasopharyngeal carcinoma using iTRAQ labeling, two-dimensional liquid chromatography, and tandem mass spectrometry. *J Histochem Cytochem* 58: 517–527.

59 Nakatani S, Wei M, Ishimura E, Kakehashi A, Mori K, et al. (2012) Proteome analysis of laser microdissected glomeruli from formalin-fixed paraffin-embedded kidneys of autopsies of diabetic patients: nephronectin is associated with the development of diabetic glomerulosclerosis. *Nephrol Dial Transplant* 27: 1889–1897.

60 Ly L, Barnett MH, Zheng YZ, Gulati T, Prineas JW, et al. (2011) Comprehensive tissue processing strategy for quantitative proteomics of formalin-fixed multiple sclerosis lesions. *J Proteome Res* 10: 4855–4868.

61 Okayama A, Miyagi Y, Oshita F, Nishi M, Nakamura Y, et al. (2014) Proteomic analysis of proteins related to prognosis of lung adenocarcinoma. *J Proteome Res* 13: 4686–4694.

62 Braakman RB, Tilanus-Linthorst MM, Liu NQ, Stingl C, Dekker LJ, et al. (2012) Optimized nLC-MS workflow for laser capture microdissected breast cancer tissue. *J Proteomics* 75: 2844–2854.

63 Liu NQ, Braakman RB, Stingl C, Luider TM, Martens JW, et al. (2012) Proteomics pipeline for biomarker discovery of laser capture microdissected breast cancer tissue. *J Mammary Gland Biol Neoplasia* 17: 155–164.

64 Liu NQ, Dekker LJ, Stingl C, Guzel C, De Marchi T, et al. (2013) Quantitative proteomic analysis of microdissected breast cancer tissues: comparison of label-free and SILAC-based quantification with shotgun, directed, and targeted MS approaches. *J Proteome Res* 12: 4627–4641.

65 Zhu J, Nie S, Wu J, Lubman DM (2013) Target proteomic profiling of frozen pancreatic CD24+ adenocarcinoma tissues by immuno-laser capture microdissection and nano-LC-MS/MS. *J Proteome Res* 12: 2791–2804.

66 Shapiro JP, Biswas S, Merchant AS, Satoskar A, Taslim C, et al. (2012) A quantitative proteomic workflow for characterization of frozen clinical biopsies: laser capture microdissection coupled with label-free mass spectrometry. *J Proteomics* 77: 433–440.

67 Maxwell GL, Hood BL, Day R, Chandran U, Kirchner D, et al. (2011) Proteomic analysis of stage I endometrial cancer tissue: identification of proteins associated with oxidative processes and inflammation. *Gynecol Oncol* 121: 586–594.

68 Hill JJ, Tremblay TL, Pen A, Li J, Robotham AC, et al. (2011) Identification of vascular breast tumor markers by laser capture microdissection and label-free LC-MS. *J Proteome Res* 10: 2479–2493.

69 Nishida Y, Aida K, Kihara M, Kobayashi T (2014) Antibody-validated proteins in inflamed islets of fulminant type 1 diabetes profiled by laser-capture microdissection followed by mass spectrometry. *PLoS One* 9: e107664.

70 O'Malley KJ, Eisermann K, Pascal LE, Parwani AV, Majima T, et al. (2014) Proteomic analysis of patient tissue reveals PSA protein in the stroma of benign prostatic hyperplasia. *Prostate* 74: 892–900.

71 Wisniewski JR, Dus K, Mann M (2013) Proteomic workflow for analysis of archival formalin-fixed and paraffin-embedded clinical samples to a depth of 10 000 proteins. *Proteomics Clin Appl* 7: 225–233.

72 Wisniewski JR, Ostasiewicz P, Dus K, Zielinska DF, Gnad F, et al. (2012) Extensive quantitative remodeling of the proteome between normal colon tissue and adenocarcinoma. *Mol Syst Biol* 8: 611.

73 Wisniewski JR, Ostasiewicz P, Mann M (2011) High recovery FASP applied to the proteomic analysis of microdissected formalin fixed paraffin embedded cancer tissues retrieves known colon cancer markers. *J Proteome Res* 10: 3040–3049.

74 Naidoo K, Jones R, Dmitrovic B, Wijesuriya N, Kocher H, et al. (2012) Proteome of formalin-fixed paraffin-embedded pancreatic ductal adenocarcinoma and lymph node metastases. *J Pathol* 226: 756–763.

75 Zhang ZQ, Li XJ, Liu GT, Xia Y, Zhang XY, et al. (2013) Identification of Annexin A1 protein expression in human gastric adenocarcinoma using proteomics and tissue microarray. *World J Gastroenterol* 19: 7795–7803.

76 Bohnenberger H, Strobel P, Mohr S, Corso J, Berg T, et al. (2015) Quantitative mass spectrometric profiling of cancer-cell proteomes derived from liquid and solid tumors. *J Vis Exp* (96): e52435.

77 Silvestri A, Calvert V, Belluco C, Lipsky M, De Maria R, et al. (2013) Protein pathway activation mapping of colorectal metastatic progression reveals metastasis-specific network alterations. *Clin Exp Metastasis* 30: 309–316.

78 Chung JY, Hewitt SM (2015) A well-based reverse-phase protein array of formalin-fixed paraffin-embedded tissue. *Methods Mol Biol* 1312: 129–139.

79 Ichikawa H, Kanda T, Kosugi S, Kawachi Y, Sasaki H, et al. (2013) Laser microdissection and two-dimensional difference gel electrophoresis reveal the role of a novel macrophage-capping protein in lymph node metastasis in gastric cancer. *J Proteome Res* 12: 3780–3791.

80 Yang J, Zhou M, Zhao R, Peng S, Luo Z, et al. (2014) Identification of candidate biomarkers for the early detection of nasopharyngeal carcinoma by quantitative proteomic analysis. *J Proteomics* 109: 162–175.

81 Sugihara Y, Taniguchi H, Kushima R, Tsuda H, Kubota D, et al. (2013) Laser microdissection and two-dimensional difference gel electrophoresis reveal proteomic intra-tumor heterogeneity in colorectal cancer. *J Proteomics* 78: 134–147.

82 Shi H, Hood KA, Hayes MT, Stubbs RS (2011) Proteomic analysis of advanced colorectal cancer by laser capture microdissection and two-dimensional difference gel electrophoresis. *J Proteomics* 75: 339–351.

83 Skvortsov S, Schafer G, Stasyk T, Fuchsberger C, Bonn GK, et al. (2011) Proteomics profiling of microdissected low- and high-grade prostate tumors identifies Lamin A as a discriminatory biomarker. *J Proteome Res* 10: 259–268.

84 Liu YF, Chen YH, Li MY, Zhang PF, Peng F, et al. (2012) Quantitative proteomic analysis identifying three annexins as lymph node metastasis-related proteins in lung adenocarcinoma. *Med Oncol* 29: 174–184.

85 Switzar L, Giera M, Niessen WM (2013) Protein digestion: an overview of the available techniques and recent developments. *J Proteome Res* 12: 1067–1077.

86 Compton PD, Zamdborg L, Thomas PM, Kelleher NL (2011) On the scalability and requirements of whole protein mass spectrometry. *Anal Chem* 83: 6868–6874.

87 Wisniewski JR, Mann M (2012) Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. *Anal Chem* 84: 2631–2637.

88 Swaney DL, Wenger CD, Coon JJ (2010) Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res* 9: 1323–1329.

89 Choudhary G, Wu SL, Shieh P, Hancock WS (2003) Multiple enzymatic digestion for enhanced sequence coverage of proteins in complex proteomic mixtures using capillary LC with ion trap MS/MS. *J Proteome Res* 2: 59–67.

90 Medzihradszky KF (2005) In-solution digestion of proteins for mass spectrometry. *Methods Enzymol* 405: 50–65.

91 Zielinska DF, Gnad F, Wisniewski JR, Mann M (2010) Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell* 141: 897–907.

92 Ostasiewicz P, Zielinska DF, Mann M, Wisniewski JR (2010) Proteome, phosphoproteome, and N-glycoproteome are quantitatively preserved in formalin-fixed paraffin-embedded tissue and analyzable by high-resolution mass spectrometry. *J Proteome Res* 9: 3688–3700.

93 Wisniewski JR (2016) Quantitative evaluation of Filter Aided Sample Preparation (FASP) and multienzyme digestion FASP protocols. *Anal Chem* 88: 5438–5443.

94 Leon IR, Schwammle V, Jensen ON, Sprenger RR (2013) Quantitative assessment of in-solution digestion efficiency identifies optimal protocols for unbiased protein analysis. *Mol Cell Proteomics* 12: 2992–3005.

95 Tanca A, Abbondio M, Pisanu S, Pagnozzi D, Uzzau S, et al. (2014) Critical comparison of sample preparation strategies for shotgun proteomic analysis of formalin-fixed, paraffin-embedded samples: insights from liver tissue. *Clin Proteomics* 11: 28.

96 Quesada-Calvo F, Bertrand V, Longuespee R, Delga A, Mazzucchelli G, et al. (2015) Comparison of two FFPE preparation methods using label-free shotgun proteomics: application to tissues of diverticulitis patients. *J Proteomics* 112: 250–261.

97 Speicher D (2007) *Overview of Proteome Analysis, in Proteome Analysis Interpreting the Genome*. Elsevier, Amsterdam.

98 Fournier ML, Gilmore JM, Martin-Brown SA, Washburn MP (2007) Multidimensional separations-based shotgun proteomics. *Chem Rev* 107: 3654–3686.

99 Scherp P, Ku G, Coleman L, Kheterpal I (2011) Gel-based and gel-free proteomic technologies. *Methods Mol Biol* 702: 163–190.

100 Eriksson J, Fenyo D (2007) Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs. *Nat Biotechnol* 25: 651–655.

101 Svensson M, Boren M, Skold K, Falth M, Sjogren B, et al. (2009) Heat stabilization of the tissue proteome: a new technology for improved proteomics. *J Proteome Res* 8: 974–981.

102 Gatto L, Vizcaino JA, Hermjakob H, Huber W, Lilley KS (2010) Organelle proteomics experimental designs and analysis. *Proteomics* 10: 3957–3969.

103 Korfali N, Wilkie GS, Swanson SK, Srsen V, de Las Heras J, et al. (2012) The nuclear envelope proteome differs notably between tissues. *Nucleus* 3: 552–564.

104 Albrethsen J, Knol JC, Piersma SR, Pham TV, de Wit M, et al. (2010) Subnuclear proteomics in colorectal cancer: identification of proteins enriched in the nuclear matrix fraction and regulation in adenoma to carcinoma progression. *Mol Cell Proteomics* 9: 988–1005.

105 Taylor SW, Fahy E, Zhang B, Glenn GM, Warnock DE, et al. (2003) Characterization of the human heart mitochondrial proteome. *Nat Biotechnol* 21: 281–286.

106 Johnson DT, Harris RA, French S, Blair PV, You J, et al. (2007) Tissue heterogeneity of the mammalian mitochondrial proteome. *Am J Physiol Cell Physiol* 292: C689–697.

107 Nielsen PA, Olsen JV, Podtelejnikov AV, Andersen JR, Mann M, et al. (2005) Proteomic mapping of brain plasma membrane proteins. *Mol Cell Proteomics* 4: 402–408.

108 Smolders K, Lombaert N, Valkenborg D, Baggerman G, Arckens L (2015) An effective plasma membrane proteomics approach for small tissue samples. *Sci Rep* 5: 10917.

109 Cox B, Emili A (2006) Tissue subcellular fractionation and protein extraction for use in mass-spectrometry-based proteomics. *Nat Protoc* 1: 1872–1878.

110 Drissi R, Dubois ML, Boisvert FM (2013) Proteomics methods for subcellular proteome analysis. *FEBS J.* 280: 5626–34.

111 Dreger M (2003) Subcellular proteomics. *Mass Spectrom Rev* 22: 27–56.

112 Lee YH, Tan HT, Chung MC (2010) Subcellular fractionation methods and strategies for proteomics. *Proteomics* 10: 3935–3956.

113 Bunger S, Roblick UJ, Habermann JK (2009) Comparison of five commercial extraction kits for subsequent membrane protein profiling. *Cytotechnology* 61: 153–159.

114 Rockstroh M, Müller SA, Jende C, Kerzhner A, von Bergen M, et al. (2011) Cell fractionation—an important tool for compartment proteomics *J Integrated Omics* 1: 135–143.

115 Jafari M, Primo V, Smejkal GB, Moskovets EV, Kuo WP, et al. (2012) Comparison of in-gel protein separation techniques commonly used for fractionation in mass spectrometry-based proteomic profiling. *Electrophoresis* 33: 2516–2526.

116 Atanassov I, Urlaub H (2013) Increased proteome coverage by combining PAGE and peptide isoelectric focusing: comparative study of gel-based separation approaches. *Proteomics* 13: 2947–2955.

117 Gallagher SR (2012) One-dimensional SDS gel electrophoresis of proteins. *Curr Protoc Mol Biol* Chapter 10: Unit 10 12A.

118 Hoogland C, Mostaguir K, Sanchez JC, Hochstrasser DF, Appel RD (2004) SWISS-2DPAGE, ten years later. *Proteomics* 4: 2352–2356.

119 Giorgianni F, Desiderio DM, Beranova-Giorgianni S (2003) Proteome analysis using isoelectric focusing in immobilized pH gradient gels followed by mass spectrometry. *Electrophoresis* 24: 253–259.

120 Cargile BJ, Sevinsky JR, Essader AS, Stephenson JL, Jr., Bundy JL (2005) Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics. *J Biomol Technol* 16: 181–189.

121 Essader AS, Cargile BJ, Bundy JL, Stephenson JL, Jr. (2005) A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. *Proteomics* 5: 24–34.

122 Wang Y, Balgley BM, Lee CS (2005) Tissue proteomics using capillary isoelectric focusing-based multidimensional separations. *Expert Rev Proteomics* 2: 659–667.

123 Guo T, Lee CS, Wang W, DeVoe DL, Balgley BM (2006) Capillary separations enabling tissue proteomics-based biomarker discovery. *Electrophoresis* 27: 3523–3532.

124 Wang Y, Rudnick PA, Evans EL, Li J, Zhuang Z, et al. (2005) Proteome analysis of microdissected tumor tissue using a capillary isoelectric focusing-based multidimensional separation platform coupled with ESI-tandem MS. *Anal Chem* 77: 6549–6556.

125 Fang X, Balgley BM, Wang W, Park DM, Lee CS (2009) Comparison of multidimensional shotgun technologies targeting tissue proteomics. *Electrophoresis* 30: 4063–4070.

126 Wolters DA, Washburn MP, Yates JR, 3rd (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem* 73: 5683–5690.

127 Florens L, Washburn MP (2006) Proteomic analysis by multidimensional protein identification technology. *Methods Mol Biol* 328: 159–175.

128 Kislinger T, Rahman K, Radulovic D, Cox B, Rossant J, et al. (2003) PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol Cell Proteomics* 2: 96–106.

129 Cagney G, Park S, Chung C, Tong B, O'Dushlaine C, et al. (2005) Human tissue profiling with multidimensional protein identification technology. *J Proteome Res* 4: 1757–1767.

130 Ralhan R, Desouza LV, Matta A, Chandra Tripathi S, Ghanny S, et al. (2009) iTRAQ-multidimensional liquid chromatography and tandem mass spectrometry-based identification of potential biomarkers of oral epithelial dysplasia and novel networks between inflammation and premalignancy. *J Proteome Res* 8: 300–309.

131 Bakalarski CE, Elias JE, Villen J, Haas W, Gerber SA, et al. (2008) The impact of peptide abundance and dynamic range on stable-isotope-based quantitative proteomic analyses. *J Proteome Res* 7: 4756–4765.

132 Kume H, Muraoka S, Kuga T, Adachi J, Narumi R, et al. (2014) Discovery of colorectal cancer biomarker candidates by membrane proteomic analysis and subsequent verification using selected reaction monitoring (SRM) and tissue microarray (TMA) analysis. *Mol Cell Proteomics* 13: 1471–1484.

133 Liu NQ, Stingl C, Look MP, Smid M, Braakman RB, et al. (2014) Comparative proteome analysis revealing an 11-protein signature for aggressive triple-negative breast cancer. *J Natl Cancer Inst* 106: djt376.

134 Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, Damoc E, et al. (2009) A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics* 8: 2759–2769.

135 Zhang B, Wang J, Wang X, Zhu J, Liu Q, et al. (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature* 513: 382–387.

136 Michalski A, Damoc E, Hauschild JP, Lange O, Wieghaus A, et al. (2011) Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics* 10: M111.011015.

137 Sun L, Zhu G, Dovichi NJ (2013) Comparison of the LTQ-Orbitrap Velos and the Q-Exactive for proteomic analysis of 1-1000 ng RAW 264.7 cell lysate digests. *Rapid Commun Mass Spectrom* 27: 157–162.

138 Welinder C, Pawlowski K, Sugihara Y, Yakovleva M, Jonsson G, et al. (2015) A protein deep sequencing evaluation of metastatic melanoma tissues. *PLoS One* 10: e0123661.

139 Noberini R, Uggetti A, Pruneri G, Minucci S, Bonaldi T (2015) Pathology tissue-quantitative mass spectrometry analysis to profile histone post-translational modification patterns in patient samples. *Mol Cell Proteomics* 15(3): 866–877.

140 Elliott MH, Smith DS, Parker CE, Borchers C (2009) Current trends in quantitative proteomics. *J Mass Spectrom* 44: 1637–1660.

141 Geiger T, Cox J, Ostasiewicz P, Wisniewski JR, Mann M (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat Methods* 7: 383–385.

142 Zhang W, Wei Y, Ignatchenko V, Li L, Sakashita S, et al. (2014) Proteomic profiles of human lung adeno and squamous cell carcinoma using super-SILAC and label-free quantification approaches. *Proteomics* 14: 795–803.

143 Shenoy A, Geiger T (2015) Super-SILAC: current trends and future perspectives. *Expert Rev Proteomics* 12: 13–19.

144 Lu P, Vogel C, Wang R, Yao X, Marcotte EM (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat Biotechnol* 25: 117–124.

145 Listgarten J, Emili A (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* 4: 419–434.

146 Callister SJ, Barry RC, Adkins JN, Johnson ET, Qian WJ, et al. (2006) Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics. *J Proteome Res* 5: 277–286.

147 Griffin NM, Yu J, Long F, Oh P, Shore S, et al. (2010) Label-free, normalized quantification of complex

mass spectrometry data for proteomic analysis. *Nat Biotechnol* 28: 83–89.

148 Latosinska A, Vougas K, Makridakis M, Klein J, Mullen W, et al. (2015) Comparative analysis of label-free and 8-Plex iTRAQ approach for quantitative tissue proteomic analysis. *PLoS One* 10(9): e0137048.

149 Kato M, Wei M, Yamano S, Kakehashi A, Tamada S, et al. (2012) DDX39 acts as a suppressor of invasion for bladder cancer. *Cancer Sci* 103: 1363–1369.

150 Li Z, Adams RM, Chourey K, Hurst GB, Hettich RL, et al. (2012) Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ Orbitrap Velos. *J Proteome Res* 11: 1582–1590.

151 Patel VJ, Thalassinos K, Slade SE, Connolly JB, Crombie A, et al. (2009) A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *J Proteome Res* 8: 3752–3759.

152 Sjodin MO, Wetterhall M, Kultima K, Artemenko K (2013) Comparative study of label and label-free techniques using shotgun proteomics for relative protein quantification. *J Chromatogr B Analyt Technol Biomed Life Sci* 928: 83–92.

153 Trinh HV, Grossmann J, Gehrig P, Roschitzki B, Schlapbach R, et al. (2013) iTRAQ-based and label-free proteomics approaches for studies of human adenovirus infections. *Int J Proteomics* 2013: 581862.

154 Wang H, Alvarez S, Hicks LM (2012) Comprehensive comparison of iTRAQ and label-free LC-based quantitative proteomics approaches using two *Chlamydomonas reinhardtii* strains of interest for biofuels engineering. *J Proteome Res* 11: 487–501.

155 Carnielli CM, Winck FV, Paes Leme AF (2015) Functional annotation and biological interpretation of proteomics data. *Biochim Biophys Acta* 1854: 46–54.

156 Malik R, Dulla K, Nigg EA, Korner R (2010) From proteome lists to biological impact—tools and strategies for the analysis of large MS data sets. *Proteomics* 10: 1270–1283.

157 Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, et al. (2015) The GOA database: gene Ontology annotation updates for 2015. *Nucleic Acids Res* 43: D1057–1063.

158 UniProt C (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43: D204–212.

159 Gaudet P, Michel PA, Zahn-Zabal M, Cusin I, Duek PD, et al. (2015) The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res* 43: D764–770.

160 Uhlen M, Bjorling E, Agaton C, Szigyarto CA, Amini B, et al. (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol Cell Proteomics* 4: 1920–1932.

161 Carbon S, Ireland A, Mungall CJ, Shu S, Marshall B, et al. (2009) AmiGO: online access to ontology and annotation data. *Bioinformatics* 25: 288–289.

162 Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, et al. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol* 4: R28.

163 Wang J, Duncan D, Shi Z, Zhang B (2013) WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res* 41: W77–83.

164 Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10: 48.

165 Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, et al. (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* 41: D808–815.

166 Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, et al. (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* 42: D358–363.

167 Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, et al. (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43: D470–478.

168 Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, et al. (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res* 40: D857–861.

169 Kamburov A, Cavill R, Ebbels TM, Herwig R, Keun HC (2011) Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* 27: 2917–2918.

170 Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30.

171 Croft D, Mundo AF, Haw R, Milacic M, Weiser J, et al. (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res* 42: D472–477.

172 Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13: 2498–2504.

173 Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 4: 44–57.

174 Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 37: 1–13.

175 Alonso R, Salavert F, Garcia-Garcia F, Carbonell-Caballero J, Bleda M, et al. (2015) Babelomics 5.0:

functional interpretation for new generations of genomic data. *Nucleic Acids Res* 43: W117–121.

176 Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res* 33: D284–288.

177 Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, et al. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14: 128.

178 Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, et al. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21: 3674–3676.

179 Berriz GF, Roth FP (2008) The Synergizer service for translating gene, protein and other biological identifiers. *Bioinformatics* 24: 2272–2273.

180 Wein SP, Cote RG, Dumousseau M, Reisinger F, Hermjakob H, et al. (2012) Improvements in the protein identifier cross-reference service. *Nucleic Acids Res* 40: W276–280.

181 Huang H, McGarvey PB, Suzek BE, Mazumder R, Zhang J, et al. (2011) A comprehensive protein-centric ID mapping service for molecular data integration. *Bioinformatics* 27: 1190–1191.

182 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29.

183 Bader GD, Cary MP, Sander C (2006) Pathguide: a pathway resource list. *Nucleic Acids Res* 34: D504–506.

184 Bhat A, Dakna M, Mischak H (2015) Integrating proteomics profiling data sets: a network perspective. *Methods Mol Biol* 1243: 237–253.

185 Niu HT, Zhang YB, Jiang HP, Cheng B, Sun G, et al. (2009) Differences in shotgun protein expression profile between superficial bladder transitional cell carcinoma and normal urothelium. *Urol Oncol* 27: 400–406.

186 Niu HT, Yang CM, Jiang G, Xu T, Cao YW, et al. (2011) Cancer stroma proteome expression profile of superficial bladder transitional cell carcinoma and biomarker discovery. *J Cancer Res Clin Oncol* 137: 1273–1282.

187 Niu H, Jiang H, Cheng B, Li X, Dong Q, et al. (2012) Stromal proteome expression profile and muscle-invasive bladder cancer research. *Cancer Cell International* 12(1): 39.

188 Niu HT, Qi XJ, Liu YQ, Cao YW, Dong Q, et al. (2013) Parallel proteomic analysis in muscle-invasive bladder transitional cell carcinoma and cancer-related stroma. *Genetics Mol Res* 12: 4251–4263.

189 Liu PF, Cao YW, Jiang HP, Wang YH, Yang XC, et al. (2014) Heterogeneity research in muscle-invasive bladder cancer based on differential protein expression analysis. *Med Oncol* 31(9): 21.

190 Oezdemir RF, Gaisa NT, Lindemann-Docter K, Gostek S, Weiskirchen R, et al. (2012) Proteomic tissue profiling for the improvement of grading of noninvasive papillary urothelial neoplasia. *Clin Biochem* 45: 7–11.

191 Niu HT, Dong Z, Jiang G, Xu T, Liu YQ, et al. (2011) Proteomics research on muscle-invasive bladder transitional cell carcinoma. *Cancer Cell International* 11(1): 17.

192 Niu HT, Yang CM, Chen B, Dong Q (2011) Biomarker research and some deduction in superficial bladder cancer cells combined with corresponding stroma. *Cancer Biomarkers* 10: 109–116.

9

Tissue MALDI Imaging

Andrew Smith[1], Niccolò Mosele[1], Vincenzo L'Imperio[2,3], Fabio Pagni[2,3], and Fulvio Magni[1]

[1] *Department of Medicine and Surgery, Proteomics and Metabolomics Unit, University of Milano-Bicocca, Monza, Italy*
[2] *Department of Medicine and Surgery, Pathology, University of Milano-Bicocca, San Gerardo Hospital, Monza, Italy*
[3] *Nephropathology Center, University of Milano-Bicocca, San Gerardo Hospital, Monza, Italy*

9.1 Introduction

Mass spectrometry (MS) is one of the most important tools for the characterization and identification of a wide range of biomolecules, including metabolites, lipids, and proteins. The study of such molecules constitutes the major -omics disciplines studied using MS techniques. In MS, analyte molecules are first ionized in the source and can be present in solid, liquid, or gaseous form, depending upon the type of ion source employed. The ionized analytes are then separated in the mass analyzer according to their physical properties, with the corresponding electrical signals then recorded by a detector. These detected signals are correlated with a particular mass-to-charge ratio (m/z). Results are then displayed in the form of a mass spectrum, with the relative intensity of each signal presented as a function of its m/z.

The concept of mass spectrometry imaging (MSI) first came to the fore nearly 50 years ago, representing a technique suitable for the analysis of elements and other small molecules. In first instances, MSI instruments employed secondary ion mass spectrometry (SIMS) technology, which was then shortly followed by the laser microprobe analyzer. Both of these techniques were capable of performing high spatial resolution surface analysis of small organic and inorganic molecules. However, it was not until the late 1990s when the research of Richard Caprioli and coworkers led to the introduction of MSI into a clinical setting, employing matrix-assisted laser desorption/ionization (MALDI) as a means of analyzing a wider range of biomolecules, including proteins, directly on intact tissue [1]. In this early body of work, Caprioli et al. were able to demonstrate the ability of MSI to localize the distribution of biomarkers within tissue, without the need for labeling. This early research subsequently led to an explosion of MSI-based studies, having a substantial impact on clinical and pharmacological research.

Currently, there are a number of MSI techniques employed in clinical studies, including SIMS, desorption electrospray ionization (DESI), laser ablation electrospray ionization (LAESI), and rapid evaporative ionization mass spectrometry (REIMS) (Table 9.1). In addition to this, there are a number of newly emerging MSI techniques that have shown promise for employment in this field of research, including liquid junction surface sampling and mass cytometry [2]. However, as a result of its widespread availability, ability to analyze proteins, and numerous other advantages (Table 9.2), MALDI remains the most commonly applied MSI technique. Given that proteins play a significant role in a large number of pathways involved in defective cellular signaling cascades, the ability to spatially resolve the localization of a number of proteins concurrently within the same section of pathological tissue can enable the detection of pathological processes and, ultimately, disease candidates. Additionally, it has also become increasingly common for lipids and metabolites to be analyzed in order to study disease mechanisms and provide complementary information that can be integrated with proteomic findings. Since its inception, MALDI-MSI has been used in a plethora of clinical-based studies, covering the fields of oncology, pathology, diagnostics, and surgery [3]. Furthermore, it has been regularly used to monitor the distribution of xenobiotics and their metabolites, establishing itself as an invaluable tool in drug distribution studies [4]. This chapter will focus on the methodological aspects underpinning on-tissue MALDI-MSI and proceed to discuss its application and relevance in clinical-based studies.

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

Table 9.1 An overview of the most commonly used ionization sources for MSI experiments.

	Ionization	Pretreatment	Analyte class	Mass range	Spatial resolution
MALDI	Laser ablation of the surface sample and desorption/ionization of analytes	Coating of the sample with a MALDI matrix solution	Metabolites, lipids, peptides, or proteins (matrix dependent)	1 Da to 500 kDa	<5–10 μm (commercial instruments)
SIMS	The sample surface is sputtered with a primary ion beam, generating secondary ions	Not required. However, a matrix/metal coating can be used to increase the yield of generated ions	Static SIMS: elements, fatty acids, and lipids	1 Da to 10 kDa	Static SIMS: >1 μm
			Dynamic SIMS: elements		Dynamic SIMS: <1 μm
DESI	Droplets are generated via an electrospray mechanism and directed toward the surface sample	None	Small metabolites (from tissue)	1 Da to 2 kDa	>100 μm
Nano-DESI	A liquid bridge samples the surface molecules which are then ionized by nano-ESI	None	Metabolites, peptides, and proteins (solvent combination dependent)	1 Da to 2 kDa	<10 μm (dependent upon the size of liquid bridge)
LAESI	Generation of gas-phase particles by laser ablation and ionized by electrospray	None	Metabolites, peptides, and proteins	50 Da to 100 kDa	MSI: <200 μm
					Cell-by-cell analysis: <50 μm
REIMS	Rapid evaporation of analyte molecules generating gas-phase ions	None	Metabolites and lipids	150 Da to 2 kDa	<500 μm
Mass cytometry	Functionalized antibodies bound to polymers containing lanthanide	Addition of functionalized antibodies	Proteins	50 Da to 250 Da	<1 μm (when coupled with a SIMS instrument)

Table 9.2 Advantages and disadvantages of MALDI-MSI.

Advantages	Disadvantages
High throughput	No intracellular information
Label-free	Matrix application can be tissue dependent, yielding variable results
Relative quantitation	Difficulty in identifying peptides directly on tissue
Applicable to a wide range of analytes	
Maintains morphological structure	
Maintains spatial localization of analytes	
Well established	

9.1.1 MALDI-MSI: General Principles

MALDI-MSI was formally introduced in 1997 by Richard Caprioli, and its use has increased exponentially in recent years to become the most widely employed MSI technique. This technique relies on the use of a MALDI matrix, which consists of small organic molecules that are designed to absorb the energy of a pulsed laser beam. These molecules commonly possess a suitable chromophore, usually in the form of an aromatic core, and it is this property of the matrix that facilitates the absorption of the UV laser energy. When this matrix is applied to the surface of a sample, it promotes the formation of a ubiquitous layer of co-crystals, which incorporates both matrix and analyte molecules in its network. This co-crystallization process, which occurs on the surface of the sample, is characterized by significant variability and is related to a number of different parameters including the choice of solvent, time of incubation, and matrix concentration [5]. When the laser beam is applied to the surface of the sample, the absorbed energy leads to rapid desorption of both the matrix and analyte crystals and subsequent ionization (Figure 9.1). This ionization process is similar to electrospray ionization (ESI) in the aspect that both techniques are capable of generating large gaseous phase ions without extensive fragmentation occurring during the procedure. This is termed "soft" ionization. The most significant difference between MALDI and ESI is that MALDI produces far fewer multiply charged ions, leading to less complex spectra, which are, ultimately, easier to interpret.

MALDI sources can be combined with a wide array of mass analyzers, including time of flight (TOF) and Fourier transfer ion cyclotron resonance (FT-ICR),

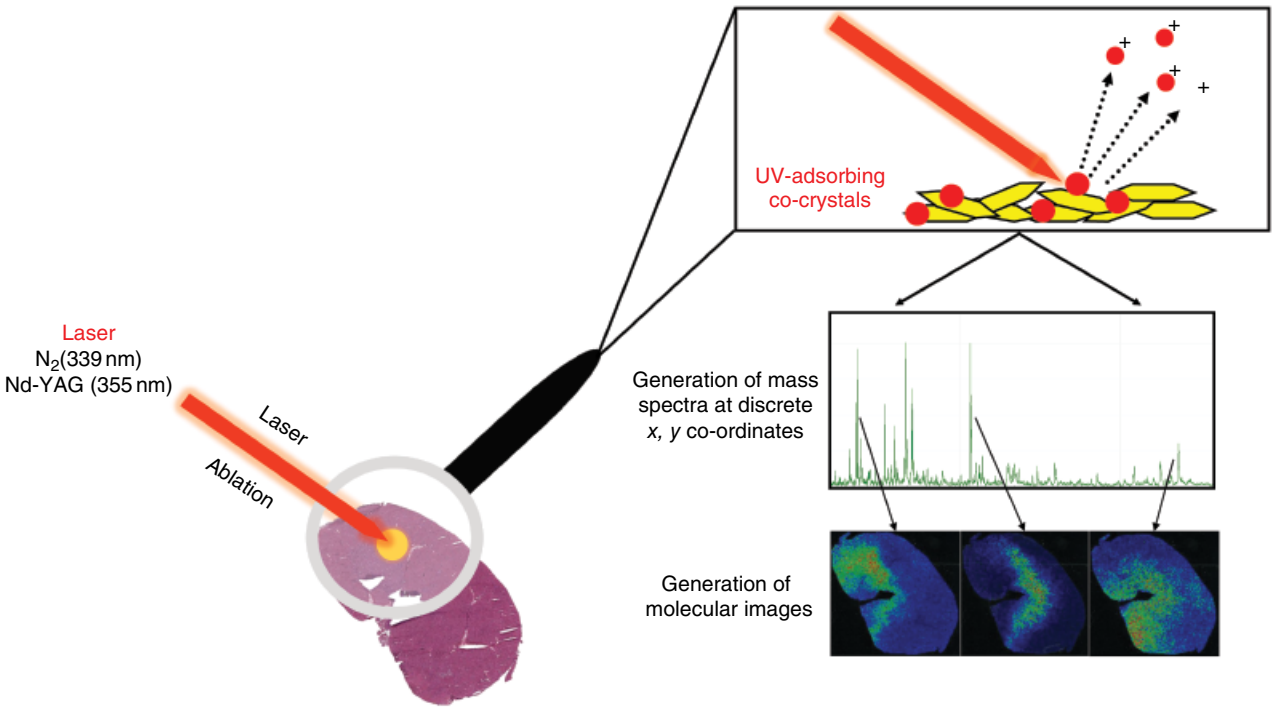


Figure 9.1 The general principles of MALDI-MSI. Laser ablation leads to the desorption and ionization of matrix and analyte ions. The detected ions yield the generation of mass spectra at discrete spatial coordinates and the spatial distribution of any of the ions present in these spectra can be visualized following the generation of a molecular image.

which are commonly employed for the analysis of intact proteins due to the wide mass range that they cover. Alternatively, multiple-stage quadrupole TOF and various forms of ion traps can be used for the analysis of smaller molecules, including metabolites, lipids, and peptides. However, coupling MALDI with TOF mass analyzers appears to be the most common approach for MALDI-MSI studies. This combination enables the analysis over a large mass range (50 Da to more than 150 kDa), the spatial resolutions higher than 20 µm, and the possibility of performing MS/MS experiments directly on tissue when TOF/TOF is employed. The ability to identify proteins of interest, by utilizing MS/MS fragmentation directly on tissue, is of paramount importance when considering the progression of this MSI technique in terms of facilitating the translation of these findings into tests that are suitable for use in a routine clinical setting.

In MALDI imaging, a mass spectrum is acquired at each desired *x,y* coordinate within a defined measurement region, which is usually related to an entire section or particular regions of interest (ROIs) present within a tissue section. Using the acquired mass spectra, the spatial distribution of any of the biomolecules present can be visualized and a molecular image of the tissue reconstructed. These molecular images can then be correlated with tissue images obtained using histological techniques. The distance between spectral acquisitions in MALDI-MSI analysis is referred the *rastering*, which is inversely proportional to the *spatial resolution* (i.e., the smaller the distance between the two raster positions, the higher the spatial resolution). MALDI-TOF instruments are capable of high-throughput MSI analysis with spatial resolutions higher than 20 µm. Although other instruments, such as TOF-SIMS, are capable of acquiring images with a higher spatial resolution (as high as 1.5 µm), they are unable to do so in the same high-throughput manner and the mass range is more limited in comparison.

Notwithstanding the rapid evolution of MALDI-MS instrumentation and sample preparation protocols [6], several technical issues related to MALDI-MSI still need to be improved, such as increased spatial resolution and sensitivity. However, next-generation instruments are beginning to address these limiting factors, not only improving spatial resolution and sensitivity but also increasing the spectral acquisition rate as well as minimizing pixel-to-pixel variability, facilitating higher quality and more robust analysis. Perhaps of greater importance is the imaging and visualization of single cells, and, in fact, when using the correct cell fixation protocols and a laser with a smaller diameter (<7 µm), this has already been shown to be possible with currently available MALDI-MSI instrumentation [7]. Continuing in

this vein, MALDI-MSI will be able to not only analyze single cells, but also potentially delve deeper and offer insights at a subcellular level. Furthermore, it will also be possible to routinely generate three-dimensional (3D) MALDI images in order to obtain a snapshot of the pathological state of an entire organ by combining MALDI-MS images of consecutive tissue sections and reconstructing a 3D representation using the appropriate (and currently available) software [8].

9.2 Experimental Procedures

9.2.1 Sample Handling: Storage, Embedding, and Sectioning

Sample handling is arguably the most critical aspect for obtaining satisfactory results from MALDI-MSI experiments, with sample collection, storage, embedding, and sectioning all to be carefully considered. The first challenging aspect is related to how the sample is treated following collection. At this initial phase, it is imperative that protein degradation is minimized and the analyte molecules are stabilized in a consistent manner. This ensures that chemical integrity of biomolecules and spatial organization of tissue structure are maintained.

Fresh samples represent the primary source of tissue for MALDI-MSI experiments and are routinely collected for this type of analysis. However, fresh samples need to be frozen directly after collection in order to stabilize the proteome by inhibiting enzymatic proteolysis. The major advantage of using fresh-frozen (FF) tissue is that it closely mimics the native state of the tissue, preserving its morphology and integrity. The freezing process here must be gentle and homogeneous in order to avoid different parts of the tissue from cooling at different rates, which can lead to the formation of ice crystals and, ultimately, tissue cracking. The most common approach involves loosely wrapping the tissue in aluminum foil and freezing in liquid nitrogen or cooled alcohol (to ~ −70°C) for approximately 1 min. Alternatively, the tissue can also be cooled in isopentane dry ice. One final solution to avoid protein degradation can be through the use of conductive heat transfer. However, it is important to check the compatibility of each tissue with this treatment, as, in some cases, tissue morphology can be altered during the process. Once stabilization has been performed, the tissue is stored at −80°C prior to MALDI-MSI analysis.

More recently, and of perhaps greater importance to employing MALDI-MSI in a clinical setting, protocols have been developed in order to facilitate the analysis of formalin-fixed paraffin-embedded (FFPE) tissue, which represents a large percentage of the patient samples collected and stored in hospitals and other medical centers, thus representing potential gold mines of information for histopathological studies involving MALDI-MSI [9]. Ultimately, the analysis of FFPE tissue enables retrospective studies with much larger cohorts of patients. This can be of particular importance when attempting to collect samples of particularly rare diseases, which would take a considerably longer period of time if attempting to obtain an appropriate number using FF specimens. In terms of sample storage, FFPE tissue can also be stored for up to 10 years at room temperature (RT).

Upon treatment with formalin, a reaction occurs between formaldehyde and the amine groups of the tissue proteins, promoting the formation of methylene bridges between amino acids. This ultimately leads to the formation of inter- and intramolecular cross-links in proteins. While at RT, lipids, nucleic acids, and molecules not cross-linked to formaldehyde can degrade, the formed cross-links promote the stability of the proteins present in the tissue and preserve the morphological structures present. However, as a result of this extensive cross-linking, specific sample preparation steps are required in order to facilitate the MALDI-MSI analysis of FFPE samples. Firstly, the paraffin in which the sample is embedded needs to be removed as it acts as a strong ion suppressant. Secondly, antigen retrieval is required in order to unmask the epitopes in order to allow for the enzymatic digestion to occur, most commonly using trypsin. This protein digestion can be completed *in situ* using a number of automated devices, which keep the localization of the peptides intact and produce highly reproducible results. The peptides generated from this enzymatic digestion can be submitted not only for MALDI-MSI analysis but also for direct MS/MS analysis when using a TOF/TOF instrument, facilitating protein identification.

The sectioning of FFPE samples is usually performed at RT, with the thickness of the sections commonly being 5 µm. Sections are then transferred into a water bath at 37°C and mounted onto a conductive surface compatible with MALDI-MSI instruments, such as indium titanium oxide (ITO) glass slides. Once the sections are mounted, excess water is removed and the slides are gently warmed at 30–37°C for a few minutes in order to ensure proper adhesion. Following this step, the sections mounted onto the ITO glasses can be stored at RT for up to 10 years prior to the MALDI-MSI preparation process; however it is preferable to analyze within the first 2 years [10].

On the contrary, the sectioning of FF tissue is usually performed in a cryostat set to approximately −20°C. Additionally, the tissue requires an embedding medium in order to ensure proper attachment of the tissue to the stage of the cryostat, improve the ease of cutting, and avoid tissue damage. However, embedding the tissue

entirely in this medium should be avoided because it leads to ion suppression. Therefore, it is recommended to add only a very small amount of the cutting medium to the bottom of the organ in order for it to adhere correctly to the stage, while the major proportion of the tissue remains uncovered for the purpose of sectioning for MALDI-MSI experiments. Optimal cutting temperature (OCT) compound is the most common polymer used in standard histological applications, but, like other polymer-based embedding media, it has been shown to be a strong ion suppressant [11]. Therefore, once the FF tissue sections have been cut and mounted onto glass slides, any remaining OCT that surrounds on and around the tissue has to be carefully and efficiently removed. Alternatively, the sample can also be embedded in gelatin or in 15% poly[*N*-(2-hydroxypropyl)methacrylamide] (pHPMA) [12], with both of these compounds being reportedly more compatible with MALDI-MSI analysis. Once the section has been mounted onto the glass slide, it is also recommended to assist the adhesion of the tissue to the slide by placing a finger on the back of the slide, where the tissue section is mounted, and this is known as *thaw mounting*.

The thickness of the section is often set to between 10 and 12 µm, representing a balance between tissue conductivity and robustness. For example, the rate of moisture evaporation is accelerated when the thickness of the section is reduced. This can be an important factor that limits proteolysis and other enzymatic activities. However, thin sections are also more fragile and more difficult to manipulate onto the slide. Once the section has been mounted onto the slide, it is recommended to dry it under vacuum in order to remove condensation from the surface of the tissue and minimize protein diffusion. At this time, if the cutting of the sections is done outside of the routine clinical workflow (where the following steps are always performed), it is advised to cut some additional consecutive sections to which immunohistological staining is applied in order to combine molecular and histological findings [13].

Prior to MALDI-MSI analysis, it is highly recommended to wash the tissue in order to remove any molecules that may interfere with the ionization of proteins (e.g., salts, lipids). Standard protocols recommend washing the tissue sequentially in increasing concentrations of EtOH, commencing with a short wash (~30 s) in 70% EtOH in order to remove cell debris and salts followed by 95 and 100% EtOH in order to fix the tissue. A solvent such as EtOH is recommended for this step as it does not promote the diffusion of proteins. However, it has been widely reported that optimization of the washing steps, based upon the chemical composition of the tissue, can lead to enhanced sensitivity of the MALDI-MSI analysis. For example, brain tissue is often associated with a high content of lipids, a strong protein ion suppressant, and washing this tissue with chloroform or xylene can improve protein detection. Conversely, a different washing protocol should be used if the intended analytes are not proteins, for example, EtOH (70%) with the addition of ammonium acetate (NH_4Ac) is recommended for the desalting of tissue prior to tissue lipidomic analysis [14].

Following these washing steps, it is again recommended to perform an additional drying step under vacuum. However, the drying time should be selected carefully based upon the thickness of the tissue and the type/percentage of solvents used in the washing procedure. Once this additional drying phase has been performed, the FF tissue is then ready for matrix application.

9.2.2 Matrix Application

Matrix deposition plays a critical role in MALDI-MSI experiments, being a major limiting factor in the lateral resolution that can be achieved. The general aim of the matrix deposition is to achieve an optimal balance between crystal dimension/shape (homogeneous and small) and maximal analyte extraction while at the same time avoiding diffusion. Thus, matrix deposition represents arguably the most crucial step in the sample preparation phase. Depending upon the target analyte of choice, a number of different matrices can be used. For example, sinapinic acid (SA) (3,5-dimethoxy-4-hydroxycinnamic acid) and α-cyano-4-hydroxycinnamic acid (α-CHCA) are most commonly the matrices of choice for the extraction of proteins, peptides, and lipids (1–20 kDa). Furthermore, ferulic acid (3-(4-hydroxy-3-methoxy-phenyl)-prop-2-enoic acid) has also been reported for the extraction of high molecular weight proteins (up to 140 kDa). For MALDI-MSP purposes, DHB (2,5-dihydroxybenzoic acid) is also commonly used for the extraction of low molecular weight proteins. However, due to the large crystal size, it is unsuitable for modern imaging applications as this large crystal size severely hampers the spatial resolution achievable. Additionally, ionic matrices, such as CHCA/aniline (CHCA/ANI) and CHCA/*N*,*N*-dimethylaniline (CHCA/DANI), have also been employed alone or in combination with other matrices in order to improve the homogeneity of the crystallization and the detection of protein signals [15]. It is also important to note the increased prevalence of metabolomic-targeted MSI analysis in clinical studies. In these instances, the matrix9AA (9-aminoacridine) is often employed and the mass spectrometer is set in negative ion mode [16].

It is also important to note that with the increased demand for higher spatial resolution images and the rapid evolution in instrument technology, new matrices have been explored. For example, Garate et al. [17] demonstrated

the use of 2-mercaptobenzothiazole (MBT) and 2,5-diaminonapthalene (DAN) as MALDI matrices that can enable the acquisition of higher spatial resolution images (with pixel sizes as low as 5 µm), both in positive and negative ion modes. Furthermore, it stressed the importance of the correct choice of matrix and crystal size with regard to the spatial resolution achievable in MALDI-MSI experiments, suggesting that higher spatial resolutions can yet be obtained with the instrumentation currently in place in many research laboratories.

Focusing on the most commonly employed matrices, the matrix concentration for SA is usually between 10–30 mg/ml and 7–20 mg/ml for CHCA. The choice of concentration is dependent upon the choice of deposition, ensuring the correct balance between minimizing analyte diffusion and maximizing analyte inclusion into the matrix co-crystal network. For example, low matrix concentration can lead to analyte diffusion, whereas high matrix concentration induces rapid crystal formation and prevents proper analyte incorporation into the crystals.

Matrices such as SA and CHCA are dissolved in a solution containing organic solvents, water, and trifluoroacetic acid (TFA). TFA is added to the matrix solution in order to assist in the MALDI ionization process, while the organic solvents facilitate optimal crystal formation. The balance between the rate of solvent evaporation and time of incubation of the analytes is highly important in order to achieve the optimal analyte extraction, and this is directly related to the ratio between the organic solvent and water. Usually, a good balance between solvent evaporation and crystal formation can be obtained with the use of a 1:1 ratio between solvent and water. The choice of organic solvent is dependent upon the physical properties of the analyte molecule. However, 50% acetonitrile (ACN) and 50% ethanol (EtOH) are most commonly employed, with a higher ratio of solvents, such as methanol (MeOH) or isopropanol, to extract more hydrophobic molecules from tissue.

It is also important to note that this crystallization process can also be affected by the particular type of tissue used, with the surfaces of some sample types not being conducive to optimal and homogeneous crystal formation. Therefore, it is imperative that the protocol for matrix deposition is optimized for each type of tissue used. Furthermore, the presence of salts, lipids, and other compounds (e.g., hemoglobin) can also impact upon the crystallization process, thus affecting spectral quality. The impact of these interfering molecules can be minimized by performing washing steps prior to matrix deposition. Additionally, other strategies can be used in order to improve spectral quality, with the addition of detergents to the matrix solution being a successful method to obtain increased sensitivity of proteins. For example, the addition of 0.05% Triton X-100 to an SA solution deposited onto brain tissue, a tissue with a high abundance of lipids, which can often lead to ion suppression of protein signals, led to an increase in intensity of 42% of the peaks present in the average spectrum compared with the deposition of SA without the detergent [18].

The choice of method for matrix deposition is dependent upon the spatial resolution desired in the MALDI-MSI experiment. However, there are a number of methods, both manual and automated, that can be employed for various purposes. For tissue profiling, manual spotting can be performed by pipetting the matrix solution directly onto the tissue ROI. This approach deposits large droplets onto the surface of the sample (a few µl), extracting analytes from a region that is a few millimeters in diameter. However, there are devices that can be used to spot matrix in an automated fashion across the entire surface of the sample or in specific regions of the tissue. Quite commonly, the droplet area is reduced to below 200 µm in diameter, thus leading to less efficient analyte extraction but enabling higher spatial resolution images (Figure 9.2). Deposition using an automated spotter offers the advantage of the droplets generally being

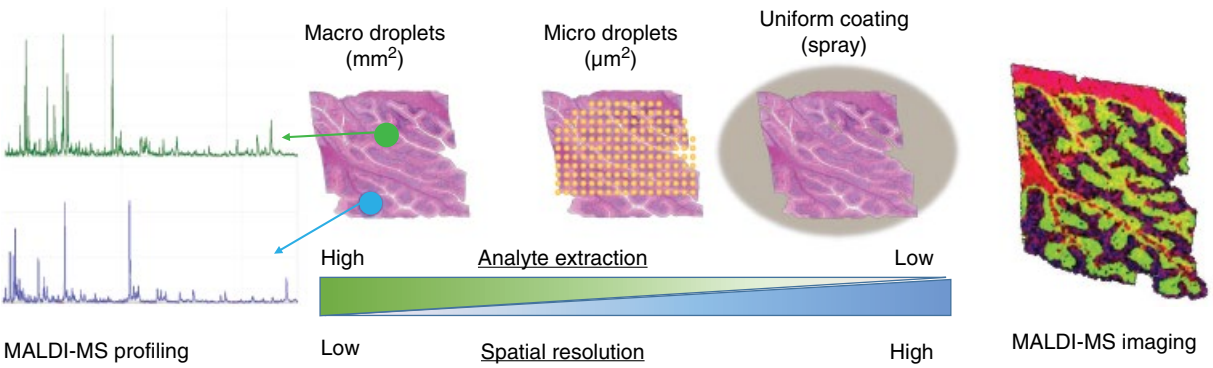


Figure 9.2 The different approaches in matrix deposition for MALDI-MS profiling and MALDI-MSI.

deposited in discrete positions. Thus, if any diffusion of the analyte occurs, then it is limited to a relatively small region and the spatial localization of the analyte is maintained. However, the disadvantage of this approach means that it can be quite difficult to obtain a completely homogeneous layer of matrix. Automated matrix spotters are often associated with piezoelectric nozzles and acoustic wave transfer. In these instances, the dimension of the droplets depends on the following factors: solvent composition and surface tension, number of layers deposited, tissue structure, and chemical composition of the analytes.

In order to achieve a more homogeneous coating of matrix, automated spraying of the matrix solution onto the surface of the sample is performed. In these instances, the spraying devices produce very small droplets that, after drying, produce a very homogeneous and thin layer of solid matrix crystals, which are less than 20 µm in diameter. A homogeneous layer of small matrix crystals offers the possibility for higher spatial resolution MALDI-MS images (<100 µm) compared to automated spotting. Additionally, spraying of the matrix is faster than automated spotting, thus reducing the sample preparation time and increasing sample throughput. In contrast to spotting, matrix spraying forms a layer of liquid on the tissue. Thus, the method must be optimized in order to avoid analyte diffusion. Automated spraying devices are commonly based upon vibrational, pneumatic, or electrospray mechanisms, providing depositions in a highly controlled and reproducible manner. More recently, Gou et al. introduced the concept of electric field-assisted matrix coating, which employs the use of a uniform static electric field that can enhance the detection of positively or negatively charged small molecule metabolites in the MALDI matrix layer [19]. Manual spraying of matrix can also be accomplished using a thin-layer chromatography (TLC) sprayer, airbrush, or pneumatic sprayer. However, these approaches are certainly less reproducible, with the crystallization being dependent upon the surrounding environment (i.e., temperature and humidity) [3].

A sublimation-based approach can also be employed in order to generate the smallest matrix crystal size possible and ultimately acquire higher spatial resolution images when using the appropriate instrumentation. Generally, this approach involves two steps: (i) the sublimation of matrix onto the surface of the sample and (ii) a rehydration/recrystallization step [20].

Depending upon the properties of the particular matrix used, the slide is heated under vacuum (~145°C at 25 mTorr for SA), which promotes sublimation onto the surface of the tissue. Then, the slide containing the tissue is placed on a petri dish containing a piece of filter paper soaked in aqueous TFA (commonly a 5% TFA solution) and the petri dish is sealed in order to create a hydration chamber. In order to achieve tissue rehydration the petri dish is placed in an oven at 85°C for 3.5 min. However, in the case of sublimation, protein extraction is not as efficient compared with other matrix deposition techniques. An additional rehydration phase can increase analyte extraction. It is challenging to optimize sublimation protocols in order to obtain a good balance between matrix crystal size and analyte extraction.

Finally, the same automated devices that are used for matrix deposition can also be used for the deposition of other solutions and derivatizing agents employed for some specific MALDI-MSI experiments [21]. Most importantly, they can also be used for the application of enzymes for on-tissue digestion in a highly homogeneous manner, and, in some instances, the enzyme application and sample incubation can also be performed using the same device (Figure 9.3).

Following MALDI-MSI analysis, the matrix can be removed from the tissue by washing sequentially in increasing concentrations of EtOH (70, 95, and 100%) for short periods of time. Once the washes have been performed, the tissue can be observed under a microscope in order to ascertain if the matrix has been entirely removed. Following this step, the tissue sections can be subjected to histological staining if desired.

9.2.3 Spectral Processing

A MALDI-MSI dataset can be visualized as a data cube in which the three dimensions are represented by m/z values, signal intensity, and spatial coordinates (Figure 9.4). Since a single MALDI-MSI analysis is composed of thousands of spectra topologically positioned in a 2D array, the output file of an MSI analysis can range from several gigabytes in size to up to more than one terabyte. This high dimensionality of the data is a challenge since it makes data management and elaboration time-consuming. Data processing requires significant computer resources.

Preprocessing steps, applied to raw data, are employed in order to remove sources of variation or noise that could lead to artifacts in the data elaboration phase and to enhance the biologically relevant information in the MALDI spectra. Preprocessing steps include baseline correction, smoothing (Figure 9.5), normalization, alignment and calibration, and peak detection.

9.2.3.1 Baseline Removal

The baseline of a spectrum is the connecting line between the data points with the lowest intensity values, on which the entire spectrum lays. Shin proposed

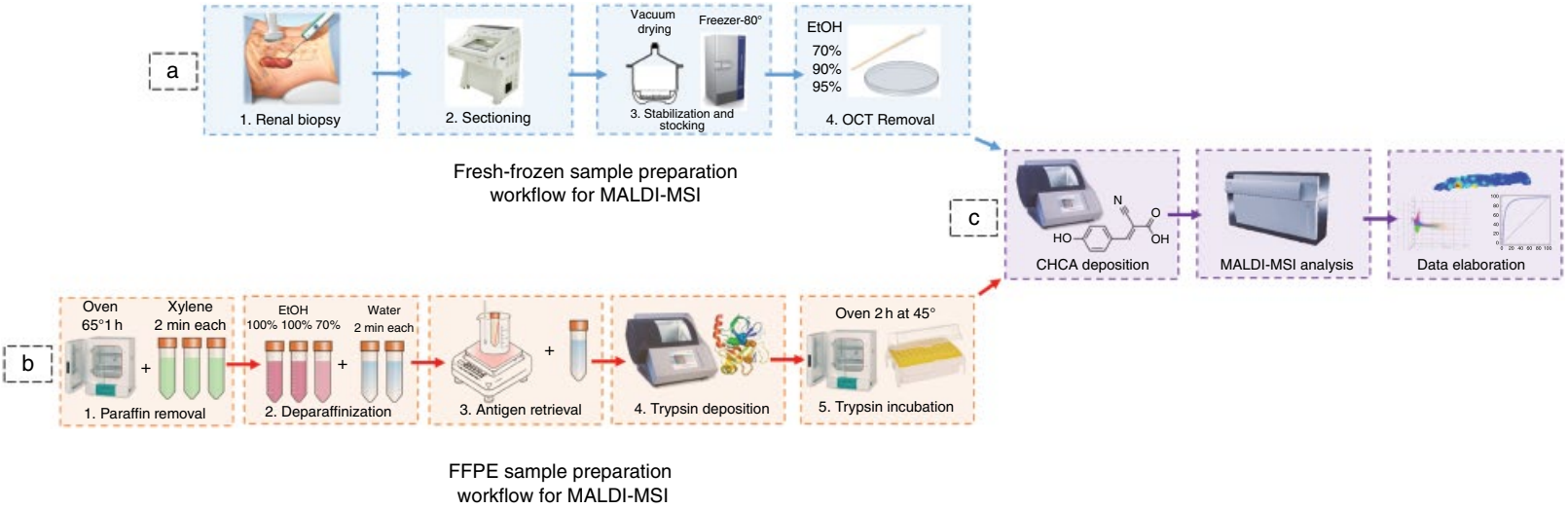


Figure 9.3 Sample preparation workflow for the MALDI-MSI proteomic analysis of fresh-frozen (a) and FFPE (b) tissues. The steps common to both workflows are highlighted in (c).

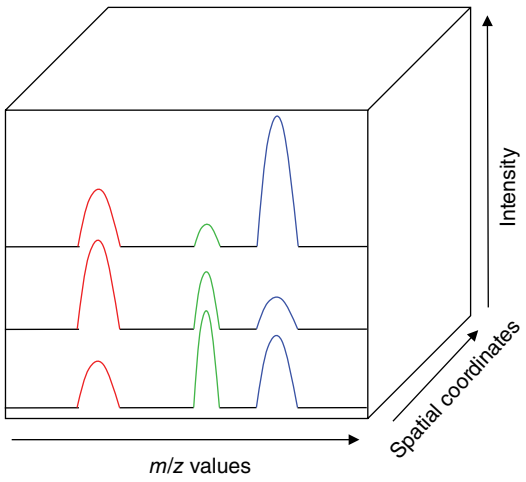


Figure 9.4 The MALDI-MSI data cube represented by *m/z* values (*x*-axis), signal intensity (*y*-axis), and spatial coordinates (*z*-axis).

that MALDI analyses may contain three discrete sources of noise [22]:

Electrical noise from the MS components
Shot noise due to the discrete nature of ion detection
Chemical background generated by impurities (matrix clusters, fragments)

Since the baseline originates from fluctuations in the background signal of the instrument, it has no biological meaning and needs to be removed. Many algorithms are employed in order to remove this contribution; the most widely used being iterative convolution and TopHat.

9.2.3.1.1 Iterative Convolution

Baseline removal through iterative convolution applies a Gaussian filter multiple times, removing spectral features (i.e., the peaks). The algorithm requires the sigma parameter, σ, in order to control the width of this Gaussian filter. The filtering iteration yields a spectrum without peaks, that is, just the baseline. The next step is to take the pointwise minimum of the trend as an estimation of the baseline, subtracting it to the spectrum. The method converges very quickly, meaning that after 15–30 iterations the output does not change.

9.2.3.1.2 TopHat Operation

The TopHat operator was derived from the theory of mathematical morphology and allows the "extraction" of peaks from an image. It is based on the principle that features of interest should stand out in a complex environment (i.e., the noise). This algorithm computes the so-called morphological opening, which in this instance is the background signals, of the spectrum and then subtracts the result from the original spectrum.

9.2.3.2 Smoothing

Signal smoothing aims to alleviate the spectral noise. It aids interpretation and the visualization of the single spectrum and can be performed using two common algorithms: Savitzky–Golay and Gaussian smoothing.

9.2.3.2.1 Savitzky–Golay

Savitzky–Golay is a digital filter that can be applied to a set of data points. It is known to be an almost universal method to improve the signal-to-noise (S/N) ratio. It achieves smoothing by fitting adjacent data points with a low-degree polynomial, taking the central point of the fitted polynomial curve as output. Since it does not distort the essential features in the spectrum, this filter tends to preserve the peak waveform and does not shift the peak positions.

Savitzky–Golay filter is very slow, but intensity loss is much lower and should be preferentially used to smooth low mass spectra where peaks are sharp.

9.2.3.2.2 Gaussian Smoothing

The degree of smoothing is again determined by the standard deviation, σ, of the spread parameter, with a larger σ implying a wider Gaussian filter and thus a greater degree of smoothing. It is a fast method, but it may cause significant intensity loss for sharp peaks. Due to these characteristics, it is usually employed to smooth high mass spectra where peaks are broader.

9.2.3.3 Spectral Normalization

A MALDI imaging dataset can be considered as a collection of independently measured spectra; for this reason, a normalization step is a crucial task in the preprocessing phase in order to compensate for the chemical and analytical differences, facilitating a fair comparison between spectra. It is an indispensable step if several sets of spectra have to be compared with each other, not only intra-analysis but also, and more importantly, inter-analysis.

Normalization is the process of multiplying a mass spectrum with an intensity-scaling factor, *f*, in order to expand or reduce the range of the intensity axis. It is able to project spectra of varying intensity onto a common intensity scale, removing variations in pixel-to-pixel intensity due to uneven matrix deposition, ion suppression, or other factors that can alter the intensity of peaks not strictly due to the actual analyte composition present at a specific spot.

Each normalization method is based on certain assumptions regarding the data, and it is necessary to carefully choose the most appropriate algorithm for a particular dataset in order to avoid generation of artifacts that do not correspond to significant biological information. There are many algorithms employed for

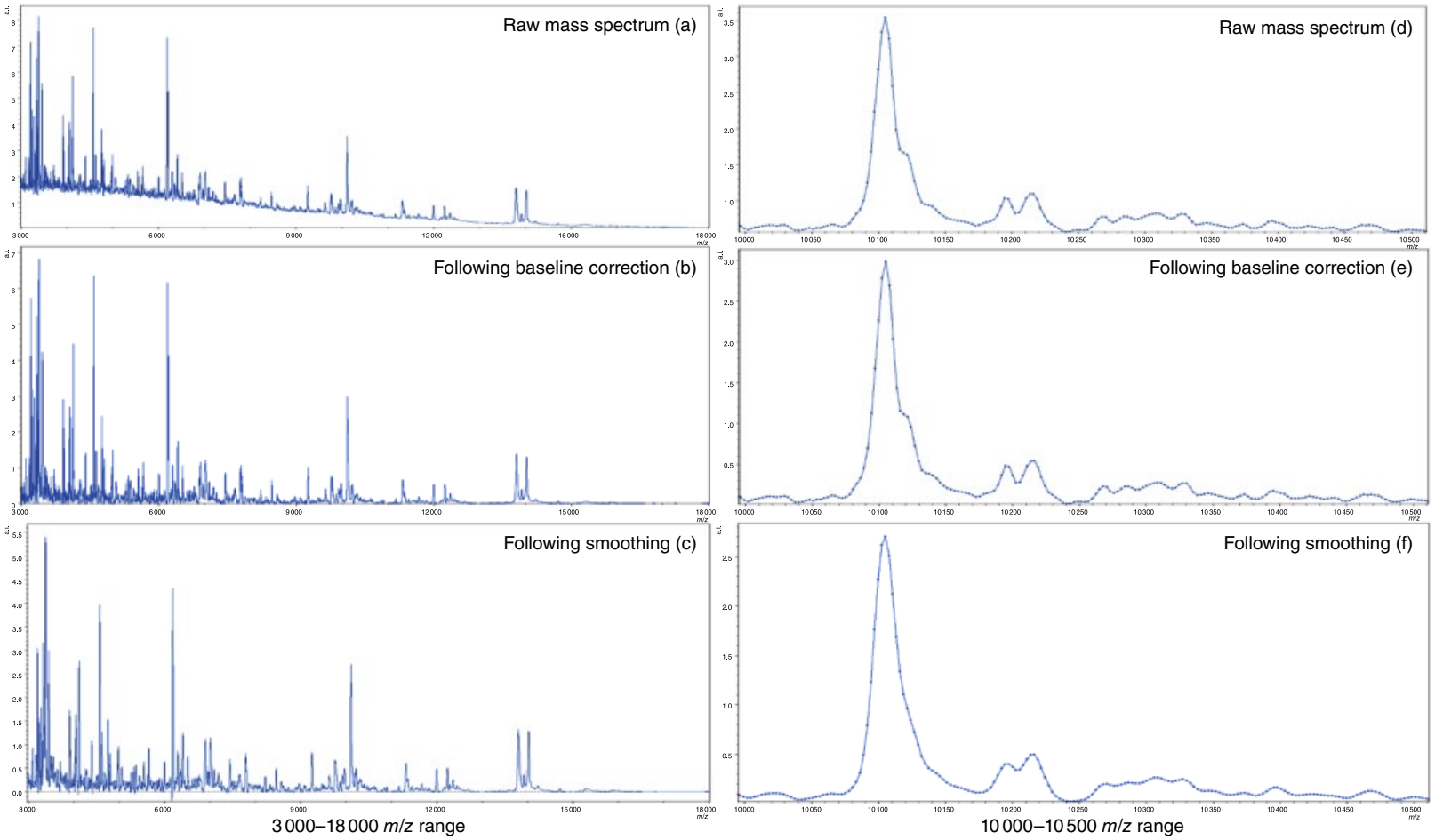


Figure 9.5 The MALDI-MS spectra preprocessing steps. Left panel, in the 3 000–18 000 *m/z* range: Raw (a), following baseline correction (b) and following smoothing (c). Right panel, in the 10 000–10 500 *m/z* range: Raw spectrum (d), following baseline correction (e) and following smoothing (f).

spectra normalization, such as total ion current (TIC), root mean square (RMS), and median.

9.2.3.3.1 Total Ion Current (TIC)

TIC normalization method divides all spectra by their total ion current (i.e., the sum of the intensities of all the peaks), yielding spectra with a common area under the curve. The assumption on which this normalization is based on is that all spectra have a similar area, defined largely by the chemical noise and only to a small extent by the peak intensities. This normalization is the most widely used and can be applied to the majority of MALDI-MSI datasets.

It is important to stress that artifacts may be created in spectra containing a single high intensity ion (e.g., insulin or hemoglobin). This peak would significantly suppress the intensity of every other peak in the spectrum after normalization. It is possible, however, to exclude such peaks from the normalization calculation, potentially solving this problem.

9.2.3.3.2 Root Mean Square (RMS)

The RMS normalization method divides all spectra by the RMS of all data points. This method is most appropriate for use with datasets containing spectra that are expected to have small variations in the peak intensities. The RMS normalization method usually leads to a very uniform distribution of intense signals.

9.2.3.3.3 Median

This normalization method divides all spectra in the dataset by the median of all data points. Median normalization is not significantly affected by the intensity or area of signals in the spectra and can therefore be used if the RMS or TIC normalization methods lead to artifacts.

Median normalization results depend on the type of noise in the spectra. If spectra do not contain a fully symmetrical noise profile, this method will generate significant artifacts.

9.2.3.4 Spectral Realignment

Spectral alignment is an optional step during the preprocessing phase of mass spectra. It is used to account for the slight shifts in the output m/z of peaks as a result of chemical noise and instrument accuracy. Spectral realignment ensures that all of these peaks are realigned to a common mass and thus enable correct spectral comparisons. Most commonly, spectra are realigned by considering the peaks of the mean spectrum as a reference, covering the entire mass range of the analysis, and by then adjusting the spectra by linear or nonlinear interpolation.

9.2.3.5 Generating an Overview Spectrum

It is useful to evaluate peaks obtained in the entire section or in specific areas. The generation of a single spectrum that efficiently represents the molecular composition of the ROI helps in achieving this goal. There are two main approaches to obtain these types of spectra: average and skyline (Figure 9.6).

9.2.3.5.1 Average Spectrum

The average spectrum is obtained simply by averaging all the intensities for each data point. This is the most used approach in the data mining performed after MALDI-MSI analyses, but it can potentially yield artifacts since it can reduce the contribution of ions that are present only in small and specific regions of the tissue.

9.2.3.5.2 Skyline Spectrum

The skyline spectrum is obtained by picking the highest value of intensity for each data point; regiospecific peaks are well represented in the total spectrum and are not eliminated as in the case of the average spectrum approach.

This method should be applied only to well-calibrated, aligned, low-noise spectra in order to avoid peak broadening and loss of accuracy.

9.2.3.6 Peak Picking

This process detects a representative set of m/z values in a group of mass spectra that significantly rise above the noise level (i.e., above a certain S/N threshold). The aim of the peak picking is to reduce the number of m/z values by discarding those values corresponding to noise signals or to nonspecific baseline. Various peak picking methods for MALDI mass spectra are available and are implemented in MS software packages: the most used algorithms are orthogonal matching pursuit (OMP) and local maximum.

Applying this process in MSI spectra containing large amounts of data takes a lot of time and uses significant computing resources. The simplest approach would be to apply the algorithm on the single mean spectrum. However, this approach results in elimination of peaks with high intensity in very discrete areas of tissue.

9.2.3.6.1 Orthogonal Matching Pursuit

OMP is a signal processing application that models each spectrum as a sum of Gaussian-shaped functions (peaks). The parameter, sigma, determines the width of the Gaussian peaks (can also be estimated automatically based on the mean spectrum). For each single spectrum, OMP selects m/z peaks that fit the Gaussian shape. This algorithm allows the user to manually set a maximum number of possible peaks.

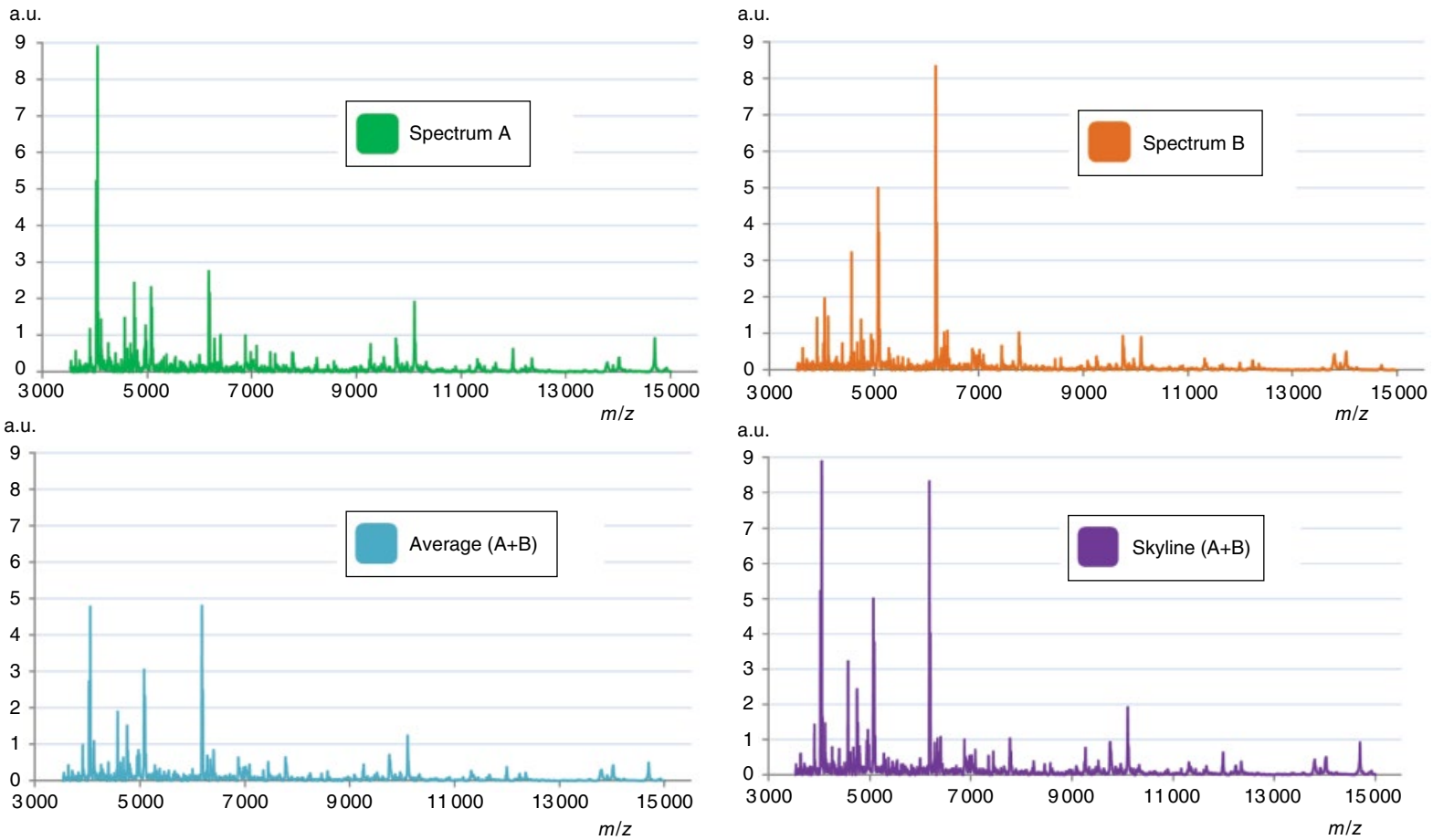


Figure 9.6 The generation of overview MALDI spectra. (top) mass spectra obtained from two separate *x*–*y* coordinates and (bottom) the resulting average and skyline spectra.

9.2.3.6.2 Local Maxima

This approach identifies all the local maxima in the spectra and marks them as peak positions (it assumes that a peak should differentiate itself from the background noise). It is the simplest approach, but the most prone to yield false positive peaks positions that are in fact correlated with the background noise.

9.2.4 Data Elaboration

On order to analyze MALDI-MSI data, conventional statistical tests can be applied (such as ANOVA, *t*-test, and *z*-test). For example, ANOVA is commonly used to detect differences between groups (i.e., peaks and therefore proteins), a typical task in research and clinical diagnosis. However, the inherent variability in MALDI spectra due to anatomically and biologically distinct regions may hinder significant conclusions. In order to improve the reliability of the statistical analysis, ROIs in spectra, which correlate with histopathological features, can be defined. Following this, both unsupervised and supervised data mining approaches can be undertaken as a means of finding biologically significant information within the dataset.

9.2.4.1 Unsupervised Data Mining

Unsupervised methods aim at revealing hidden structures in unlabelled data and can be applied without any prior knowledge of the data structure [23].

Examples of unsupervised analysis are hierarchical clustering, principal component analysis (PCA), and bisecting K-mean.

9.2.4.1.1 Hierarchical Clustering

Hierarchical clustering is a data mining method that consists in grouping several data subsets into clusters based on their similarity and then building a hierarchy, exploiting intra-cluster differences. The algorithm is able to evaluate dissimilarities by calculating the actual distance between two spectra according to different metrics (Euclidian distance, Manhattan distance).

The output is a dendrogram representing a hierarchical tree in which similar spectra are clustered under a single node. In a MALDI-MSI dataset, it is possible to plot the spatial distribution of the clusters identified by this analysis, which can then be correlated with the histological image.

It can be performed in a bottom-up (each observation starts in its own cluster) or top-down (all observations start in one cluster) approach. The only downside is that this approach may use significant computer resources since it requires the creation of a distance matrix of size n^2 (where n is the number of spectra).

9.2.4.1.2 Bisecting K-Means

Bisecting K-means is a divisive clustering algorithm that combines K-means and hierarchical clustering: it iteratively splits the data into two maximally different clusters (bisectioning) that are then further separated into subclusters according to similarities in their data points (K-mean). The subclustering is achieved by selecting K points as the initial centroids, assigning spectra to the closest centroid and then computing the centroid of each cluster again until a convergence is reached.

9.2.4.1.3 Principal Component Analysis (PCA)

PCA is the most commonly used component analysis method for MALDI imaging data representation. Due to the complexity of the information (i.e., the number of variables) enclosed in a single mass spectrum, it is impossible to visualize the entire dataset in an N-dimensional space.

PCA is capable of reducing the dimensionality of a dataset while retaining the majority of the information contained in the data. Since many variables often contain redundant information, it is possible to replace a group of variables with a single, more informative variable (component) by a linear combination of the single variables. PCA differs from the other variable transformations employed in statistics since the data itself determines the transformation vectors.

From a technical standpoint this analysis employs an orthogonal transformation to convert a set of observations into a set of values of linearly uncorrelated variables (principal component). Orthogonal means that every component is uncorrelated with the preceding components and these new variables are used to plot the data distribution. The variables are ordered by their variance (with the first component accounting for the highest possible variance).

Using PCA, one can represent the full dataset with a few score images corresponding to first principal components. These score images reveal spatial structures hidden in the dataset by showing prominent spatial patterns (high intensity regions) [24].

9.2.4.2 Supervised Data Mining

9.2.4.2.1 Receiver Operating Characteristic (ROC)

A receiver operating characteristic (ROC) curve illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (sensitivity) against the false positive rate (1 specificity) at various threshold settings.

It is a univariate measurement used to assess the ability of a single peak or of a classifier based on several peaks to differentiate between two populations.

The area under the ROC curve (AUC) measures the discrimination quality in the interval between 0.5 and 1.0. A perfect discrimination would yield an AUC equal to 1 or 0. The closer the AUC to 0.5, the less useful the m/z value, and the closer it is to 1.0, the more suitable the m/z value is to be used as a univariate criterion [25].

9.2.5 Correlating MALDI-MS Images with Pathology

Particular pathological ROIs present within the tissue can be well defined using traditional histopathology stains, with the most widely used being hematoxylin and eosin (H&E), cresyl violet, methylene blue, toluidine blue, DAPI, and/or immunohistochemistry, directing the analysis toward obtaining region-specific molecular signatures. There are two different staining approaches that can be currently used: staining on the same section of tissue used for MALDI-MSI analysis, both pre- and post-analyses, or staining on consecutive tissue sections. While performing staining on the same tissue used for MALDI-MSI analysis enables unambiguous correlation of MALDI-MS images with histological images, it can be potentially hampered by a loss in integrity of the tissue following analysis and/or removal of the MALDI matrix. Conversely, correlating with consecutive tissue sections can avoid the aforementioned issue; however, it is not always certain that the adjacent tissue sections will be identical. If histological staining is to be performed prior to MALDI-MSI analysis, dyes such as cresyl violet or methylene blue are preferable since HE dyes interfere with the analysis and affect spectral quality [26]. The resulting histological images can then later be scanned with an optical scanner and stored in a database. Optimal results can be achieved in microscopic resolution with an MSI compatible MIRAX SCAN instrument (Carl Zeiss); however, other scanners that achieve image resolutions greater than 10 000 dpi can also be used. It is also important to note that using the appropriate digital platforms, such as Aperio Spectrum, which are now becoming commonplace in clinical centers in order to facilitate the sharing of digital slides, a pathologist can annotate these scanned images electronically in order to highlight ROIs [27]. This can enable the correct interpretation of the slides by individuals who are not experts in the field of histopathology and, ultimately, increase the throughput of the analysis.

Various software packages are currently available and can be used to exploit the unique capability to correlate molecular and histological images. SCiLs Lab 2014 software enables the importation of histological images and can automatically search for particular m/z markers that are co-localized with the histopathological annotations. Furthermore, recent developments in instrumentation have further facilitated the correlation of MALDI-MS images with histology. For example, Shimadzu Corporation has recently introduced a novel imaging mass microscope (*iMScope*) that combines an optical microscope for the visualization of high-resolution images with a hybrid ion trap TOF mass spectrometer with a MALDI source. This novel instrument visualizes the distribution of molecules in a scanned tissue sample at atmospheric pressure.

9.3 Applications in Clinical Research

MALDI-MSI is a highly flexible platform and has been successfully employed in numerous studies, ranging all the way from the study of human diseases to forensic science. Furthermore, there has been an emerging trend toward combining MALDI-MSI with clinical imaging techniques, such as magnetic resonance imaging (MRI) [28], in a multimodal manner, highlighting the rapidly evolving nature of this approach. However, perhaps of greatest clinical relevance is the role of MALDI-MSI in the study of cancer biology, with studies targeting breast [29], colon [30], lung, ovarian [31], prostate [32], and thyroid cancers [33] being widely published. Furthermore, MALDI-MSI approaches in cancer biology have also been applied to novel sample types, such as cytological smears taken from thyroid via fine-needle aspiration biopsy (FNAB) [33, 34], further highlighting the flexibility of this technique. In all of these studies, the primary objective has been to discriminate cancer tissue from the normal and/or tumor margin regions and to classify different grades of cancer at a molecular level. Of particular note, Balluff and colleagues studied tissue taken from gastric and breast carcinoma patients [29]. They were able to study phenotypic intratumor heterogeneity, identifying different regions within tumor tissue that appeared homogeneous using traditional histological techniques. Using elegant spatial segmentation and multivariate analysis methods, they were able to identify tumor subpopulations, within histologically homogeneous regions, that were associated with changes in the levels of DEFA-1 and histone H2A. Moreover, by combining this information with clinical data obtained from the patients studied, they were able to predict the survival rate of patients based upon the number of observed phenotypic tumor subpopulations. This is a powerful example of how MALDI-MSI can not only support traditional histological analyses but also potentially provide additional, and clinically significant, information that was previously not possible.

The combination of MSI and histology is now extensively used for pharmacological research due to the capability

of this technique to simultaneously monitor the distribution of a drug and its metabolites [35]. In addition, with suitable calibration curves, it is possible to obtain semiquantitative measurements of drug compound concentrations [36]. Such studies can be used in the preclinical phase, filtering out lead compounds that are shown to accumulate in nonspecific regions of tissue and/or generate toxic metabolites [37]. Furthermore, MALDI-MSI techniques are now being applied in order to image the spatial localization of drugs and their metabolites within 3D cell cultures, enabling more detailed information related to the site of action [38].

As with proteomic applications, studies are increasingly correlating MALDI-MSI with traditional histology, focusing on the accumulation of cancer drugs in heterogeneous tumor tissue environments. One example of this approach was the study of microvascularization effects of numerous anticancer drugs in tumor tissue [39]. In another example, MALDI-MSI was used to monitor the distribution of a targeted medicine, vemurafenib, for malignant melanoma to metastatic lymph nodes tumors [40]. The study provided evidence that the drug specifically targeted proto-oncogene BRAF, a gene that promotes an expression of the serine/threonine-protein kinase B-Raf, that is ultimately responsible for sending signals that are related to cell growth, and mutations of this gene have been shown to be implicated in cancer. The ability to monitor the distribution of a drug to histologically specific regions provides a greater understanding of the mode of action of drugs within particular disease environments while highlighting whether or not a drug targets the intended site. This can provide an insight into the potential success, or failure, of a developing drug and help drive the pharmaceutical industry toward the development of personalized drug therapies.

MALDI-MSI has also become an important tool in the investigation of renal diseases, studying proteins/peptides, lipids, and drugs in both animals and humans. One approach involved the isolation of glomeruli from rats with focal segmental glomerulosclerosis (FSGS) via laser capture microdissection (LCM) and analyzed using MALDI-MSP [41]. The authors demonstrated that they were able to generate proteomic patterns of sclerotic and nonsclerotic glomeruli within FSGS. However, they also noted that the proteomic patterns of nonsclerotic glomeruli were more similar with those of sclerotic glomeruli than with those of completely healthy glomeruli, postulating that that there is an early activation of sclerotic processes occurring at the molecular level.

Furthermore, the pathogenesis of IgA nephropathy (IgAN) was investigated in a mouse model that spontaneously develops mesangioproliferative lesions with IgA deposition, comparable to the human disease [42]. The molecular distribution of a number of lipids was mapped in the hyper-IgA (HIGA) murine kidneys using MALDI-MSI. Interestingly, a number of lipids were found to be over-expressed in the cortical region of the HIGA kidney, with respect to controls, for example, *O*-phosphatidylcholine, PC(O-16:0/22:6) and PC(O-18:1/22:6).

MALDI-MSI was applied recently to the study of primary glomerulonephritis (GN) in humans. Mainini et al. subjected to MALDI-MSI analysis renal tissue obtained by biopsy. Interestingly, it was determined that the glomeruli and tubules of healthy tissue presented similar proteomic profiles. However, in the case of primary GN, glomeruli and tubules presented different protein profiles. Furthermore, altered protein expression compared to controls was evident between different types of primary GN, such as membranous glomerulonephritis (MGN) and minimal change disease (MCD). Finally, GN tubules even without morphological evidence of the disease showed a different protein profile compared with controls. Thus, it is possible to detect early molecular alterations of the disease that are not accessible by traditional histological methods. This feasibility study highlighted the potential role that MALDI-MSI could play in the detection of diagnostic biomarkers associated with primary GN [24].

This body of work was further expanded upon by Smith et al. in 2016, applying MALDI-MSI to bioptic renal tissue taken from patients with the most frequent glomerular kidney diseases (GKDs): FSGS, IgAN, and MGN [43]. Firstly, this technique was able to generate molecular signatures capable of distinguishing between normal kidney and pathological GN, with specific signals representing potential biomarkers of CKD progression. Furthermore, specific disease-related signatures for FSGS and IgAN were detected. Among the specific FSGS-related signatures, one protein was identified as α-1-antitrypsin and, upon validation with the antibody, was shown to be localized to the podocytes within sclerotic glomeruli. This work showed a promising application of MALDI-MSI in the study of GN, highlighting a number of potential biomarkers of CKD progression.

In order to bring such MALDI-MSI information into a clinical setting, these findings should be correlated with information obtained from the urinary proteome or peptidome in order to highlight whether tissue-derived proteins of glomerular diseases can also be detected in urine, a biological fluid that can be easily collected. If successful, such biomarkers could be translated into less-invasive diagnostic or prognostic tools, which is the ultimate goal of many clinical proteomics applications.

References

1 Caprioli RM, Farmer TB, Gile J. Molecular imaging of biological samples: localization of peptides and proteins using MALDI-TOF MS. *Anal Chem*. 1997;69(23):4751–4760.

2 Bandura DR, Baranov VI, Ornatsky OI, et al. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal Chem*. 2009;81(16):6813–6822.

3 Chughtai K, Heeren RM. Mass spectrometric imaging for biomedical tissue analysis. *Chem Rev*. 2010;110(5):3237–3277.

4 Cornett DS, Frappier SL, Caprioli RM. MALDI-FTICR imaging mass spectrometry of drugs and metabolites in tissue. *Anal Chem*. 2008;80(14):5648–5653.

5 Mainini V, Lalowski M, Gotsopoulos A, et al. MALDI-imaging mass spectrometry on tissues. *Methods Mol Biol*. 2015;1243:139–164.

6 Thomas A, Chaurand P. Advances in tissue section preparation for MALDI imaging MS. *Bioanalysis*. 2014;6(7):967–982.

7 Boggio KJ, Obasuyi E, Sugino K, et al. Recent advances in single-cell MALDI mass spectrometry imaging and potential clinical impact. *Expert Rev Proteomics*. 2011;8(5):591–604.

8 Römpp A, Spengler B. Mass spectrometry imaging with high resolution in mass and space. *Histochem Cell Biol*. 2013;139(6):759–783.

9 Longuespée R, Fléron M, Pottier C, et al. Tissue proteomics for the next decade Towards a molecular dimension in histology. *OMICS*. 2014;18(9):539–552.

10 Lemaire R, Desmons A, Tabet JC, et al. Direct analysis and MALDI imaging of formalin-fixed, paraffin-embedded tissue sections. *J Proteome Res*. 2007;6(4):1295–1305.

11 Norris JL, Caprioli RM. Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chem Rev*. 2013;113(4):2309–2342.

12 Zaima N, Hayasaka T, Goto-Inoue N, et al. Matrix-assisted laser desorption/ionization imaging mass spectrometry. *Int J Mol Sci*. 2010;11(12):5040–5055.

13 Schwartz SA, Reyzer ML, Caprioli RM. Direct tissue analysis using matrix-assisted laser desorption/ionization mass spectrometry: practical aspects of sample preparation. *J Mass Spectrom*. 2003;38(7):699–708.

14 Wang H-YJ, Liu CB, Wu H-W. A simple desalting method for direct MALDI mass spectrometry profiling of tissue lipids. *J Lipid Res*. 2011;52(4):840–849.

15 Calvano CD, Carulli S, Palmisano F. Aniline/alpha-cyano-4-hydroxycinnamic acid is a highly versatile ionic liquid for matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun Mass Spectrom*. 2009;23(11):1659–1668.

16 Fagerer SR, Nielsen S, Ibánez A, et al. Matrix-assisted laser desorption/ionization matrices for negative mode metabolomics. *Eur J Mass Spectrom*. 2013;19(1):39–47.

17 Garate J, Fernández R, Lage S, et al. Imaging mass spectrometry increased resolution using 2-mercaptobenzothiazole and 2,5-diaminonaphtalene matrices: application to lipid distribution in human colon. *Anal Bioanal Chem*. 2015;407(16):4697–4708.

18 Leinweber BD, Tsaprailis G, Monks TJ, et al. Improved MALDI-TOF imaging yields increased protein signals at high molecular mass. *J Am Soc Mass Spectrom*. 2009;20(1):89–95.

19 Guo S, Yammin W, Zhou D, et al. Electric field-assisted matrix coating method enhances the detection of small molecule metabolites for mass spectrometry imaging. *Anal Chem*. 2015;87(12):5860–5965.

20 Yang J, Caprioli RM. Matrix sublimation/recrystallization for imaging proteins by mass spectrometry at high spatial resolution. *Anal Chem*. 2011;83(14):5728–5734.

21 Flinders B, Morrell J, Marshall P et al. The use of hydrazine-based derivatizaion reagents for improved sensitivity and detection of carbonyl containing compounds using MALDI-MSI. *Anal Bioanal Chem.* 2015;407:2085–2094.

22 Shin H, Mutlu M, Koomen JM, et al. Parametric power spectral density analysis of noise from instrumentation in MALDI TOF mass spectrometry. *Cancer Inform.* 2007;3:219–230.

23 Dziuda DM. *Data Mining for Genomics and Proteomics*. 1st Ed. Wiley-Interscience, Hoboken;2010.

24 Mainini V, Pagni F, Ferrario F, et al. MALDI imaging mass spectrometry in glomerulonephritis: feasibility study. *Histopathology*. 2014;64(6):901–906.

25 Swets JA. Measuring the accuracy of diagnostic systems. *Science*. 1998;240(4857):1285–1293.

26 Chaurand P, Schwartz SA, Billheimer D, et al. Integrating histology and imaging mass spectrometry. *Anal Chem*. 2004;76(4):1145–1155.

27 Krishnamurthy S, Mathews K, McClure S, et al. Multi-institutional comparison of whole slide digital imaging and optical microscopy for interpretation of hematoxylin-eosin-stained breast tissue sections. *Arch Pathol Lab Med*. 2013;137(12):1733–1739.

28 Attia AS, Schroeder KA, Seeley EH, et al. Monitoring the inflammatory response to infection through integration of MALDI-MSI and MRI. *Cell Host Microbe*. 2012;11(6):664–673.

29 Balluff B, Frese CK, Maier SK, et al. De novo discovery of phenotypic intratumour heterogeneity using imaging mass spectrometry. *J Pathol*. 2015;235(1):3–13.

30 Mirnezami R, Spagou K, Vorkas PA, et al. Chemical mapping of the colorectal cancer microenvironment via MALDI imaging mass spectrometry (MALDI-MSI) reveals novel cancer-associated field effects. *Mol Oncol*. 2014;8(1):39–49.

31 Kang S, Shim HS, Lee JS, et al. Molecular proteomics imaging of tumor interfaces by mass spectrometry. *J Proteome Res*. 2010;9(2):1157–1164.

32 Steurer S, Borkowski C, Odinga S, et al. MALDI mass spectrometric imaging based identification of clinically relevant signals in prostate cancer using large-scale tissue microarrays. *Int J Cancer*. 2013;133(4):920–928.

33 Pagni F, Mainini V, Garancini M, et al. Proteomics for the diagnosis of thyroid lesions: preliminary report. *Cytopathology*. 2015;26:318–324. doi:10.1111/cyt.12166.

34 Mainini V, Pagni F, Garancini M, et al. An alternative approach in endocrine pathology research: MALDI-IMS in papillary thyroid carcinoma. *Endocr Pathol*. 2013;24(4):250–253.

35 Ait-Belkacem R, Sellami L, Villard C, et al. Mass spectrometry imaging is moving toward drug protein co-localization. *Trends Biotechnol*. 2012;30(9):466–474.

36 Reyzer ML, Hsieh Y, Ng K, et al. Direct analysis of drug candidates in tissue by matrix-assisted laser desorption/ionization mass spectrometry. *J Mass Spectrom*. 2003;38(10):1081–1092.

37 Aichler M, Walch A. MALDI imaging mass spectrometry: current frontiers and perspectives in pathology research and practice. *Lab Invest*. 2015;95(4):422–431.

38 Liu X, Hummon AB. Mass spectrometry imaging of therapeutics from animal models to three-dimensional cell cultures. *Anal Chem*. 2015;87(19):9508–9519.

39 Buck A, Halbritter S, Späth C, et al. Distribution and quantification of irinotecan and its active metabolite SN-38 in colon cancer murine model systems using MALDI MSI. *Anal Bioanal Chem*. 2015;407(8):2107–2116.

40 Sugihara Y, Végvári Á, Welinder C, et al. A new look at drugs targeting malignant melanoma—an application for mass spectrometry imaging. *Proteomics*. 2014;14(17–18):1963–1970.

41 Xu BJ, Shyr Y, Liang X, et al. Proteomic patterns and prediction of glomerulosclerosis and its mechanisms. *J Am Soc Nephrol*. 2005;16(10):2967–2975.

42 Kaneko Y, Obata Y, Nishino T, et al. Imaging mass spectrometry analysis reveals an altered lipid distribution pattern in the tubular areas of hyper-IgA murine kidneys. *Exp Mol Pathol*. 2011;91(2):614–621.

43 Smith A, L'Imperio V, De Sio G, et al. α-1-antitrypsin detected by MALDI-Imaging in the study of glomerulonephritis: its relevance in chronic kidney disease progression. *Proteomics*. 2016 Jun;16(11–12):1759–1766.

10

Metabolomics of Body Fluids

Ryan B. Gil and Silke Heinzmann

Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environment Health, Neuherberg, Germany

10.1 Introduction to Metabolomics

Metabolomics is the investigation of the metabolite composition of a cell, tissue section, or biological [1]. Metabolites are defined as small molecules (typically <1000 Da), which are transformed during cellular metabolism [2]. Over the past 100 years, intensive research efforts were dedicated in understanding biochemical pathways. This knowledge was extensive, but fragmented due to the nature of biochemical experiments of the past, which often focused on specific features of a specific enzymatic reaction [3]. These types of experiments could not reflect the changes taking place in metabolism as a whole. An example is diabetes, with blood glucose being the trademark metabolite. Technologies and computational instruments are now in place for a broader look into the complete ensemble of metabolites that are affected by disease.

While single metabolite markers may change, little can be deduced about why these single alterations have happened; therefore it is important to move away from the single marker practice and move toward global metabolic approaches for individual and population health management [3]. This way a fingerprint of many different metabolites can be used to provide a deeper perspective of a patient's health. A paradigm shift of such as this will likely open the way to global metabolite profiling.

Today, metabolomics as a screening tool is clearly the missing link in current healthcare practices [3] and could be greatly utilized to avoid trial-and-error therapies. As well, it could be used to monitor the side effects from such therapies. In more wealthy and developed countries, the incidence of diseases resulting from perturbations (e.g., diet/blood pressure) in metabolism is steadily increasing [4]. In addition, the causes of many common diseases or conditions, such as hypertension, diabetes, and chronic kidney disease (CKD), are metabolic imbalances [3].

The terms metabonomics and metabolomics are often used interchangeably. Metabonomics refers to the measurement of the global metabolomic response of a living organism to some kind of stimuli (e.g., genetic or environmental) [5]. Metabolomics refers to the analytical description of the samples taken from such organisms (e.g., urine metabolome, tissue metabolome) [5]. In contrast to other -omics fields, such as genomics and proteomics that are directly subject to epigenetic regulation and posttranslational modifications, metabolites generally serves as direct markers of biochemical activity [2]. Therefore, within the -omics cascade, metabolomics is in direct relation to phenotype (Figure 10.1 *-omics cascade*), meaning metabolites can act as direct markers of metabolic health and are easier to correlate to a respective phenotype [2, 6] as they are the endpoints of metabolism. However one should not consider the metabolome completely immune to regulatory changes. Additionally, metabolomics can offer a real-time assessment of an organism's phenotype [6].

This can be done through targeted or nontargeted approaches. Targeted metabolomics is generally driven by a particular hypothesis. In this case a researcher will attempt to measure specific metabolites related to one or maybe more biological pathways. This approach could then be most suited for pharmacokinetic-type studies, where specific drug metabolism is of interest [2]. Alternatively, nontargeted metabolomics is global in nature and is designed to be as unbiased as possible in metabolites measured. Ideally, researchers would seek to measure the entire metabolome of an organism to produce data that comprehensively represents the whole metabolome [2]. Both approaches have their respective limitations and challenges, but one can argue that nontargeted metabolomics is best suited to drive modern

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

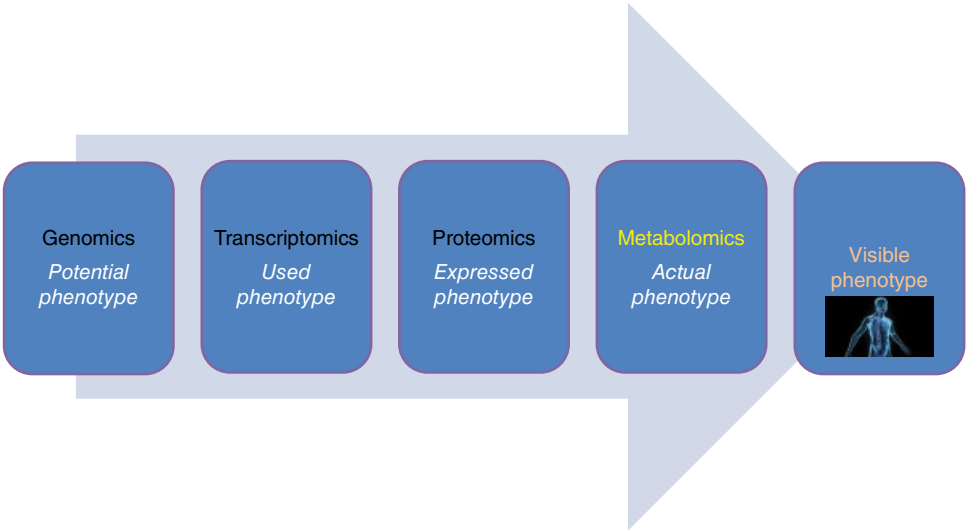


Figure 10.1 The -omics progression toward phenotype characterization. Four major -omics fields build from the previous to describe the phenotype. As can been seen, metabolomics is the most distal of the -omics fields and therefore can most accurately describe the true phenotype.

medicine away from the single-biomarker philosophy. Advancement in personalized medicine and drug discovery could most likely come through metabolomics [7, 8].

10.2 Analytical Techniques

As mentioned earlier, metabolomics studies can follow two main approaches: targeted and nontargeted [6]. The nature of the study, sample type selected, and analytes of interest should determine the type of instrumentation selected. High throughput and robustness are necessary if metabolomics is to be clinically practical, and with regard to kidney disease diagnostics and therapy, sample types should focus on noninvasive biofluids. Biopsy tissue would also be extremely useful in metabolomics, especially for human CKD etiology.

For general sample handling of biobanks, an extensive review has been written by Bernini et al. [9]. However it is necessary for one when describing metabolomics in a clinical setting to briefly discuss sample selection, handling, and storage with particular attention to the specialized needs for quality metabolomics studies.

For general subject selection, most metabolomics studies will take into consideration factors such as age, gender, and diet. Beyond the aforementioned, other factors such as ethnicity, body mass index (BMI), geographical location, and lifestyle (e.g., level of exercise) should also be considered [10]. Perhaps a factor not normally considered is the levels at which healthy controls consume over-the-counter pharmaceuticals (e.g., paracetamol) and other products such as herbal and dietary supplements [10]. These could have profound effects on the metabolome and introduce substantial bias to the data [4].

The identification and quantification of metabolites cannot be determined from genetic or biochemical assays alone. It is for this reason that metabolomics relies on sophisticated instrumentation. The two main analytical instruments used in general metabolomics are nuclear magnetic resonance (NMR) and mass spectrometry (MS). Both have their advantages and disadvantages and will be further discussed in detail. A diagram of the general workflow for metabolomics studies is illustrated in Figure 10.2 (*metabolomics workflow*).

10.2.1 NMR

NMR is a reliable and robust technique for the analysis of the human metabolome that can currently measure metabolites down to micromolar ranges [11]. Here, only a brief and simple description of the instrumentation can be given. The NMR phenomenon can be described as the absorption of electromagnetic radiation by a given atomic nuclei in a magnetic field. The spin state of a nucleus can be altered by a specific radiofrequency (RF), and this alteration produces a measurable signal. The frequency needed for absorption is based on three parameters: the type of nucleus, the chemical environment of the nucleus, and the strength of the magnetic field applied. Therefore, for a given magnetic field, each nucleus will absorb a slightly different RF. These factors are expressed as chemical shift (δ) in ppm, which is a horizontal scale that describes the difference in RF required

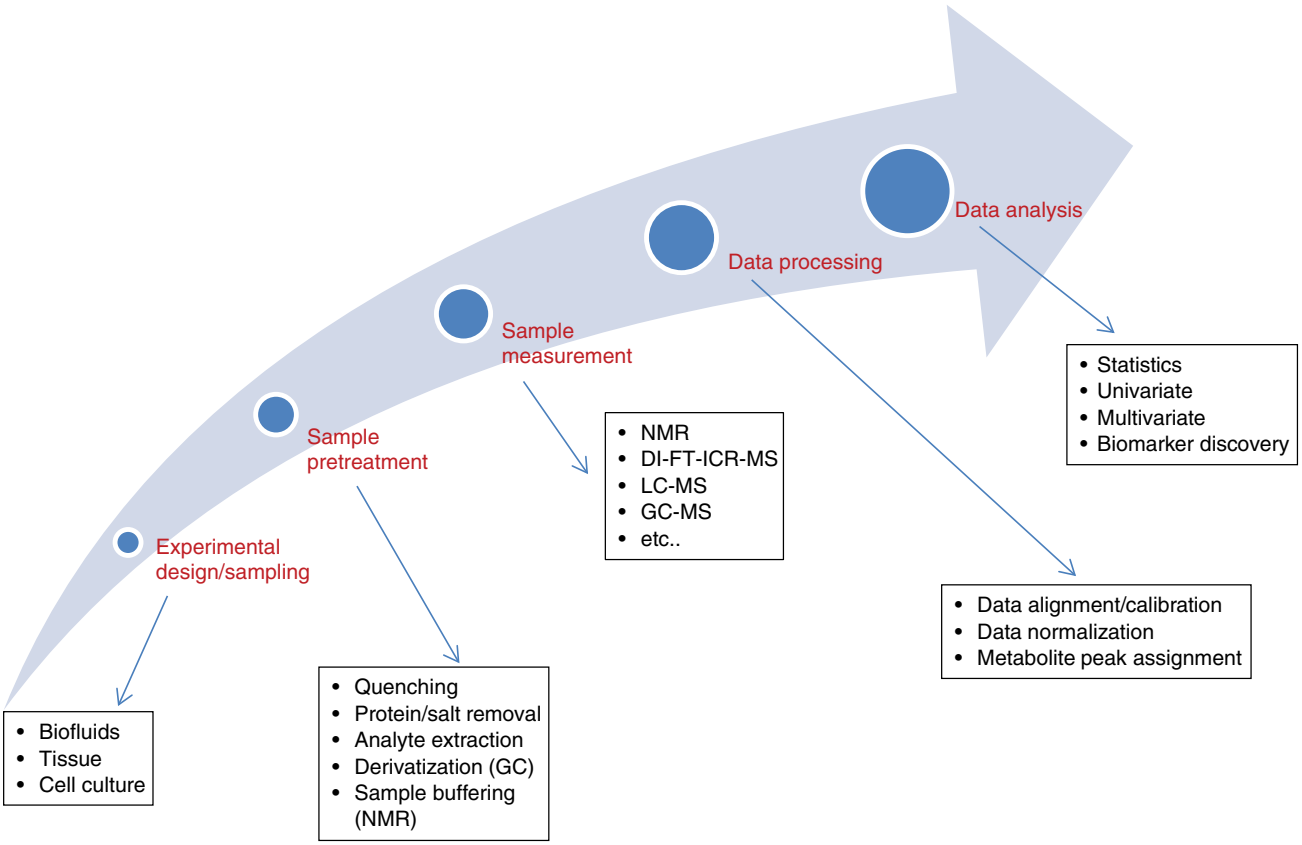


Figure 10.2 Metabolomics workflow.

from one nucleus to a standard compound. Therefore, the chemical shift of a nucleus describes the character of a respective nucleus. For example, if a proton is attached to a methyl group or a carboxyl group, it will have a different chemical shift [12].

NMR can allow for the acquisition of thousand distinct peaks in biofluids and has potential to detect and quantify hundreds of metabolite compounds [10, 13]. Fundamentally, it offers a top-down approach metabolomics analysis, as all metabolites are present and measured by NMR-active atoms, mainly ^{1}H and ^{13}C [3]. It is also nondestructive, allowing the same samples to be reanalyzed in different ways if needed. Conveniently, NMR sample preparations do not require extensive analyte extraction protocols [10, 11]; however such approaches could be utilized if desired.

NMR spectral information is therefore well suited for further chemometric analysis in a nontargeted approach, as sample preparations for NMR generally do not need to remove metabolites from the originally sample matrix [11]. Nevertheless, specific NMR pulse sequences can be used to investigate subsets of metabolites if needed [11]. As a result NMR may be considered as the best analytical techniques for metabolic profiling and screening of human urine in a nontargeted fashion. Unfortunately, it is initially expensive and requires skilled personnel to both operate and interpret data. A sample ^{1}H NMR spectrum can be seen in Figure 10.3, with metabolite peaks annotated.

10.2.1.1 Sample Preparation for Urine

NMR analysis of urine is well established in metabolomics, but variations in a urine metabolomics profile exist due to diet, drugs, overall health, and urine sample handling [11]. Therefore it is advisable to collect many samples over an extended time, which can give a more comprehensive metabolomics profile of a patient's urine. A single urine sample will unlikely be a complete picture of an individual's metabolic profile [10, 14, 15].

Even though human urine is typically sterile, preanalytical variation in the urine metabolome can arise from contaminating human or bacterial cells, which if

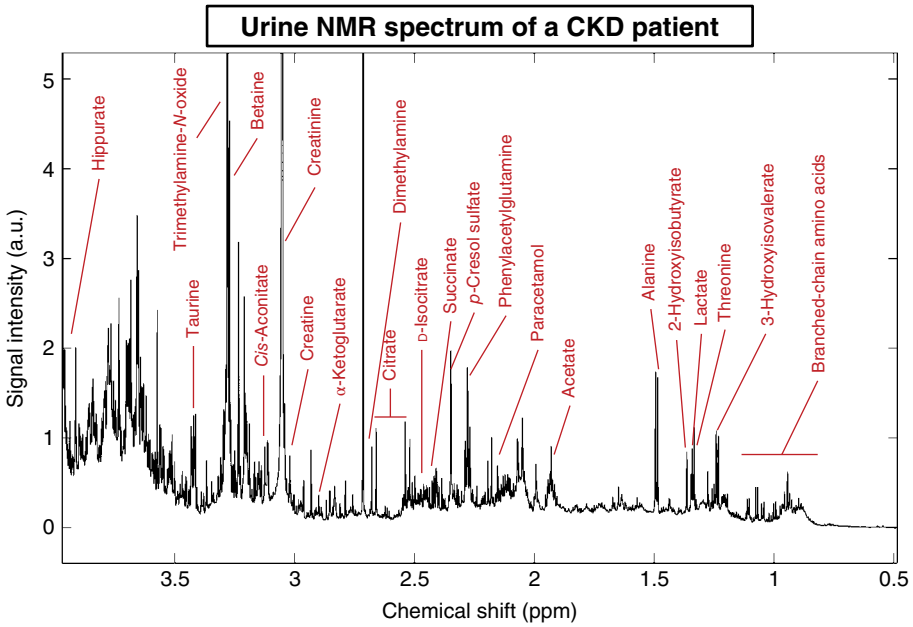


Figure 10.3 Example of a partial ^{1}H NMR of urine with selected metabolite annotations.

lysis and/or add secretions will contaminate the metabolome of the pure urine [10]. It can therefore be recommended that sodium azide be added before storage at 3 mM and also to have the clinician take the sample midstream [10, 16]. Additionally it could be suggested to use a 0.20 µm filter to remove contamination [9, 10]. However, it has been shown by Lauridsen et al. that when samples are stored at −25°C or below, there is minimal change in NMR urine metabolome with or without addition of preserving agents [17]. This group also reported that freeze-drying urine and reconstituting in D_2O (pH 7.4) results in the disappearance of creatinine CH_2 signals at δ 4.06 due to deuteration [17]. Another group has shown that with rat urine, the metabolome remains stable when stored at < −20°C for up to 2 years or 14 days when kept at 4°C [18]. Additionally, freeze–thawing appears to not have significant effects on structural integrity of a number of metabolites such as urea, citrate, and creatinine, with up to five freeze–thaw cycles [18]. However, one should always be cautious when considering freeze–thawing effects on the entire metabolome.

There are two main factors that can have a profound effect on the spectral analysis of the urine metabolome in NMR. These factors are pH and divalent cation concentrations, specifically Ca^{2+} and Mg^{2+} [10, 19, 20]. Both of these can have profound effects on metabolite signals, especially metabolites with a pK_a close to physiological conditions or with multiple ionizable groups [10, 19, 20]. These factors will contribute to what is known as "positional noise," [21] meaning that the chemical shift of metabolites can vary due to the aforementioned variations in the urine matrix composition. Positional noise will therefore bring problems in computational analysis of data, especially in the search for biomarkers.

Significant work has been done to buffer urine for NMR analysis and to create a consensus within the community, but variations do exist in practice. Urine samples are generally now mixed with a phosphate buffering system (pH 7.4) [19, 20, 22]. Phosphate buffer has now become standard practice, but there is not complete consensus on the appropriate phosphate concentration. Xiao et al. [19] systematically tested various phosphate buffer concentrations and found that a working stock of 150 mM K_2HPO_4/NaH_2PO_4 at a urine–buffer ratio of 10:1 was best for healthy urine samples [19]. This differed from conventional buffer protocols that ask for a urine–buffer ratio of 2:1 (0.2 M NaH_2PO_4/Na_2HPO_4 (pH 7.4)) [23]. The differences of the two buffer systems appear subtle, but potassium ions have higher water solubility relative to sodium, allowing for a more concentrated buffer and therefore less dilution of samples. One should also be aware that increasing salt concentration into the sample can affect the signal-to-noise (S/N) ratio [19] of NMR analysis.

The second important variable in urine is the variable inter-sample concentrations of Ca^{2+} and Mg^{2+}. This has been an acknowledged nuance with urine NMR studies, and only some research has been done to systematically address the problem and come to a general consensus. These divalent cations can form complexes with some metabolites, such as citrate and histidine, which create changes in the electrochemical environment of neighboring protons [17, 20]. Various techniques have been

proposed to neutralize these effects [20, 24], but thus far it appears to be an unresolved issue of measuring the urine metabolome via ^{1}H NMR.

EDTA has been proposed as a chelator of these metal ions [24]; however, the introduction of EDTA will add significant peaks for EDTA itself and its various complexes, distorting the spectra and overpowering the signals of interesting metabolites. Perhaps the most promising is a proposal of Jiang et al., [20] which suggests that an additional buffering step using potassium fluoride (KF), prior to phosphate buffering, can remove metal ions from the urine matrix [20]. In this technique, Ca^{2+} and Mg^{2+} bind fluoride with greater affinity than the metabolites of interest. Additionally, the KF solution does not introduce additional signals to the spectra. Overall, this method improves metabolite signal positioning on the chemical shift axis, especially with regard to citrate, limiting the inter-sample "positional noise" of metabolite signals.

10.2.1.2 Sample Preparation for Blood

For the preparation of blood serum, blood is collected into tubes with no additives. Blood plasma requires the addition of Li-heparin or EDTA (BD vacutainers) and is generally collected in 8 ml aliquots containing these anticoagulants [11]. Therefore it should be considered that when doing plasma NMR analysis with EDTA, a high-intensity NMR peak from EDTA and its complexes with Mg^{2+} and Ca^{2+} ions will appear. This EDTA signal will overlap the smaller signals of real metabolites, complicating the evaluation of NMR spectra [11]. Additionally, the separation of blood cells from the plasma should ideally be within 30 min of sample collection, and care should be taken in the centrifugation process. Do not exceed 1600 g and spin for only 15 min at 4 °C. Long-term storage of supernatant should be at −40°C [22]. Plasma or serum can then be mixed with 0.9% NaCl (w/v) in D_2O at a ratio of 1 : 2 (sample/saline) [22] and transferred into glass NMR tubes for analysis. TSP should not be used as a reference standard in samples with high protein content (plasma, serum) as this compound will bind to proteins and only a broadened chemical shift will be visible in the NMR spectrum. Furthermore, calibration of the plasma or serum spectra is generally done relative to beta-glucose (δ 4.64). Formic acid can also be used as an alternative internal reference (δ 8.45) [22].

10.2.1.3 Sample Preparation for Tissue

Tissue extracts such as kidney biopsies should be homogenized in organic and aqueous solvents (1 : 1) and then centrifuged to remove cellular debris. The supernatant must then be lyophilized as the solvents will disrupt the NMR analysis [22]. At this point there are options in extraction/homogenization methods depending on whether polar or nonpolar metabolites are of interest. For polar metabolites one can choose to use perchloric acid. Following homogenization the solution must be neutralized to pH 7.4 and then lyophilized [22]. Alternatively, one can use a liquid–liquid extraction (LLE) with methanol and chloroform to extract nonpolar metabolites. A detailed stepwise procedure is well described by Beckonert et al. [22] and can be further referenced if needed. No matter which extraction method was chosen, lyophilized samples should be reconstituted in NMR buffer with an internal reference standard [22].

10.2.1.4 Instrumental Setup

Several aspects should be considered when doing a global, nontargeted, and high-throughput NMR metabolomics study. Of utmost importance is the development of a standard operation procedure for sample preparation and assurance that instrumental parameters of the NMR are kept constant throughout the study analysis.

Plasma consists of 90% water, and urine has even higher. To allow a sensitive measurement of many metabolites in these biofluids and reduce interference and signal overload from water signals, H_2O proton signals must be suppressed via specific suppression methods in the NMR acquisition [22].

Internal reference compounds must also be used such as 3-trimethylsilypropionic acid (TSP), which gives a reference point for the chemical shift, from which other signals can be referenced to [10]. Other reference standards are 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) or tetramethylsilane (TMS) for use with organic solvents [10, 11, 22].

The temperature setting of the instrument is an important consideration. Urine is generally measured at 300–303 K and plasma/serum at 310 K. Temperature is particularly important for plasma/serum as large proteins, lipids, and lipoprotein form molecular aggregates to small molecules depending on temperature. Therefore the temperature should be kept around physiological levels [11]. Also, moieties containing amide groups can form hydrogen bonds when temperatures vary, which will cause variations in chemical shift [10].

Locking of the NMR via the introduction of the so-called field-frequency lock and the locking substance D_2O (for aqueous samples or MeOD as an organic solvent) compensates for slight drifts of the magnetic field. Furthermore, the magnetic field is slightly inhomogeneous along the whole magnet. These variations especially around the sample can be adjusted by shimming. This procedure is automated in modern NMR consoles, but manual shimming is recommended to maximize spectral quality [10, 11]. Shimming is assessed by reviewing the width of the reference peak (e.g., TSP/DSS) at half height and should not exceed 1.0 Hz [10]. This variable

is especially important in large sample sets and should be routinely monitored for quality control.

Depending on the biochemical question at hand, various pulse sequences have evolved. While one-pulse sequences (zgpr) and one-dimensional (1D) nuclear Overhauser enhancement spectroscopy (NOESY) presat (noesypr1d) give an overview of both small and large molecules in a given sample, CPMG presat (cpmgpr) focuses only on small molecules, and a diffusion-edited experiment (ledbpgppr2s1d) allows the measurement of only large molecules such as proteins and lipids without the need of preceding sample extraction. More details on the pulsing methods can be found in the review by Beckonert et al. [22].

2D experiments can also be set up with NMR and are mainly used to aid structural elucidation. J-resolved experiments (jresgpprqf) show the multiplicity of each peak and reduce peak overlap. Correlation spectroscopy (COSY) and total correlation spectroscopy (TOCSY) experiments can reveal connections between signals that are connected via spin couplings. Heteronuclear experiments such as the heteronuclear single quantum coherence (HSQC) experiment gives the ^{1}H and neighboring ^{13}C chemical shift. The latter experiment is also used in recent 2D NMR metabolomics approaches, with the advantage of substantial reduction in peak overlap and therefore a great potential for the detection of more metabolites and increased quantitative accuracy [25].

10.2.2 MS

MS offers relatively high sensitivity, combined with good selectivity. It also offers information on chemical structure via accurate mass, isotope distribution patterns, and characteristic fragment ions [26]. Today there are a variety of MS technologies that all have their respective advantages and disadvantages. These 3 main technologies of MS include ionization techniques, mass analyzers, and coupling techniques. Therefore, depending on the nature of the study and metabolites of interest, one will select a specific platform. However, a compromise between sensitivity, selectivity, cost, and speed is usually made [26].

10.2.2.1 Ionization

Electrospray ionization (ESI) is a soft ionization technique that has become a key ionization method of biological material over the past several decades. In a clinical setting, femtomole quantities of analytes in microliter sample volumes can be studied via this technique with excellent sensitivity and robustness [27]. This method uses electrical energy to assist in the transition of ions from the liquid phase to gaseous phase prior to spectral analysis. First, a fine spray of sample droplets is charged via capillary voltage. A nebulizing gas (e.g., nitrogen) assists in the formation of droplets and enhances sample flow rate. The charged droplets then exit the electrospray tip and across a pressure potential gradient [27]. Then, the solvent is gradually evaporated via heating temperature and nitrogen gas. The droplets are reduced in size till the Rayleigh stability limit is reached, at which a critical point is reached and the droplets of ions undergo Coulomb fission, releasing the ions into the gaseous phase [27]. Ions are then taken up into the sampling cone and accelerated into the mass analyzer. Bruins [28] can offer a detailed review of the technique. ESI can be performed in both negative and positive modes, generating negative and positive ion metabolites, respectively.

Similar to ESI, atmospheric pressure chemical ionization (APCI) is a soft technique that preserves the structural integrity of the analytes while achieving efficient levels of ionization [29]. APCI was first reported by Horning et al. over 30 year ago and has been well integrated into techniques such as liquid chromatography (LC)-MS [29]. With this technique a buffer gas is ionized by a beam of electrons accelerated in a high electron field. Then in series of reactions, which depends on the buffer gas composition, reagent ions are formed. The efficiency in the formation of reagent ions is a direct measure of the analyte ionization efficiency. However, limited efficiency in reagent ionization formation is one of the main drawbacks of this method [29]. Therefore in recent years atmospheric pressure photoionization (APPI) has been also developed to improve the technique [30]. This method is designed to directly photoionize organic compounds with high energy photons, bypassing the need for reagent ion formation.

Matrix-assisted laser desorption/ionization (MALDI) is also a widely used soft ionization technique in practice. However it is not normally used in conventional metabolomics, but rather in tissue imaging techniques [31], which has been discussed in a previous chapter.

The main ionization methods used in metabolomics are ESI and APCI. In general, ESI is suitable for semipolar and polar compounds, while APCI is recommended for neutral or less polar compounds [26]. Fairly recently Nordström et al. compared ESI, APCI, and a technique described as multimode (MM) ionization [32]. What can be determined from this study is that there is not a clear-cut, one-size-fits-all approach with ionization and metabolomics. When looking at total ion count produced by the two methods, APCI and ESI are comparable; however it appears that when measuring in negative mode, ESI may be more efficient [32]. One aspect that should be considered when trying to measure in positive mode is the formation of strong Na^{2+} adducts, which is not typically seen in APCI+ mode [32, 33]. There is difference as well in selectivity in positive mode between

the two methods [32]; therefore it is reasonable to select a method that fits to specific interests or use a combination of the two, if high throughput is not a concern.

10.2.2.2 Mass Analyzers

Metabolomic samples can be loaded into mass analyzers either by direct injection (DI) or coupled to a separation technique. Direct injection MS (DI-MS) methods of crude sample mixtures without coupled chromatographic separations are well suited for large clinical screening studies [26], but measurements can be heavily influenced by the matrix of the samples. However, the matrix of samples can be cleaned with improved sample extraction methods. Under these circumstances DI-MS can provide a global nontargeted metabolomic "fingerprinting" of sample sets [26, 34].

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) is an important and reliable ultra-high-resolution instrument for DI analysis. With resolution of (>1 000 000) and mass accuracy (<1 ppm), it offers a great way to efficiently and effectively examine the metabolic fingerprint of the sample [26]. The main functional part of FT-ICR-MS is the measuring cell (Penning trap). More than 10^6 ions can be trapped in the magnetic field, and then by applying an RF electric field, rotating ions induce image charges in detection electrodes, which are then amplified. The high resolution and mass accuracy is only possible with a complex understanding of ion motion dynamics and taking into account ion–ion interactions [35]. A detail description of the theory behind the instrumentation can be found in recent reviews (Nikolaev et al. [35] or Marshall et al. [36]).

As an example of the utility of FT-ICR-MS, Han et al. were able to detect 570 distinct metabolite features in mouse serum by monoisotopic mass within a range of m/z 90–570 [37]. In addition, they observed numerous metabolites that clustered around a single nominal mass, indicating that chromatography prior to mass analysis may not be even necessary with the given instrumental setup [37]. Furthermore, with proper internal calibration, FT-ICR-MS has shown to have mass accuracies within 0.2 ppm for most metabolites and 0.65 ppm for all metabolites. At this level of accuracy, when combining with metabolite databases and computational molecular formula techniques, most measured metabolites could be identified by mass alone without fragmentation [37]. The FT-ICR-MS has great potential for being used routinely in urine metabolomics studies.

The Orbitrap is a recent mass analyzer that, similar to FT-ICR, uses an electrostatic field to trap ions [26]. A more detailed explanation of orbital trapping can be found in the manuscript of Makarov [38]. Typical resolving power of Orbitrap mass spectrometers is about 150 000 and mass accuracy of 1–5 ppm [26], and they have been used in various studies of metabolism. The lower resolution compared with FT-ICR is compensated for by a more compact and simple design [38].

As part of the ion trapping family, multiple-pass time-of-flight (TOF) MS instruments have been used as 3D and linear ion traps for DI-MS analysis [26]. In linear trapping, ions are introduced and trapped between two ion mirrors coaxially located to the ion beam. In order to load the ions, one mirror is switched off and switched on before the ion can exit the trap [38]. Resolving power is typically around 6 000–17 000 and mass accuracy can be (<5 ppm) [34]. However, relative to the previously mentioned ion traps, TOF instruments have limited resolving power and mass accuracy, which has limited their role in DI-MS studies [26] and their ability to distinguish isobaric ions [34]. This instrument is mainly utilized with a coupled chromatography technique.

DI-MS has been applied in a variety of metabolomics studies. These instrumental methods are particularly useful in global metabolic fingerprinting techniques where case versus control can be distinguished by review of the total measurable metabolome rather than by a single metabolic biomarker. The main challenge for this type of instrumentation is its susceptibility to ion suppression/enhancement due to the lack of sample pretreatment. With the case of urine or blood plasma for CKD studies, high salt concentration will result in salt adduct formation. Furthermore, the formation of ion products and the differentiation of isomers can be a real challenge in post-analysis data evaluation [26, 34].

It should also be mentioned that DI of samples with high protein content can also rapidly deteriorate mechanical components of the instrument [34]. This can obviously have deleterious effects on the hardware and data quality for any large sample set analysis. However such problems can be reduced with sample preparation/metabolite extraction methods. Again, one must always compromise between sensitivity, selectivity, cost, and speed.

10.2.2.3 Coupled Separation Methods

Coupling chromatography separation to MS has been widely used in metabolomics both online and offline. Chromatographic separation can provide advantages such as reduction of matrix effects, ionization suppression and separation of isomers; while adding additional orthogonal data (i.e., retention time). Also important is the added ability to more accurately quantify individual metabolites [26]. Three main separation technologies are used with MS: gas chromatography (GC), LC, and capillary electrophoresis (CE).

GC has often been coupled to single quadruple MS detectors, with low cost, high sensitivity, and wide dynamic range. However, this method suffers from slower

scan rates and run times (40–60 min/sample) as well as reduced mass accuracy relative to TOF analyzers (<10 min/sample) [26, 34]. On the other hand, reliability, robustness, and affordability have reserved a place for GC/quadrupole-MS in many analytical laboratories.

GC-MS has one prerequisite, which is the need for volatile, thermally stable analytes [26, 34]. This means that highly polar metabolites usually require a derivatization step to replace a functional group with a more stable moiety. Finding the right method for a specific sample can be complicated and often artifacts can be formed that complicates data interpretation [34]. With reference to CKD studies, biofluid sample preparation for GC often includes a lyophilization step, resulting in concentrated sample matrix components that can cause column and detector overload [34, 39]. Other complications such as incomplete compound identification are common in global nontargeted metabolomics of urine. Other factors like spectral overlapping, incomplete separation, poor S/N, and spectral artifacts distort results [13]. However, 179 compounds (89 unique) have been reported using GC-MS [13].

LC is a good complement to GC as it can handle polar, nonvolatile compounds and has been integrated into targeted and nontargeted metabolomics approaches. In addition to compound separation, LC can reduce ion suppression and decrease background noise. Two main methods are widely used for metabolomics: normal phase (NP) and reverse phase (RP) [26, 34, 40]. Most commonly found are C_{18} and C_8 RP columns [26]. Conventional C_{18} columns use a particle size of 3–5 μm, which can have insufficient separation and relatively poor resolution with complex mixtures such as biofluids [34]. Newer technology has given way to ultra-performance liquid chromatography (UPLC), which utilizes <2 μm particles. Analysis of rat urine using this method increased resolution and detection limits [41].

RP is a standard separation tool for moderately polar to nonpolar analytes; unfortunately highly polar compounds are not retained and are eluted within the initial void volume [34]. NP columns provide just the opposite kind of interactions with analytes. With NP, highly polar metabolites are retained; therefore this method is well suited for biofluids such as urine [26]. Hydrophilic interaction liquid chromatography (HILIC); a form of NP, is now becoming a widely used technique for urine metabolomics. A key characteristic of HILIC is that the mobile phase is a water-miscible solvent such as methanol [42]. This method relies on a stationary phase of sulfoalkylbetaine with zwitterionic properties and is capable of binding water so that the interface with metabolites separates compounds based on polarity and charge. The nature of the mobile phase also makes it suitable for coupling to ESI-MS [42].

CE is another separation technique utilized in biofluid metabolomics. This technique offers the separation of metabolites based on charge and size. One major advantage is the short analysis time and small sample volume requirement, at just 1–20 nl [40]. CE-MS has been used for both targeted and nontargeted analysis, detecting a wide range of metabolites [40]. Hybrid techniques had also been created from CE-based separations, such as capillary electrochromatography (CEC), which uses capillary columns that are packed with LC stationary phase materials [34]. This technique has been reported to have been coupled with MS to analyze compounds such as proteins and peptides, as well as amino acids and carbohydrates [43].

One general concern with metabolomics is the lack of complete consensus on a standardized method, and currently researchers typically customize a method for their particular needs. However, some work has been done to optimize analytical methods specifically for metabolomics [44]. In 2009, Büscher et al. designed a systematic study to review the three main coupled MS platforms with a quantitative metabolic focus [44]. They used a reference mixture of 75 metabolite compound groups with distinct molecular weights (MW), coming from a wide range of biochemical pathways. The three platforms were LC, GC, and CE, all of which were coupled to TOF-MS detectors [44]. Within each platform, two variations of separation where used. For GC evaluation, both trimethylsilyl (TMS) and *tert*-butyldimethylsilyl (TBDMS) derivatization methods were used. For LC, ion pairing and HILIC were used, and with CE, cationic and anionic separations were performed. They found that of the 75 compound groups, 72 could be measured on at least one platform, while 33 could be measured by all 3 [44].

In general terms, sensitivity, time, and reproducibility were evaluated, and of the three platforms, it was determined that LC-TOF-MS was the best option for coverage and robustness. If a second platform could be integrated, it was shown that GC-TOF-MS would complement the performance of LC best. CE-TOF-MS was found to have comparable sensitivity and separation but lacked robustness, as retention times of CE vary as much as 3% from sample to sample [44]. Matrix effects were observed in the three platforms; thus the use of radiolabeled internal standards (i.e., ^{13}C biomass) is recommended for proper quantification [44]. It should be noted that no DI-MS techniques were evaluated here. See Table 10.1 for an overall comparison of analytical techniques.

10.2.2.4 MS Sample Pretreatment Techniques

Sample collection for a metabolomics analysis is challenging due to the rapid dynamic changes that characterize the human metabolome. Therefore, care should be given

Table 10.1 A general comparison of analytical techniques in metabolomics.

Platform	Strengths	Weaknesses
Nuclear magnetic resonance (NMR) spectroscopy	• Highly robust • Nondestruction of sample • Relatively simple sample preparations • Compatible with automation • Short sample run time potential • Quantitative and qualitative • Variety of pulse sequences can be utilized for focused analyses • 2D NMR for structural elucidation	• Relatively insensitive when compared with MS • High initial cost • Peak overlap reduces metabolite resolution
Direct injection mass spectrometry (DI-MS)	• Low sample volume needed • High sensitivity • Ion fragmentation • Compatible with automation • Robust quantification • Limited carryover or cross-contamination • Sample run time	• Issues with specificity • Unable to separate isomers and isobars alone (this can be mitigated with super high-resolution instruments) (i.e., FT-ICR-MS) • Stable isotope-labeled standards needed for absolute quantification • Can require time-consuming sample preparation
Liquid chromatography–mass spectrometry (LC-MS)	• High sensitivity • Good reproducibility • Good for nonvolatile compounds • Can be optimized for polar or nonpolar separation	• Larger sample analysis times • Lower resolution • Large quantities of solvents needed • Can be susceptible to batch effects
Gas chromatography mass–spectrometry (GC-MS)	• Comparable sensitivity • Good reproducibility • Good resolution • Faster analysis run times well suited for volatile compounds, as well as amino acids, organic acid, and lipids • Low initial cost	• Derivatization steps needed, which is costly and time consuming • Relatively good resolving power • Not well suited for high throughput

and steps taken to ensure that any kind of potential biochemical activity is halted upon sample collection. This is known as quenching [45], which relies on rapid inactivation of enzymatic activity. Methods for this can vary depending on the sample type, but aside from the host organism itself, microbial quenching should also be considered. A more detailed outline of quenching techniques is described in a publication by Mushtaq et al. [45].

A sample pretreatment step is essential in clinical metabolomics, as biological samples such as plasma, serum, cell cultures, and tissue are protein and salt rich [46]. Both protein and salts will interfere with metabolite detection and lead to poor data quality, especially for larger automated studies. Healthy urine generally contains low protein amounts, so it can simply be centrifuged and diluted before DI. In contrast, urine from CKD patients usually has a higher protein concentration, can contain blood cells, and has variable osmolarity; therefore a more intensive sample preparation step is needed [47].

Here we will discuss deproteinization and extraction methods that will efficiently remove proteins and salts, but preserve the metabolites of interest. It should be noted that any kind of sample preparation will result in some analyte loss [46]. Sample preparation methods are much like what has been discussed thus far, in that a compromise must always be considered depending on the nature of the study and analytes of interest.

10.2.3 Protein Removal (PPT)

The simplest way for protein removal in biofluids would be to dilute the sample in an organic solvent, centrifuge, and then remove the supernatant for analysis. However, organic solvents will not remove phospholipids and salts that will cause ion suppression [46]. This is also the case with ultrafiltration methods [48]; therefore, one should consider a more robust technique to remove proteins and salts together, especially with blood-derived samples.

10.2.4 LLE

LLE enables the separation of polar and nonpolar metabolites into aqueous and organic phases. Each phase can then be analyzed separately and this can be a great method for nontargeted studies. As this method is time consuming and requires standardized handling practice [46], new LLE technologies have been developed to make this method considerably faster than conventional techniques. The introduction of supported liquid extraction (SLE) plates (96-well) has dramatically decreased the time needed for LLE and can potentially have similar extraction capabilities as traditional methods. However, these plates have primarily been used in the analysis of pharmaceutical compounds in drug development studies [49, 50]. Despite the limited uses thus far, the SLE plates do appear to have recovery rates up to around 93% [50] and removal rate of 99% of phospholipids [49].

What is important to remember about these plates is the choice of solvent. Solvents must be water immiscible, such as methyl *tert*-butyl ether, chloroform, or ethyl acetate. These solvents are low on the polarity index scale [51], meaning that extractions with these methods will be focused on moderate to nonpolar metabolites. Application for these plates may be best for plasma/serum or perhaps urine from end-stage renal failure (ESRD) patients with high protein concentration. A more detailed review of applications for SLE plates can be found in the manuscript by Raterink et al. [46] as well as from manufactures.

10.2.5 Solid-Phase Extraction (SPE)

Solid-phase extraction (SPE) is a well-established extraction method for the enrichment of analytes and removal of interfering compounds. Like with LC, there are varieties in solid-phase material that have varying types of chemical interactions with analytes. These interactions can include weak/strong cation/anion, reverse phase (C_{18} or C_8), and HILIC. However, due to the selectivity of the various sorbents, obtaining a wide range of metabolite coverage is challenging. Therefore these methods may not be optimal for global untargeted studies [46].

HILIC-SPE may be the most interesting approach for analyzing urine samples, as it allows enrichment of polar metabolites that are abundant in urine. A drawback of HILIC-SPE, and SPE in general, is that this method can be laboratory intensive and requires a considerable amount of time for sample preparation [52]. There are now SPE 96-well plates available [46] that can add speed and automation, but currently HILIC-SPE is available only in single cartridge form. Therefore, HILIC-SPE may be suitable for lower-throughput studies.

Questions have been posed about the reliability of metabolomic data across different analytical platforms and laboratories. A recent study by Martin et al. in 2014 addressed just this issue [44]. In this large interlaboratory study, 5 NMR instruments and 11 LC-MS instruments (Orbitrap, TOF, QTOF) where used to investigate the robustness of nontargeted methods [53]. In test 1, urine samples from adult volunteers, spiked and non-spiked with 32 metabolite standards, were analyzed. In test 2, plasma of rats with normal diet and a supplemented diet were analyzed. All samples were measured across all platforms [53]. Researchers concluded that there is high convergence in the spectral information produced, regardless of instrument, standardization, and deconvolution methods used. Methods to identify and match individual metabolites are being explored more intensively [53]. What studies like this reveal is that nontargeted metabolomic techniques can be used to generate hypotheses relevant to CKD research. These methods will only increase in robustness, as instrumentation and computational methods are perfected and tailored for the tasks.

10.3 Statistical Tools and Systems Integration

Methods for handling and interpretation of metabolomics data can be as diverse as the instrumentation that generates the data. This depends on the nature of the study, which can be broken down into two major approaches.

Targeted metabolomics is the quantitative analysis of a defined group of metabolites that are involved in one or a few related metabolic pathways. Examples are the analysis of all biogenic amino acids or metabolites from the tricarboxylic acid (TCA) cycle. This method in particular is best suited for robust quantification of specific known metabolites of interest, and protocols need to be adopted or newly developed for each set of metabolite targets [34, 54]. Quantification in metabolomics is critical for understanding biological processes; however this can be challenging. A major obstacle is that a metabolite signal is dependent not only on its concentration but also on its structure and the nature of the sample matrix [26]. For this reason, it can be said that the absolute quantification of metabolites is a "slow-lane" [6, 55] approach. This method could also be described as hypothesis driven [54], such as the use of creatinine in the determination of eGFR.

The alternative method would be nontargeted acquisition of metabolomics data from biological samples to classify changes between sample groups and to look for sample clustering [34]. This will utilize NMR and high-resolution MS instrumentation [34]. This method is unbiased, global, nontargeted, and well suited for

hypothesis generation. This approach has also been coined as "fast lane" [6, 55]. It gives the unique opportunity to discover novel and previously unexpected metabolites; while in parallel covering a large set of common metabolites.

10.3.1 Post-Measurement Spectral Processing

Prior to any statistical analysis, spectral processing must be performed, whether the data is generated from NMR or MS. Data processing can include techniques like spectral alignment and normalization [1]. This guarantees that the data obtained from a particular feature are the same throughout the sample set and are generally arranged in a feature quantification matrix (FQM) [1]. Furthermore, metabolomics data is asymmetrical; therefore spectral processing techniques like mean centering and scaling are needed [54].

10.3.2 Spectral Alignment

Spectral alignment is a key process in metabolomics studies that involves multiple samples. Peaks or features of the same metabolite may differ in position on the x-axis due to nonlinear shifts and matrix affects [19]. In NMR, this is due to ionic conditions, pH, or protein content of the biological sample. In coupled MS techniques, it can be due to changes in retention time associated with variation in the solid phases of chromatography [56].

Spectral alignment methods can be classified in two ways. The first technique known as warping is a nonlinear transformation of the x-axis in NMR and MS data in order to maximize the correlation between spectra. Alignments are done by either stretching or shrinking segments of the spectra in order to maximize correlation [1]. The two most common techniques of warping are correlation optimized warping (COW) and dynamic time warping (DTW) [1], and a more detailed description of these methods can be found in Tomasi et al. [57].

Segmental alignments, on the other hand, apply a unified shift to all spectral points [1]. One of the most used algorithms is the *icoshift* algorithm [58], which is based on the convergence of a reference signal using fast Fourier transform (FFT) and automatic segmentation methods [21]. Computational bias can be introduced during this process [59], and therefore care should be taken when using a reference based alignment, but nevertheless, segment alignment has been proven to be more effective than warping techniques, regardless of instrumentation used [58–60].

Other methods such as recursive segment-wise peak alignment (RSPA) [21] have also been developed to increase interpretability and robustness of spectral data, primarily for NMR data. This method reduces peak positional noise in spectra by refining a segment of reference and testing spectra in a top-down fashion, further subdividing as needed to maximize alignment. Then aligned spectral segments can be rejoined [21], as shown in Figure 10.4 (*alignment methods diagram*).

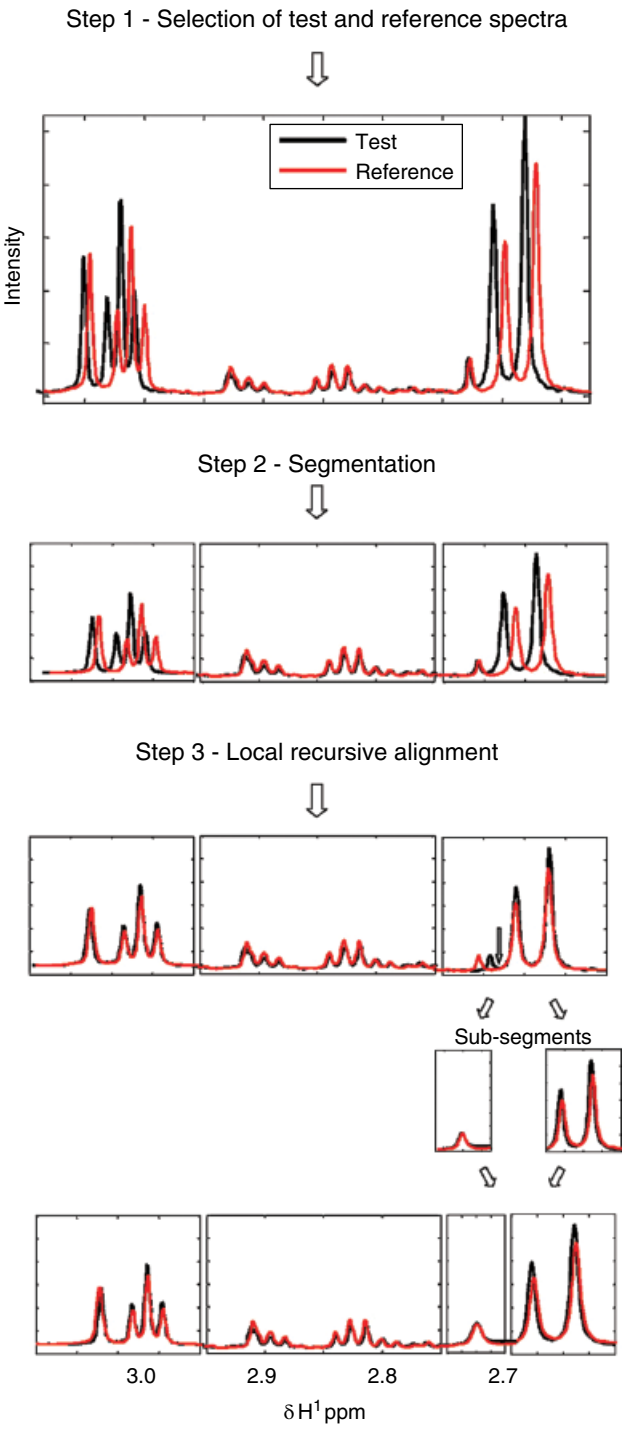


Figure 10.4 RSPA alignment scheme. *Source:* Reprinted (adapted) with permission for Veselkov et al. [21]. Copyright 2009 American Chemical Society.

10.3.3 Normalization and Scaling

In order to accurately quantify features in metabolomics data, a normalization step is required. The goal of normalization is to remove variation in overall metabolite concentration between samples introduced by various physiological factors [1]. For example, in urine this could be the patient hydration status or perhaps fasting/feeding states with regard to plasma samples. Many methods exist for normalization [61]. Two of the most common are (i) normalization to single endogenous stable metabolites (i.e., creatinine) and (ii) total spectral area under the curve (AUC) normalization [1]. It should be mentioned here that even though creatinine normalization is frequently used in routine clinical chemistry, it is not recommended for CKD studies, as creatinine excretion levels vary with disease progression.

A review by Kohl et al., in which normalization methods were compared on samples from autosomal dominant polycystic kidney disease (ADPKD) patients, revealed that probabilistic quotient normalization (PQN) to be the most robust of the methods under review [61]. This method is based on the calculation of the most probable dilution factor by examining the distribution of the quotients of the amplitudes of a test spectrum with those of a reference spectrum [62]. The reference spectrum can be defined as single spectrum of the study, a "golden" reference spectrum from a database, or a calculated median or mean spectrum [62].

Each metabolite has a different mean value and different margin of variation. To ensure that low-abundant and high-abundant metabolite contribute equally to metabolomics data, the matrix needs to be mean centered and scaled or transformed. The most commonly applied scaling methods in metabolomics are autoscaling, Pareto scaling, range scaling, and vast scaling [63]. An extensive review of different methods was done by van der Berg et al. [63].

10.3.4 Peak Versus Feature Detection

Deconvolution strategies can reduce the complexity of data by removing spectral noise, multiply charged species, cluster, and adducts [34]. Deconvolution can also make quantification of metabolomics data easier and can be useful when metabolite signals overlap. However one must have some prior knowledge of the compounds present in a sample [1].

The purpose of feature detection is to identify and quantify features present on a spectrum. However this approach has drawbacks such as when features overlap, especially in NMR data [1]. Also, due to poor spectral alignments, these methods simply do not perform efficiently as peak-based methods [1].

Peak-based detection methods use algorithms to analyze each sample spectrum independently [1, 64]. In the first step, the spectra are smoothed, a process that requires significant computer resources. In the second step, metabolite peaks are identified using one or more detection thresholds. These thresholds can be parameters such as S/N ratios, peak intensity, or a frequency filter in which the peak must show on a certain percentage of spectra to be significant [1, 65].

10.3.5 Data Analysis

Once the dataset is properly preprocessed by methods just previously discussed, univariate and multivariate statistical tools can be applied. What tools used will be influenced by the nature of the study and its design, whether it is targeted or nontargeted in nature.

Generally speaking, univariate statistical methods analyze metabolomic features independently [1] and are used for targeted biomarker discovery when the chemical class of the compound of interest is known [34, 54]. This slow -lane approach works toward a single metabolite marker detection and absolute quantification. Such quantitative data can offer information to answer clinical questions such as the determination of eGFR based on creatinine measurements [6, 55]. Univariate methods for this include ANOVA, Mann–Whitney U test, Student's t-test, Wilcoxon signed-rank test, and logistic regression [1, 34, 54].

In contrast to univariate methods, multivariate methods take into account all metabolite features simultaneously [1] and attempt to identify patters and clusters in metabolite classes related to pathological features. This can be particularly important, because a single biomarker may not be specific to a single disease [54]. This can be considered a fast-lane approach, because metabolite identification may not even be necessary to distinguish case and control. Here one can simply determine the health status of a patient, relying only on spectral information [6, 55]. Furthermore, these approaches can be great tools for hypothesis generation. Multivariate pattern recognition tools can follow two approaches: supervised and unsupervised analyses [1, 34].

10.3.6 Unsupervised

Unsupervised tools like principal component analysis (PCA) and hierarchical clustering analysis (HCA) are good ways in which one can summarize complex metabolomics data and create an overview of the data structure. These methods provide effective ways to reduce high-dimensional data into fewer dimensions and allow a first look at patterns and sample cluster characteristics of the whole dataset [1].

PCA is perhaps the most popular of unsupervised tools. It is based on the linear transformation of metabolic features into groups of orthogonal variables called principal components. After principal components have been assigned, loading and score vectors are obtained. Loading vectors represent the principal components that correspond to the individual contribution of each variable to the principal component. Score vectors represent the projection of each sample into the new orthogonal matrix and represent each sample to the principal component [1, 66]. When this is plotted, variable components in multidimensional data can be easily visualized in either 2D or 3D plots [34]. This can also be a great tool for the detection of outliners or for assessing the impact of technical issues on datasets.

HCA can complement PCA, in that this tool can detect nonlinear trends in data that would otherwise be missed by PCA [1]. HCA provides a powerful tool to visualize clustering features in metabolomics data based on similarity/dissimilarity of features in a distance plot [67]. Self-organizing maps (SOM) [68] are also applied to metabolomics datasets as a great way to visualize phenotypes as well as prioritize metabolites of interest based on similarities [1].

10.3.7 Supervised

Supervised methods, which can build from knowledge provided by unsupervised tools, are used in CKD metabolomics to correlate known phenotypic variables of choice with metabolic data. This allows the elucidation of potential discriminating features in the dataset.

Partial least squares (PLS) analysis is a commonly used tool for supervised statistical analysis of metabolomics data. As PCA seeks to find the maximal variation in datasets for separation, PLS seeks to extract variation in the data that is related to a specified sample class or a variable of discrimination [69]. In other words, a feature coefficient (loadings) of PLS components represents a measure of the contribution of that feature to discrimination of sample groups [1]. PLS can be used as a regression analysis [1], for example, to relate metabolites to CKD stage or age associations, or as PLS-DA (discriminate analysis) as mentioned previously, to discriminate between sample groups (e.g., healthy and CKD) or even as a predictor of groups [69]. Figure 10.5 show a comparison of PCA and PLS.

An extension to the classical PLS method, O-PLS splits up the data variation into the variance of interest and the orthogonal part of the data that is unrelated to the parameter of interest [70]. This simplifies data interpretation [69]. In such supervised approaches, the validity of the model needs to be assessed to avoid overfitting of the model. This can be done by carrying out a sevenfold cross-validation, which retains the explained variance in relation to the group descriptor (R^2Y, which should be close to 1) and the predictive ability of the model (Q^2Y, which should be >0). Visualization of both PLS and O-PLS analysis can be done via plotting the scores plot, similar to the previously described PCA scores plot. In the case of NMR data, a straightforward technique has been developed by Cloarec et al. where the loading plots are visualized as pseudo-NMR spectra and the R^2 values represented by a color coding system that highlights all correlating features represented by the variable weights and their direction of correlation shown as a covariance plot [71]. Figure 10.5 illustrates such plots from an O-PLS-DA approach in CKD.

In addition, an NMR data analysis technique known as statistical total correlation spectroscopy (STOSCY) [71] aids metabolite identification via highlighting metabolites that have correlating features in the NMR spectrum, and therefore have structural similarities, or are generated from the same biological pathway. Then if combined with O-PLS-DA, one can highlight peaks with strong correlation to a determined parameter or sample class (i.e., CKD stage) (see Figure 10.5). This can show negative and positive correlations simultaneously of all peaks in the pseudo-spectra.

This is not a comprehensive review of statistical methods in metabolomics and only some selected methods that are popularly used were presented. A detailed review of the topic was written by Bartel et al. [69].

10.3.8 Spectral Databases and Metabolite Identification

Currently the most common method for peak assignment in metabolomics is to query the feature against a known database; whether NMR or MS data is being analyzed. Perhaps the most extensive metabolomics database is the Human Metabolome Database (HMDB), which stores more than 40 000 metabolites with chemical descriptions, metadata, and MS/NMR spectra [1]. In MS-based studies certain peaks of interest can be search based on the m/z and whether it was measured in positive or negative mode. Also ionized adducts (H, Na, K, Cl, etc.) can also be considered with a certain ppm range of error [1]. Regarding high-resolution instruments such as FT-ICR-MS, certain filtering programs such as NetCalc have been developed by Tziotis et al. This algorithm can make preliminary annotations of between 40 and 60% of datasets [72], which can dramatically speed up the peak annotation process. In NMR data, similar database queries are used. Software programs do exist such as MetaboHunter [73] and FOCUS [59] that can aid metabolite identification in an automated manner.

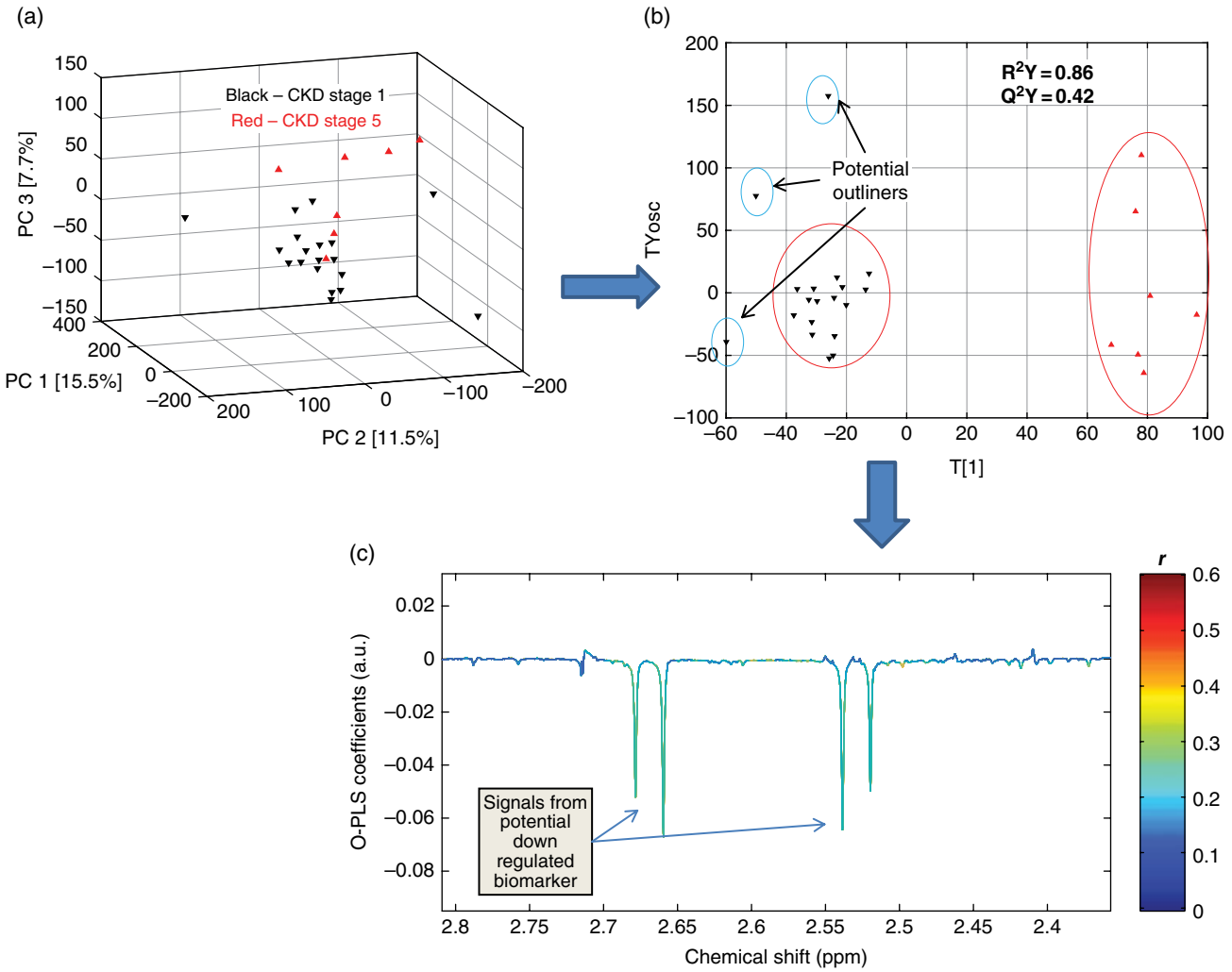


Figure 10.5 (a) PCA plot of urine samples from CKD patients. (b) Orthogonal partial least squares discriminant analysis (O-PLS-DA) scores plot of the same sample set with CKD stage 1 and 5 being the discriminating factor. (c) O-PLS-DA loading plot. The color bar represents the correlation coefficient of the corresponding variable in discriminating between the groups. On the *y*-axis, the O-PLS covariance measures the degree and direction of a given peak with that of variable. The layout of the plot resembles the original metric of the NMR spectra, meaning stereochemistry of the plot is preserved and is therefore straightforward to interpret for an NMR spectroscopist. This supervised analysis allows a more targeted search for biomarkers of a certain predetermined variable.

10.3.9 Pathway Analysis

It is only very recently that metabolite relationships and pathway analysis could be done in a single measurement. Groups of metabolites deriving from the sample biochemical pathway can now be analyzed due to the vast amounts of data being generated from metabolomics studies. Databases for pathway analysis are now routinely used for the incorporation of metabolomics data into the context of systems biology. Some of the main databases are Kyoto Encyclopedia of Genes and Genomes (KEGG), the Small Molecule Pathway Database (SMPDB), and MetaCyc, just to name a few [1]. This topic cannot be fully covered in this chapter, but methods for pathway analysis known as metabolite set enrichment analysis (MSEA) and gene set enrichment analysis (GSEA) are some of the main approaches to pathway analysis [1, 74]. Further information on MSEA can be found in the review from Khatri et al. [74].

10.3.10 Validation and Performance Assessment

Statistical methods for validation of findings are critical to biomarker discovery [54] and for therapeutic target detection in toxicometabolomics [75]. Sensitivity and specificity of the prediction model must be cross-validated (e.g., sevenfold cross-validation on PLS models) to ensure the model is not overfitted by use of independent validation sets with healthy control and the disease group.

If also needed, an additional related disease control group with similar phenotype and disease can be used [54]. The most used performance method for model testing is receiver operating characteristic (ROC), which is a nonparametric procedure comparing specificity against sensitivity according to a specific boundary [1, 76].

Common pitfalls such as poor data filtering and normalization techniques should be considered prior to performance assessment [76], and it should also be noted that statistical analysis of data can never compensate for poor study design. Care should be given to randomization, specimen collection, handling, and storage. Additionally, adequate sample size is required in order to make statistical inferences and validate biomarkers [54].

10.3.11 Application into Systems Biology

Integrating metabolomics data with other -omics data for drug discovery is a challenge of the future. This topic cannot be discussed here in detail, but it is recognized that in health and disease, metabolomics must find a place within systems biology. Metabolomics data can be utilized in a "systems pharmacology" approach [77], where multiple therapeutic targets can be analyzed and reviewed for effectiveness. The workflow for this approach incorporates full -omics modeling, with emphasis on maximum correlation between gene expression [77] and metabolic flux [78]. This analysis results in generation of predictive models that can lead to therapeutic approaches within the context of personalized medicine. Personalized medicine, albeit a term now more frequently used, is still loosely defined. Different definitions of personalized medicine are currently being used [79], probably in part due to the different specialized fields of molecular biology. The various -omics fields rightfully have some claim, but can a single -omics technique be considered a global personalized medical approach as a stand-alone field of research?

In the context of metabolomics, a personalized metabolic assessment of a patient cannot be based on a single biomarker [3], such as creatinine clearance. Standing alone this fails to answer questions about why kidneys are failing in the first place. Currently in clinical healthcare management, as well as research, single biomarker endpoint compounds are being used for a variety of applications [3]. This approach may be sufficient for bacterial infections or toxicity, but is it effective in assessing complex diseases, like CKD?

The current model for disease diagnosis and prognosis should be reconsidered as global nontargeted methods yield viable alternatives. This should also be the case with metabolomics and CKD. One can then foresee a future when a person in a healthy state has their metabolome assessed (e.g., NMR/MS) and then periodically monitored or when disease is suspected. If abnormalities are noticed by a nontargeted metabolic analysis, a targeted analysis could then be implemented to verify changes in metabolite biomarkers. Combined with metabolic flux approaches [80–82] and other -omics data (e.g., genomics, proteomics), total body metabolic assessment could be available for clinicians in order to help them implement targeted therapies.

10.4 Metabolomics in CKD

Mankind has been using metabolites for diagnostic purposes since ancient times by tasting urine for glucose, an example of targeted metabolomics [5]. Characterized by a persistently low GFR [83], CKD is diagnosed if GFR is less than 60 ml/min/1.73 m^2 for 3 months or more [84]. GFR is estimated from serum creatinine and MDRD or Cockcroft–Gault formula [84]. CKD can also be diagnosed via albumin-to-creatinine ratio >30 mg/g in two of three spot urine specimens [84], as well as through serum cystatin C measurements. Perhaps a panel of metabolites could offer a better and more personalized diagnosis of CKD. Recent studies [85–88] have revealed potential biomarkers of CKD and related diseases and have added valuable data to metabolomics databases. Perhaps most important is the search for early-stage biomarkers [87] that will give clinicians a head start on treatments in order to halt CKD progression. Also important are biomarkers indicative of other CKD-related diseases such as type 2 diabetes [85] and stage-related progression biomarkers [86, 88]. In this section the focus is on several interesting topics in relation to specific metabolic pathways that are impaired in CKD.

10.4.1 Uremic Toxins and New Biomarkers of eGFR and CKD Stage

Uremic syndrome is the progressive retention of compounds that normally are excreted by healthy kidneys. These compounds are known as uremic toxins (UTs) which they negatively interact with physiological functions of the body [89, 90]. Vanholder et al. [89] in a comprehensive review of past studies, created an in-depth, systematic overview of 55 publications. This is important especially when concerning ESRD and the need to have improved dialysis technology to remove UTs [89]. In their publication [89], a database of 857 publications between 1966 and 2002, was considered, with all data being taken from plasma/serum concentrations. The intent was to create a database of median/mean uremic concentrations (C_U) reported and compare those values with a normal uremic concentration (C_N). In addition,

care was taken to subdivide UTs into three major classes: (i) small solutes (MW < 500 Da) with no known protein binding, (ii) solutes with known or likely protein binding, and (iii) medium MW molecules (≥500 Da) [89]. This work offers a useful database for future metabolomics studies.

In recent years analytical techniques, mainly NMR and MS, have been key in detecting and quantifying UTs and their relation to CKD and its progression. Toyohara et al. [91] were able to detect 64 UTs via CE-MS in plasma and had significantly altered concentration as eGFR decreased. These results identify a number of uremic compounds that may predict deteriorating renal function and provide diagnostic information for therapies [91].

Previously, the same group demonstrated the accumulation of UTs during ESRD due to inactivation of the SLCO4C1 organic anion transporter. This transporter has been shown to excrete UTs, and its inactivation leads to the accumulation of UTs as CKD progresses [92]. In this study 41 CKD patents were assessed for eGFR by MDRD, and Spearman's rank correlation was calculated for UT detected via CE-MS [91]. Among their finding are some UTs that may be used for early detection of CKD and correlated well with conventional eGFR measurements. Some of these UTs include 1-methyladenosine, *N*-acetylglucosamine, gamma-butyrobetaine, sebacic acid, *cis*-aconitate, and homovanillate [91]. However an admitted drawback of this study was that metabolomics data were not adjusted and correlated with age, gender, BMI, and lifestyle factors [91].

10.4.2 Dimethylarginine

One UT that deserves special attention is asymmetric dimethylarginine (ADMA). This amino acid inhibits nitric oxide (NO) synthase and in high concentrations can cause a signification decrease in NO production. Kidney damage and endothelial dysfunction are associated with decreased NO production [93, 94]. Ravani et al. [94] in a cohort of 131 patients showed the potential for ADMA to be a predictor of morbidity and CKD progression to ESRD. Their findings showed plasma ADMA to be inversely related to GFR. A follow-up study showed ADMA to predict morbidity independent of hemoglobin, GFR, and proteinuria [94]. However limitations of this study were that it was a single-center study, with limited number of participants and high average age (71 years) [94]. Despite these limitations, the study indicates the potential of metabolites for predicting CKD progression.

Another study including 227 patients aged between 18 and 65 years old, with nondiabetic CKD showed significant correlations of ADMA with serum creatinine ($r = 0.595$), GFR ($r = -0.591$), parathyroid hormone ($r = 0.586$), hemoglobin ($r = 0.336$), age ($r = 0.281$), proteinuria ($r = 0.184$), and uric acid ($r = 0.177$; all P 0.01) [95].

ADMA was also shown to be a useful biomarker in diabetic-related CKD. In two independent studies with diverse cohorts ($n > 200$), results indicate that plasma and/or serum levels of AMDA could accurately predict CKD progression in patients with either type 1 or 2 diabetes [96, 97].

10.4.3 *p*-Cresol Sulfate (PCS)

p-Cresol (4-methylphenol) is a protein-bound solute retained in the body as renal failure ensues [89, 98]. This UT derives from bacterial tyrosine fermentation in the large intestine and has been shown *in vitro* and clinically to have toxic effects [90, 98]. Additionally, *p*-cresol undergoes a detoxification step in the colonic mucosa and liver via conjugation with sulfur or glucuronic acid derivatives [99, 100].

Most attention at first was given to *p*-cresol alone, as most studies used strong acid for sample deproteinization. However, when deproteinization was done by methanol, virtually no *p*-cresol was detected in serum of dialysis patients. Instead, high concentration of PCS was measured [99, 101]. This specific derivative, PCS, is far more abundant in the body than *p*-cresol glucuronide, or even *p*-cresol, and is known to be associated with increases in microparticle release, an indicator of endothelial damage [99, 100]. A previous study also showed PCS to increase free radical production in leukocytes [102], so when combined with endothelial damage [99, 100], it is easy to see the role of PCS in the cardiovascular morbidity of ESRD patients.

10.4.4 Indoxyl Sulfate (IS)

Similar to PCS, indoxyl sulfate (IS) also originates from protein fermentation in the large intestine. Microbiota from the colon metabolizes tryptophan into indole, which is then hydroxylated into 3-hydroxyindole. Further sulfonation of the product in the liver yields IS, which is similar to PCS in that the majority of the compound is bound to albumin [100].

Wu et al. [103] have shown a correlation of both UTs with CKD progression and morbidity. In an observational study of 268 patients with different stages of CKD and follow-up of 21 ± 3 months, both IS and PCS were shown to correlate with eGFR (r −0.72, $P < 0.001$) and (r −0.64, $P < 0.001$), respectively [103]. Of the 268 patients, 35 (13.1%) had CKD progression (defined as a decrease in eGFR > 50%), and 14 (5.2%) died. Univariate Cox regression analysis showed that high serum PCS

levels were associated with CKD progression and mortality; independent of variables such as age, gender, and various other health parameters such as serum creatinine and even serum IS levels [103].

It may be challenging for individual metabolites to predict CKD progression, but metabolomics could offer a set of correlating metabolites that has better predictive power than single metabolites alone. The studies on PCS and IS indicate the importance of global nontargeted metabolomics approaches as there are metabolic exchanges between the intestine and the kidney, which need further investigation.

10.4.5 Gut Microbiota

Evidence of the intestine–kidney relationship has only recently materialized. Previous views of the intestine as a largely trivial, uninfluential organ limited to its function as a digestion and absorption organ are now contested, since intestinal microbiota have been linked to numerous diseases such as metabolic syndrome, CVD, and CKD. It is possible that the gut microbiota is the common denominator for many of these chronic illnesses [104]. The intestine is involved in uremia primarily through the production of protein-bound UTs such as IS and PCS. Other phenols, indoles, hippurate, and trimethylamine-*N*-oxide (TMAO) are all generated metabolically from precursors through microbial fermentation [104]. Hippurate is a metabolite derived from polyphenol fermentation [105], while TMAO is an end product of choline metabolism linked to CVD risk [106].

Evidence suggests that the intestinal microbiota–uremia relationship is bidirectional. In one direction, dietary food intake influences the contribution of substrate to the gut microbial metabolism, which can promote the growth of bacteria that are specific to the production of UTs [104, 107]. This can result in a global shift in the gut microbiome over time, leading to abnormal UT production. Lastly, this can contribute to abnormal bowel motility and ultimately absorption impairment. Prolonged colon transit deprives microbial species in downstream colonic regions of nutrients, causing spatial relocation of microbial species that otherwise would not be there or even overgrowth in certain regions [104, 108]. In total, gut microbial health is a leading factor in UT accumulation in the body; therefore gut microbial metabolomics can give valuable insight to early-stage detection of CKD.

In the other direction, uremia itself can cause spatial and metabolic modifications to gut microbiota, with pathophysiological ramifications. Firstly, bacterial species with metabolism in favor of UT fermentation will begin to outnumber others [104, 109]. Secondly, in uremia there is a translocation of bacteria in the intestinal tract, which can even include migration of certain species into the jejunum, as mentioned earlier [104, 108]. Lastly, there is evidence of the loss of the intestinal protective barrier in cases of uremia. These could lead to a perpetuated "leakage" of toxins into the bloodstream and further contribute to uremic syndrome [104, 109].

In total, uremia causes a series of intestinal changes—physically and biochemically—which leads to increased inflammation. It is commonly known that inflammation is one of the main factors in all types of chronic illness, including CKD. In light of this, only few studies have been performed on the composition of gut microbiota specifically in CKD [110, 111]. Given that inflammation in CKD is a multifactorial phenotype [111], global nontargeted metabolomics can offer much insight into the condition, as it is affected by the changing gut microbiota and its relationship with the CKD patient.

Wikoff et al. [112] showed the importance of the gut microbiome in a study of plasma and urine metabolomics. Looking at the serum of a germ-free (GF) mouse model with normal kidney function, they were able to show differences in hundreds of features using ESI-TOF-MS. Additionally, about 10% of detectable features in both wild-type (WT) and GF mice varied in concentration by at least 50% [112]. This study also showed the effects of the microbiome on indole-containing compounds, sulfonated metabolites, and other conjugated metabolites such as hippurate [112]. There were also several glycine conjugates exclusively identified in the serum of WT mice, which included cinnamoylglycine, a metabolite that at the time was not listed in any database. Also detected was phenylpropionylglycine, a metabolite that is most likely a product of the conjugation of glycine and phenylpropionic acid, which is a known metabolic product of anaerobic bacteria [112]. Building on this principle using metabolomics, Wikoff et al. [113] demonstrated through Oat1/SLc22a6 knockout mice that disruption of this key organic anion transporter resulted in the accumulation of many previously mentioned UTs that are known to be associated with CKD [113].

The gut metabolome in the analysis of CKD plays an integral role. A complete systems approach to this topic will generate a better understanding of total health of CKD patients. As shown in a systems approach study of renal failure, Mishima et al. [114] demonstrated the effects that renal failure can have on the gut bacterial populations. An adenine-induced renal failure (ReF) mouse model was used to demonstrate the interactions of the intestine, intestinal microbiota, and the kidneys. As kidneys failed, significant changes were seen in the physical and histological properties of fecal weight/number and overall intestinal transit in the small intestine and colon [114]. Kidney molecular changes also accrued, measured via immunohistochemistry and qPCR. As one would

expect, increases in fibrosis and inflammation-related gene markers were also observed [114].

Furthermore, changes were observed in gut microbiota measured with 454 pyrosequencing techniques of 16S rRNA genes. Alterations of gut bacterial populations were observed in RT mice, most interestingly in the family *Lactobacillaceae* and genus *Prevotella* [114]. The microbiota genomics data was then interpreted in parallel with plasma metabolomics data generated by CE-TOF-MS. Increases in notable gut UTs were seen, including IS, PCS, and hippurate. Also observed were increased concentrations of ADMA and cholate [114]. Plasma increase of TCA metabolites, namely, citrate, *cis*-aconitate, fumarate, and malate [114], was also detected.

To summarize, the previous three studies mentioned demonstrate the utility of metabolomics and illustrate its role in a systems approach for understanding chronic diseases such as CKD. Although these studies were performed in mouse models, they demonstrate the proof of concept. This concept is that CKD can be described in a multi-omics systems biology approach and that the intestine and gut microbiome may lead to deeper understanding of the origins of CKD.

10.4.6 Osmolytes

Among many factors that may contribute to CKD, hypertension and hyperosmolarity could reveal new and novel metabolomics biomarkers, with treatment options. Regular dehydration can cause plasma osmolarity to rise, requiring the body to retain water by producing urine that is increased in both specific gravity and osmolarity. The physiological response to elevated plasma osmolarity is the activation of two pathways: vasopressin synthesis and the fructokinase pathway [115]. The latter is interesting from a metabolomics perspective. Hyperosmolarity increases the activation of aldose reductase and induces conversion of glucose into sorbitol, which is an important osmolyte for protecting kidney tissue while in hyperosmotic environments [115]. The isomer of the enzyme fructokinase C (KHK-C) metabolizes fructose rapidly and results in transient depletion of intracellular phosphate and ATP. This process leads to oxidative stress, inflammation, and uric acid generation. KHK-C is mainly located in the liver and small intestine but is also expressed in the proximal tubules, mostly concentrated in the S3 segment [115].

What makes sorbitol metabolism even more interesting is its properties as a protective osmolyte. The importance of sorbitol for mammalian renal medullary cells has been known for some time now. It was shown using a PAP-HT25 cell line that sorbitol was used primarily as an osmolyte when cells were put under osmotic stress. When the metabolic production of sorbitol was disrupted, cell growth was also inhibited [116]. Sorbitol is one of several organic osmolytes that are produced to protect cells from hyperosmotic conditions. These include another carbohydrate (myo-inositol), amino acids (glycine, taurine), the methylamine, betaine, and glycerophosphorylcholine (GPC) [117].

Osmolytes are important in the health of kidneys for several reasons. First, during acute short-term water loss, cells utilize inorganic ions as metabolically "cheap" osmolytes. However, if water stress is long term, the cell will turn to organic osmolytes to protect the integrity of its proteins and enzymes [117, 118]. These osmolytes and their respective pathways could also provide other avenues to diagnose CKD and determine its progression.

The physiological topics discussed previously were meant to offer a brief overview of metabolic conditions that could relate to CKD while also being measureable by metabolomic techniques. It goes without saying that other metabolic alternations to the human physiology could also contribute to CKD and related renal diseases and metabolomics could be a very useful tool in generating data for their investigation.

10.5 Conclusions

In this chapter we have presented the utility of metabolomics studies in relation to CKD. There is great potential for biomarker discovery, as well as the development of personalized medicine and clinical screenings. There are some drawbacks to metabolomics, especially in nontargeted approaches. Firstly, there is not one single method for measuring the human metabolome in its totality. The development of such a method when considering extraction, separation, and the needed high-tech instrumentation to measure all metabolites is still a great challenge. Measurement of the total metabolome will probably only be possible via a multi-sample preparation and multi-platform analysis with post-analysis data fusion into a single comprehensive metabolite overview matrix. This requires the development of efficient methods for data integration. Various aspects of these two drawbacks have already been addressed in this review. It should be reiterated that metabolomics is the most distal measure of biochemical reactivity in an organism. Therefore, diet, lifestyle, and other environmental factors can have profound effect on the metabolome [4, 119], so researchers must always consider this issue when designing studies and reviewing data.

There is great utility in an optimized nontargeted metabolomics study. A total metabolomics master map and its alterations in CKD progression are of obvious use. Unlike the previous generation of metabolic research,

modern-day metabolomics has the advantage to measure vast numbers of different metabolites simultaneously. When integrated to a high-throughput system, an extremely insightful and practical tool will be at the disposal of clinicians. In addition to finding therapeutic targets and biomarkers, global metabolomics has the ability to measure the metabolic response of patients to a certain drug or treatment. Thus the new field of pharmacometabolomics will offer exciting novel perspectives in clinical research [54, 75]. As instrumentation and techniques improve, metabolomics will have an important role in diseases such as CKD.

References

1 Alonso A, Marsal S, Julia A. Analytical methods in untargeted metabolomics: state of the art in 2015. *Front Bioeng Biotechnol.* 2015;3:23.

2 Patti GJ, Yanes O, Siuzdak G. Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol.* 2012;13:263–269.

3 German J, Hammock B, Watkins SM. NIH public access. *Metabolomics*. 2005;1(1):3–9.

4 Holmes E, Loo RL, Stamler J, et al. Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*. 2008;453(7193):396–400.

5 Nicholson JK, Lindon JC. Metabonomics. *Nature.* 2008;455(23):1054–1056.

6 Wettersten HI, Weiss RH. Applications of metabolomics for kidney disease research. *Organogenesis*. 2013;9(1):11–18.

7 Holmes E, Wilson ID, Nicholson JK. Metabolic phenotyping in health and disease. *Cell*. 2008;134(September 5):714–717.

8 Robertson DG, Frevert U. Metabolomics in drug discovery and development. *Clin Pharmacol Ther*. 2013;94(5):559–561.

9 Bernini P, Bertini I, Luchinat C, Nincheri P, Staderini S, Turano P. Standard operating procedures for pre-analytical handling of blood and urine for metabolomic studies and biobanks. *J Biomol NMR.* 2011;49:231–243.

10 Emwas A-H, Luchinat C, Turano P, et al. Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. *Metabolomics*. 2014;11(4):872–894.

11 Dona AC, Jimenez B, Schäfer H, et al. Precision high-throughput proton NMR spectroscopy of human urine, serum, and plasma for large-scale metabolic phenotyping. *Anal Chem*. 2014;86(19):9887–9894.

12 Levitt MH. *Spin Dynamics: Basics of Nuclear Magnetic Resonance*; 2001. John Wiley & Sons, Inc., New York.

13 Bouatra S, Aziat F, Mandal R, et al. The human urine metabolome. *PLoS One*. 2013;8(9):e73076.

14 Assfalg M, Bertini I, Colangiuli D, et al. Evidence of different metabolic phenotypes in humans. *PNAS*. 2008;105(5):1420–1424.

15 Bernini P, Bertini I, Luchinat C, et al. Individual human phenotypes in metabolic space and time. *J Proteome Res*. 2009;8(9):4264–4271.

16 Grison S, Favé G, Maillot M, et al. Metabolomics identifies a biological response to chronic low-dose natural uranium contamination in urine samples. *Metabolomics*. 2013;9(6):1168–1180.

17 Lauridsen M, Hansen SH, Jaroszewski JW, Cornett C. Human urine as test material in 1H NMR-based metabonomics: recommendations for sample preparation and storage. *Anal Chem*. 2007;79(3):1181–1186.

18 Schreier C, Kremer W, Huber F, et al. Reproducibility of NMR analysis of urine samples: impact of sample preparation, storage conditions, and animal health status. *Biomed Res Int.* 2013;2013:878374.

19 Xiao C, Hao F, Qin X, Wang Y, Tang H. An optimized buffer system for NMR-based urinary metabonomics with effective pH control, chemical shift consistency and dilution minimization. *Analyst*. 2009;134(5):916–925.

20 Jiang L, Huang J, Wang Y, Tang H. Eliminating the dication-induced intersample chemical-shift variations for NMR-based biofluid metabonomic analysis. *Analyst*. 2012;137(18):4209.

21 Veselkov KA, Lindon JC, Ebbels TMD, et al. Recursive segment-wise peak alignment of biological 1H NMR spectra for improved metabolic biomarker recovery. *Anal Chem*. 2009;81(1):56–66.

22 Beckonert O, Keun HC, Ebbels TMD, et al. Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nat Protoc*. 2007;2(11):2692–2703.

23 Holmes E, Nicholls AW, Lindon JC, et al. Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chem Res Toxicol*. 2000;13(6):471–478.

24 Asiago VM, Nagana Gowda GA, Zhang S, Shanaiah N, Clark J, Raftery D. Use of EDTA to minimize ionic strength dependent frequency shifts in the 1H NMR spectra of urine. *Metabolomics*. 2008;4(4):328–336.

25 Mavel S, Nadal-Desbarats L, Blasco H, et al. 1H-13C NMR-based urine metabolic profiling in autism spectrum disorders. *Talanta.* 2013;114:95–102.

26 Lei Z, Huhman DV, Sumner LW. Mass spectrometry strategies in metabolomics. *J Biol Chem*. 2011;286(29):25435–25442.

27 Ho CS, Lam CWK, Chan MHM, et al. Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin Biochem Rev*. 2003;24(1):3–12.

28 Bruins AP. Mechanistic aspects of electrospray ionization. *J Chromatogr A*. 1998;794(1–2):345–357.

29 Andrade FJ, Shelley JT, Wetzel WC, et al. Atmospheric pressure chemical ionization source. 1. Ionization of compounds in the gas phase atmospheric pressure chemical ionization source. 1. Ionization of compounds in the gas phase. *Anal Chem*. 2008;80(8):2646–2653.

30 Robb DB, Covey TR, Bruins AP. Atmospheric pressure photoionization: an ionization method for liquid chromatography—mass spectrometry. *Anal Chem*. 2000;72(15):3653–3659.

31 Walch A, Rauser S, Deininger SO, Höfler H. MALDI imaging mass spectrometry for direct tissue analysis: a new frontier for molecular histology. *Histochem Cell Biol*. 2008;130(3):421–434.

32 Nordström A, Want E, Northen T, Lehtiö J, Siuzdak G. Multiple ionization mass spectrometry strategy used to reveal the complexity of metabolomics. *Anal Chem*. 2008;80(2):421–429.

33 Thurman EM, Ferrer I, Barcelo D, Survey USG, Place QC. Choosing between Atmospheric pressure chemical ionization and electrospray ionization interfaces for the HPLC/MS analysis of pesticides. *Anal Chem*. 2001;73(22):5441–5449.

34 Dettmer K, Aronov P, Hammock B. Mass spectrometry-based metabolomics. *Mass Spectrom Rev*. 2007;26(1): 57–78.

35 Nikolaev EN, Kostyukevich YI, Vladimirov GN. Fourier transform ion cyclotron resonance (FT ICR) mass spectrometry: theory and simulations. *Mass Spec Rev*. 2014;35(2):219–258.

36 Marshall AG, Hendrickson CL, Jackson GS. Fourier transform ion cyclotron resonance mass spectrometry: a primer. *Mass Spectrom Rev*. 1998;17(1):1–35.

37 Han J, Danell RM, Patel JR, et al. Towards high-throughput metabolomics using ultrahigh-field Fourier transform ion cyclotron resonance mass spectrometry. *Metabolomics*. 2008;4(2):128–140.

38 Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem*. 2000;72(6):1156–1162.

39 Fu XW, Iga M, Kimura M, Yamaguchi S. Simplified screening for organic acidemia using GC/MS and dried urine filter paper: a study on neonatal mass screening. *Early Hum Dev*. 2000;58(1):41–55.

40 Zhang A, Sun H, Wang P, Han Y, Wang X. Modern analytical techniques in metabolomics analysis. *Analyst*. 2012;137(2):293.

41 Wilson ID, Nicholson JK, Castro-Perez J, et al. High resolution "ultra performance" liquid chromatography coupled to oa TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *J Proteome Res*. 2005;4(2):591–598.

42 Cubbon S, Amtonio C, Wilson J, Thomas-Oates J. Metabolomic applications of HILIC-LC-MS. *Mass Spec Rev*. 2009;26(3):451–466.

43 Klampfl CW. Review coupling of capillary electrochromatography to mass spectrometry. *J Chromatogr A*. 2004;1044(1–2):131–144.

44 Büscher JM, Czernik D, Ewald JC, Sauer U, Zamboni N. Cross-platform comparison of methods for quantitative metabolomics of primary metabolism cross-platform comparison of methods for quantitative metabolomics of primary metabolism. *Metab Clin Exp*. 2009;81(6):2135–2143.

45 Mushtaq MY, Choi YH, Verpoorte R, Wilson EG. Extraction for metabolomics: access to the metabolome. *Phytochem Anal*. 2014;25(4):291–306.

46 Raterink R-J, Lindenburg PW, Vreeken RJ, Ramautar R, Hankemeier T. Recent developments in sample-pretreatment techniques for mass spectrometry-based metabolomics. *TrAC Trends Anal Chem.* 2014;61:157–167.

47 Want EJ, Wilson ID, Gika H, et al. Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc*. 2010;5(6):1005–1018.

48 Hyötyläinen T. Critical evaluation of sample pretreatment techniques. *Anal Bioanal Chem*. 2009;394(3):743–758.

49 Jiang H, Cao H, Zhang Y, Fast DM. Systematic evaluation of supported liquid extraction in reducing matrix effect and improving extraction efficiency in LC-MS/MS based bioanalysis for 10 model pharmaceutical compounds. *J Chromatogr B Anal Technol Biomed Life Sci.* 2012;891–892:71–80.

50 Rao RN, Prasad KG, Kumar KVS, Ramesh B. Diatomaceous earth supported liquid extraction and LC-MS/MS determination of elvitegravir and ritonavir in rat plasma: application to a pharmacokinetic study. *Anal Methods*. 2013;5(23):6693.

51 Snyder LR. Classification of the solvent properties of common liquids. *J Chromatogr A*. 1974;92(2):223–230.

52 Van Damme T, Lachová M, Lynen F, Szucs R, Sandra P. Solid-phase extraction based on hydrophilic interaction liquid chromatography with acetone as eluent for eliminating matrix effects in the analysis of biological fluids by LC-MS. *Anal Bioanal Chem*. 2014;406(2):401–407.

53 Martin J-C, Maillot M, Mazerolles G, et al. Can we trust untargeted metabolomics? Results of the metabo-ring initiative, a large-scale, multi-instrument inter-laboratory study. *Metabolomics*. 2014;11(4):807–821.

54 Weiss RH, Kim K. Metabolomics in the study of kidney diseases. *Nat Rev Nephrol.* 2012;8:22–33.

55 Kind T, Tolstikov V, Fiehn O, Weiss RH. A comprehensive urinary metabolomic approach for identifying kidney cancer. *Anal Biochem*. 2007;363(2):185–195.

56 Burton L, Ivosev G, Tate S, Impey G, Wingate J, Bonner R. Instrumental and experimental effects in LC-MS-based metabolomics. *J Chromatogr B Anal Technol Biomed Life Sci*. 2008;871(2):227–235.

57 Tomasi G, Van Den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *J Chemometr*. 2004;18(5):231–241.

58 Savorani F, Tomasi G, Engelsen SB. icoshift: a versatile tool for the rapid alignment of 1D NMR spectra. *J Magn Reson*. 2010;202(2):190–202.

59 Alonso A, Rodr MA, Vinaixa M, et al. Focus: a robust work flow for one-dimensional NMR spectral analysis. *Anal Chem*. 2013;86:116–1169.

60 Giskeødegård GF, Bloemberg TG, Postma G, et al. Alignment of high resolution magic angle spinning magnetic resonance spectra using warping methods. *Anal Chim Acta*. 2010;683(1):1–11.

61 Kohl SM, Klein MS, Hochrein J, Oefner PJ, Spang R, Gronwald W. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics.* 2012;8:146–160.

62 Dieterle F, Ross A, Schlotterbeck G, Senn H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics. *Anal Chem*. 2006;78(13):4281–4290.

63 van den Berg RA, Hoefsloot HCJ, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics.* 2006;7:142.

64 Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics.* 2010;11:395.

65 Yang C, He Z, Yu W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinformatics.* 2009;10:4.

66 Bro R, Smilde AK. Principal component analysis. *Anal Methods*. 2014;6(9):2812.

67 Tikunov Y, Lommen A, de Vos CHR, et al. A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol*. 2005;139:1125–1137.

68 Haddad I, Hiller K, Frimmersdorf E, Benkert B, Schomburg D, Jahn D. An emergent self-organizing map based analysis pipeline for comparative metabolome studies. *In Silico Biol*. 2009;9(4):163–178.

69 Bartel J, Krumsiek J, Theis FJ. Statistical methods for the analysis of high-throughput metabolomics data. *Comput Struct Biotechnol J.* 2013;4(5):1–9.

70 Wiklund S, Johansson E, Sjöström L, et al. Visualization of GC/TOF-MS-based metabolomics data for identification of biochemically interesting compounds using OPLS class models. *Anal Chem*. 2008;80(1):115–122.

71 Cloarec O, Dumas M, Craig A, et al. Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic H NMR data sets. *Anal Chem*. 2005;77(5):1282–1289.

72 Forcisi S, Moritz F, Kanawati B, Tziotis D, Lehmann R, Schmitt-Kopplin P. Liquid chromatography-mass spectrometry in metabolomics research: mass analyzers in ultra high pressure liquid chromatography coupling. *J Chromatogr A*. 2013;1292:51–65.

73 Tulpan D, Léger S, Belliveau L, Culf A, Čuperlović-Culf M. MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures. *BMC Bioinformatics*. 2011;12(1):400.

74 Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*. 2012;8(2):e1002375.

75 Bouhifd M, Hartung T, Hogberg HT, Kleensang A, Zhao L. Review: toxicometabolomics. *J Appl Toxicol*. 2013;33(12):1365–1383.

76 Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics*. 2013;9(2):280–299.

77 Kell DB, Goodacre R. Metabolomics and systems pharmacology: why and how to model the human metabolic network for drug discovery. *Drug Discov Today*. 2014;19(2):171–182.

78 Zamboni N, Fendt S-M, Rühl M, Sauer U. (13)C-based metabolic flux analysis. *Nat Protoc*. 2009;4(6):878–892.

79 Redekop WK, Mladsi D. The faces of personalized medicine: a framework for understanding its meaning and scope. *Value Heal*. 2013;16(6 suppl):4–9.

80 Zamboni N, Saghatelian A, Patti GJ. Defining the metabolome: size, flux, and regulation. *Mol Cell*. 2015;58(4):699–706.

81 Gerstl MP, Ruckerbauer DE, Mattanovich D, Jungreuthmayer C, Zanghellini J. Metabolomics integrated elementary flux mode analysis in large metabolic networks. *Sci Rep.* 2015;5:8930.

82 Wasylenko TM, Stephanopoulos G. Metabolomic and 13 C-metabolic flux analysis of a xylose-consuming Saccharomyces cerevisiae strain expressing xylose isomerase. *Biotechnol Bioeng*. 2015;112(3):470–483.

83 Yu HT. Progression of chronic renal failure. *Arch Intern Med*. 2003;163(12):1417–1429.

84 Levey AS, Eckardt K-U, Tsukamoto Y, et al. Definition and classification of chronic kidney disease: a position statement from kidney disease: improving global outcomes (KDIGO). *Kidney Int*. 2005;67(6):2089–2100.

85 Pena MJ, de Zeeuw D, Mischak H, et al. Prognostic clinical and molecular biomarkers of renal disease in type 2 diabetes. *Nephrol Dial Transplant*. 2015;30(suppl 4):iv86–iv95.

86 Shah VO, Townsend RR, Feldman HI, Pappan KL, Kensicki E, Vander Jagt DL. Plasma metabolomic profiles in different stages of CKD. *Clin J Am Soc Nephrol*. 2013;8(3):363–370.

87 Rhee EP, Clish CB, Ghorbani A, et al. A combined epidemiologic and metabolomic approach improves CKD prediction. *J Am Soc Nephrol*. 2013;24(8):1330–1338.

88 Nkuipou-Kenfack E, Duranton F, Gayrard N, et al. Assessment of metabolomic and proteomic biomarkers in detection and prognosis of progression of renal function in chronic kidney disease. *PLoS One*. 2014;9(5):e96955.

89 Vanholder R, De Smet R, Glorieux G, et al. Review on uremic toxins: classification, concentration, and interindividual variability. *Kidney Int*. 2003;63(5):1934–1943.

90 Vanholder R, Laecke S, Glorieux G. What is new in uremic toxicity? *Pediatr Nephrol*. 2008;23(8):1211–1221.

91 Toyohara T, Akiyama Y, Suzuki T, et al. Metabolomic profiling of uremic solutes in CKD patients. *Hypertens Res*. 2010;33(9):944–952.

92 Toyohara T, Suzuki T, Morimoto R, et al. SLCO4C1 transporter eliminates uremic toxins and attenuates hypertension and renal inflammation. *Ther Res*. 2009;31(9):1221–1223.

93 Fassett RG, Venuthurupalli SK, Gobe GC, Coombes JS, Cooper MA, Hoy WE. Biomarkers in chronic kidney disease: a review. *Kidney Int*. 2011;80(8):806–821.

94 Ravani P, Tripepi G, Malberti F, Testa S, Mallamaci F, Zoccali C. Asymmetrical dimethylarginine predicts progression to dialysis and death in patients with chronic kidney disease: a competing risks modeling approach. *J Am Soc Nephrol*. 2005;16(8):2449–2455.

95 Fliser D, Kronenberg F, Kielstein JT, et al. Asymmetric dimethylarginine and progression of chronic kidney disease: the mild to moderate kidney disease study. *J Am Soc Nephrol*. 2005;16(8):2456–2461.

96 Lajer M. Dimethylarginine (ADMA) predicts cardiovascular morbidity and mortality in type 1 diabetic patients with diabetic. *Diabetes Care*. 2008;31(4):747–752.

97 Hanai K, Babazono T, Nyumura I, et al. Asymmetric dimethylarginine is closely associated with the development and progression of nephropathy in patients with type 2 diabetes. *Nephrol Dial Transplant*. 2009;24(6):1884–1888.

98 de Loor H Gas chromatographic–mass spectrometric analysis for measurement of p-cresol and its conjugated metabolites in uremic and normal serum. *Clin Chem*. 2005;51(8):1533–1535.

99 Vanholder R, Bammens B, De Loor H, et al. Warning: the unfortunate end of p-cresol as a uraemic toxin. *Nephrol Dial Transplant*. 2011;26(5):1464–1467.

100 Meijers BKI, Evenepoel P. The gut-kidney axis: indoxyl sulfate, p-cresyl sulfate and CKD progression. *Nephrol Dial Transplant*. 2011;26(3):759–761.

101 Martinez AW, Recht NS, Hostetter TH, Meyer TW. Removal of P-cresol sulfate by hemodialysis. *J Am Soc Nephrol*. 2005;16(11):3430–3436.

102 Schepers E, Meert N, Glorieux G, Goeman J, Van der Eycken J, Vanholder R. P-cresylsulphate, the main in vivo metabolite of p-cresol, activates leucocyte free radical production. *Nephrol Dial Transplant*. 2007;22(2):592–596.

103 Wu IW, Hsu KH, Lee CC, et al. P-cresyl sulphate and indoxyl sulphate predict progression of chronic kidney disease. *Nephrol Dial Transplant*. 2011;26(3):938–947.

104 Vanholder R, Glorieux G. The intestine and the kidneys: a bad marriage can be hazardous. *Clin Kidney J*. 2015;8(2):168–179.

105 Lees HJ, Swann JR, Wilson ID, Nicholson JK, Holmes E. Hippurate: the natural history of a mammalian—microbial cometabolite. *J Proteome Res*. 2013;12:1527–1546.

106 Wang Z, Klipfell E, Bennett BJ, et al. Gut flora metabolism of phosphatidylcholine promotes cardiovascular disease. *Nature*. 2011;472(7341):57–63.

107 Schepers E, Glorieux G, Vanholder R. The gut: the forgotten organ in Uremia? *Blood Purif*. 2010;29(2):130–136.

108 Strid H, Simren M, Stotzer P-O, Ringström G, Abrahamsson H, Björnsson ES. Patients with chronic renal failure have abnormal small intestinal motility and a high prevalence of small intestinal bacterial overgrowth. *Digestion*. 2003;67(3):129–137.

109 Vaziri ND, Wong J, Pahl M, et al. Chronic kidney disease alters intestinal microbial flora. *Kidney Int*. 2012;83(2):308–315.

110 Mafra D, Lobo JC, Barros AF, Koppe L, Vaziri ND, Fouque D. Role of altered intestinal microbiota in systemic inflammation and cardiovascular disease in chronic kidney disease. *Future Microbiol*. 2014;9(3):399–410.

111 Mafra D, Fouque D. Gut microbiota and inflammation in chronic kidney disease patients. *Clin Kidney J*. 2015;8(3):332–334.

112 Wikoff WR, Anfora AT, Liu J, et al. Metabolomics analysis reveals large effects of gut microflora on mammalian blood metabolites. *Proc Natl Acad Sci U S A*. 2009;106(10):3698–3703.

113 Wikoff WR, Nagle MA, Kouznetsova VL, Tsigelny IF, Nigam SK. Key structural features for substrate binding to organic anion transporter 1 (Oat1; slc22a6) identified by global untargeted metabolomics of Oat1null plasma. *J Proteome Res*. 2012;29(6):997–1003.

114 Mishima E, Fukuda S, Shima H, et al. Alteration of the intestinal environment by lubiprostone is associated with amelioration of adenine-induced CKD. *J Am Soc Nephrol.* 2015;26:1787–1794.

115 Johnson RJ, Rodriguez-Iturbe B, Roncal-Jimenez C, et al. Hyperosmolarity drives hypertension and CKD-water and salt revisited. *Nat Rev Nephrol*. 2014;10(7):415–420.

116 Yancey PH, Burg MB, Bagnasco SM. Effects of NaCl, glucose, and aldose reductase inhibitors on cloning efficiency of renal medullary cells. *Am J Physiol*. 1990;258(1 Pt 1):C156–C163.

117 Yancey PH. Water stress, osmolytes and proteins. *Am Zool.* 2001;41(4):699–709.

118 Neuhofer W, Beck F-X. Response of renal medullary cells to hypertonic stress. *Contrib Nephrol.* 2005;148:21–34.

119 Heinzmann SS, Merrifield CA, Rezzi S, et al. Stability and robustness of human metabolic phenotypes in response to sequential food challenges. *J Proteome Res*. 2012;11(2):643–655.

11

Statistical Inference in High-Dimensional Omics Data

Eleni-Ioanna Delatola[1] *and Mohammed Dakna*[2]

[1] *Systems Biology Ireland, University College Dublin, Dublin, Ireland*
[2] *Mosaiques Diagnostics GmbH, Hannover, Germany*

11.1 Introduction

In the last two decades, advanced laboratory instruments were developed to decipher the structure and function of genes, proteins, and other constituents of the cellular molecular networks and their alterations in diseases. All these high-throughput technologies typically generate high-dimensional omics datasets with thousands of molecular signals by analyzing few samples. The data structure requires hence the development of appropriate algorithms for efficient analysis. The complexity of biomedical problems represents a natural obstacle to the generation of new knowledge for diagnosis, prediction, and monitoring of the disease from the measured datasets. To accomplish that goal, the noisy raw data generated by different instruments must be filtered and normalized across samples to remove batch effects and technical variability. This low-level processing usually generates a set of $n \times m$ ($n \ll m$) data matrices with n being the number of samples and m the number of the measured biological entities (genes, proteins, metabolites, etc.) that can be used for downstream statistical/bioinformatical data mining [1, 2]. Usually the data matrix contains many features that are irrelevant to the problem under study. Thus, it is important to perform feature selection and ranking that eventually leads to the construction of classification rules. The stability of the selected feature is also of great importance in order to allow for unique biological function assignment to the selected features. The correlation of the feature set to clinical outcome should be assessed as it determines the added value of such a set for clinical decision making. For fixed number of samples n, using different technologies (next-generation sequencing (NGS), proteomics, metabolomics), different feature sets are generated (e.g., m_g genes leading to the $n \times m_g$ gene expression matrix and m_p proteins leading to $n \times m_p$ protein expression matrix). The main issue that statistics has to solve is to combine the different data matrices in order to facilitate knowledge extraction.

We present an overview of general data processing steps such as feature selection, sample classification, and data integration that may be necessary due to the complexity of human diseases. Omics data may be also analyzed in the context of molecular networks to detect meaningful biological targets and understand disease processes or to infer the sequence or the molecular structure of biologically relevant molecular entities. These important issues are however out of the scope of the present chapter.

11.2 From Raw Data to Expression Matrices

Genomics data are related mainly to the collection of DNA sequences and their transcripts. The information about the sequence may be used to derive position-specific person-to-person variations that are known as single nucleotide polymorphism (SNP). This information is useful as a genetic biomarker for predicting susceptibility to different diseases. More frequently, for diagnosis and prognosis of diseases, the relative abundance of transcripts representing the level of gene expression in cells is compared between healthy and diseased patients. Microarray technology was the major tool used to monitor genomic and transcriptomic expression levels of genes in a given organism. With the development of NGS technology, researchers have started to take advantage of this new method for gene expression experiments as it provides a more detailed analysis of the transcriptome. Genomics and trascriptomics data are usually signal outputs from electronic and/or optical devices. System noise caused by operational variability and probe performance variability complicates the interpretation of

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

these data. But usually the manufacturers of the devices provide tools and steps to remove the systematic noise and enhance the confidence in the instrument output. Nevertheless it is advisable to check for batch effects and other sources for systematic noise in the data [2–4].

Proteomics and metabolomics data consist of measuring absolute or relative abundances of proteins and metabolites in biological fluids or tissues. The abundances of these molecules are measured by mass spectrometry (MS) sometimes coupled to chromatographic separation to lower the high complexity level of analytes in the sample under study. Mass spectrometers record peak intensity (abundance) for ionized molecules characterized by their mass-to-charge ratio (m/z). Compared with proteins, metabolites are smaller molecules in size and mass, and for their analysis nuclear magnetic resonance (NMR) spectroscopy is also used. The MS and NMR spectra are usually very noisy. To obtain the final features and to significantly reduce the feature space, the recalibrated and baseline corrected spectra are further processed using a peak picking procedure [5, 6]. Therefore a peak in a spectrum indicates a local maximum in the signal with a specific width related to the local position. In addition, criteria such as a signal-to-noise (SNR) ratio are used to separate small artifacts from a real peak. Peak picking, in general, is a complex processing task. Aspects such as the resolution of the device and appropriate noise estimation are taken into account [5, 7].

Genomics, proteomics, and metabolomics data of similar samples are not always similarly quantified. An important issue is data normalization to avoid systematic bias or batch effects on the samples that may jeopardize downstream statistical analysis of the data. There are actually a plethora of reported false discoveries that may be traced back to incorrect handling of raw data and misuse of statistical methods. In Ref. [8], for example, an overview about this topic can be found. The goal of the normalization step is indeed to adjust for the effects that are due to variations in the technologies rather than the biology. We refer to the literature for further details on this topic [9–16].

11.3 Brief Introduction R and Bioconductor

The R framework for statistical computing [17] has been well established in the field of bioinformatics and features a variety of tools to perform omics data analysis [18]. The main online repositories for R packages are the Comprehensive R Archive Network (CRAN) and Bioconductor [19], which currently contain 7983 and 1104 packages, respectively. Bioconductor organizes its packages in software, annotation, and experimental data, whereas CRAN provides a task list that groups the packages according to their relevance to a given task (e.g., Bayesian inference or meta-analysis). Both repositories provide a search functionality that allows the user to browse these repositories easily. However, it is hard to judge which package is the right one for a given task just by the package name and short description. Table 11.1 summarizes some of the R packages used in this chapter.

11.4 Feature Selection

As already mentioned, a typical characteristic of omics data is the very high number of simultaneously measured variables. This usually exceeds by many orders of magnitude the number of the available samples. The difficulty in analyzing high-dimensional data is due to two effects that make it difficult to detect the dependence between the response variable and the collection of the covariates. Firstly, high-dimensional spaces have

Table 11.1 R packages dedicated to different omics tasks such as feature selection, sample classification, and data integration.

Name	Description	Repository
GeneSelector	Variable selection	Bioconductor
iCluster+	Integrative Clustering	Bioconductor
mixOmics	Data integration (CCA,PLS,PCA)	CRAN
omicade4	MCIA and CIA	Bioconductor
PMA	Sparse CCA	CRAN
pwOmics	Pathway-based data integration of omics data	Bioconductor
RGCCA	Regularized CCA with variable selection	CRAN
Caret	Classification and regression training	CRAN

geometrical properties that are counterintuitive "known as curse of dimensionality" and significantly different from the properties that can be observed in two- or three-dimensional spaces. Secondly, traditional statistical data analysis tools are most often designed having in mind intuitive properties and examples in low-dimensional spaces. To reduce the number of covariates, mainly two prominent approaches exist in the data mining literature. One is variable selection, where one guesses that among all the available covariates, only a few are truly related to the response and all others are redundant and have no real explanatory effect. The second approach to reduce the number of covariates is the so-called projective dimension reduction [20]. Here one assumes that the response variable relates to only a few linear combinations of the many covariates. Thus, it is possible that all the covariates contain information, but the information can be represented by a few linear combinations. Prominent examples of this approach are principal component analysis (PCA) [21] and partial least squares (PLS) [22]. PLS, like PCA, search for linear combinations of input features that optimize the variance of the data. Unlike PCA, PLS achieves this optimization and maximizes the correlation of the transformed features and the class variable [23]. In this chapter we limit our discussion to variable selection (see Ref. [24] for a review). The advantage of this approach is that the feature may well be assigned to biological function, whereas in the projective approach the derived combination of feature lacks direct interpretation. In a typical omics study, a key challenge is to detect potentially significant features.

Differentially expressed features across control and case groups are distributed among a large set of variables. This type of analysis is referred to as multiple hypothesis testing. To give an example, let us suppose that we perform independent tests using $\alpha = 0.05$ as the critical significance level. The probability for a single test to come to a nonsignificant (that is a correct conclusion) result is hence $1 - 0.05 = 0.95$ (95%). Since n tests are independent, the probability that all these n tests to correctly reject the n null hypothesis is simply given by the product of the single results, that is, $0.95 \cdots 0.95 = 0.95^n$. The probability of wrongly rejecting at least one of the n null hypothesis is given by $1 - 0.95^n$. Thus, if our experiment performs 100 tests on 100 biomarkers, the error probability is given by $1 - 0.95^{100} = 0.99408$. In other words, we are almost sure that by performing 100 tests on 100 features, at least one of the declared significant findings will be a false positive. Thus, corrections or adjustments have to be made to these values to reduce the possibility of spurious results. The multiple hypothesis testing problem can be addressed through the estimation of the family-wise error rate (FWER) and the false discovery rate (FDR) [25].

Similar to differential expression, the correlation between a feature and a clinical outcome reflects the level of association between the two variables that might be of interest. As in statistical hypothesis testing, such an association can be computed using parametric and nonparametric techniques. The former assumes that the variables can be jointly modeled with a normal distribution. The latter does not make this assumption and is based on the idea of comparing the value ranks of the variables. The correlation with outcome is of particular interest for clinical diagnostics as the question there is always how the omics signatures add to potentially existing and routinely measured classical clinical parameters specific to a disease [26–29].

Another important property of a feature selection method is stability that refers to robustness of the selected features to perturbations in the data. Stability might be more important for knowledge discovery as in biomarker discovery than in constructing accurate classifiers as several feature subsets may lead to optimal classifier. Due to its importance for the biological interpretation of the identified features, stability has been the focus of extensive research [30, 31].

As an example we used the Bioconductor package GeneSelector to analyze urinary peptidomics and metabolomics data that were collected from patients with early and late chronic kidney disease (CKD) (the clinical and demographic data are presented in detail in Section 11.6). This package implements several variable selection and ranking resampling methods for assessing the stability of the features that can discriminate the two patient groups. For power issues we use here the full dataset (training plus test sets). Peptidomic dataset and six different ranking statistics (fold change to Wilcoxon) results are shown in Table 11.2.

From the previous table we see that the peptide number 1078 (collagen-alpha-1(III) chain) is ranked among the top 10 peptides in all selection methods. However its position in the top 10 list varies from 1 under the shrinkage method to 6 under the fold change method, similarly for the metabolome dataset we have (Table 11.3).

It is obvious that the metabolite number 226 (symmetric dimethylarginine) is ranked among the top 10 peptides in all selection methods. However its position in the top 10 list varies from 1 under the Wilcoxon method to 10 under the fold change method. The package GeneSelector provides several methods for assessing the stability for the chosen feature by perturbing the original dataset. Bootstrap sampling, jackknife, and label swap are by far the mostly used methods. Using 50 bootstrap resamples from the original proteomic dataset yielded the following results for peptide number 1220. It was ranked on the first position in 18 replicates and ranked on the second position in 11 (out of 50). The same approach for the

Table 11.2 Top 10 ranking of the 1828 peptides in the proteomic dataset under different feature selection methods.

Rank	Fold change	ordinaryT	Limma	FoxDimmicT	ShrinkageT	Wilcoxon
1	1067	1220	1220	1033	1078	1220
2	1033	1078	1078	1067	1219	1078
3	1217	1221	1221	1218	1031	1221
4	1218	41	41	1217	597	1217
5	1005	1219	1219	1078	1217	1222
6	1078	1067	1067	1005	1067	1827
7	1117	1031	1031	907	1073	1031
8	522	1217	1217	1190	1220	503
9	947	597	597	1117	503	1603
10	360	1003	1003	1149	1190	1602

Each number corresponds to a peptide.

Table 11.3 Top 10 ranking of the 227 metabolites in the metabolites dataset under different feature selection methods.

Rank	Fold change	ordinaryT	Limma	FoxDimmicT	ShrinkageT	Wilcoxon
1	6	2	1	6	85	226
2	7	1	2	7	6	2
3	58	7	7	58	7	1
4	18	85	85	18	226	85
5	34	226	226	34	18	109
6	13	14	14	226	84	84
7	24	6	6	13	58	24
8	193	181	181	193	13	225
9	209	222	222	209	181	219
10	226	183	183	57	165	14

Each number corresponds to a metabolite.

metabolomics data determined that metabolite number 226 was ranked on the first position in 20 replicates and ranked on the second position in 6 (out of 50). The stability of these features may be a strong indication that they play important roles in the development of kidney disease.

11.5 Sample Classification

A major goal of every omics experiment and subsequent data analysis pipelines is predicting patient medical status or disease evolution. The gathered information is formulated as a classifier for supporting medical decision making. There are two main types of classification. In unsupervised classification, the measured omics profiles are used for patient categorization without any additional information. In contrast to that, in supervised classification, sample information such as class membership is used to train the classifier for deducing a prediction rule. For sample classification using omics data, the three frequently occurring tasks have been centered on how such data may contribute to class comparison, class membership prediction, and class discovery [32, 33].

Classification may be generally formulated as seeking for relations between the x_{ij} entries of the aforementioned $n \times m$ data matrix (this element denotes the expression level of the jth analyte (feature) in the ith sample) and an outcome y_i to deduce the pattern underlying the data and generalize the obtained information to non-analyzed samples. Often classification implies that the outcome is a two-level factor, that is, $y_i = 1$ for disease and $y_i = 0$ for healthy samples. However most of the developed methods can be extended to situations where the outcome is a continuous variable (in this context one

refers to the relation as regression, e.g., a survival time analysis) or a multiclass problem (e.g., disease with k stages). Many statistical/machine learning algorithms are available for classification, and it is well known that a single method cannot be applied universally. Among the most prominent algorithms used for omics data classification are k-nearest neighbors (kNN) [34], logistic regression [35], nearest shrunken centroid method (NSC) [36], k-top-scoring pair (kTSP) [37], the linear and quadratic discriminant analysis [38] (LDA and QDA) (if the classes have diagonal covariance matrices, these methods reduce to DLDA and DQDA), classification trees, support vector machines (SVM) [39], and neural networks (NN) [40]. It has been a common practice to apply some of these machine learning methods combined with several feature selection algorithms to publicly available omics datasets and compare the resulting classification performance [25, 41]. Recently, the combination of many simple classifiers using ensemble methods is increasingly used in omics and may lead to better classification rules [42]. Prominent examples for such approaches are the bagging [43], random forests [44], and boosting [45–47].

After a classifier has been adopted, the next step is to decide which validation strategy will be used to assess the classifier's performance. A straightforward strategy, for instance, is to randomly split the samples into two disjoint sets called training and validation sets. The training data will be used to deduce the association between expression levels and the outcome, while the validation data will be used to assess the classifier's generalization ability. If the sample size is not large enough to put aside a validation dataset, it is usual to evaluate the performance of classifiers based on cross-validation (CV) including leave-one-out (LOO) CV as a special case, repeated splitting into training and test datasets, or bootstrap sampling. Details of CV are well elaborated in Refs. [48–50]. When the sample size is too small, using CV may however be overoptimistic in assessing classifier performance [51–54].

To give an example here we use the proteomics dataset and require that a peptide must be present in a least 50% of either cases or controls. This results in a reduction of the number of peptides from 1828 to 607. We then use caret package (short for classification and regression training) that contains several functions to streamline the variable selection and model training process for complex regression and classification problems. There are many modeling functions in R with a plethora of syntaxes. The caret package however provides a uniform interface for the functions of other packages, as well as a way to standardize common tasks (such parameter tuning and variable importance). For the proteomic dataset we used as classifier a recursive feature elimination support vector machine (RFFSVM) with a basis radial function kernel (RBF-kernel). As error estimation we use a 10-fold CV with five repeats. The data was split in a test set of fixed size ($N = 16$) and a training set of variable size ($N = 18, 25, 30$). Table 11.4 summarizes the classification performance as measured by the area under the curve.

Table 11.4 Classifier performance as a function of training set size.

N = size of train data	AUC_cross-validation	AUC_test_set
18	0.903	0.768
25	0.875	0.817
33	0.836	0.828

From the table we see that for small sizes of the training data ($N = 18$), the CV estimates are overoptimistic. But for training sample size of 33, the CV and test set estimates are comparable.

11.6 Real Data Example

To give an example of the statistical analysis of omics data, we refer to a recently published proteomics and metabolomics study on CKD [55]. The dataset is comprised of 49 urine samples with metabolites and peptides quantified by MS. Using different thresholds for estimated glomerular filtration rate (eGFR), the data was divided into the "early CKD" group containing patients with high eGFRs ($59.9 \pm 16.5\,\mathrm{ml/min/1.73\,m^2}$), whereas the "advanced CKD" group contained patients with low eGFRs ($8.9 \pm 4.5\,\mathrm{ml/min/1.73\,m^2}$). Follow-up information about the patient outcome was available. For testing different hypothesis about the data such as whether the combination of the proteomic and metabolic profile provide better correlation with the kidney function, different CKD classifier scores were generated and correlated with eGFR at baseline and follow-up eGFR to assess the prediction of the progression of the renal function. Table 11.5 summarizes the study design and patient characteristics.

11.7 Multi-Platform Data Integration

Data integration plays a crucial role in deciphering the function of biological systems. The analysis of high-throughput data can yield more information when performed on a global rather than on an individual scale. Several reviews on statistical methods in data integration of omics data are available [55–60]. The aforementioned list is far from being exhaustive, since this is an active field of research. For reviews on data integration in plant biology and CKD, the reader should refer to Refs. [56]

Table 11.5 Demographic and clinical data.

	Training set			
	"Early CKD"	"Advanced CKD"	*p*-values	Test set
n	10	10		29
Age (years)	65.9 ± 10.9	70.7 ± 9.8	0.2767	73.3 ± 9.0
Gender (M/F)	7/3	7/3		17/12
Baseline eGFR (ml/min/1.73 m^2)	59.9 ± 16.5	8.9 ± 4.5	<0.0001	29.5 ± 15.6
Follow-up eGFR (ml/min/1.73 m^2)	61.2 ± 26.2	8.7 ± 3.1	0.0025	28.1 ± 14.5
BMI (kg/m^2)	31.5 ± 5.9	29 ± 4.7	0.3085	29.7 ± 6.7
Serum creatinine (μmol/l)	110.7 ± 27.1	473.7 ± 162.2	<0.0001	232.4 ± 136.7
Serum albumin (g/l)	41.6 ± 2.4	35.5 ± 3.7	0.0004	38.5 ± 3.1
CRP (mg/l)	3.4 ± 3.0	4.9 ± 4.4	0.3848	4.4 ± 3.9

and [57], respectively. Recently applications of genomic and proteomic data fusion strategies have been reported in Ref. [58]. In Ref. [59] the fusion of proteomic and metabolomics profiles was investigated.

According to Ritchie et al. [60], data integration methods can be divided into multistage and meta-dimensional analysis. In the case of the former, analysis is performed on a stepwise manner. However, in the case of the latter, inference is conducted across all omics data simultaneously. In this chapter, we will focus on meta-dimensional analysis methods. However, we will also provide a summary of the multistage methods.

Multistage analysis can be divided into two subcategories: the genomic variation and the domain knowledge-guided approaches. A characteristic example of the first is the triangle principle [61]. SNP data whose correlation is statistically significant with respect to the phenotype are used as the basis of the analysis. In turn, the correlation of the SNPS as defined in the previous step with other omics data is tested. Finally, the association of a subset of the omics data (only the statistically significant are used) with the phenotype was tested. Methods belonging to this category are eQTL and mQTL methods (QTL stands for quantitative trait locus).

Domain knowledge-guided approaches refer to pathway and biological network-based integration. To this end, different methods use information provided by databases like KEGG, Reactome, and BioCarta. Examples that fall in this category are Ingenuity Pathway Analysis, Cytoscape [62], IMPaLA [63], pwOmics [64], PaintOmics [65], and ENViz [66].

Meta-dimensional integration can be performed at an early (concatenation based), intermediate, and late stage (model based). A schematic representation of this distinction can be seen in Figure 11.1. Details on these methods will be given in the following subsections.

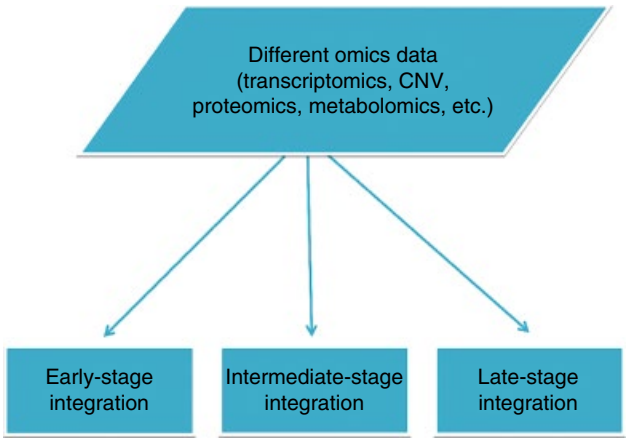


Figure 11.1 Different methods for data integration.

11.7.1 Early-Stage Integration

Early-stage integration, using concatenation-based methods, is performed by merging the different omics data into a single matrix. Since high-throughput data are measured on different scales, a major challenge posed is data transformation. As one would expect, this could lead to loss of information. Hence, the resulting model would not fully capture the dynamics of the data. Methods belonging in this class are SVM and regularization methods like LASSO. Examples of these methods provide specific information [67, 68].

11.7.2 Late-Stage Integration

In late-stage integration, or model-based integration, each omics dataset is modeled individually. Inference is conducted by weighting the fitted values from each model. Naïve Bayes, random forests, and in general ensemble classifiers are popular methods used for this

integration method. Moreover, PLS where the response variables are dummy variables offer an alternative and belong to model-based integration methods [69]. A paragraph dedicated on PLS will be given later on this section [60]. However it is recommended that this class of methods should be used only in the occasion that early- and intermediate-stage integration is not possible.

11.7.3 Intermediate-Stage Integration

In intermediate-stage integration, a joint model is constructed in order to account for all the different omics data. In this category, a rough distinction could be made in matrix factorization/kernel transformation (transformation-based integration), unsupervised methods, and network-based integration. In this review, we will be covering the first two categories. Readers interested in network-based integration are referred to Refs. [70, 71] and references therein.

11.7.4 Intermediate-Stage Integration: Matrix Factorization

Matrix factorization is defined as the decomposition of a matrix to a product of lower rank matrices. This is done for computational purposes in order to identify structural data features that were not initially apparent. A way of solving this problem is via factor models. Probably, the most popular methods in this framework are iCluster [72] and iCluster+ [73]. iCluster is a clustering algorithm that combines the idea of dimension reduction via PCA and latent factor models. The purpose of this method is to perform integrative clustering of omics data measured on the same samples and on a continuous scale while at the same time performing dimension reduction. An extension of the method is iCluster+ [73], which performs integrative clustering of mixed data type datasets (continuous and discrete).

Zhao et al. [74] reviewed different regularization techniques used in the integrative genomics framework. Data integration techniques require the use of penalties to deal with "the large number of variables, small sample size" problem. Without the use of such penalties, matrix manipulation would be an impossible task.

Canonical correlation analysis (CCA) is an exploratory statistical method that allows the analysis of the correlations that exist between two or more sets of variables. In the case of two datasets, let $\boldsymbol{X}$ be $n \times p$ matrix and $\boldsymbol{Y}$ be a $n \times q$ matrix where $n << p$ and $n << q$. Due to the high dimensionality of the problem, matrix inversion is not feasible. Tenenhaus et al. extend the idea of the regularized generalized canonical correlation analysis (RRCCA) [75, 76] to account for variable selection when there are more than three omics datasets measured on the same samples. RGCCA is the respective R package for this method.

Another method with a scope similar to CCA is co-inertia analysis (CIA). CIA aims in identifying trends and co-relationships in two different datasets. Initially, it was developed for the analysis of ecological data. In the genomics setting, it was introduced by Culhane et al. [77] and extended by Meng et al. [78] to account for multiple datasets (MCIA). While in CCA one needs to apply some sort of regularization when dealing with high-dimensional data, this is not the case for CIA and MCIA. omicade4 is the dedicated R package.

A widely used method in the field of data integration is PLS. As in the case of the two previous methods, PLS is an explorative method that aims to maximize the covariance between two or more datasets measured on the same samples. One of the drawbacks of PLS is that it fails to capture the relationships between different datasets under the presence of systematic variation. To tackle this problem, a two-way orthogonal PLS model was introduced by Trygg et al. [79].

11.7.5 Intermediate-Stage Integration: Unsupervised Methods

In the previous paragraphs, transformation-based integration methods were presented. Some of the methods also belong to the family of unsupervised methods. iCluster [72] and iCluster+ [73] are some examples. However, these models employ dimension reduction techniques. For instance, iCluster [72] combines the idea of PCA and latent variable models.

11.8 Discussion and Further Challenges

Extracting useful information from the nowadays available biological data is a challenging task that aims as deciphering the biological functions of a system at the molecular level and should open the perspective for applying high-throughput bioanalytical methods for diagnostic and prognostic tasks in the so-called personalized medicine approach. Machine learning and data mining extends traditional statistical techniques to handle problems with much higher dimensionality.

Feature selection is a key step in data mining. Choosing the important features (genes, proteins, metabolites) is essential for the discovery of important disease-related biomarkers. Although feature selection methods can help in choosing from large numbers of features the ones related to the condition studied, the results generated tend to be unstable and thus cannot be reproduced in other experiments. This has triggered the search for methods that measure the stability of feature ranking

aiming at enhancing the reproducibility of the findings. Despite being interesting for its own, for example, for biological function discovery, feature selection methods are usually combined or are part of classification procures that should enable to characterize a subject based on a personalized signature aiming to increase our understanding of disease genesis and progression and, in final consequence, to improve diagnosis and treatment options. Different classification procedures have been designed, and several CV methods have been developed for accessing their performance. In the clinical setup however, an external validation independent dataset might be mandatory.

Until recent times, different omics profiling technologies were used as single sources for knowledge discovery. There is however increasing need in the emerging field of joint analysis of omics data from genomics, transcriptomics, proteomics, and metabolomics in order to better explain biological interactions at the systems level to extract associations and causalities between different omics levels, for example, how gene modulation and activity is related to protein expression.

For clinical applications of omics datasets, the traditional approaches to medical statistics must be adopted and expanded to accommodate the high-dimensional data. New questions arise about study designs with surrogate biomarkers, such as the required sample sizes [80], the additional predictive power of the omics signatures [26, 28, 29, 81], and the prediction of patient survival and time to event using high-dimensional omics data [82, 83].

As a prospective it is now well accepted that the challenges posed by analysis of the high-dimensional omics data will continue to catalyze the development and modification of algorithms from statistics, machine learning, and information theory to improve, for example, the fundamental bioinformatics problems such as sequence alignment [84–86]. Another important issue is the emerging network biology [87] that attempts to use the different omics datasets from several sources into a biologically meaningful framework suitable for joint analysis using established methods network analysis methods [88–90].

An open problem that many omics dataset suffers from is the high percentage of missing values that may originate due to technical or biological conditions [91, 92]. This issue must be always taken into account as statistical methods such as CCA, PLS, PCA, PLS-DA, and PCA-DA are very sensitive to missing data to the extent that may render the downstream analysis results unreliable. Machine learning algorithms such as the SVM tend to learn the pattern of missing entries in the data and are somehow more robust to this issue [93].

References

1 Morris, J.S., et al., Statistical contributions to proteomic research. *Methods Mol Biol*, 2010. 641: p. 143–166.
2 Dobbin, K.K., et al., Characterizing dye bias in microarray experiments. *Bioinformatics*, 2005. 21(10): p. 2430–2437.
3 Scharpf, R.B., et al., A multilevel model to address batch effects in copy number estimation using SNP arrays. *Biostatistics*, 2011. 12(1): p. 33–50.
4 Leek, J.T., et al., Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 2010. 11(10): p. 733–739.
5 Jeffries, N., Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 2005. 21(14): p. 3066–3073.
6 Antoniadis, A., J. Bigot, and S. Lambert-Lacroix, Peak detection and alignment for mass spectrometry data. *Journal de la Société Française de Statistique*, 2010. 151(1): p. 20.
7 Baggerly, K.A., K.R. Coombes, and E.S. Neeley, Run batch effects potentially compromise the usefulness of genomic signatures for ovarian cancer. *J Clin Oncol*, 2008. 26(7): p. 1186–1187.
8 Castaldi, P.J., I.J. Dahabreh, and J.P.A. Ioannidis, An empirical assessment of validation practices for molecular classifiers. *Brief Bioinform*, 2011. 12(3): p. 189–202.
9 Quackenbush, J., Microarray data normalization and transformation. *Nat Genet*, 2002. 32: p. 496–501.
10 Smyth, G.K. and T. Speed, Normalization of cDNA microarray data. *Methods*, 2003. 31(4): p. 265–273.
11 Alfassi, Z.B., On the normalization of a mass spectrum for comparison of two spectra. *J Am Soc Mass Spectrom*, 2004. 15(3): p. 385–387.
12 Fujita, A., et al., Evaluating different methods of microarray data normalization. *BMC Bioinform*, 2006. 7(1): p. 469.
13 Karpievitch, Y., et al., A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics*, 2009. 25(16): p. 2028–2034.
14 Griffin, N.M., et al., Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. *Nat Biotechnol*, 2010. 28(1): p. 83–89.
15 Wu, H., R.A. Irizarry, and H.C. Bravo, Intensity normalization improves color calling in SOLiD sequencing. *Nat Methods*, 2010. 7(5): p. 336–337.
16 Diamandis, E.P., Point: proteomic patterns in biological fluids: do they represent the future of cancer diagnostics? *Clin Chem*, 2003. 49(8): p. 1272–1275.

17 R Development Core Team, *R: A Language and Environment for Statistical Computing*. 2015: R Foundation for Statistical Computing, Vienna.
18 Gentleman, R., *R Programming for Bioinformatics*. 2008: Chapman & Hall/CRC, Boca Raton. p. 328.
19 Gentleman, R.C., et al., Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 2004. 5(10): p. R80.
20 Ma, Y. and L. Zhu, A review on dimension reduction. *Int Statist Rev*, 2013. 81(1): p. 134–150.
21 Ringner, M., What is principal component analysis? *Nat Biotech*, 2008. 26(3): p. 303–304.
22 Boulesteix, A.L., PLS dimension reduction for classification with microarray data. *Stat Appl Genet Mol Biol*, 2004. 3: p. Article33.
23 Hilario, M. and A. Kalousis, Approaches to dimensionality reduction in proteomic biomarker studies. *Brief Bioinform*, 2008. 9(2): p. 102–118.
24 Saeys, Y., I. Inza, and P. Larranaga, A review of feature selection techniques in bioinformatics. *Bioinformatics*, 2007. 23(19): p. 2507–2517.
25 Dudoit, S. and M.J. Van Der Laan, *Multiple Testing Procedures with Applications to Genomics*, Springer Series in Statistics. 2008: Springer, New York.
26 Boulesteix, A.-L. and W. Sauerbrei, Added predictive value of high-throughput molecular data to clinical data and its validation. *Brief Bioinform*, 2011. 12(3): p. 215–229.
27 Truntzer, C., D. Maucort-Boulch, and P. Roy, Comparative optimism in models involving both classical clinical and gene expression information. *BMC Bioinform*, 2008. 9(1): p. 434.
28 Truntzer, C., et al., Comparison of classification methods that combine clinical data and high-dimensional mass spectrometry data. *BMC Bioinform*, 2014. 15: p. 385.
29 Boulesteix, A.-L. and T. Hothorn, Testing the additional predictive value of high-dimensional molecular data. *BMC Bioinform*, 2010. 11(1): p. 78.
30 Boulesteix, A.-L. and M. Slawski, Stability and aggregation of ranked gene lists. *Brief Bioinform*, 2009. 10(5): p. 556–568.
31 He, Z. and W. Yu, Stable feature selection for biomarker discovery. *Comput Biol Chem*, 2010. 34(4): p. 215–225.
32 Simon, R., Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol*, 2005. 23(29): p. 7332–7341.
33 Simon, R., Development and validation of therapeutically relevant multi-gene biomarker classifiers. *J Natl Cancer Inst*, 2005. 97(12): p. 866–867.
34 Cover, T. and P. Hart, Nearest neighbor pattern classification. *IEEE Trans Inform Theory*, 1967. 13(1): p. 21–27.
35 Hosmer, D.W. and S. Lemeshow, *Applied Logistic Regression*, Wiley Series in Probability and Statistics. 3rd ed. 2013: Wiley-Interscience Publication, Chichester.
36 Tibshirani, R., et al., Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci U S A*, 2002. 99(10): p. 6567–6572.
37 Tan, A.C., et al., Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 2005. 21(20): p. 3896–3904.
38 Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics. 2009: Springer, New York.
39 Vapnik, V., *Statistical Learning Theory*. 1998: John Wiley & Sons, Inc., New York.
40 Bishop, C.M., *Neural Networks for Pattern Recognition*, Neural Networks for Pattern Recognition. 1995: Oxford University Press, Oxford.
41 Wu, B., et al., Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, 2003. 19(13): p. 1636–1643.
42 Pengyi, Y., et al., A review of ensemble methods in bioinformatics. *Curr Bioinform*, 2010. 5(4): p. 296–308.
43 Breiman, L., Bagging predictors. *Mach Learn*, 1996. 24(2): p. 123–140.
44 Breiman, L., Random forests. *Mach Learn*, 2001. 45(1): p. 5–32.
45 Hothorn, T. and P. Bühlmann, Model-based boosting in high dimensions. *Bioinformatics*, 2006. 22(22): p. 2828–2829.
46 Bühlmann, P. and T. Hothorn, Boosting algorithms: regularization, prediction and model fitting. *Statist Sci*, 2007. 22(3): p. 477–505.
47 Buhlmann, P., Boosting for high-dimensional linear models. *Ann Stat*, 2006. 34(2): p. 559–583.
48 Boulesteix, A.L., et al., Evaluating microarray-based classifiers: an overview. *Cancer Inform*, 2008. 6: p. 77–97.
49 Bernau, C., et al., Cross-study validation for the assessment of prediction algorithms. *Bioinformatics*, 2014. 30(12): p. i105–i112.
50 Simon, R.M., et al., Using cross-validation to evaluate predictive accuracy of survival risk classifiers based on high-dimensional data. *Brief Bioinform*, 2011. 12(3): p. 203–214.
51 Braga-Neto, U.M., A. Zollanvari, and E.R. Dougherty, Cross-validation under separate sampling: strong bias and how to correct it. *Bioinformatics*, 2014. 30(23): p. 3349–3359.
52 Molinaro, A.M., R. Simon, and R.M. Pfeiffer, Prediction error estimation: a comparison of resampling methods. *Bioinformatics*, 2005. 21(15): p. 3301–3307.
53 Varma, S. and R. Simon, Bias in error estimation when using cross-validation for model selection. *BMC Bioinform*, 2006. 7(1): p. 91.

54 Braga-Neto, U.M. and E.R. Dougherty, Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 2004. 20(3): p. 374–380.

55 Nkuipou-Kenfack, E., et al., Assessment of metabolomic and proteomic biomarkers in detection and prognosis of progression of renal function in chronic kidney disease. *PLoS One*, 2014. 9(5): p. e96955.

56 Yang, D.F., et al., Transcriptomics, proteomics, and metabolomics to reveal mechanisms underlying plant secondary metabolism. *Eng Life Sci*, 2014. 14(5): p. 456–466.

57 Cisek, K., et al., The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. *Nephrol Dial Transplant*, 2016. 31: p. 2003–2011.

58 Blanchet, L., et al., Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC Bioinform*, 2011. 12(1): p. 254.

59 Oberbach, A., et al., Combined proteomic and metabolomic profiling of serum reveals association of the complement system with obesity and identifies novel markers of body fat mass changes. *J Proteome Res*, 2011. 10(10): p. 4769–4788.

60 Ritchie, M.D., et al., Methods of integrating data to uncover genotype-phenotype interactions. *Nat Rev Genetics*, 2015. 16(2): p. 85–97.

61 Holzinger, E.R. and M.D. Ritchie, Integrating heterogeneous high-throughput data for meta-dimensional pharmacogenomics and disease-related studies. *Pharmacogenomics*, 2012. 13(2): p. 213–222.

62 Shannon, P., et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 2003. 13(11): p. 2498–2504.

63 Kamburov, A., et al., Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics*, 2011. 27(20): p. 2917–2918.

64 Wachter, A. and T. Beissbarth, pwOmics: an R package for pathway-based integration of time-series omics data using public database knowledge. *Bioinformatics*, 2015. 31(18): p. 3072–3074.

65 Garcia-Alcalde, F., et al., Paintomics: a web based tool for the joint visualization of transcriptomics and metabolomics data. *Bioinformatics*, 2011. 27(1): p. 137–139.

66 Steinfeld, I., et al., ENViz: a Cytoscape App for integrated statistical analysis and visualization of sample-matched data with multiple data types. *Bioinformatics*, 2015. 31(10): p. 1683–1685.

67 Fridley, B.L., et al., A Bayesian integrative genomic model for pathway analysis of complex traits. *Genet Epidemiol*, 2012. 36(4): p. 352–359.

68 Mankoo, P.K., et al., Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One*, 2011. 6(11): p. e24709.

69 Doeswijk, T.G., et al., On the increase of predictive performance with high-level data fusion. *Anal Chim Acta*, 2011. 705(1–2): p. 41–47.

70 Gligorijevic, V. and N. Przulj, Methods for biological data integration: perspectives and challenges. *J R Soc Interface*, 2015. 12(112): pii: 20150571.

71 Wei, Y., Integrative analyses of cancer data: a review from a statistical perspective. *Cancer Inform*, 2015. 14(Suppl 2): p. 173–181.

72 Shen, R.L., A.B. Olshen, and M. Ladanyi, Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 2009. 25(22): p. 2906–2912.

73 Shen, R.L., S.J. Wang, and Q.X. Mo, Sparse integrative clustering of multiple omics data sets. *Ann Appl Statist*, 2013. 7(1): p. 269–294.

74 Zhao, Q., et al., Integrative analysis of '-omics' data using penalty functions. *Wiley Interdiscip Rev Comput Stat*, 2015. 7(1): p. 99–108.

75 Tenenhaus, A., et al., Variable selection for generalized canonical correlation analysis. *Biostatistics*, 2014. 15(3): p. 569–583.

76 Tenenhaus, A. and M. Tenenhaus, Regularized generalized canonical correlation analysis. *Psychometrika*, 2011. 76(2): p. 257–284.

77 Culhane, A.C., G. Perriere, and D.G. Higgins, Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinform*, 2003. 4: p. 59.

78 Meng, C., et al., A multivariate approach to the integration of multi-omics datasets. *BMC Bioinform*, 2014. 15: p. 162.

79 Trygg, J. and S. Wold, O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J Chemometrics*, 2003. 17(1): p. 53–64.

80 Cai, G., X. Lin, and K. Lee, Sample size determination with false discovery rate adjustment for experiments with high-dimensional data. *Statist Biopharm Res*, 2010. 2(2): p. 165–174.

81 Boulesteix, A.-L., C. Porzelius, and M. Daumer, Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics*, 2008. 24(15): p. 1698–1706.

82 Mittal, S., et al., High-dimensional, massive sample-size Cox proportional hazards regression for survival analysis. *Biostatistics*, 2014. 15: p. 207–221.

83 Stratford, J.K., et al., A six-gene signature predicts survival of patients with localized pancreatic ductal adenocarcinoma. *PLoS Med*, 2010. 7(7): p. e1000307.

84 Song, K., et al., New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Brief Bioinform*, 2014. 15(3): p. 343–353.

85 Schwende, I. and T.D. Pham, Pattern recognition and probabilistic measures in alignment-free sequence analysis. *Brief Bioinform*, 2014. 15(3): p. 354–368.

86 Vinga, S., Information theory applications for biological sequence analysis. *Brief Bioinform*, 2014. 15(3): p. 376–389.

87 Bebek, G., et al., Network biology methods integrating biological data for translational science. *Brief Bioinform*, 2012. 13(4): p. 446–459.

88 Albert, R. and A.-L. Barabási, Statistical mechanics of complex networks. *Rev Mod Phys*, 2002. 74(1): p. 47–97.

89 Needham, C.J., et al., A primer on learning in Bayesian networks for computational biology. *PLoS Comput Biol*, 2007. 3(8): p. e129.

90 Alon, U., Network motifs: theory and experimental approaches. *Nat Rev Genet*, 2007. 8(6): p. 450–461.

91 Gromski, P., et al., Influence of missing values substitutes on multivariate analysis of metabolomics data. *Metabolites*, 2014. 4(2): p. 433.

92 Hrydziuszko, O. and M. Viant, Missing values in mass spectrometry based metabolomics: an undervalued step in the data processing pipeline. *Metabolomics*, 2012. 8(Suppl 1): p. 161–174.

93 Garcia-Laencina, P.J., et al., Pattern classification with missing data: a review. *Neural Comput Appl*, 2010. 19(2): p. 263–282.

12

Epidemiological Applications in -Omics Approaches

Elena Critselis[1] *and Hiddo Lambers Heerspink*[2]

[1] *Proteomics Laboratory, Biotechnology Division, Biomedical Research Foundation of the Academy of Athens, Athens, Greece*
[2] *Department of Clinical Pharmacy and Pharmacology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*

12.1 Overview: Importance of Study Design and Methodology

One of the strategic objectives of -omics research is to identify biomarkers relating to the diagnosis of diseases, risk of health conditions, and/or prediction of adverse health outcomes. Such biomarkers may be pivotal in identifying the risk for manifesting disease (**diagnostic biomarkers**), predicting disease progression (**prognostic biomarkers**), and/or predicting response to clinical management and therapy (**predictive biomarkers**) [1]. As a result, it is envisioned that biomarkers may potentially be applied routinely both in primary and secondary disease prevention strategies and in clinical practice.

During the recent decade disproportionate efforts have been devoted to developing -omics analytical tools as opposed to clinically oriented biomarker discovery or, even more so, clinical applications [2]. Albeit limited, assessments of several biomarkers, particularly in relation to noncommunicable chronic diseases, have been undertaken in population-based studies [3]. However, several initial promising findings arising from analytical -omics research have been subsequently deemed unreliable, not reproducible, and/or biased [4–6]. This has rendered considerable skepticism, among both investigators and clinicians alike, as to whether the "-omics" revolution is able to achieve its true potential [7]. Recently, basic and clinical researchers agree that for -omics investigations to achieve its aspired potential, including the implementation of personalized medicine in routine clinical practice, fundamental aspects relating to the study design and statistical analyses in clinical -omics investigations must be dully addressed [1, 8, 9].

Issues regarding the epidemiological and statistical methods in the design and analysis of clinical -omics studies should be detailed for biomarkers evaluated for either identifying the presence or predicting the progression of chronic noncommunicable diseases. Moreover, the evaluation methods of biomarkers, as assessment indicators, for identifying health conditions will be addressed. It is aspired that by identifying the key issues that remain to be elucidated in this field, the interdisciplinary communication between basic researchers, clinical epidemiologists, and clinicians will be facilitated. Thus, future directions, including the development of new methodologies, may be implemented to further enhance the efficiency and findings of clinical -omics research [6].

12.2 Principles of Hypothesis Testing

12.2.1 Definition of Research Hypotheses and Clinical Questions

A primary issue that must be addressed prior to the initiation of clinical -omics investigations is the elucidation of clearly defined research hypotheses and clinical queries to be evaluated [10, 11]. Clinical queries may relate to screening healthy individuals or patients for the primary or secondary prevention of disease, respectively [2]. In particular, a clear description of the clinical query to be addressed, including definitions of the patient groups, comparison groups, exposures, and clinical outcomes of interest, as well as their clinical pertinence, must be stated [1]. It is advocated that once all aforementioned components of a comprehensive research hypothesis have been defined, the selection of the appropriate study design is simplified. Specifically, research hypotheses relating to the prognostic value of biomarkers require the adoption of longitudinal cohort studies [12], while those relating to the diagnostic value of markers may be additionally assessed through case–control or cross-sectional studies, albeit with notable methodological challenges [6].

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

However, particularly in analytical -omics research, broad unspecific hypotheses are often expressed so as to allow for flexibility in assessing serial biomarkers of specific protein pathways [10]. Moreover, high-throughput analyses usually follow a mode of serial hypothesis generation [6]. The implications of deriving multiple models, overfitting of data, and consequent implications on findings are detailed in following sections. In order to avoid such pitfalls and their effects on reproducibility and reliability of findings, it is recommended that a single hypothesis is predefined and accordingly evaluated in clinical -omics research. Based on the research question to be addressed, the hypothesis should be clearly defined based on the PECO (patient group–exposure (biomarker) under assessment–comparison group–outcome) or PICO (patient group–intervention–comparison group–outcome) principles. To this effect, the hypothesis tested is most proximally correlated to the type of biomarker which will be assessed.

12.2.2 Hypothesis Testing in Relation to Types of Biomarkers Under Assessment

The three most common types of biomarkers evaluated include (i) diagnostic biomarkers, (ii) prognostic biomarkers, and (iii) predictive biomarkers. In particular, **diagnostic biomarkers** aim to identify patients with disease, regardless of health outcomes [13]. For example, Metzger et al. [14] utilized CE-MS platforms to validate a biomarker pattern, consisting of 20 urinary polypeptides, which detected acute kidney injury up to 5 days earlier than serum creatinine. Particularly in noncommunicable chronic diseases, detection of disease may be targeted either among incident (i.e., newly diagnosed) or recurrent disease (i.e., prevalent cases). For example, Jahn et al. [15] utilized CE-MS platforms to identify a biomarker signature (consisting of proSAAS, apolipoprotein J, neurosecretory protein VGF, phospholemman, and chromogranin A) in cerebrospinal fluid for detecting incident Alzheimer's disease.

Prognostic biomarkers aim to identify patients with differing risks of a specific health outcome, such as disease recurrence, progression, and/or death. Hence, prognostic biomarkers constitute a baseline patient characteristic, independent of therapy or treatment, which can be applied to categorize patients according to their risk of a predefined health outcome. For example, Liu et al. [16] utilized 2D-MS technology to identify a novel biomarker, annexin A3, which predicts lymph node metastasis in lung cancer adenocarcinoma patients. Similarly, Ottervald et al. [17] identified a biomarker panel, consisting of 10 proteins, in cerebrospinal fluid that predicted relapse/remission in 70% of multiple sclerosis patients. Hence, prognostic biomarkers can be applied to estimate the likelihood of a health outcome, but cannot guide clinicians' choice of a particular treatment scheme [13].

Finally, **predictive biomarkers** aim to predict the differential outcome of a particular therapy or treatment. Hence, predictive biomarkers constitute a baseline patient characteristic that categorizes patients by their degree of response to a particular treatment. For example, Melmer et al. [18] employed proteomics approaches to identify the plasma protein afamin, which was associated with therapeutic response and survival following platinum-based chemotherapy in advanced ovarian cancer patients. Thus, predictive biomarkers are used to guide the optimal choice of treatment in patient populations [19].

12.3 Selection of Appropriate Epidemiological Study Design for Hypothesis Testing

The selection of an appropriate epidemiological study design to evaluate a research hypothesis is of quintessential importance [10]. Several long-standing clinical epidemiological study designs for the optimal assessment of putative biomarkers of chronic noncommunicable diseases have been documented (see Figure 12.1), while further specifications and guidance notes for clinical -omics applications have been proposed in recent years [20]. Appropriate applications of study design can render the evidence necessary to address a research hypothesis while concomitantly diminishing systematic bias and consequently augmenting the quality of research findings [10]. It is noteworthy that the selection of unsuitable study designs may potentially introduce sources of bias, which seldom can be corrected post hoc by statistical approaches [21]. Hence, the methodological study design (including the strategy for comparison, selection of subjects and outcomes of interest, and foresight for potential sources of confounding and bias) [22] is pivotal in securing the quality and reliability of study findings [6].

As aforementioned, a well-defined research hypothesis will for all practical purposes delineate the appropriate study design to be adopted. Epidemiological studies follow a hierarchy of design based on the level of evidence to be accumulated (see Figure 12.2). However, it is of note that, particularly in relation to biomarker research, the sequential application of study designs (i.e., hierarchical establishment of the level of evidence) is not necessarily linear [23], but rather in accordance with the clinical query to be addressed.

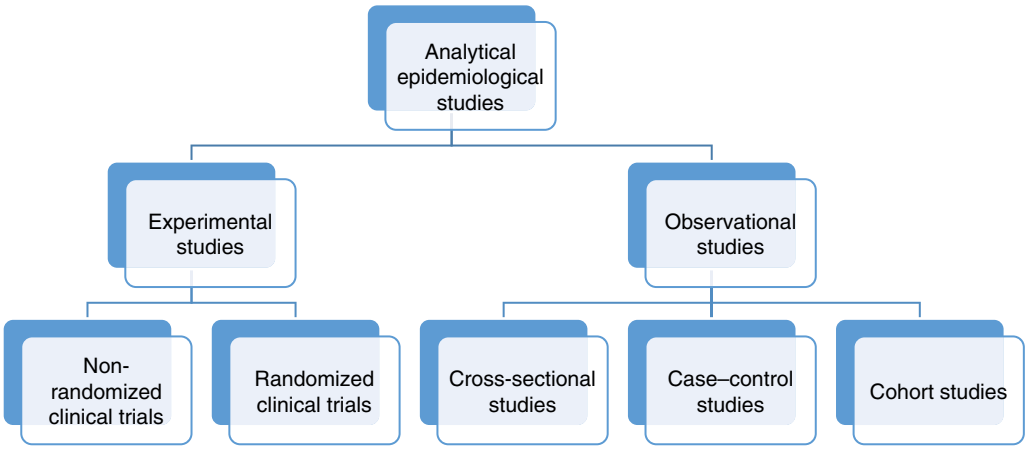


Figure 12.1 Types of epidemiological studies utilized in -omics research. The types of analytical epidemiological studies most frequently applied in -omics research are presented. Observational studies are most commonly used for the discovery and validation of diagnostic and prognostic biomarkers, while experimental studies are most frequently employed for evaluating the utility of predictive biomarkers.

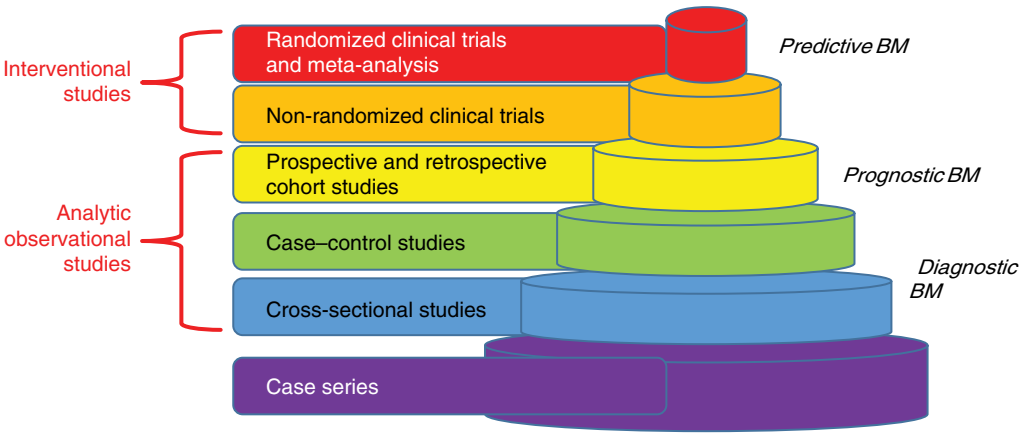


Figure 12.2 Hierarchy of research design. The hierarchy of observational and interventional epidemiological study designs, in relation to the type of biomarker (BM) under evaluation, is displayed. Optimal study designs for the evaluation of diagnostic, prognostic, and predictive biomarkers (BM) are annotated.

12.4 Types of Epidemiological Study Designs

12.4.1 Observational Studies

12.4.1.1 Cross-Sectional Studies

Cross-sectional studies are most often preferred for evaluating diagnostic biomarkers in discovery and test sets. Cross-sectional studies evaluate the applicability of biomarkers in either a single or sequential time frame. The presence of disease according to clinical and/or laboratory criteria, the presence of diagnostic biomarker in patient samples, and the mediating effects of confounding factors are assessed concomitantly (see Figure 12.3). As a result, though, the temporal association between factors under investigation cannot be confirmed. Therefore, initial findings arising from cross-sectional studies often require to be validated in robust longitudinal study designs, such as prospective cohort studies. Even so, cross-sectional studies may render insights regarding the occurrence of putative diagnostic biomarkers in a patient population while circumventing potential confounding effects of disease progression upon biomarker values [24]. In particular, cross-sectional studies are typically conducted in the general population and/or outpatient settings, particularly for prevalent diseases.

The attributes of cross-sectional studies include that they can be readily performed in the general population or outpatient settings, as well as that they are rapid, easy to implement, and financially sound. This methodology also enables the assessment of multiple biomarkers and/or various disease stages concomitantly. Hence, cross-sectional studies are particularly preferred in the

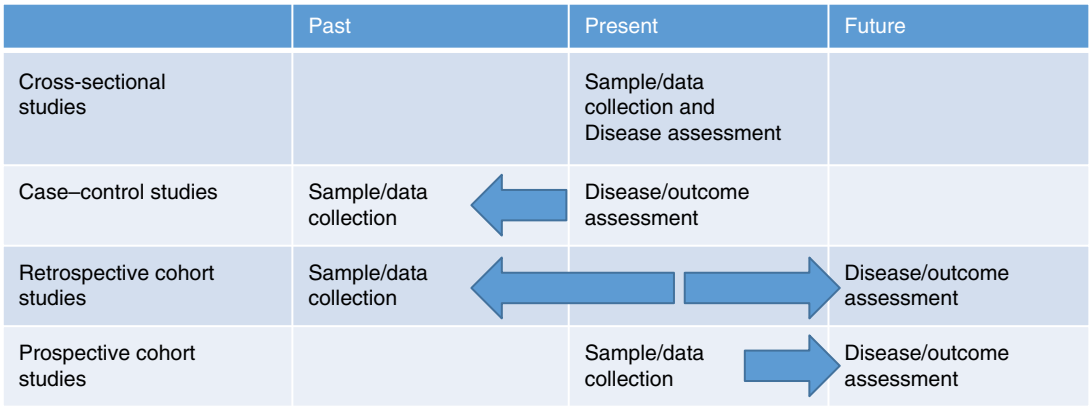


Figure 12.3 Temporal association between collecting sample and clinical information in relation to assessing the presence of disease and/or patient outcomes of interest, according to type of epidemiological study design applied.

discovery phase of diagnostic biomarkers. However, the utility of cross-sectional studies is limited, due to statistical power, when assessing rare diseases. Moreover, cross-sectional studies are particularly susceptible to confounding effects, potentially inadvertently leading to erroneous conclusions, also known as ecological fallacies. Therefore, initial findings regarding the utility of diagnostic biomarkers should be further validated in longitudinal investigations.

12.4.1.2 Case-Control Studies

Previous reviews have made suggestions on improving epidemiological design and translational potential of clinical -omics investigations [2, 25]. However, several ongoing studies still adopt cross-sectional designs, although they have inherent methodological challenges in establishing biomarker–disease associations [26], as previously detailed. Consequently, (nested) case–control studies are increasingly applied instead.

Case–control studies are often applied for the **discovery and/or test set of prognostic biomarkers**. These are characterized by the retrospective assessment of biomarkers, which predict the likelihood of health outcomes. Hence, they require the prior collection of **prediagnostic** samples. Within this context, biobank samples are most often used.

Case–control study populations are selected based on the presence or absence of the disease of interest, according to international disease classification systems (see Figure 12.3). The requirements for a proper study design including targeted context of use, selection of well-characterized cohorts, appropriate statistical analysis, and application of standardized protocols for sample collections have been highlighted [27]. Appropriate controls are most often selected based on the intended clinical context of use, following frequency or individual matching for potential confounding factors. It is noteworthy that particularly for the assessment of potential markers for secondary prevention of diseases, healthy controls do not constitute an appropriate comparison group [1]. Even so, hospital-based participants may be subjected to other extraneous factors (i.e., concomitant therapies, adjuvant tests and treatments, etc.), which may affect blood protein composition, and consequently biomarker assay values [6]. In every event, though, random selection of both cases and controls is preferable. Standardized clinical conditions and outcomes should be evaluated with documented criteria, and potential misclassifications should be avoided. For the analysis, matched cases and control patient groups are compared based on prediagnostic biomarker positivity.

The strengths of case–control studies include that they are efficient and cost-effective for evaluating prognostic biomarkers in discovery and/or test phases, particularly when examining rare diseases. Since study sample recruitment is conducted based on patient accrual, case–control studies are an excellent option for assessing prognostic biomarkers for outcomes of rare diseases, particularly in tertiary healthcare settings. Finally, the design of case–control studies allows for the concomitant testing of multiple biomarkers and thus is efficient for biomarker discovery. However, appropriate controls are often difficult to define, as they are based on the intended clinical context of biomarker use [28]. For example, in evaluating the diagnostic performance of a biomarker for detecting recurrent urinary bladder cancer, it is more appropriate to utilize urinary bladder cancer patients without recurrence, as opposed to healthy controls [29]. Additionally, case–control studies rely on previously collected biological samples, most often being either prediagnostic or biobank samples, which, though, often have not been systematically collected for the purposes of -omics studies. In practice, several limitations arise from the necessary use of such samples. First, use of

prediagnostic samples often predicates the induction of a population bias, limiting the generalizability and external validity of study findings. Second, the potential utility of prediagnostic samples depends on the quality of medical and/or hospital records for assessing potential confounding and/or mediating treatment effects. Finally, biobank samples are often not readily available to the scientific community. Moreover, the potential utility of such samples is often impeded by insufficient clinical information for assessing confounding effects. In light of these limitations of utilizing biobank samples, suggestions for improvement have been documented, including the development of suitable multidisciplinary panels for evaluating promising evidence arising from initial biomarker discovery and validation investigations, as well as consequent recommended procedures for requesting samples from biobanks to be utilized in further investigations [24, 30].

12.4.1.3 Cohort Studies

The predictive value of putative prognostic biomarkers is most appropriately assessed in prospective cohort studies. The assessment of robust randomly selected population samples is preferable to that of convenience samples, all other factors being held as equal [6, 24]. Hard endpoints of disease (i.e., patient-oriented outcomes, including disease recurrence, progression, or death), rather than surrogate outcomes, should be adopted [1, 2, 24]. Sufficient information regarding demographic and clinical characteristics of study participants should be documented [1]. Particular attention must be devoted to the evaluation of potential confounding effects of factors, which may influence either the associations assessed or determination of protein signaling per se, since such effects may introduce notable bias and render erroneous interpretations of findings [21]. However, cohort studies require a notable magnitude of human and financial resources, as well as time allotment. The necessity for the allocation of such resources may be minimized through the utilization of retrospective cohort studies, that is, facilitation through the proteome analyses of already collected biobank samples [25]. However, this option is often deferred due to practical limitations in the retrieval of adequately numbered and appropriately stored samples, as well as suboptimal documentation of clinical characteristics and potential confounding factors [11]. Particularly in the case of prognostic biomarkers for bladder cancer, while proteomic studies have revealed multiple candidate biomarkers, extensive validation of findings in large cohorts is generally missing [31].

The implementation of a prospective cohort study requires the assimilation of either (i) a **healthy cohort**, wherein biomarker sampling is conducted prior to the occurrence of disease and follow-up is conducted to detect the occurrence of adverse health outcomes (i.e., disease recurrence, progression, or death), or (ii) a **patient (inception) cohort**, wherein biomarker sampling is conducted either at diagnosis or early disease stages and follow-up is conducted to detect the disease outcomes of interest. As aforementioned, cohort studies may also be retrospective in nature, utilizing previously collected samples (i.e., biobank samples) in inception cohorts (see Figure 12.3).

Whether prospective or retrospective in nature, cohort studies hold several attributes, justifying their high ranking in the hierarchy of epidemiological study designs. First, they are highly appropriate for the validation of prognostic biomarkers. Moreover, particularly prospective cohort studies allow for study investigators to define the methods of assessment of confounding factors. Finally, multiple disease and patient-oriented outcomes can be assessed in a single study population. However, cohort studies require extensive time as well as financial and human resources to be implemented. This limitation is particularly evident when assessing biomarkers in relation to rare diseases and/or chronic diseases with lengthy disease latency periods. In addition, suboptimal patient follow-up (i.e., <80% complete follow-up) severely hinders the quality of the cohort study and validity of findings.

12.4.1.4 Health Economics Assessment

Particularly in settings inflicted by financial crises and related healthcare-associated austerity measures [32], long-term adoption of biomarkers in clinical practice requires the prior evaluation of their cost-effectiveness in healthcare settings [24, 33]. Such analyses should account not only for diverse sources of costs associated with hospitalizations but also for patient outcomes and adverse events [2, 11]. While certain -omics platforms have progressed remarkably through this pipeline, and consequently achieved widespread adoption in clinical practice (i.e., next-generation sequencing utilized for prenatal screening), such evaluations are limited in more recently developed -omics platforms, namely, that of proteomics, primarily as a result of not having yet accomplished such research progress to date [34].

12.5 Selection of Appropriate Statistical Analyses for Hypothesis Testing

Appropriate statistical analyses serve as an integral component of quality assurance of clinical -omics studies [10]. During recent years, a multitude of approaches

have been used in developing appropriate bioinformatics tools and data analysis procedures for analyzing basic proteomics research data [35, 36]. The appropriate use of statistical tests diminishes the possibility of the inadvertently false interpretations of the processes investigated [35].

The primary adversity to be addressed in clinical -omics research is limited sample size and consequent non-normally distributed variables. To this effect, testing the normality of distribution of variables is recommended by means of either the Shapiro–Wilk test or the Kolmogorov–Smirnov test [35]. Alternatively, log transformation of variable values may be applied prior to testing the normality of distribution [37]. For normally distributed variables, Student's *t*-test or two-way ANOVA (or repeated measures ANOVA, depending on the study design) may be adopted to compare continuous variables between two groups. In the event of non-normal distributions, the nonparametric Wilcoxon–Mann–Whitney test provides a robust method for comparing two population groups instead [10, 35]. In pending subgroup sample sizes, categorical variables may be compared with either the chi-squared or Fisher exact tests.

A second adversity is that multiple hypothesis testing is often employed in -omics research [37]. Consequently, multiple testing correction methods, including Bonferroni correction and false discovery rate (FDR), should be dully applied to diminish the overall type I error rate and false positive findings [10]. In the former correction method, the unadjusted *p*-values are multiplied (i.e., corrected) by the total number of tests performed. Alternatively, the FDR, which is a less conservative correction method, can be used instead.

It is of paramount importance to clarify that the probability of both false negative findings is most likely to occur among underpowered studies [1]. Hence, power calculations must be conducted to determine the required study population size, wherein the significance level is adjusted accordingly to the number of proteins tested [37]. The selection of sample size should be justified on the basis of recommended calculations of statistical power [1, 10]. It is of note that due to the importance of information related to outliers and variability, pooling of samples so as to enhance statistical power is not recommended [1].

Finally, clinical -omics applications are high-dimensional marker assays, thus challenging traditional methods of statistical analyses [25]. They are commonly identified as "high p, small n" studies (i.e., high number of variables in relation to the number of clinical samples) [8]. As a result, appropriate statistical methods are necessary [37]. Univariate and multivariate approaches are often adopted, as they may be proximally applied in classical epidemiological designs. Depending on both the study design and nature of the variables of interest, linear or logistic regression models may be applied for assessing the association with continuous or binary outcome variables of interest, respectively. Specifically, quantitative biomarkers are assessed with linear regression models, while qualitative biomarkers are evaluated using logistic regression models. Time-dependent outcomes may be assessed with Cox regression models and Kaplan–Meier curves [8]. The aforementioned methods allow for the adjustment of potential covariates, which is essential for interpreting plausible associations [28].

Alternatively, the high dimensionality of -omics data may be reduced by principal component analysis (PCA). PCA renders several benefits since it can accommodate for continuous and categorical variables and can be applied in settings where the number of variables may exceed the number of observations (i.e., "high p, small n" settings). However, the primary limitation of PCA is that it assumes that the relevant models are representative of the diversity inherent in the dataset. Even so, it is increasingly adopted as the mean for analyzing clinical -omics data [8].

Finally, the added value of biomarkers must be established in order to facilitate their uptake in clinical practice [2]. Hence, predictive scores, including demographic characteristics, clinical variables, and the biomarker of interest, could be applied in order to evaluate the net sensitivity and specificity of such models. Thus, such an approach would provide the evidence of at least incremental superiority of biomarkers under assessment in comparison with current medical practices [24].

12.6 Summary

For -omics investigations to achieve their aspired potential, fundamental aspects relating to the study design and statistical analyses in clinical -omics investigations must be dully addressed. The most appropriate epidemiological study designs for assessing diagnostic and prognostic biomarkers include cross-sectional, case–control, and cohort studies. Modes of their optimal implementation, as well as potential pitfalls, have been detailed. Statistical approaches for evaluating biomarker utility depend primarily on the research hypothesis tested, type of study design applied, and quantitative and/or qualitative nature of the biomarker under assessment. It is foreseen that application of the aforementioned principles will enhance the clinical utility of -omics research investigations.

References

1 Mischak, H., Allmaier, G., Apweiler, R., Attwood, T., Baumann, M., Benigni, A., Bennett, S. E., Bischoff, R., Bongcam-Rudloff, E., Capasso, G., Coon, J. J., D'Haese, P., Dominiczak, A. F., Dakna, M., Dihazi, H., Ehrich, J. H., Fernandez-Llama, P., Fliser, D., Frokiaer, J., Garin, J., Girolami, M., Hancock, W. S., Haubitz, M., Hochstrasser, D., Holman, R. R., Ioannidis, J. P., Jankowski, J., Julian, B. A., Klein, J. B., Kolch, W., Luider, T., Massy, Z., Mattes, W. B., Molina, F., Monsarrat, B., Novak, J., Peter, K., Rossing, P., Sanchez-Carbayo, M., Schanstra, J. P., Semmes, O. J., Spasovski, G., Theodorescu, D., Thongboonkerd, V., Vanholder, R., Veenstra, T. D., Weissinger, E., Yamamoto, T. & Vlahou, A. 2010. Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med*, 2, 46 ps42.

2 Ioannidis, J. P. 2011. A roadmap for successful applications of clinical proteomics. *Proteomics Clin Appl*, 5, 241–247.

3 Rossing, K., Mischak, H., Dakna, M., Zurbig, P., Novak, J., Julian, B. A., Good, D. M., Coon, J. J., Tarnow, L. & Rossing, P. 2008. Urinary proteomics in diabetes and CKD. *J Am Soc Nephrol*, 19, 1283–1290.

4 Petricoin, E. F., 3rd, Ornstein, D. K., Paweletz, C. P., Ardekani, A., Hackett, P. S., Hitt, B. A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C. B., Levine, P. J., Linehan, W. M., Emmert-Buck, M. R., Steinberg, S. M., Kohn, E. C. & Liotta, L. A. 2002. Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst*, 94, 1576–1578.

5 Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C. & Liotta, L. A. 2002. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, 359, 572–577.

6 Ransohoff, D. F. 2007. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J Clin Epidemiol*, 60, 1205–1219.

7 Di Meo, A., Diamandis, E. P., Rodriguez, H., Hoofnagle, A. N., Ioannidis, J. & Lopez, M. 2014. What is wrong with clinical proteomics? *Clin Chem*, 60, 1258–1266.

8 Chadeau-Hyam, M., Campanella, G., Jombart, T., Bottolo, L., Portengen, L., Vineis, P., Liquet, B. & Vermeulen, R. C. 2013. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environ Mol Mutagen*, 54, 542–557.

9 Hu, J., Coombes, K. R., Morris, J. S. & Baggerly, K. A. 2005. The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. *Brief Funct Genomic Proteomic*, 3, 322–331.

10 Cairns, D. A. 2011. Statistical issues in quality control of proteomic analyses: good experimental design and planning. *Proteomics*, 11, 1037–1048.

11 Ioannidis, J. P. & Khoury, M. J. 2011. Improving validation practices in "omics" research. *Science*, 334, 1230–1232.

12 Concato, J. 2001. Challenges in prognostic analysis. *Cancer*, 91, 1607–1614.

13 Coresh, J., Selvin, E., Stevens, L. A., Manzi, J., Kusek, J. W., Eggers, P., Van Lente, F. & Levey, A. S. 2007. Prevalence of chronic kidney disease in the United States. *JAMA*, 298, 2038–2047.

14 Metzger, J., Kirsch, T., Schiffer, E., Ulger, P., Mentes, E., Brand, K., Weissinger, E. M., Haubitz, M., Mischak, H. & Herget-Rosenthal, S. 2010. Urinary excretion of twenty peptides forms an early and accurate diagnostic pattern of acute kidney injury. *Kidney Int*, 78, 1252–1262.

15 Jahn, H., Wittke, S., Zurbig, P., Raedler, T. J., Arlt, S., Kellmann, M., Mullen, W., Eichenlaub, M., Mischak, H. & Wiedemann, K. 2011. Peptide fingerprinting of Alzheimer's disease in cerebrospinal fluid: identification and prospective evaluation of new synaptic biomarkers. *PLoS One*, 6, e26540.

16 Liu, Y. F., Xiao, Z. Q., Li, M. X., Li, M. Y., Zhang, P. F., Li, C., Li, F., Chen, Y. H., Yi, H., Yao, H. X. & Chen, Z. C. 2009. Quantitative proteome analysis reveals annexin A3 as a novel biomarker in lung adenocarcinoma. *J Pathol*, 217, 54–64.

17 Ottervald, J., Franzen, B., Nilsson, K., Andersson, L. I., Khademi, M., Eriksson, B., Kjellstrom, S., Marko-Varga, G., Vegvari, A., Harris, R. A., Laurell, T., Miliotis, T., Matusevicius, D., Salter, H., Ferm, M. & Olsson, T. 2010. Multiple sclerosis: identification and clinical evaluation of novel CSF biomarkers. *J Proteomics*, 73, 1117–1132.

18 Melmer, A., Fineder, L., Lamina, C., Kollerits, B., Dieplinger, B., Braicu, I., Sehouli, J., Cadron, I., Vergote, I., Mahner, S., Zeimet, A. G., Castillo-Tong, D. C., Ebenbichler, C. F., Zeillinger, R. & Dieplinger, H. 2013. Plasma concentrations of the vitamin E-binding protein afamin are associated with overall and progression-free survival and platinum sensitivity in serous ovarian cancer—a study by the OVCAD consortium. *Gynecol Oncol*, 128, 38–43.

19 Buyse, M., Michiels, S., Sargent, D. J., Grothey, A., Matheson, A. & De Gramont, A. 2011. Integrating biomarkers in clinical trials. *Expert Rev Mol Diagn*, 11, 171–182.

20 Oberg, A. L. & Vitek, O. 2009. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J Proteome Res*, 8, 2144–2156.

21 Ransohoff, D. F. 2005. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer*, 5, 142–149.

22 Pepe, M. S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*, New York, Oxford University Press, p. 168–173.

23 Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M. & Yasui, Y. 2001. Phases of biomarker development for early detection of cancer. *JNCI*, 93, 1054–6101.

24 Mischak, H., Ioannidis, J. P., Argiles, A., Attwood, T. K., Bongcam-Rudloff, E., Broenstrup, M., Charonis, A., Chrousos, G. P., Delles, C., Dominiczak, A., Dylag, T., Ehrich, J., Egido, J., Findeisen, P., Jankowski, J., Johnson, R. W., Julien, B. A., Lankisch, T., Leung, H. Y., Maahs, D., Magni, F., Manns, M. P., Manolis, E., Mayer, G., Navis, G., Novak, J., Ortiz, A., Persson, F., Peter, K., Riese, H. H., Rossing, P., Sattar, N., Spasovski, G., Thongboonkerd, V., Vanholder, R., Schanstra, J. P. & Vlahou, A. 2012. Implementation of proteomic biomarkers: making it work. *Eur J Clin Invest*, 42, 1027–1036.

25 Bonassi, S., Taioli, E. & Vermeulen, R. 2013. Omics in population studies: a molecular epidemiology perspective. *Environ Mol Mutagen*, 54, 455–460.

26 Ransohoff, D. F. 2002. Challenges and opportunities in evaluating diagnostic tests. *J Clin Epidemiol*, 55, 1178–1182.

27 Vlahou, A. 2013. Network views for personalized medicine. *Proteomics Clin Appl*, 7, 384–387.

28 Mischak, H., Critselis, E., Hanash, S., Gallagher, W. M., Vlahou, A. & Ioannidis, J. P. 2015. Epidemiologic design and analysis for proteomic studies: a primer on -omic technologies. *Am J Epidemiol*, 181, 635–647.

29 Frantzi, M., Van Kessel, K. E., Zwarthoff, E. C., Marquez, M., Rava, M., Malats, N., Merseburger, A. S., Katafigiotis, I., Stravodimos, K., Mullen, W., Zoidakis, J., Makridakis, M., Pejchinovski, M., Critselis, E., Lichtinghagen, R., Brand, K., Dakna, M., Roubelakis, M. G., Theodorescu, D., Vlahou, A., Mischak, H. & Anagnou, N. P. 2016. Development and validation of urine-based peptide biomarker panels for detecting bladder cancer in a multi-center study. *Clin Cancer Res*, 15, 4077–4086.

30 Labaer, J. 2012. Improving international research with clinical specimens: 5 achievable objectives. *J Proteome Res*, 11, 5592–5601.

31 Frantzi, M., Latosinska, A., Fluhe, L., Hupe, M. C., Critselis, E., Kramer, M. W., Merseburger, A. S., Mischak, H. & Vlahou, A. 2015. Developing proteomic biomarkers for bladder cancer: towards clinical application. *Nat Rev Urol*, 12, 317–330.

32 Karanikolos, M., Mladovsky, P., Cylus, J., Thomson, S., Basu, S., Stuckler, D., Mackenbach, J. P. & Mckee, M. 2013. Financial crisis, austerity, and health in Europe. *Lancet*, 381, 1323–1331.

33 Horvath, A. R., Lord, S. J., Stjohn, A., Sandberg, S., Cobbaert, C. M., Lorenz, S., Monaghan, P. J., Verhagen-Kamerbeek, W. D., Ebert, C. & Bossuyt, P. M. 2014. From biomarkers to medical tests: the changing landscape of test evaluation. *Clin Chim Acta*, 427, 49–57.

34 Anderson, N. L. 2010. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem*, 56, 177–185.

35 Biron, D. G., Brun, C., Lefevre, T., Lebarbenchon, C., Loxdale, H. D., Chevenet, F., Brizard, J. P. & Thomas, F. 2006. The pitfalls of proteomics experiments without the correct use of bioinformatics tools. *Proteomics*, 6, 5577–5596.

36 Mischak, H., Apweiler, R., Banks, R. E., Conaway, M., Coon, J., Dominiczak, A., Ehrich, J. H., Fliser, D., Girolami, M., Hermjakob, H., Hochstrasser, D., Jankowski, J., Julian, B. A., Kolch, W., Massy, Z. A., Neususess, C., Novak, J., Peter, K., Rossing, K., Schanstra, J., Semmes, O. J., Theodorescu, D., Thongboonkerd, V., Weissinger, E. M., Van Eyk, J. E. & Yamamoto, T. 2007. Clinical proteomics: a need to define the field and to begin to set adequate standards. *Proteomics Clin Appl*, 1, 148–156.

37 Urfer, W., Grzegorczyk, M. & Jung, K. 2006. Statistics for proteomics: a review of tools for analyzing experimental data. *Proteomics*, 6 Suppl 2, 48–55.

Part II

Progressing Towards Systems Medicine

13

Introduction into the Concept of Systems Medicine

Stella Logotheti[1] *and Walter Kolch*[2]

[1] *Proteomics Laboratory, Biomedical Research Foundation, Academy of Athens, Athens, Greece*
[2] *Systems Biology Ireland and Conway Institute of Biomolecular & Biomedical Research; and School of Medicine, University College Dublin, Dublin, Ireland*

13.1 Medicine of the Twenty-First Century: From Empirical Medicine and Personalized Medicine to Systems Medicine

Until the mid-twentieth century, healthcare was experience based. The so-called empirical therapy relied mainly on observational data and the clinical experience of healthcare practitioners who, in the absence of complete or accurate molecular information on a disease's underlying mechanism, prescribed therapeutic solutions using practice-derived guidelines. Empirical therapy was applied based on symptoms occurrence and often before the confirmation of a definitive diagnosis, in order to avoid patient management delays, which might worsen the patient's disease progression. For instance, wide-spectrum antibiotics were given to a person before the specific bacterium causing an infection is known. Fighting an infection sooner rather than later is important to minimize morbidity, risk, and complications, so there is value in getting started with the symptomatic information available rather than waiting for accurate information. These treatment decisions were guided by observations of drug efficacy and safety on whole populations rather than individuals. The main decision maker was the treating physician who was called upon to take the initiative against an anticipated and likely cause of a manifested disease. However, putting a diagnosis on limited causal evidence and estimating risk-to-benefit ratio for a drug prescription was often akin to Damocles' sword hanging over the healthcare practitioners' heads. Drug overprescription, resistance to therapy, difficulties with nonresponsive patients, side effects, drug interactions, and implications of other patient-specific conditions were common issues.

Triggered by the revolution of molecular biology in the second half of the twentieth century, the concept of personalized therapy came to the fore. Personalized therapy introduced the idea of "the right treatment, for the right patient, at the right time" [1]. The cornerstone for this approach has been the significant advancement in the understanding of the molecular mechanisms that underlie several diseases, mainly promoted by the decoding of the complete human genetic code. It was realized that each patient has a characteristic genetic background that can affect the disease incidence, the course of the disease, and the response to therapy. It was also realized that several diseases can be asymptomatic and do not manifest until it is too late for effective patient management. Thus, early detection of a disease based on reliable biomarkers could prevent disease incidence and progression and lead to better prognosis and outcome. The ambition of personalized medicine is to tailor diagnosis and treatment to each patient within a population. For final decision making, the physician receives assistance by input from molecular diagnostics. Recently, interactions among the clinicians, academia, industry, and the pharmacopoeia have been developed to this end. The aim of personalized medicine also implies that the patients become more involved in their treatment, in terms of prevention and prediction. In recent years, genome-wide association clinical studies in large cohorts of patients have shed light on disease aspects and response to drugs in correlation with patients' genetic background. This was anticipated to lead to safer decisions for patient management [2].

However, there was another aspect that has largely gone unnoticed in the concept of personalized therapy: the fact that each individual patient in addition to having a unique genetic background also has a unique history of how he/she grew up, to which environment he/she was

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

exposed, what his/her dietary habits were, what his/her lifestyle is, and so on. As a consequence each individual presents differences in cell milieu, microbiome, host-defense interactions, immune status, epigenome, and so on. Even the socioeconomic factors and culture may affect an individual's clinical condition, outcome prediction, and response to therapy. These factors overall create a unique external and internal environment for each patient, which affect the vulnerability to a disease, the course of the disease, and the severity of a disease. They are also associated with comorbidities as well as the individual's response to drugs and relative side effects. It was soon realized that personalized genotypes are associated with personalized phenotypes [3]. Therefore, comprehensive approaches in the setting of each disease are needed. This is a gap that systems medicine is anticipated to fill in at the dawn of the twenty-first century. It is an effort to understand the complex interactions within the human body in light of a patient's genetic background, behavior, and environment. In order to achieve this goal, wide cross-talks and networking of clinicians, academia, industry, pharmacopoeia, and the patients are needed [4].

At first glance, such systemic approach requires too many parameters (such as genetic, epigenetic, biochemical, immunological, environmental, developmental, and social) to be defined and correlated on several levels, which make the approach seem a Herculean task. However, looking at it more closely, we understand that many of these parameters in relation to a disease setting have already been defined in each individual field of genetics, epigenetics, epidemiology, immunology, and so on. The challenge is to correlate and integrate these diverse parameters at the interface of the different scientific fields in a comprehensive and systemic way that would further enable individualized prevention and treatment optimization for a specific disease. Advancements in research in multiple disciplines during the last decades have provided most of the pieces of the puzzle. Importantly, the now affordable sequencing of individual human genomes has provided the opportunity to gain detailed information on each patient [5]. The current challenge is how all these pieces will be combined and integrated in order to develop a novel comprehensive strategy for each patient's diagnosis, and treatment.

The advanced -omics technologies in combination with modeling and computational approaches will play a catalytic role toward establishing this novel strategy for patient treatment [3]. Systems medicine aims at the management of a disease before the establishment of symptoms and at the maintenance of wellness in the population. Similar to personalized medicine, it takes into account the genomic characteristics of each patient in order to produce therapeutic solutions tailored to the needs of each individual. However, in this case, individuality of each patient is defined based on high-throughput measurements of many other molecular entities in addition to genome profiling. Molecular data are integrated with clinical information and environmental parameters to produce a comprehensive profile for each patient. Another novelty is the participation and interactivity of patients with clinicians to generate information, influence decision making, and participate in the maintenance of their wellness through lifestyle adaptations [6] (Figure 13.1).

13.2 The Emerging Concept of Systems Medicine

13.2.1 The Need for Establishment of Systems Medicine and the Field of Application

At the dawn of the twenty-first century, the world is facing an unprecedented situation of aging population. In particular, it is projected that the population of <5 years old will be outnumbered by the population of >65 years old before 2020. This means that we will soon have more old people than children and more people at extreme old age than ever before. Population aging is a powerful and transforming demographic force, which will define the future trends of the current healthcare approaches. As both the proportion of older people and the length of life increase throughout the world, key questions arise. Issues that have emerged are whether population aging can be accompanied by longer periods of good health, well-being, social engagement, and productivity or it will be associated with more illness, disability, and dependency. Other issues are how aging will affect healthcare and social costs, if a physical and social infrastructure can be established that might foster better health and well-being in older age, and how population aging will differentially affect low-income, developing countries as compared with the industrialized and developed ones [7].

This situation is aggravated by an increase in the prevalence and burden of several chronic diseases that are largely age related, for example, diabetes, cancer, cardiovascular diseases, and chronic respiratory diseases. In such diseases, knowledge of the background and the overall history of a patient, in addition to their genetic and epigenetic background, is a prerequisite for better and sustainable disease management.

In addition to these noncommunicable diseases, there is a recurrence of communicable diseases due to emergence of resistant strains of bacteria and viruses that can now spread quicker than ever before due to the increased movement and migration of large populations. Population mobility vastly facilitates the emergence of

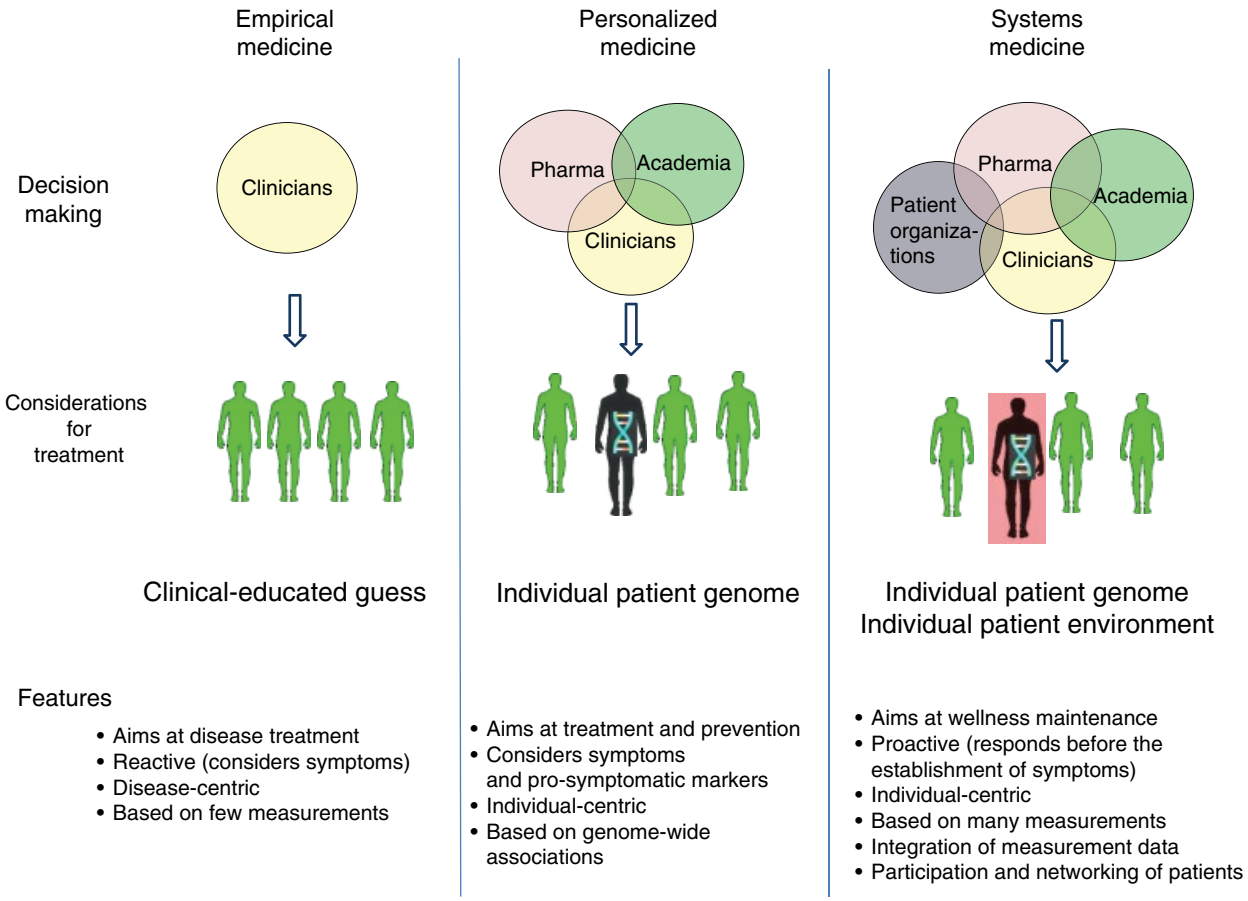


Figure 13.1 Overview of the empirical medicine, personalized medicine, and the emerging systems medicine.

resistant pathogen strains and the rapid evolution of diseases from epidemics to pandemics [7].

Personalized therapy approaches alone are insufficient to address the emerging healthcare challenges [8]. Except for the increased need for personalized and predictive therapy, which is being addressed by personalized therapy, these trends create an extra need of reenforcing preventive measures and proactive patient participation in healthcare. To do so, appropriate data systems and research capacities need to be developed, aiming to monitor and understand patterns and relationships. Better research coordination is also needed, which could unveil the most cost-effective ways for improving healthcare and wellness in countries at different stages of economic development and with varying resources [8]. Toward this purpose systems biology has been recruited to fill in the gaps of personalized therapy.

13.2.2 Bridging the Gap: From Systems Biology to Systems Medicine

During the last 15 years, the successful emergence of systems biology as a research field in its own right has revolutionized our understanding of the complexity of the human body and the diseases we develop. Systems biology is an interdisciplinary approach that focuses on complex interactions within biological systems in holistic rather than a reductionistic manner. Overall, it is a coordinated effort to understand the properties of a system as a whole instead of focusing on the properties of its individual components. This comes from the realization that the properties of an individual component may be significantly influenced by its environment. For instance, until now, the focus was on characterizing the properties of individual molecules, assuming that a molecule has specific functions and is associated with certain phenotypes. However, this molecule is part of a complex molecular network, and its effect on the cellular phenotype is highly influenced by other components of the network. Hence, its properties as an individual unit are not the same when it is viewed as a part of a larger network [9]. Biological systems are now viewed as collections of networks operating at multiple levels, ranging from molecules, cells, tissues organisms, and populations [10]. To this end, systems biology aims at the computational and mathematical modeling of complex

biological networks across all levels in order to unveil properties of cells, tissues, and organisms functioning as a system.

This concept possibly signals the end of an era of reductionism, which had prevailed during the previous decades and was characterized by elucidating functions of molecules or pathways in isolation [11]. Breaking down a complex problem into smaller and simpler units has facilitated the analysis of its individual parameters. This effort has now reached a plateau. Major diseases are complex and multifactorial, and, thus, reductionism struggles to provide solutions. Instead, viewing these diseases through a systems biology lens might be a more accurate perspective on the rules that dictate their pathogenesis and appropriate management. This is the main reason why reductionism tends to be gradually replaced by systemic approaches, especially in the study of chronic diseases [11].

Biomedical sciences research now moves from a reductionist approach to a systemic approach and attempts to understand pathophysiology in an integrative manner. To this end, the rapidly increasing amounts of high-throughput "big data" and other relevant quantitative biological/medical data that are becoming available and accessible are exploited. For extracting and mining meaningful information from these data and make the most of it in the context of a complex disease management, concepts of a wide variety of sciences, such as mathematics, physics, and engineering, are being "co-opted" into the biological sciences. This attempt also requires cooperation of experts at the interface of these disciplines.

Technological advancements of the previous century have produced adequate new knowledge both in preclinical and in clinical setting. The new century's challenge is finding ways to organize and integrate knowledge and information and to establish overarching guidelines for cooperation among multiple disciplines and stakeholders for comprehensive disease management, including prevention, prediction, and therapy. Years of research have generated detailed information on all levels of organization of biological components, from molecules and cells to ecosystems. The time has come for integration of the available information in order to understand how all these components work together as systems [10].

Systems medicine is a newly emerging area aiming to produce a conceptual and theoretical framework for the interpretation and implementation of the rules that govern this new way of organizing biomedical information. A systems approach to healthcare that will be facilitated by multidisciplinary collaboration and networking among academia, the clinic, the pharmacopoeia, and the patients is key for its successful implementation and the achievement of a paradigm shift in healthcare. Systems medicine can build on the successes in the field of systems biology that defines the human body as the multidimensional ensemble of networks. In other words, systems medicine perceives the human body as an onion-layered arrangement of networks-within-networks and attempts to comprehend the rules governing their collective behavior [12].

In Europe, the significant potential of systems medicine has been recognized since 2004, and 73 health projects for research, training, and systems biology infrastructure have already been funded [8].

13.2.3 Attempting a Definition

Systems medicine is a rapidly changing field still in its infancy. As is often the case, there is no assigned way to define it. Its definition has been recently discussed by the Coordinating Systems Medicine across Europe (CasyM) panel, that is, Europe's official multidisciplinary consortium that intents to develop an implementation strategy for systems medicine. As a result, the European expert panel came up with the following definition for systems medicine:

> Systems Medicine is the implementation of Systems Biology approaches in medical concepts, research and practice. This involves iterative and reciprocal feedback between clinical investigations and practice with computational, statistical and mathematical multiscale analysis and modeling of pathogenetic mechanisms, disease progression and remission, disease spread and cure, treatment responses and adverse events as well as disease prevention both at the epidemiological and individual patient level. As an outcome, Systems Medicine aims at a measurable improvement of patient health through systems-based approaches and practice.
>
> *(https://www.casym.eu/lw_resource/datapool/_items/item_328/roadmap_1.0.pdf)*

13.2.4 The Network-Within-a-Network Approach in Systems Medicine

Complex systems can be graphically represented as networks. The components represent the nodes of the network, whereas their interassociations form the links or edges. Networks in technological, social, and biological systems have common designs that are governed by fundamental and quantifiable organizing principles [13]. For cellular networks, genes, proteins, and metabolites are represented as nodes and the interactions among

them as links. Within networks, a fraction of the nodes have multiple links and serve as pivotal hubs that can exert large individual effects. The vast majority of nodes have few links and present milder effects if individually modulated. An additional significant characteristic of networks is that functionally related nodes tend to be highly interconnected and co-localize in networks, thereby forming modules that participate in common biological functions [14].

Systems medicine takes into account that a human is a dynamic network within other networks. Networks are formed at every level of biological organization (molecular, cellular, organ, individual, and social/environmental). This means that each upper level consists of networks that are formed in lower levels. Molecules form networks within cells. Cells communicate and interact with each other and form networks (i.e., tissues) within organs. Organs form networks (i.e., organ systems) within the human body. Each human interacts with other humans and constitutes a node of social and environmental networks (Figure 13.2). Each of these networks has highly dynamic rather than static features, which greatly influences how each human will react in the context of a pathological condition [15].

13.2.4.1 Great Expectations for Systems Medicine: The P4 Vision

In 2004, the pioneers of systems medicine have set forth the so-called P4 vision. P4 medicine is a healthcare system that is predictive, preventive, personalized, and participatory. It is expected that systems medicine will catalyze the development of this emerging field, resulting in a paradigm shift in healthcare.

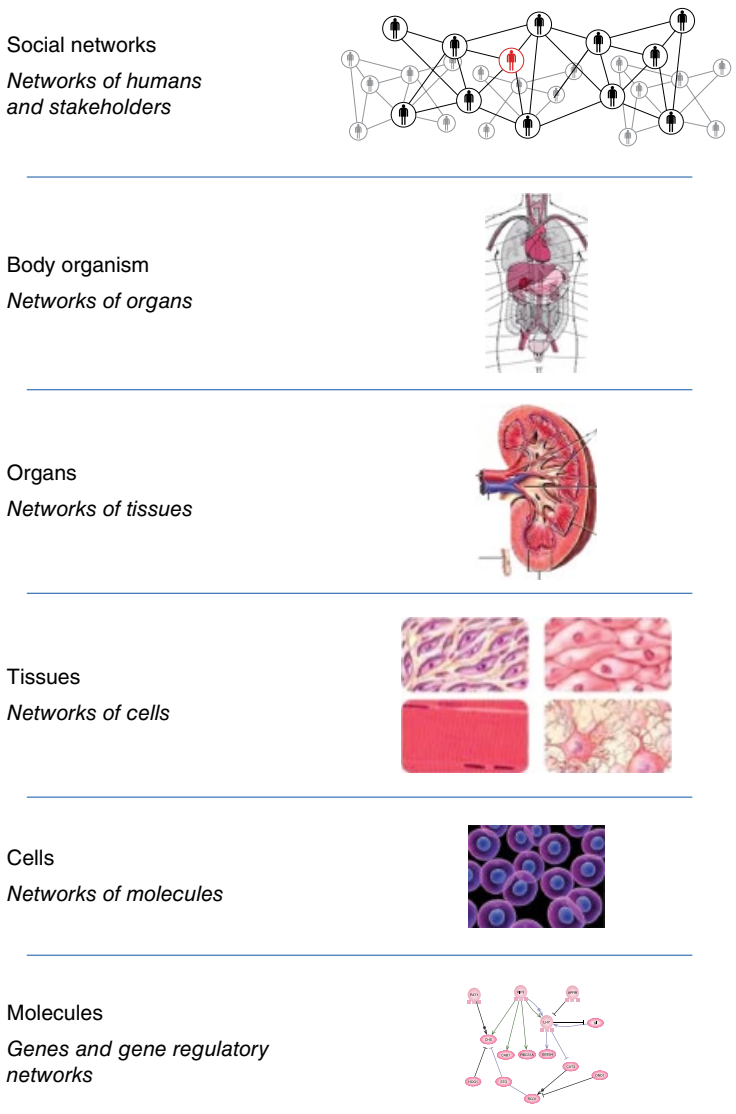


Figure 13.2 The network-within-a-network concept as a cornerstone of the systems medicine approach. *Source:* Vogt et al. [15]. https://link.springer.com/article/10.1007/s11019-016-9683-8. Licensed under CC BY 4.0.

In detail, three recent developments have converged to give rise to the concept of P4 medicine:

The increasing ability of systems biology and systems medicine to unveil the biological complexity of a disease

The information revolution, which has facilitated collecting, storing, integrating, meta-analyzing, and communicating medical information and patient data at a large scale

Access to disease-related information for everyone involved in disease management, including healthcare professionals, patients, healthy individuals, scientists, clinicians, industry, and policy makers

The enormous amounts of data produced during the previous years must now be integrated to achieve a comprehensive understanding of a human disease. To this end, it has been proposed that detailed molecular profiles along with epidemiological parameters should be obtained for as many people as possible. Then, these data could be analyzed using advanced computational tools. This will create new models that will shed light on how all elements in the biological networks interact to produce healthy and disease states. These models are anticipated not only to decode the "black box" of complex diseases but also to quantify what it means to be healthy [10].

P4 medicine is comprised of four elements that are anticipated to be significantly advanced by the collection, storage, accessibility, and exploitation of medical information across multiple levels:

a) *Prediction*: If as much information as possible from several disciplines is collected and systemically evaluated, it could be exploited in such a way so as to assist the detection of a disease at an earlier stage. This poses several advantages: the disease is easier and more cost-effective manageable, the drugs are deployed in more effective way reducing side effects, the patient outcomes are better, and the health economics are improved.
b) *Prevention*: Exploitation of information could contribute to avoiding the disease in the population, by, for example, decreasing exposure to the associated risk factors. This is anticipated to overall decrease disease incidence and prevalence in the general population.
c) *Personalization*: By unveiling which biological networks are perturbed in diseases, new therapeutic targets can be selected more wisely and can be tailored to the needs of individual patients. To understand what each patient needs though, the generation of "personal data clouds" is required. This means that for each patient, data should be collected and stored over time, regarding their genome, blood tests, lifestyle, epidemiological data, activity and stress levels, transcriptome, metabolome, microbiome, traditional medical records, and so on. These comprehensive datasets constitute a database of personalized information about each person's health and disease condition [10].
d) *Participation*: The new component of P4 medicine compared with the concept of personalized therapy is the participation of citizens. Systems medicine can be expanded out from hospitals and clinics into homes, workplaces, and schools. The citizens themselves are concerned about their health, judging by the fact that one in three Americans have gone online to investigate a medical condition. This concern creates an interaction between citizens and the healthcare system with regard to the exchange of information. Citizens can provide information that can be accumulated to a central "data pool" and be combined with other data in this pool. In turn, the citizens can retrieve personalized information that will enable them to either pro-act or alter their habits toward a lifestyle that will benefit their personal well-being. On the one hand, the participatory parameter is a monitoring of each individual health status provided by each individual per se, which will assist systems medicine to understand disease complexity. On the other hand, this effort ensures the citizens' right to be informed, concerned, and proactive in real time about their health status [10].

13.2.4.2 How Systems Medicine Will Transform Healthcare

P4 medicine is anticipated to induce transformations that will form the basis for a new healthcare system. The major transformations are as follows [10]:

1) Clinical studies will be transformed into population-based studies. Instead of relying on data from limited test cohorts, data of the whole population will be used, aided by sophisticated computational analysis for generating more comprehensive models of disease.
2) Diagnosis and treatment based on symptoms will be replaced by treatment choices based on the cellular and molecular patient profiles. The optimization of parameters for personalization therapy, through acquiring and analyzing more data corresponding to each individual patient, is anticipated to improve efforts for treating the right patient with the right drug at the right time accruing better results and far greater cost-effectiveness.
3) Basic and applied sciences will be integrated with clinical practice, prevention, and wellness care.

4) Healthcare is not restricted to the clinic. Instead, it includes the active preservation and enhancement of wellness by individuals in their homes and workplaces.
5) A new wellness industry is emerging that will effect economic growth in the twenty-first century.

13.2.4.3 The Five Pillars of Systems Medicine

Managing a complex disease within the concept of systems medicine by correlating large datasets all-with-all is an appealing goal, but, at the same time, challenging to achieve. However, the recent advances of bioinformatics and systems biology can facilitate such efforts. In general, the pioneers of systems medicine have defined five pillars that could support the handling of large datasets for the development of systems medicine [10]. These are as follows:

State-of-the-art technologies are available and accessible for generating data regarding health and disease states for each person at multiple levels (from their molecular profiling to their behavior within populations).

A digital infrastructure is emerging to link participating discovery science and clinical institutions, as well as patients/consumers, and healthcare providers.

Personalized data clouds that provide information about multiple aspects of each individual, such as molecular profiling, social and demographic parameters, medical history, and genetic and phenotypic characteristics.

New analytic techniques and technologies that can derive actionable knowledge from the data.

Systems models for understanding the health status of each individual in terms of dynamic network states.

13.2.4.4 The Stakeholders of Systems Medicine

The concept of systems medicine necessitates the involvement and interaction of several social entities, that is, the academia, the clinicians, the industry, the funders, the citizens, and the policy makers. One major difference of systems medicine from personalized medicine with regard to the involved stakeholders is that it foresees a more active participation for the patients. In personalized medicine, academia provides the knowledge and the tools for effective disease targeting, prediction, and prevention in the preclinical setting, in close collaboration with the clinicians. Trials are being designed in the clinical setting and run by experienced clinicians, who work together with sponsors (i.e., pharmaceutical industry, biotechnology companies, R&D companies). Successful therapeutic approaches and interventions as well as novel diagnostic markers are then being carried forward into medical practice. Policy makers were taking the new advances into account to modify the existing healthcare frames and legislation, for example, by approving novel diagnostic tests for routine examination, including new molecular entities as approved drugs [16].

Systems medicine relies largely on the inclusion and engagement of citizens to this effort. Instead of treating patients as "trial subjects" according to the Helsinki Declaration [17], it considers citizens as important "collaborators" in this effort. In systems medicine, citizens are essential components of the stakeholders' network, whose interaction with other components is anticipated to dynamically contribute to the transformation of healthcare toward more efficient, safe, and cost-effective methods. Active, informed, and networked patients have the ability to both provide the large datasets that are essential to power healthcare innovation and reduce incidence and prevalence of diseases in the population by being educated how to choose health beneficial lifestyle decisions [10].

13.2.4.5 The Key Areas for Successful Implementation

Healthcare systems across countries present heterogeneity, which might pose some obstacles for the implementation of systems medicine across European countries in a consistent manner. Therefore, in order to successfully implement systems medicine, it is important to exchange experience in developing new systems medicine infrastructures to connect the initiatives and to harmonize their activities. Establishment of stable local networks that continuously bring together the key stakeholders (patients, care managers, care personnel, technology providers, entrepreneurs, policy makers, and regulators), generation of reference sites to identify and overcome local barriers and challenges, exchange of good practice and transfer of knowledge among reference sites, and adaptation of current legal, financial, regulatory, incentive, and educational frameworks are crucial steps toward this goal [18].

The CASyM consortium in their last workshop defined 10 key areas that have to be enhanced in order to facilitate implementation of systems medicine in the heterogeneous landscape of European healthcare systems.

13.2.4.6 Improvement of the Design of Clinical Trials

Due to their current design, clinical trials often lack high-quality, accessible, and standardized datasets. Data on each patient are produced by analysis of their samples and frequent monitoring of their health condition. However, these data are produced in an inconsistent manner, are not readily accessible, and require meta-analyses. It is anticipated that phases I, II, III, and IV will be redesigned in order to favor a large-scale, patient-oriented approach. Actions that are suggested to be taken toward this goal include targeting of pathways instead of individual molecules, evaluating comorbidities

and drug interactions, considering adaptive trial design, development of *in silico* clinical trial design, access to datasets, development of robust tools for integration of datasets using user-friendly interfaces for the convenience of translational researchers, and categorization of modeling approaches [8].

13.2.4.7 Development of Methodology and Technology, with Emphasis on Modeling

Data produced over the years using a wide range of technologies and sources should be integrated in a comprehensive and plausible manner. Computational models could be a useful aid in this effort, offering multiscale models. The large amount of different data and data types as well as the many possible ways to combine them render data integration difficult by current, mainly regression-based, models. Advanced computational methods of analysis and modeling are needed. Despite research advances, it remains a sad fact that the vast majority of investigational drugs that present satisfactory efficacy and safety profiles upon preclinical testing are not consistently successful during the transition "from bench to bedside," thus leading to unacceptably high and expensive failure rates of clinical trials. Establishing and introducing appropriate models is anticipated to reduce the inconsistency between the preclinical and clinical outcomes. The future challenge of mathematicians and bioinformaticians working in the context of systems medicine would be to produce reliable models from *in vitro* and *in vivo* preclinical studies, which could better predict outcomes in the clinical setting. For instance, drug response and disease risk could be better predicted upon successful integration of data from the preclinical and the clinical setting [8].

13.2.4.8 Generation of Data

Large amounts of data are required for systems medicine. Data generation should rely on the systematic deployment of high-throughput methods, aim to address a clinically relevant research question, be based on suitable information from clinical sample analysis and patient medical records, be appropriate for generation of predictive models, and be properly validated [8].

13.2.4.9 Investment on Technological Infrastructure

Exploitation of knowledge from different disciplines requires state-of-the-art infrastructure to support data handling, storage, sharing, and access. This needs to be accompanied by the establishment of standards for the assessment of the quality of data and of mathematical models used. It also requires core data services and management, as well as generation of a reliable archive and technical support [8].

13.2.4.10 Improvement of Patient Stratification

The vast majority of clinical trials fail due to insufficient efficacy or safety. This necessitates better stratification of patients to be enrolled into trials based on integrated molecular and clinical patient profiles. Stratification of patients based on their genomic, proteomic, and/or metabolomics profile could enable tailoring of treatment to their personal parameters. In this context, systems medicine can contribute to finding specific combinations of disease-associated genes per patient. The -omics technologies, which enable the generation, the integration, and the interpretation of high-throughput data for each patient and for the whole patient population, are anticipated to play a crucial role in stratification. Instead of focusing on expression of individual disease-associated genes for patient stratification, as was the case until now with largely insufficient power, stratification will be based on the expression of specific combinations of highly interconnected disease-relate genes, proteins, and metabolites organized into networks. Analysis of high-throughput patient molecular data is anticipated to reveal such combinations (which have been defined as "modules"), which characterize disease processes and can be targeted accordingly [8].

13.2.4.11 Cooperation with the Industry

The term "industry" may include a wide range of companies, from multinational pharmaceutical enterprises to technology-driven personal diagnostics operations. In the context of systems medicine, emphasis is on public–private partnerships. These should combine basic research, translational research, clinical research, and healthcare aspects. They also have to focus on the need of the patient. Their interaction is anticipated to determine the gaps and technological challenges that future R&D should address. Additionally, in light of the high failure rates of clinical trials, the industry has hesitated in making large-scale investments in innovative approaches for disease management. The introduction of a systemic framework could produce a more robust proof-of-concept portfolio, which may benefit industry and encourage further investment to innovation [8].

13.2.4.12 Defining Ethical and Regulatory Frameworks

The gathering and handling of this amount of personal information raises ethical, social, regulatory, and financial issues. First of all, patient participation in the systems medicine effort requires improving and promoting health literacy and education. Secondly, a pan-European legal framework and corresponding educational programs are required to protect patient privacy and data. In addition, collection and availability of data able to predict illness and disease will significantly influence current models of health insurance. This will inevitably necessitate revisiting

current healthcare insurance schemes and establishing a novel framework for health insurance and insurers, aiming to ensure healthcare for all [8].

13.2.4.13 Multidisciplinary Training

For decades, the prevailing trend for scientists and clinicians was reductionism and overspecialization. They had to be focused on certain topics in order to address them more thoroughly. In this decade, the rise of systems biology has favored the exact opposite trend of taking a look at the "big picture." This requires a scientist or clinician to have training in a wide range of disciplines in order to be able to integrate data in a comprehensive manner. This ambition is reminiscent of the Renaissance era, where the comprehension of scientific concepts in a global manner was the guiding beacon that sought rules and laws universally applicable across a wide range of disciplines. A systems-oriented scientist is, of course, not tasking the scientist or clinician to possess all knowledge from all disciplines. Instead, he/she is expected to understand the principles through which he/she can access and use multidisciplinary information from several databases. The task is to educate the future generation to be able to collaborate with peers from different fields in their everyday routine work. A reductionist- to- systemic transition in the way a scientist understands medicine or a clinician understands science requires the introduction of appropriate training at several levels of education. In terms of medical education, we need to create a coherent link between the preclinical disciplines (chemistry, biochemistry, cellular and molecular biology, statistics, and anatomy), which should be complemented by training on networks, statistics, data handling, and modeling. This type of training is very important for the medical doctors and clinical practitioners who are in the front line of patient management and are directly involved in the diagnosis and treatment of diseases. These practitioners would greatly benefit by familiarization with genomics, data integration, bioinformatics, and "-omics" technologies, in order to be able to incorporate them into their work. Therefore, continuing medical education is anticipated to play a key role toward this goal. Educational information and training programs for all career stages, aided by web-based and e-learning modules, will be the cornerstones of continuing education. Such training programs can be designed to be flexible and custom made for each practitioner, for example, modular, "study-at-own-pace," and cost-effective. They should also be applicable to a variety of background of healthcare professionals, from clinicians to paramedics and nurses [8]. The creation of such courses is challenging not only due to the students' different backgrounds but also because systems biology relies in a wide variety of disciplines, which no student can fully master. Therefore, there is no default solution for teaching topics and techniques. Instead, it seems to be much more important to inspire students to have a motivation on pursuing multidisciplinary information on biological systems and search for models that can be used to explore them. For the future scientist, the ability to estimate the complexity of a system, the understanding of the medical problems, and the generation and handling of relevant information will become more important than overspecializing in techniques of systems analysis [19].

13.3 Networking Among All Key Stakeholders

The successful implementation of systems medicine will be impossible without efficient networking and communication among all the involved stakeholders, including patient organizations, academia, industry, medical practitioners, and policy makers. Each stakeholder plays a vital role in the translation of research into novel predictive tools, prevention measures, and personalized targeted therapies. Continuous discussion and interaction among all stakeholders through frequent conferences, workshops, and training events is required, and initial actions toward this goal have already been taken. Toward this goal, an extensive stakeholder advisory board (http://www.healthydietforhealthylife.eu/index.php/organisation/stakeholder-advisory-board) has already been established. It includes a panel of experts from several disciplines as well as representatives of the initiative for implementation of systems medicine across Europe. Clinical and patient organization representatives have been recruited into all aspects of these networks, so that all stakeholders have a voice. The main purposes of such networks will be (i) addressing the needs of clinicians and patients and (ii) suggesting ways of exploitation of both existing and novel infrastructures to meet these needs. To date, several collaborative initiatives and synergies have been launched among stakeholders across Europe, for example, the European Strategy Forum on Research Infrastructures, the Infrastructure for Systems Biology Europe, and Biobanking and Biomolecular Resources Research Infrastructure. Other initiatives aim at raising awareness about financial opportunities to investors and politicians [8].

13.4 Coordinated European Efforts for Dissemination and Implementation

The establishment of a European society of systems medicine, with the participation of representatives of all stakeholder groups, will facilitate efficient implementation of

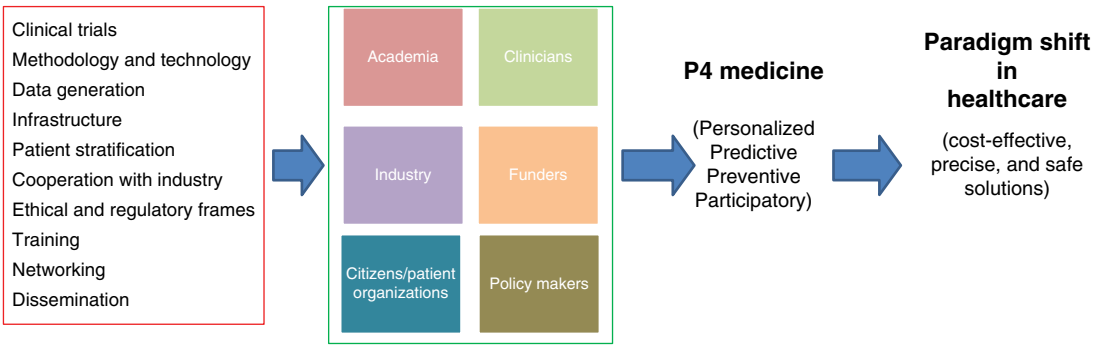


Figure 13.3 Systems medicine toward implementation of P4 medicine and a paradigm shift in healthcare.

systems medicine across Europe. This society has been recently established (http://www.eisbm.org/projects/easym/). It aims to implement actions toward integration of efforts across Europe under the umbrella of systems medicine. They comprise (i) engagement of public funders to pan-European initiatives, (ii) development of systems biology-oriented pan-European research programs, (iii) recruitment of private funders, (iv) establishment of extended networks among clinics and research centers, and (v) inclusion of the P4 medicine in the European Union's agenda, which will be translated to coordinated support at several levels (scientific, regulatory, legal, politic, clinical). Stable, sustainable cooperations between major research centers of systems biology, computational research centers, and academic clinics across Europe will also enhance this effort. It aims to establish an integrative and open community of researchers and clinicians, as well as creation of a communication platform for researchers implicated in system medicine [8].

Similar initiatives in the context of precision medicine and improvement of personalized medicine have also been announced by the White House in the United States, with special emphasis on the treatment of cancer and diabetes. The initiative includes the involvement of the National Institutes of Health toward the implementation of P4 medicine [20, 21]. Figure 13.3 describes how the emphasis on key areas of systems medicine, orchestrated by participation of the network of stakeholders, will lead to the implementation of P4 medicine and to a paradigm shift in future healthcare.

13.5 The Contributions of Academia in Systems Medicine

Academia plays a decisive role in the changing landscape of personalized therapy toward systems medicine. It is anticipated to make major contributions to producing and handling multiple levels of high-throughput data. It will also have a major role in solving the technical difficulties that emerge when generating quantitative datasets for large numbers of system variables across different levels of organization. Another major challenge is the development of efficient tools that can handle the wide variety of available data and mine the disease-relevant data subsets. Overcoming these obstacles will subsequently lead to the development of models that can be used in clinical practice for disease prevention and prediction, as well as optimization of personalized therapy. Thus, the main contributions of academia in systems medicine will be:

a) The generation and management of "big data" from each patient: "-omics" and "multi-omics" technologies that will be the cornerstones of this effort. Toward data management the EU has taken decisive strides by funding relevant infrastructures for the handling of big data, such as ELIXIR (www.elixir-europe.org).
b) The integration of "big data," which is largely based on bioinformatics approaches but has to proceed toward computational network modeling.
c) The development of robust predictive disease models based on integration of "big data." Computational modeling relies largely on mathematics and computational biology [22]. However, novel animal models to simulate the disease could also emerge in the molecular biology research setting to complement data accordingly [23].

Figure 13.4 summarizes these contributions of academia in systems medicine.

13.6 Data Generation: Omics Technologies

The advancement of high-throughput analyses has provided the opportunity to collect and process information about a sample in a high-throughput manner at specific

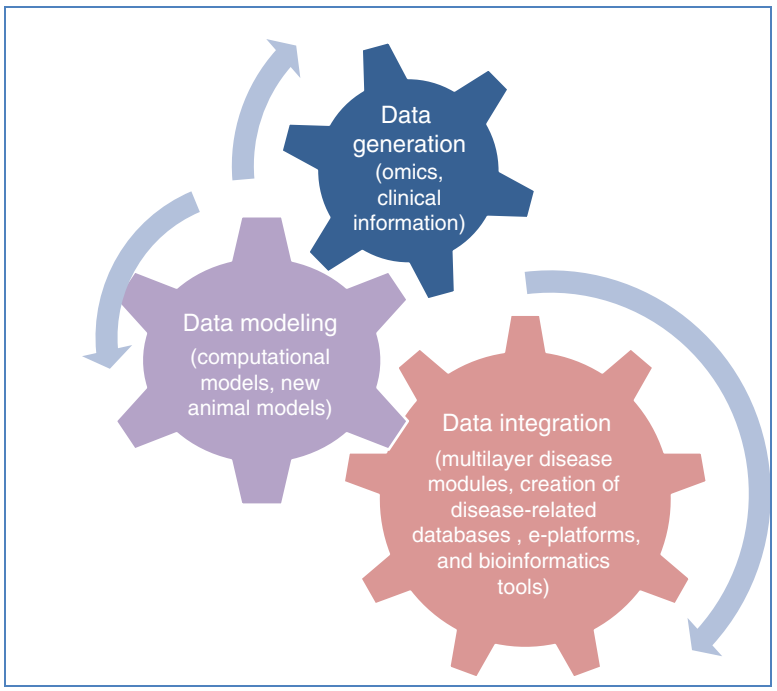


Figure 13.4 The main contributions of academia in systems medicine involve generation, integration, and modeling of disease-related data in an interassociated and interlocking manner.

time points. "Omics" technologies can provide systems-level information of all genes or gene products for any sample, for example, for genome (genomics) and epigenome (epigenomics), coding and noncoding RNA transcripts (transcriptomics), protein products (proteomics), lipids (lipidomics), and metabolites (metabolomics). Notably, the term "omics" is expanding to include other systems, such as all microbes (microbiome), environmental exposures (exposome), or even all diseases (diseasome). High-throughput analysis of patient samples, followed by appropriate data integration, can be exploited for comprehensive understanding of diseases in terms of systems medicine [24].

The most widely used "omics" technologies are genomics and transcriptomics. Other popular, though challenging, omics technologies are proteomics and metabolomics. These technologies are discussed in detail in other chapters.

13.7 Data Integration: Identifying Disease Modules and Multilayer Disease Modules

"Omics" approaches can identify genes/gene products/metabolites that are more commonly deregulated in a certain disease. However, due to organismal complexity, this information alone does not automatically highlight the most appropriate molecules to be used as diagnostic markers and therapeutic targets. These only can be identified by systems-level approaches. Integrated multiomics approaches are preferred over omics analysis, which relies on only a single data type. Combining multiple data types and several levels of high-throughput information compensates for missing or unreliable information attributed to a single data type. In addition, a pathway or gene that is confirmed by multiple sources of high-throughput evidence is less prone to false positive predictions. Moreover, by integrating all different levels of genetic, genomic, and proteomic information, a more comprehensive disease-specific description that simulates more accurately the disease complexity can be formed. Data can be integrated by two main approaches: the multistage analysis, which integrates information following a stepwise or hierarchical analysis approach, and meta-dimensional analysis, which integrates multiple different data types to develop a multivariate model that is associated with one outcome or phenotype [25].

In the context of a disease, there is a significant reason for integrating -omics data: it has been observed that disease-associated genes tend to form networks of functionally related genes, which are termed "disease modules." These disease modules can facilitate identifying the organization and prioritization of the disease-associated genes collected by high-throughput analyses. They can also provide insights on the mechanisms and the pathways underlying the disease. In addition, disease models can uncover novel disease genes, biomarkers, or therapeutic targets, which were not initially identified by single-omics analyses. The general principles of networks apply to disease modules. For instance, alteration of hub genes is likely to have large effects, whereas alterations of

the many genes with few links will probably have small effects. Therapeutic targeting of a hub gene is more likely to be effective than targeting a gene with few interactions. However, this increases the risk of off-target effects. Identification of disease modules can improve the selection of the most druggable components of the disease module, which would be efficient against the disease, while at the same time avoiding side effects due to unwanted interactions with molecules outside the disease module. This approach is anticipated to optimize the efficacy and safety of therapeutic targeting [26].

More importantly, the disease modules are multilayered. This means that the relative network module components do not belong to only one category of molecules, for example, proteins, and therefore cannot be unveiled by using only a single data type. The disease modules consist of components of several molecule categories, such as transcription factors, noncoding RNAs, gene products, and other modifiers of gene function (e.g., epigenetic factors), which dynamically interact with each other. The integration of high-throughput data from multiple levels, both molecular and clinical, can lead to generation of multilayered disease modules, which will be more informative on the disease pathogenesis. For example, defining an appropriate multilayered disease module that is based on the integration of clinical and molecular information could lead to the establishment of an optimal combination of clinical examination/molecular diagnostics for the earlier prediction of disease outcome for individual patients.

Another consideration when defining multilayered disease modules is that they dynamically change over time, reflecting disease progression. The network components may be re-wired over time, in relation with disease outcome. Integrating data produced at several time points can facilitate the stepwise monitoring of disease course [24, 26]. Overall, an important challenge of systems medicine is to develop robust methods able to integrate all the clinical and molecular -omics information.

This step of the process necessitates the development of user-friendly and accessible computational tools for the collection, storage, and handling of information. Additionally, the establishment of platforms and databases for knowledge management will facilitate efforts for the integration of "big data" [23].

13.8 Modeling: Computational and Animal Disease Models for Understanding the Systemic Context of a Disease

The integration of clinical and molecular data in regard to a disease across all relevant levels of organization will generate disease models that can be exploited as tools for further understanding a disease and generating testable predictions to improve therapy and prevention. Modeling includes molecular modeling (e.g., drug/vaccine target predictions, protein–protein interactions), modeling of subcellular processes (e.g., linking signaling pathways to phenotypes), cell-based modeling (e.g., cell–cell interactions, interactions of cells with their microenvironment), tissue/organ modeling (e.g., biomechanical models, formation, and maintenance of tissue architecture), and body–systems-level modeling (e.g., pharmacokinetic/pharmacodynamic predictions, overall survival predictions). Samples and experimental models cannot produce all data for all parameters that constitute a disease. Thus, computational modeling will be recruited to fill in the gaps derived by incomplete experimental data through extrapolation and simulation approaches. For generating multiscale models to reflect the complexity across all levels that are involved to a disease, from molecular/cellular to organismal/environmental, several actions have to be taken. These include the development of computational tools and algorithms for robust multiscale simulations; the development of mathematical approaches to analyze and multiscale models in terms of parameter evaluation, sensitivity analysis, identifiability analysis, and image analysis; and the establishment of workflows for modeling, including computational tools that facilitate data management, model construction, and analysis and approaches to investigate the interplay between the environment and cell response and integrate relative data to unified and comprehensive disease models [27].

The computational models may unveil novel associations of genes with diseases. This could further guide the development of novel animal models in the preclinical setting, complementary to the existing ones for more thoroughly understanding of disease pathogenesis. For instance, if high-throughput data analysis indicates that specific under noticed or overlooked genes are strongly implicated in a disease module, then corresponding knockout/knockin animal models could be generated for the validation and further investigation of this aspect [23].

13.9 Examples and Success Stories of Systems Medicine-Based Approaches

Systems medicine is at the beginning of an exciting and challenging road. Systems medicine-based examples are emerging in the context of several chronic diseases. These efforts are in several stages of preclinical or clinical testing and aim to improve prognostic methods or targeted therapeutic interventions against serious chronic diseases. For instance, genome and whole-exome

sequencing methods have been used in the clinic for the earlier treatment of neurodevelopmental disorders. Transcriptomics has been successfully recruited for the prognosis, classification, and stratification of breast cancer. In a similar manner, high-throughput gene expression profiling has been recruited for early prediction and therapeutic targeting in hepatocellular cancer, colorectal cancer, and glioma. Metabolomics are exploited in clinical studies for prediction and targeting of Alzheimer's disease [24]. Moreover, systems-based approaches have been adopted in the setting of chronic obstructive pulmonary disease (COPD). COPD patients display different phenotypes as a result of a complex interaction between various genetic, environmental, and lifestyle factors. This phenotypic complexity is being analyzed in large datasets, in correlation with functional genomics assays, using computational biology approaches. The management of COPD is now steering toward an integrative and systemic approach, focusing on proteomics and metabolomics. This strategy aims to identify disease subclusters in order to improve the development of more effective therapies [28]. Another complex chronic condition currently being addressed using comprehensive systems medicine-based strategies is intestinal bowel disease (http://www.sysmedibd.eu/).

One of the first European projects that adopted a comprehensive systems medicine approach is the Mechanisms of the Development of ALLergy (MeDALL) (EU FP7-CP-IP; Project No: 261357; 2010-2015) consortium for the study of allergies. The results of this project have been published recently and provide the first proof of concept for the feasibility of systems medicine for the successful management of complex diseases. MeDALL (http://cordis.europa.eu/result/rcn/175936_en.html) was based on a systems medicine approach carried out by a networked panel of experts in 54 European sites, which linked epidemiological, clinical, and basic research data using a stepwise, large-scale, and integrative approach. A knowledge management platform was developed, where all partners deposited patient data and information, as well as experimental and computational tools. The partners had open sharing and access rights to the platform. Overall, the database integrates historical and newly collected data from around 44 000 participants reporting 398 clinical and phenotypic attributes and 160 different follow-ups at 25 different time points between pregnancy and age 20, as well as information about available blood samples. Samples have been stored in the individual biobanks of the different partners of the consortium. The database also includes information on allergy-associated genes based on literature reviews and automated text mining. These data were integrated with molecular data on protein–protein interactions, transcriptional regulation, miRNA regulation, and signaling pathways from publicly available databases. Omics data produced or made available in this database includes 23 000 historical genome-wide association studies, 9 500 epigenetics, 2 000 proteomics, 750 transcriptomics, and IgE microarrays on 4 000 subjects, as well as individual estimates of ambient air pollution exposure (10 000 children) using computer modeling methods. Ethics considerations were also taken into account by establishing a dedicated website to address practical information on regulatory issues for exchanging biological samples and relevant data among the partners of the consortium.

Moreover, a highly sensitive and reliable allergen chip tool for detection of 170 allergen molecules was developed and applied in the clinic for early detection of allergic immune response, within the context of this consortium. For addressing bioinformatics aspects, machine learning methods were applied using epidemiological and clinical data from large patient datasets, and a bioinformatic model of multimorbidity of allergic diseases was developed. The population-based studies in patient cohorts were complemented with experimental animal studies and development of novel mouse models for the study of allergy.

Exploitation of the large datasets of this consortium led to novel findings. In detail, allergic multimorbidities and IgE polysensitization were found to be correlated with the persistence or severity of allergic diseases. These parameters were confirmed as novel means for differential diagnosis. The integration of multimorbidities and polysensitization parameters in patients has resulted in reclassification of allergic diseases. This helped to improve the understanding of genetic and epigenetic mechanisms of allergy, as well as to better manage allergic diseases. Ethics and gender were considered. The results of this study were translated to clinical activities. In detail, the MeDALL data were successfully exploited to improve the stratification of allergic preschool children using multimorbidity and IgE polysensitization as predictive markers of disease persistence. In addition, it was realized that allergy burden can be reduced with relatively simple means on national level. Finland and Norway have considered the results of this consortium for reforming their Allergy Health Programme. These reformations are anticipated to serve as exemplifiers for other countries in the future [23].

13.10 Limitations, Considerations, and Future Challenges

Biological systems are complex. Systems medicine is a coordinated attempt to address this complexity with systems-driven, integrative, cross-disciplinary, and

milestone-driven platforms and methods. Individual investigators and their laboratories will play an important role in deciphering the complex details of the broad overview that are obtained by systems biology and systems medicine. The ultimate objectives are to improve healthcare, reduce the cost of healthcare, and stimulate innovation. The development of systems medicine presents the potential to lead to a paradigm shift in healthcare. However, this effort also faces limitations and challenges, in several levels of its implementation, including academic, clinical, socioeconomical, and ethical.

Challenges in academia involve significant methodological problems, such as the accurate detection of signals when numerous variables are measured, as well as the discrimination of the informative data from background noise. Other problems related with the clinical setting include the combinations of omics data from patient samples, which are needed for diagnostic and therapeutic purposes, especially given the fact that this technology is expensive. Analyses of the cost-effectiveness of omics platforms for individualized medicine should be initiated. Novel software for diagnostic classification is needed, based on the integration of clinical information and molecular profiling for a large number of patients. In terms of networking among stakeholders, multidisciplinary collaborations that include clinicians, representatives from patient organizations, experts in genomics and bioinformatics, participants from pharmaceutical and biotechnological industries, and healthcare and academic leaders are required. The interaction among representatives of different backgrounds is a challenging task and requires open communication channels in order to be effective [24].

Ethics considerations include the final decision making on therapy and prevention measures. Personalized therapeutic approaches need to be patient oriented. This means that in everyday clinical practice the patient's objectives, preferences, and values as well as the available economic resources have to be taken into account [29]. However, in a process which is based on interactions among several stakeholders, boundaries are not clear. Therefore, the extent of individual responsibility in therapeutic decision making should be clearly defined. Ethics concerns for the implementation of systems medicine include the risks of "eclipsing environmental factors" in the context of prediction; exerting reproductive control and sliding to "eugenic practice" in the context of prevention; reducing a patient's identity to his or her genes and establishing a rigid genetic social hierarchy, in terms of personalization; and poor decision making as well as a misconception that "patients have a moral responsibility to become well" [30]. There is also the concern about the confidentiality of information. A prerequisite of systems medicine is collecting, electronically recording, and exploiting a large amount of data on any individual of the whole population. Increased accessibility and transferability of these data is also needed among stakeholders. Therefore, policies that apply to personalized therapy should be accordingly modified. Patient data information should be protected so as to both ensure the trust of patients and support interoperable health information exchange [31].

Last but not least, the socioeconomic benefits of systems medicine need to be quantified in order to evaluate the effectiveness of this novel concept. Indicators should be defined to measure the impact on quality of life, reduction of disease burden, and health economics. A positive impact could justify additional investments [18].

An old Indian story talks about three blind men who came across an elephant. Each of the blind men touched a different part of the elephant and gave a description of what he believed an elephant was. The person who touched the elephant's trunk claimed the elephant to be a snake. The person who touched the elephant's leg declared the elephant to be a tree trunk. The person who touched the elephant's ear identified the elephant to be a sail. On the one hand, based on the blind men's confined level of interaction with the elephant, their observations made sense. On the other hand, if they had collaborated and holistically studied the elephant, its true structure might have started to become apparent. However, we should keep in mind that in this story all the men had been blind and just discussing their observations pose a risk of creating more confusion than clarity. Insightful interaction among them would be necessary for the meaningful integration of their conclusions to comprehend the global picture of the elephant. Understanding complex systems such as the human body can benefit from collaboration, but appropriate interactions and interdisciplinary networking are an indispensable parameter for the successful outcome of this effort [6].

References

1 Schork, N. J. 2015. Personalized medicine: time for one-person trials. *Nature*, 520, 609–611.

2 Auffray, C., Caulfield, T., Griffin, J. L., Khoury, M. J., Lupski, J. R. & Schwab, M. 2016. From genomic medicine to precision medicine: highlights of 2015. *Genome Med*, 8, 12.

3 Snyder, M., Weissman, S. & Gerstein, M. 2009. Personal phenotypes to go with personal genomes. *Mol Syst Biol*, 5, 273.

4 Boissel, J. P., Auffray, C., Noble, D., Hood, L. & Boissel, F. H. 2015. Bridging systems medicine and patient needs. *CPT Pharmacometrics Syst Pharmacol*, 4, e00026.

5 Mardis, E. R. 2006. Anticipating the 1,000 dollar genome. *Genome Biol*, 7, 112.
6 Hood, L. 2013. Systems biology and p4 medicine: past, present, and future. *Rambam Maimonides Med J*, 4, e0012.
7 WHO 2011. Global Health and Aging. NIH Publication no. 11-7737. WHO, Geneva.
8 Casym Consortium 2014. The CASyM roadmap: implementation of systems medicine across Europe. CASyM administrative Office on behalf of the CASyM consortium Project Management Jülich, Forschungszentrum Jülich GmbH, Germany.
9 Chuang, H. Y., Hofree, M. & Ideker, T. 2010. A decade of systems biology. *Annu Rev Cell Dev Biol*, 26, 721–744.
10 Flores, M., Glusman, G., Brogaard, K., Price, N. D. & Hood, L. 2013. P4 medicine: how systems medicine will transform the healthcare sector and society. *Pers Med*, 10, 565–576.
11 Federoff, H. J. & Gostin, L. O. 2009. Evolving from reductionism to holism: is there a future for systems medicine? *JAMA*, 302, 994–996.
12 Vandamme, D., Fitzmaurice, W., Kholodenko, B. & Kolch, W. 2013. Systems medicine: helping us understand the complexity of disease. *QJM*, 106, 891–895.
13 Barabasi, A. L. & Albert, R. 1999. Emergence of scaling in random networks. *Science*, 286, 509–512.
14 Mitra, K., Carvunis, A. R., Ramesh, S. K. & Ideker, T. 2013. Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet*, 14, 719–732.
15 Vogt, H., Hofmann, B. & Getz, L. 2016. The new holism: P4 systems medicine and the medicalization of health and life itself. *Med Health Care Philos*, 19(2), 307–323.
16 Chan, I. S. & Ginsburg, G. S. 2011. Personalized medicine: progress and promise. *Annu Rev Genomics Hum Genet*, 12, 217–244.
17 World Medical Association 2013. Declaration of Helsinki ethical principles for medical research involving human subjects. *JAMA*, 310, 2191–2194.
18 Kirschner, M., Bauch, A., Agusti, A., Hilke, S., Merk, S., Pison, C., Roldan, J., Seidenath, B., Wilken, M., Wouters, E. F., Mewes, H. W., Heumann, K. & Maier, D. 2015. Implementing systems medicine within healthcare. *Genome Med*, 7, 102.
19 Voit, E. O., Newstetter, W. C. & Kemp, M. L. 2012. A feel for systems. *Mol Syst Biol*, 8, 609.
20 Collins, F. S. & Varmus, H. 2015. A new initiative on precision medicine. *N Engl J Med*, 372, 793–795.
21 Fox, J. L. 2015. Obama catapults patient-empowered Precision Medicine. *Nat Biotechnol*, 33, 325.
22 Chen, R. & Snyder, M. 2012. Systems biology: personalized medicine for the future? *Curr Opin Pharmacol*, 12, 623–628.
23 Bousquet, J., Anto, J. M., Akdis, M., Auffray, C., Keil, T., Momas, I., Postma, D., Valenta, R., Wickman, M., Cambon-Thomsen, A., Haahtela, T., Lambrecht, B. N., Lodrup-Carlsen, K., Koppelman, G. H., Sunyer, J., Zuberbier, T., Annesi-Maesano, I., Arno, A., Bindslev-Jensen, C., De Carlo, G., Forastiere, F., Heinrich, J., Kowalski, M. L., Maier, D., Melen, E., Palkonen, S., Smit, H. A., Standl, M., Wright, J., Arsanoj, A., Benet, M., Balardini, N., Garcia-Aymerich, J., Gehring, U., Guerra, S., Hohman, C., Kull, I., Lupinek, C., Pinart, M., Skrindo, I., Westman, M., Smagghe, D., AKDIS, C., Albang, R., Anastasova, V., Anderson, N., Bachert, C., Ballereau, S., Ballester, F., Basagana, X., Bedbrook, A., Bergstrom, A., Von Berg, A., Brunekreef, B., Burte, E., Carlsen, K. H., Chatzi, L., Coquet, J. M., Curin, M., Demoly, P., Eller, E., Fantini, M. P., Gerhard, B., Hammad, H., Von Hertzen, L., Hovland, V., Jacquemin, B., Just, J., Keller, T., Kerkhof, M., Kiss, R., Kogevinas, M., Koletzko, S., Lau, S., Lehmann, I., Lemonnier, N., Mceachan, R., Makela, M., Mestres, J., Minina, E., Mowinckel, P., Nadif, R., Nawijn, M., Oddie, S., Pellet, J., Pin, I., Porta, D., Ranciere, F., Rial-Sebbag, A., Saes, Y., Schuijs, M. J., Siroux, V., Tischer, C. G., Torrent, M., Varraso, R., De Vocht, J., Wenger, K., Wieser, S. & Xu, C. 2016. Paving the way of systems biology and precision medicine in allergic diseases: the MeDALL success story. *Allergy*, 71(11), 1513–1525.
24 Benson, M. 2016. Clinical implications of omics and systems medicine: focus on predictive and individualized treatment. *J Intern Med*, 279, 229–240.
25 Ritchie, M. D., De Andrade, M. & Kuivaniemi, H. 2015. The foundation of precision medicine: integration of electronic health records with genomics through basic, clinical, and translational research. *Front Genet*, 6, 104.
26 Gustafsson, M., Nestor, C. E., Zhang, H., Barabasi, A. L., Baranzini, S., Brunak, S., Chung, K. F., Federoff, H. J., Gavin, A. C., Meehan, R. R., Picotti, P., Pujana, M. A., Rajewsky, N., Smith, K. G., Sterk, P. J., Villoslada, P. & Benson, M. 2014. Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome Med*, 6, 82.
27 Wolkenhauer, O., Auffray, C., Brass, O., Clairambault, J., Deutsch, A., Drasdo, D., Gervasio, F., Preziosi, L., Maini, P., Marciniak-Czochra, A., Kossow, C., Kuepfer, L., Rateitschak, K., Ramis-Conde, I., Ribba, B., Schuppert, A., Smallwood, R., Stamatakos, G., Winter, F. & Byrne, H. 2014. Enabling multiscale modeling in systems medicine. *Genome Med*, 6, 21.
28 Lococo, F., Cesario, A., Del Bufalo, A., Ciarrocchi, A., Prinzi, G., Mina, M., Bonassi, S. & Russo, P. 2015. Novel therapeutic strategy in the management of COPD: a systems medicine approach. *Curr Med Chem*, 22, 3655–3675.
29 Sacristan, J. A. 2013. Patient-centered medicine and patient-oriented research: improving health

outcomes for individual patients. *BMC Med Inform Decis Mak*, 13, 6.

30 Juengst, E. T., Settersten, R. A., JR., Fishman, J. R. & Mcgowan, M. L. 2012. After the revolution? Ethical and social challenges in "personalized genomic medicine". *Pers Med*, 9, 429–439.

31 Mcguire, A. L., Fisher, R., Cusenza, P., Hudson, K., Rothstein, M. A., Mcgraw, D., Matteson, S., Glaser, J. & Henley, D. E. 2008. Confidentiality, privacy, and security of genetic and genomic test information in electronic health records: points to consider. *Genet Med*, 10, 495–499.

14

Knowledge Discovery and Data Mining

Magdalena Krochmal[1] *and Holger Husi*[2,3]

[1] *Proteomics Laboratory, Biomedical Research Foundation Academy of Athens, Athens, Greece*
[2] *Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK*
[3] *Department of Diabetes and Cardiovascular Science, Centre for Health Science, University of the Highlands and Islands, Inverness, UK*

14.1 Introduction

Data mining (DM) is an interdisciplinary area focusing upon methodologies for extracting knowledge from data using various tools such as database systems, statistics, machine learning (ML), and data visualization [1]. Analysis of big datasets is centered on finding repeating patterns and systematic relations—a task that can be achieved only by using sophisticated algorithms and computing power. The goal of DM is detection, interpretation, and ultimately prediction of qualitative or quantitative patterns in data, and for this reason DM techniques are commonly used in business, healthcare, economics, and scientific research [2].

The ongoing growth of scientific data, caused by rapid technological improvements and common usage of high-throughput technologies, created the need for effective data handling and analysis [3]. Because significant progress has been made in generating, collecting, storing, and managing information, DM became an important tool in research. Employment of database systems into research workflow resulted in easier information retrieval and management and hence empowered the development of more advanced data analysis techniques. This is particularly important in life sciences, given the high complexity and multidimensionality of biological data and growing trend toward integrative analysis and modeling [4]. DM can facilitate discovery of biomedical knowledge to support clinical and administrative decisions, as well as generate novel scientific hypotheses from large experimental data, clinical databases, and biomedical literature [2].

DM is often referred to as knowledge discovery in databases (KDD) and is in fact the core of a multistep process leading to a comprehensive data analysis and knowledge extraction. In this chapter, we will describe the knowledge discovery processes with focus on DM methodologies and their growing applications in scientific research. The terminology commonly employed in this field is summarized in Table 14.1.

14.2 Knowledge Discovery Process

The interdisciplinary field of DM has grown in popularity over the last few decades, as big data analysis has proven to be beneficial in many areas, from business to science. KDD, a multilevel process designed for automatic exploratory analysis and modeling of large-scale data repositories, is often referred to as DM. In fact, KDD consists of several steps aimed at identifying novel, intrinsic, and characteristic patterns from large and complex datasets. DM is the core of the knowledge discovery workflow, preceded by data preparation and preprocessing and followed by extensive patterns evaluation [5]. DM frameworks are divided into two groups: theory-oriented (databases, statistics, ML, etc.) and process-oriented (Fayyad's, CRISP, etc.). In theory-oriented framework, we can distinguish different approaches to discovery. Database approach is based on the query concept that assumes that established theory is an essential tool in the discovery process. Both statistical and ML approaches rely solely on the data, by either fitting model to data or allowing data to suggest the model, respectively. On the other hand, process-oriented framework addresses the issues of viewing DM as an interactive and iterative process.

Along the DM evolutionary path, standards and good practices have been established and collected in the form of a specific set of methodologies, named *cross-industry standard processes for data mining* (CRISP-DM) [6]. They are defined by six crucial steps of knowledge discovery processes and guide analysts through the entire procedure.

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

Table 14.1 Data mining terminologies.

• *Supervised learning*—A DM method used to build prediction models based on prior knowledge, for example, classification
• *Unsupervised learning*—A condition whereby a prediction model is built based on a discovered structure, where the grouping is not known, for example, clustering algorithms
• *Training set*—A portion of data used to learn/create a model
• *Test set*—A subset of data used for model validation and assessment of classifiers' performance
• *Cross-validation set*—A subset of data used for classifier enhancement and error estimation
• *Attribute*—Field, variable, feature, and table column
• *Target value/attribute*—A value that has to be predicted
• *Categorical attribute*—An attribute where the values correspond to discrete categories
• *Discretization*—Grouping of related values together under a single value, quantitative data into qualitative data
• *Generalization*—Ability to produce reasonable outputs for inputs not encountered during the training
• *Overfitting*—A problem in modeling, when a function is too closely fitted to a training set, thus does not generalize well when input values come from external datasets

Common terms frequently used in the process of data mining are listed.

CRISP-DM steps include the following:

1) *Business understanding*—Understanding and defining of business goals and the actual goals of DM. With regard to research, this step defines the scientific questions to be answered through the application of DM.
2) *Data understanding*—Familiarization with the data and the application domain by exploring and defining the relevant prior knowledge. Importantly, due to the high complexity, DM of biological data requires deep understanding of the research domain, biological processes, and relations reflected in the data.
3) *Data preparation through data cleaning and preprocessing*—Creating the relevant data subset through data selection, as well as finding of useful properties/attributes, generating new attributes, defining appropriate attribute values, and/or value discretization.
4) *DM*—The most important step of this process, which is concerned with choosing the most appropriate DM tools—from the available tools for summarization, classification, regression, association, and clustering—and searching for patterns or models of interest.
5) *Evaluation and interpretation of results*—Aided by pattern/model visualization, transformation, and removal of redundant patterns.
6) *Deployment*—The use of the discovered knowledge.

All steps of KDD are illustrated in Figure 14.1 and will be described in detail in this chapter. The KDD process is iterative and interactive; thus multiple refinements can be performed at each step of the cycle. The ultimate goals are either verification of a hypothesis, which relies solely on users input, or discovery (further subdivided into prediction and description), where new insights can be generated by the system.

Although the applications of DM techniques in research and healthcare are relatively recent, they hold great potential for prediction, diagnosis, and treatment by discovery and elucidation of patterns and processes that cannot be directly concluded from experimental data [2]. There are several research domains where learning algorithms are being used: genomics, proteomics, microarrays, systems biology, evolution, and text mining (TM) [7]. Possible research applications are presented in Figure 14.2. Moreover, healthcare domain benefits from the application of DM through support of medical decision making (such as diagnosis process or treatment choice) and improvement in administrative management (demographic trends, insurance, and quality assurance).

14.2.1 Defining the Concept and Goals

The initial step in the KDD process is defining the purpose and the end objective of the analysis. This preparatory step is needed to characterize the data, algorithms, and

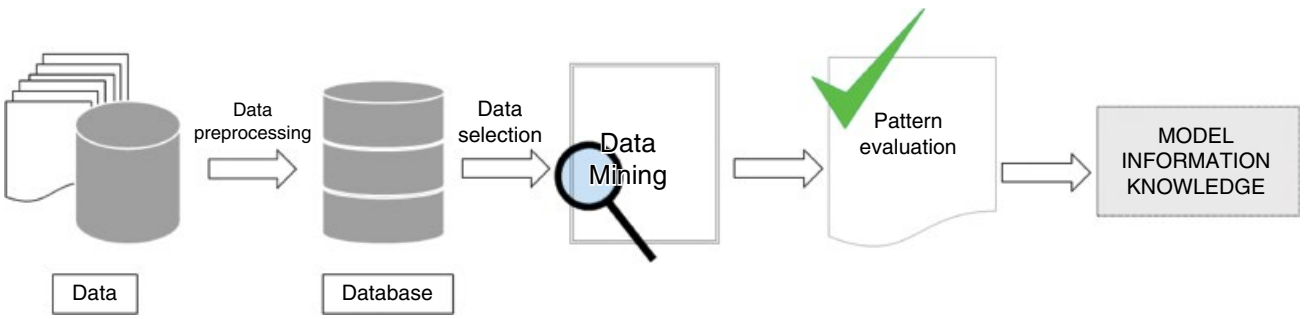


Figure 14.1 The end-to-end knowledge discovery (KDD) process. This figure shows the flow of information from the acquisition stage through to the final output via numerous steps.

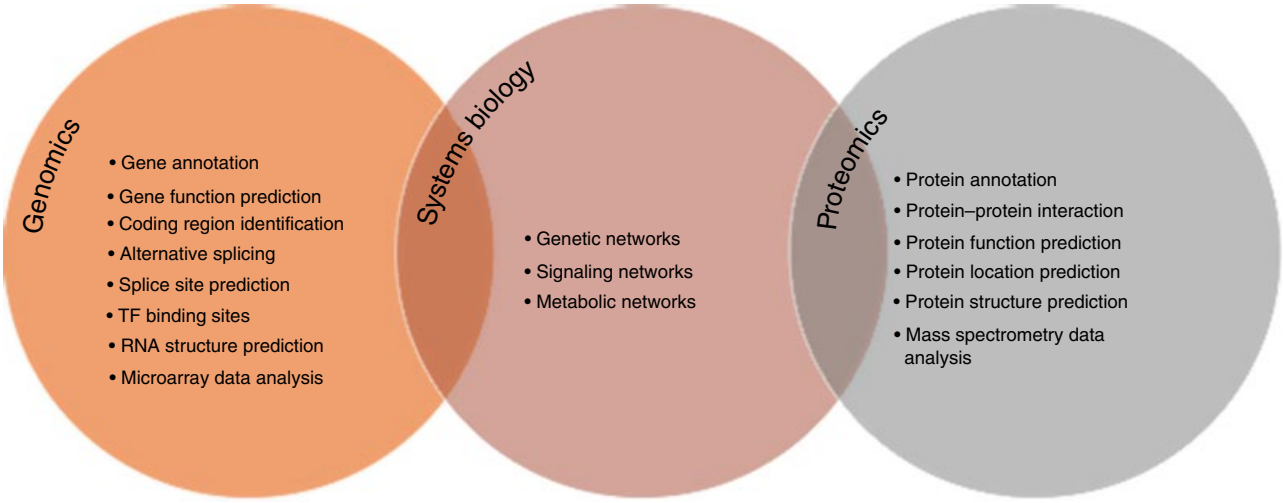


Figure 14.2 Possible research applications of KDD techniques.

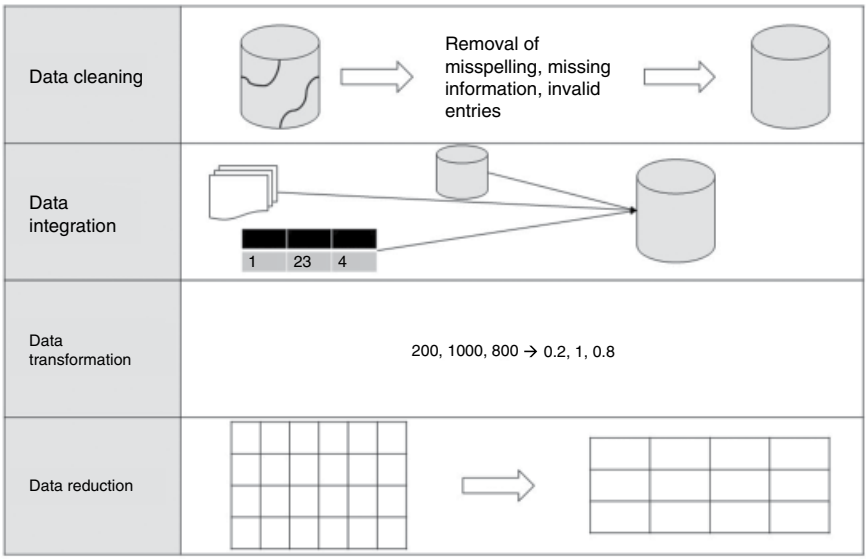


Figure 14.3 Forms of data preprocessing. Examples of data handling routines generally used in the transformation of acquired raw data prior to modeling are shown.

desired outputs and is vital in order to direct DM efforts. Without specific goals, DM may fail to meet desired expectations and satisfactory outcome.

14.2.2 Data Preparation/Preprocessing

Data preparation is a major step in the KDD process, as the quality of a constructed model is highly dependent on the data quality. It can often be the most demanding and laborious of all stages, due to the "raw" data heterogeneity, complexity, and existing uncertainties caused by noise, inconsistencies, and missing information. Therefore, preprocessing is required to adjust the data to the requirements of DM algorithm. Preprocessing can be divided into operations such as data collection, understanding, and preparation that include cleaning, integration, transformation, and reduction (Figure 14.3), which are essential to perform in order to achieve consistent, high confidence results.

Once the purpose of the analysis is established and data scope is characterized, the consecutive phase is data collection and understanding. Often, datasets are derived from different sources, and as a consequence, some discrepancies can exist regarding common identifiers, formats, relationships between attributes, and so on. Hence, a comprehensive knowledge of existing variations is needed in order to identify preparation steps and therefore obtain unified data. Data cleansing consists of actions aimed at detection of incomplete, inaccurate records in the dataset and dealing with those inconsistencies by corrections, deletions, or even predictions of missing values. Consecutively, datasets are consolidated,

and a transformation step is applied to convert data from the source format to match the destination data system. This can involve the following tasks:

- Smoothing (noise reduction/removal, binning, clustering, regression)
- Aggregation (summarization)
- Generalization (simplification, hierarchical organization)
- Normalization (scaling)
- Attribute construction (new attributes constructed from the existing ones, support for mining process)

Subsequently, the reduction phase is used to simplify/aggregate/compress data for easier model generation while still retaining critical information. It is especially useful when the analyzed dataset is large and various exploratory analysis solutions are being tested. One of the most common approaches used for the size reduction is sampling, that is, random selection of a smaller subset of the initial dataset. Ultimately, processed data are loaded into the database repositories. From there they can be used by mining algorithms in the modeling process.

14.2.3 Database Systems

Considering the huge amount of information, especially when referred to multidimensional biological data, employment of databases to support the mining process is necessary. Database systems are designed to enable efficient storage, maintenance, exploration, and exchange of deposited data; thus their application in the research field is now common [8]. Regardless of the type of applied enterprise data storage (e.g., SQL Server, Oracle, IBM DB2, Sybase, MySQL, PostgreSQL), all of them serve the same purpose, which is storing and enabling easy information retrieval. The main characteristic feature of databases is the storing of data in a strictly organized way. The logical structure of the database is determined by the data model, which affects the manner of how information will be stored, formatted, and manipulated. There are several architectural data models used in database creation, for example, relational, hierarchical, or network. Among them, the most popular is the relational model, which relies on the mathematical theory of sets and is used to build relational databases [9]. For the clarity of the chapter, database models and characteristics will not be further described.

In the recent years, biological databases have grown on popularity as their application significantly facilitates information retrieval and bioinformatics analysis [10]. They cover various research topics and systemize general knowledge about genome, proteome, and RNAs, as well as known biological pathways and molecule interactions. Moreover, numerous databases cover more specialized subjects such as -omics studies in disease research [11]. Often, these data repositories are publicly available online and serve the research community by providing up-to-date, manually, and/or computationally curated data. Given the wealth of the knowledge combined with the easy access to information that databases provide, both wet-lab scientists and *in silico* researchers can benefit from this resources [12].

14.2.4 Data Mining Tasks and Methods

DM is at the heart of knowledge discovery process, as in this phase, through application of suitable algorithms, model is generated from previously selected and preprocessed data. The DM tasks can be divided into predictive (e.g., classification or regression) and descriptive (clustering or association analysis) (Figure 14.4). The general idea is that initial input data is subjected to an analysis aiming at the discovery of generalized patterns to create useful knowledge. This scenario is often called *inductive learning*, where the model is derived from training data and should have a predictive ability when applied to new data. Major DM tasks, that is, classification, regression, and clustering, are graphically represented in Figure 14.5.

DM tasks can be performed using various methods, developed in the field of statistics or using ML algorithms. Determination of which approach to use in the process requires primary knowledge about the data being analyzed as well as the expected result. In general, statistics are especially useful in the analysis of continuous datasets. On the other hand, ML methods originated from categorical data analysis (e.g., binary outcome), which makes them especially useful for dealing with the descriptive data features [13].

A general workflow applied for model generation and learning starts with data separation into training, validation, and testing sets. The training set, containing most of the data (~80% according to good practice), is used to create the model. Consecutively, a small validation set (~20% of input data) is used to enhance the performance of the classifier and estimate the error. Lastly, the model is applied to make predictions against the test set. An evaluation of the model is very important for the prediction of how well it will perform in the future and is a vital step due to the common DM problem of overfitting. This phenomenon appears when the model is too strongly fitted to the training set and, as a result, performs with a poor generalization and prediction. In contrary, underfitting occurs when a model cannot capture the complexity of the data. Both issues can be avoided by applying cross-validation steps during model development. Finally, optimization methods are employed to select the best performing classifier.

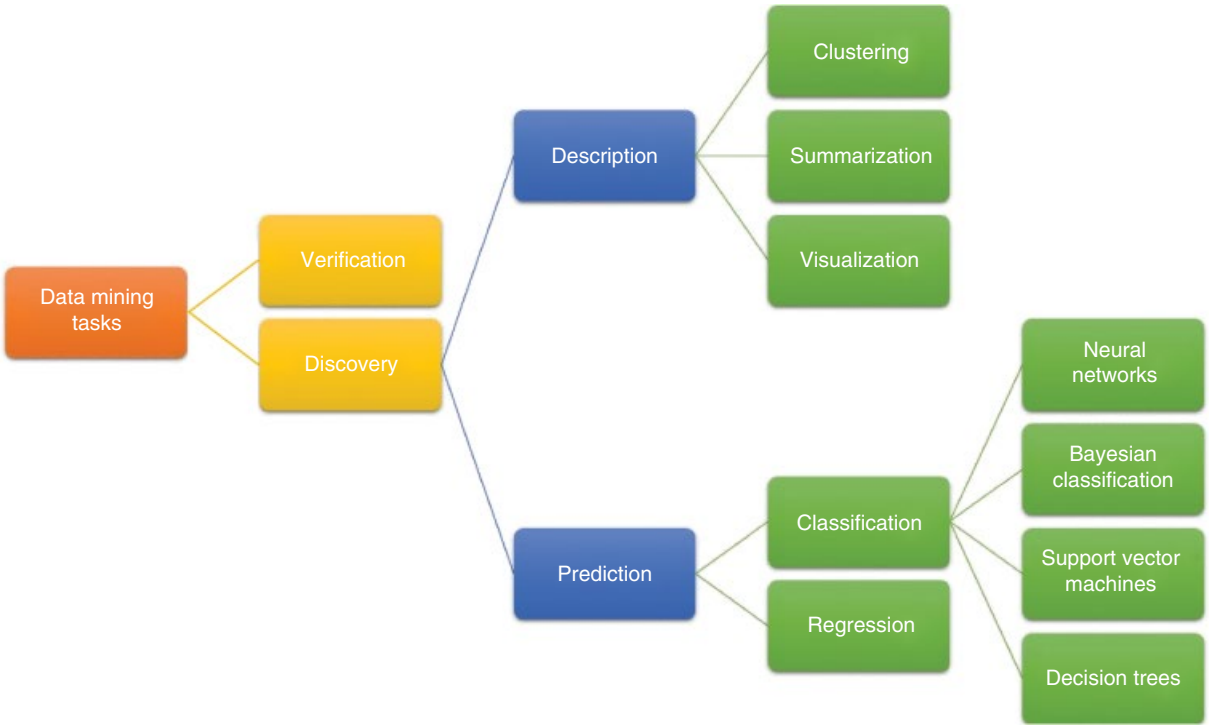


Figure 14.4 Data mining tasks. The decision tree of various data mining methods is shown. *Source:* Maimon and Rokach [5]. Reproduced with permission of Springer.

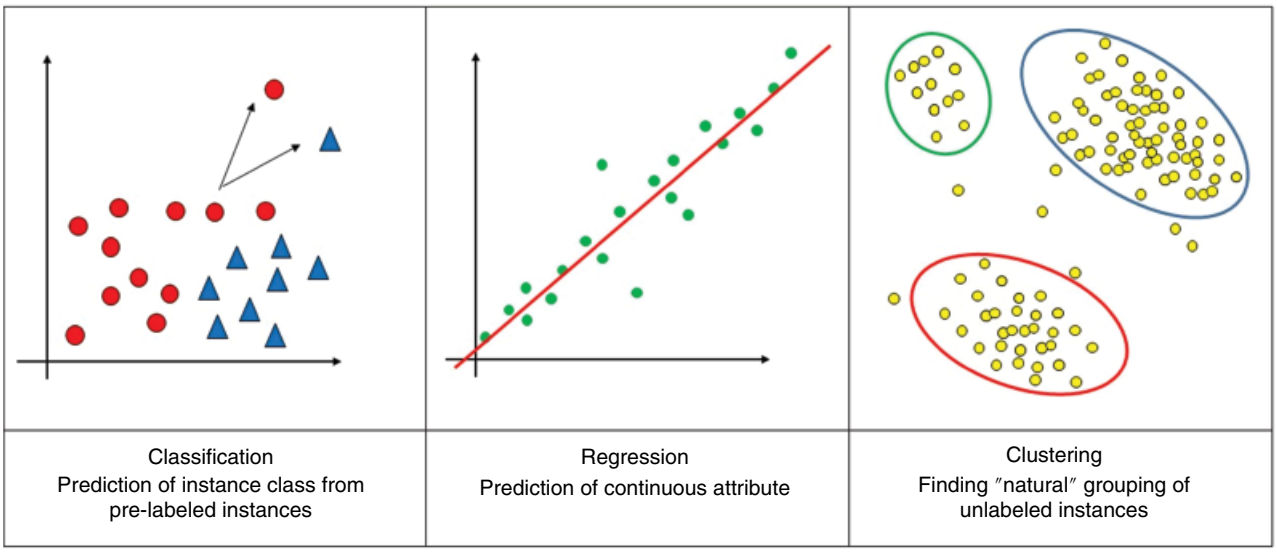


Figure 14.5 Data mining applications. Commonly employed tasks in data mining such as classification, regression analysis, or clustering are depicted.

Mining methods can be distinguished based on the learning form with different degrees of supervision, that is, supervised, semi/weakly supervised, and unsupervised [14]. In supervised learning (also called direct DM) data labels (classes) are known in advance, so the prediction model is supposed to classify new cases based on the "imposed" knowledge. Regression and classification fall into this category. Unsupervised systems are not fed with example classes, so the discovery process aims at creating classification pattern, which is data driven. The prediction is then based on discovered relationships and the structure of the underlying data. Additionally, semi

and weakly supervised systems use only a subset of training examples, enabling the algorithms to take unlabeled data into account and create own clusters. The most commonly used algorithms from both statistics and ML fields will be described in this section.

14.2.4.1 Statistics

Statistics is a subdiscipline of mathematics that is applied in DM to create probabilistic models. Statistical inference relies on quantification, collection, analysis, summarization, and interpretation of data. Imposed mathematical rigor made statistical methods a highly valued and trusted analysis tools. Among them, major approaches used in DM tasks are correlation, association analysis, regression, and clustering.

14.2.4.1.1 Classification

Classification is a cognitive process designed to organize and structure the knowledge about the world. Classification models are widely used in research fields, as they aim at creating classes that can distinguish different categories of data, for example, benign versus malignant cancer patients. A model is built upon discovered relationships between the data attributes and tested on so-called prediction sets. The goal of classification is to predict a certain outcome given an input. Predictive model is an output of processing algorithm, which discovers relationships between attributes in the training set. The algorithm is evaluated on the prediction accuracy of the discrete target attribute. This ranking is conducted using test set data, which was not used for the initial modeling. Performance of the classifier is measured as a percentage of positive hits against the total number of predictions [2]. Classic example of statistical classification is Iris flower dataset, which consisted of 50 samples from each of three species of iris (*Iris setosa*, *Iris virginica*, and *Iris versicolor*). Four features were measured from each sample: the length and the width of the sepals and petals. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other [15].

14.2.4.1.2 Correlation

A correlation indicates the strength and relationship between two random variables or two sets of data. Based on the dependence between the variables, predictive relationships can be elucidated; thus this method is commonly used in research. For example, there are many publications that prove correlation of kidney function with kidney size (measured by echo sonography) [16], blood pressure [17], and thyroid dysfunction [18]. In biomarker research it is common to investigate the correlation between renal function and a single or a group of biomarkers, for example, the urinary NGAL level is increased in children with ureteropelvic junction obstruction [19] or the correlation between cystatin C and the glomerular filtration rate (GFR) [20].

14.2.4.1.3 Regression

The regression function is used to predict numerical values; thus it is sometimes referred to as "classification with continuous classes." The constructed model is a best fit function to observational datasets provided as input. Several types of regression processes exist, such as linear, nonlinear, and multivariate types. The quality of the model is evaluated with statistical measures such as root mean squared error (RMSE) and the mean absolute error (MAE).

Regression modeling has many applications in trend analysis, time series predictions, biomedical, and drug response modeling. The regression analysis seeks to determine the values of parameters of a function that fits best to a set of observational data. There are a variety of different types of regression in statistics, but the basic idea is that a created model maps values from predictors in such a way that the lowest error occurs in making a prediction. We can distinguish the following:

- *Linear regression*—Where the relationship between predictor and target is linear.
- *Multivariate linear regression*—There are multiple predictors to determine the target value.
- *Nonlinear regression*—Relationship between predictor and target cannot be approximated with a straight line.
- *Logistic regression*—Predictor is categorical (dichotomous).

Regression analysis is widely used in research to generate predictors of patient outcome, for example, the study of Suzuki et al. presented the model of hearing outcomes of patients with sudden hearing loss. Based on selected prognostic factors such as age, days from onset to treatment, initial hearing level, and known hearing indices from the study cohort group, they were able to estimate the hearing prognosis for new patients [21]. Worth mentioning is the educational paper of Leffondre et al. where applicability of different regression models is compared (based on exemplary chronic kidney disease (CKD) patient cohort), in order to familiarize the reader with good practices and pitfalls when performing this analysis [22].

14.2.4.1.4 Association Analysis

Association rules help uncover hidden relationships between data through analysis of frequent patterns that appear in the large dataset and are rendered statistically dependent. There is a large similarity between correlation and association, but the former is considered more rigidly defined [23]. Among different measures of association, some of the most commonly used are Pearson's product moment correlation coefficient, Kendall's tau, or Spearman's rho.

Studies aiming at associations' discovery in kidney research are, for example, linking kidney conditions with common genetic variants responsible for Mendelian diseases proved that single nucleotide polymorphisms (SNPs) are not associated with eGFR [24]. In the study of Lu et al. [25], authors investigated if age and BMI influence progressive loss of kidney function. Analysis of a large cohort of veterans using Cox model showed that a BMI above 30 kg/m^2 and older age are associated with rapid loss of kidney function in patients with eGFR of at least 60 ml/min per 1.73 m^2. Moreover, best clinical outcomes are expected for patients with BMI of at least 25 kg/m^2 but less than 30 kg/m^2. Association analysis was also performed in the study of Ginsberg et al. to investigate the risk of adverse safety events in patients with advanced CKD. Observational study data was used to determine co-occurrence in events that can result in disease progression, revealing some with frequent association between disparate episodes, such as falling or severe dizziness among diabetic patients, often accompanied by hypoglycemia [26].

14.2.4.1.5 Clustering

Clustering is a common task used in statistical data analysis, which groups similar objects (discrete or continues attributes) in the set of so-called clusters. Created model can be used for mapping of new instances to initially established clusters. It is an example of unsupervised learning method, as grouping is performed on unlabeled data, contrary to classification. Usually, the points are in high-dimensional space (have multiple parameters/attributes describing them), and similarity between them is defined as a distance measure. Euclidean distance is one of the possible ways to estimate similarity (distance) between two points (e.g., samples) in a multidimensional space. It is defined as the square root of the sum of the squares of the differences between the corresponding coordinates of the points and is denoted by equation ((*x*, *y*) and (*a*, *b*) are coordinates of the points in the plane)

$$\operatorname{dist}\big((x,y),(a,b)\big)=\sqrt{(x-a)^2+(y-b)^2}$$

Clustering can be performed using various algorithms:

- *Hierarchical methods*—Build a tree of clusters, with root being a collection of all features and leaves containing one single element.
  - Agglomerative (bottom-up) algorithms—Iteratively "nearest" clusters (points) are combined together.
  - Divisive (top-down) algorithms—Initial cluster is iteratively split into smaller sub-clusters.
- *Partitioning methods*—Map a collection of elements (e.g., genes, proteins) into $k \geq 2$ clusters, aiming to maximize a particular criterion.
  - *K*-means methods—Initiated by k random centroids in Euclidian space, points are assigned to the cluster clusters and position of centroid (center of a cluster is updated).
  - Self-organizing maps (SOM).
- *Fuzzy clustering*—Form of soft clustering, that is, one element can be assigned to more than one cluster.

Clustering in DM can be applied in the preprocessing step to support other algorithms such as correlation or classification, but it can also be a stand-alone tool in the analysis. There are numerous examples of clustering usage in a variety of fields like engineering, computer sciences, life sciences, or economy [27], as well as in exploratory data analysis, for example, for species categorization and determination of interspecies relationships [28], image clustering, or pattern recognition. Image clustering can support the process of image annotation, indexing, and segmentation aiming at disease identification (e.g., cancer classification based on histology image) [29]. For example, in the field of breast cancer, image clustering largely supports risk stratification and diagnosis of patients [30]. In the work by Vivona et al., clustering algorithm successfully identified microcalcifications on mammograms, which are pathological breast lesions, generally difficult to identify due to its small size (0.1–1.0 mm) and poor contrast [31]. In terms of pattern recognition, in the study of Zang et al., hierarchical clustering was performed to analyze treatment data of over 8000 CKD patients in order to gain knowledge about commonly used clinical treatment patterns. Based on the results, six different patient classes were recognized that differed with regard to comorbidities, progression level, demographics, sex, and age. This approach shows that electronic health records (EHR) analysis can be beneficial in elucidating best suited treatment regimes, as well as applications in multifactorial disease studies such as CKD [32].

14.2.4.2 Machine Learning

The field of ML is a combination of computational science and artificial intelligence (AI) focused on data-driven model and pattern discovery, where computers are given the ability to learn without being explicitly programmed. In ML, a model is built using a statistical theory based either on training data or past experience. Importantly, the system can learn and improve through experience. An important characteristic of ML models is the ability of generalization, which is the ability to truly reproduce the output of the function [7].

ML consists of variety of approaches to tackle DM tasks. They can be categorized into two main groups:

- Symbolic approaches
  - Decision trees
  - Logical representation

- Statistical approaches
  - Bayesian classifiers
  - Neural networks (NN)
  - Support vector machines (SVM)

Remarkably, some algorithms are a mixture of both approaches. Additionally, ML algorithms can be divided into supervised and unsupervised learning methods.

14.2.4.2.1 Decision Trees

Tree-based methods are widely used in predictive modeling, due to their simplicity and good predictive power. Decision trees are a popular way of visual representation of data by splitting it into a comprehensive set of branch-like segments. Each tree consists of decision nodes and leaf nodes (representing classification or decision) and originates from the root node at the top of the tree. The decision tree analysis can be used to represent both categorical and numerical data, thus giving rise to either classification trees, which work with discrete variables, or regression trees, which assign continuous values for prediction [33].

Among the decision tree family, the most popular algorithms are ID3, C4.5, CART, CHAID, and MARS. In the basic model (ID3 algorithm), a tree is constructed using a top-down approach, where the first attribute is the root. The creation of a decision tree relies on iterative classification of instances, sorting them with regard to how well they classify the training set. The measure describing how well a particular attribute separates the set is called *information gain*. In the same manner, descendants of the root are created with possible attribute values forming tree branches [33]. At each step, splitting criterion is selected. If so-called *greedy* selection algorithm is used, a locally optimal choice is made, and there is no backtracking of former decisions. On the contrary, a *non-greedy* approach can be used to constructing globally optimal multivariate decision trees, explicitly considering all decisions in the tree concurrently [34].

Decision trees are widely used to support decision making, as they offer a clear visual representation of the selection process and can be easily transformed into a set of "IF" rules. For this reason, they can be applied in numerous research areas. For example, a decision tree was developed to guide clinical studies in the field of drug development through classification of different membrane transporters, which determined safety and efficacy profiles of drugs [35]. Additionally, the tree classification can be helpful when mining patient health records for diagnosis prediction, patient stratification, treatment options, and general prevention. In the recent paper by Huang et al. [36], classification methods were explored for the early prediction of diabetic nephropathy (DN) among diabetic patients. As DN is the leading cause of progressive renal failure among diabetics, the authors performed a decision tree modeling based on genetic and clinical characteristics of diabetic patients' cohort.

14.2.4.2.2 Bayesian Classifiers

The Bayesian classification algorithm is a statistical approach based on Bayes' probability theorem. In a nutshell, the algorithm allows us to predict a class, given a set of features using probability. As an example, we could predict whether a fruit is an apple, orange, or banana (class) based on its features, for example, color, shape, and so on. Formally, the Bayesian rule calculates a conditional probability of an event (*posterior*) given the probability of initial belief (*prior*) with relation to a *likelihood* probability of a new hypothesis. The naïve Bayes classifier is the most basic approach used for classification tasks. It offers great conceptual and implementational simplicity, as it assumes independency of an attribute value of other existing attributes [37]. A naïve Bayesian classifier offers a surprisingly excellent performance and despite its unrealistic assumption, as attribute values are not always independent, often outperforms more sophisticated classifiers [38].

Graphical models encoding a joint probabilistic distribution among the attributes are called Bayesian networks (BN). They can rely on Bayesian statistics, but not exclusively. These graphical structures help in gaining knowledge about uncertain domains. In BN each node in the graph represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables [39]. There are several advantages over other existing analysis tools that BN have to offer. First of all, they can handle missing data without producing an incorrect interpretation due to an independency of variables. What is more, a network can help with an exploration of causal relationships in the data and as a result, improve the prognosis. Since an implementation is straightforward, it is easy to combine existing knowledge with data and avoid the problem of overfitting [40].

BN classifiers are widely used in research field, facilitating systems biology modeling. Wang et al. present utility of this classifier in -omics data integration aiming at novel discoveries in mechanisms of complex diseases. Proposed BN classifier effectively predicted protein–protein interactions and grouping of proteins based on function, resulting in selection of potential biomarkers for hepatocellular carcinoma (HCC) [41]. Another interesting application was use of BN to prioritize patient clinical data in the study by Singh et al. Proposed classifier accurately classified (93.50% of accuracy) radiology reports marking them either as

"high" or "low" priority, which can support clinicians with reviewing process of medical documents and improve patient safety through faster feedback in case of worrisome results [42].

14.2.4.2.3 Artificial Neural Networks

Artificial neural network (ANN) is a classification algorithm inspired by the way a biological nervous system processes information. The artificial neuron model reflects the behavior of natural neuron processing patterns. When a signal is received by neural synapses located on the dendrites and a certain threshold, which is needed to activate the neuron, is surpassed, then the stimulated neuron emits a signal through the axon and therefore can activate other neurons. In ANN models, this complexity is highly abstracted. A simple NN consists of three interconnected layers: an input layer (similar to synapses) receiving the signal that is further multiplied by weights according to its strength, a hidden layer representing the activation function (threshold), and an output layer that returns the processing output. ANN can have many topologies depending on the complexity of the task. Nevertheless, the process of training involves adjusting the weights (values) of the functions connecting each layer to the best relationship model between the input.

The strength of an NN approach lies in the capability of learning from experience, which can help solve problems that traditional computational and statistical methods cannot handle. They are commonly used for mining patterns and trends that are too complex to be discovered without the employment of substantial computational power. There are two main types using NN architecture: feedforward networks, such as single-/multilayer perceptions where information moves in one direction without any feedback involved, or recurrent (feedback) networks, which allow neurons to send feedback signals to each other through the employment of loops and cycles in the model, for example, Kohonen's SOM.

ANN can be applied to various research fields to support decision making or classification and screening tasks, such as in disease prediction and diagnosis [43–46], as well as disease or patient stratification [47–49].

14.2.4.2.4 Support Vector Machines

The SVM algorithm is an example of a supervised training method used for classification and regression analysis. The concept of this method is to map data points (*input vectors*) in the input space to a high-dimensional feature space. The values extracted from the dataset that serve as an input to SVM algorithms are called features. In some cases, feature selection is required prior to a classifier construction, as some attributes might be considered redundant for the learning process [50]. In the feature space, a decision hyperplane is constructed, which ensures good separation of classes and high generalization ability. SVM algorithm attempts to find a linear separator of data points by maximizing the margin between the two classes. As accurate linear separation might not exist for many multidimensional datasets, SVM based on kernel functions are used to construct nonlinear separators [51].

The SVM approach emerged as a good alternative to classic analysis methods. It has proven effective for disease diagnosis and prediction, especially in the field of neuroimaging for neurological and psychiatric disease predictions [50]. Apart from imaging data, SVM are used in many case–control studies aimed at disease detection, as an example in the field of peptidomics where a classifier (CKD273) has been developed to assess the stages of CKD [52]. In the study by Marom et al. [53], the SVM algorithm was employed to detect if a simple breath test can serve for early detection of CKD.

14.2.4.3 Text Mining

TM is a topic that utilizes many DM approaches for comprehensive text analysis to be used for exploration, discovery, classification, or prediction purposes. Contrary to databases, where information is presented in an organized and informative manner, textual data has no structured components, which makes the querying and analysis a real challenge. Additionally, the volume of data considered for information retrieval is huge, which further impedes the TM process. Effective TM is of special interest in biomedical research, as evidenced by the growing number of publications concerning this topic and possible applications (Figure 14.6). As novel findings are disseminated among the scientific community in the form of manuscripts, extraction of valuable information that can be used for further modeling needs to be performed manually or with the help of TM tools. However, the rapidly generated literature is difficult to keep up with, and following distinct research areas in the traditional manner is becoming nearly impossible, because of the lack of scalability of manual curation [54].

In the search for new knowledge, many DM techniques, combined with natural language processing, have been employed and have substantially improved over the past decade [55,56]. The application of TM in the field of systems biology not only enabled automatic detection and annotation of molecules such as genes or proteins but also empowered more advanced discoveries of biological events and interactions between biological components [57]. Consequently, TM supports the development of many biological databases through automated processes of data collection [58–61]. It is also popular in studies where electronic patient records

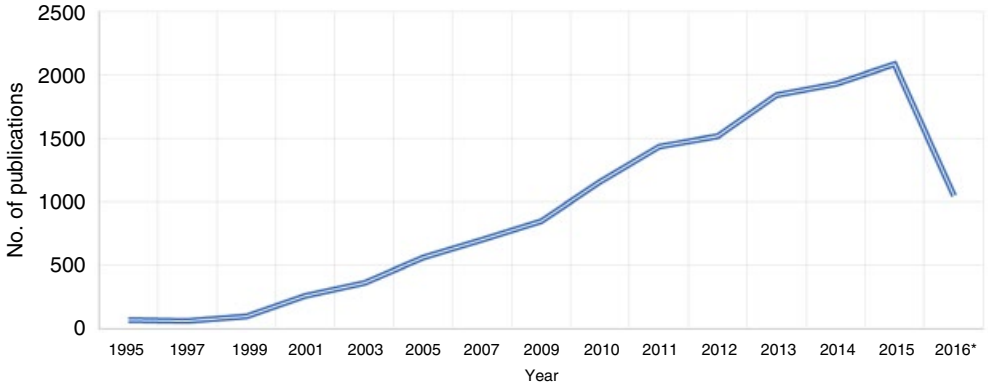


Figure 14.6 The rise of text mining approaches in science and discovery from 1995 to 2015. Information retrieved from PubMed based on "text mining OR literature mining" keywords. *search performed July 2016.

(EPR) are screened in search for disease patterns or diagnostic tools [62,63]

14.2.5 Pattern Evaluation

The final step of the knowledge discovery process is the evaluation of an obtained model and its optimization and interpretation. As previously mentioned, this assessment is needed to determine how well the classifier performs and if an under- or overfitting problem exists. There are several methods to evaluate the model by measuring various indicators such as error rate, squared error, likelihood, information gain, cost function, or margins. For this purpose, a validation set (*holdout pattern*) is isolated from the data collection; nevertheless in the case of a small sample size where a separate subset of data cannot be selected, other validation patterns can be employed [64]:

- *k-Fold cross-validation*—The dataset is divided into *k* subsets where one of the subsets is used as the validation set and the remainder is used for training. The process of model evaluation is performed for different validation sets, and an error is produced as an average of errors from *k*-iterations.
- *Leave-one-out*—The model judgment is calculated for each example from the dataset being used as a test set, and then the average is calculated for all judgment values.
- *Bootstrapping*—From the original dataset, a sample subset is selected, and a model is trained and evaluated by a calculation of training, testing, and final errors that are averages of every iterations.

Based on the given error of the predictor, the existence of under- or overfitting can be discovered. An additional use is in optimization methods that are developed to tune learning parameters of a model to choose the best scoring classifier.

14.3 Data Mining in Scientific Applications

Scientific DM is probably the most challenging and exciting area of bioinformatics. Significant technological advancements made in the field of high-throughput -omics, that is, genomics, transcriptomics, proteomics, and metabolomics, directed current efforts toward designing software able to handle the analysis of the continuing flow of experimentally generated data [65]. DM approaches have been used to support traditional statistical techniques to address "big data" challenges, such as accounting for the large dimensionality and complexity of biological data [66]. Growing interest in DM techniques in research can be noticed through rapidly increasing number of scientific publications concerning these topics (Figure 14.7).

Given the diversity of DM algorithms, distinct views on DM have been taken by researchers. Commonly, DM is considered as an induction, that is, generalization of data through exploration of possible patterns and replacement of low-level values (such as numeric value of age) by higher-level concept values (e.g., young, old, age intervals). Similarly, DM is also presented as an approximation process, in which exact information is simplified in an attempt to identify hidden structure of the data. Among database system community, DM is perceived as intelligently querying a dataset [3]. Therefore, there is no definite recommendation when deciding what DM approach to use for the exploration of the scientific data. Diversity of available options makes it difficult for a non-experienced researcher to choose the most appropriate method. However, some algorithms have been used more frequently with regard to the specific type of -omics data analyzed and will be presented here.

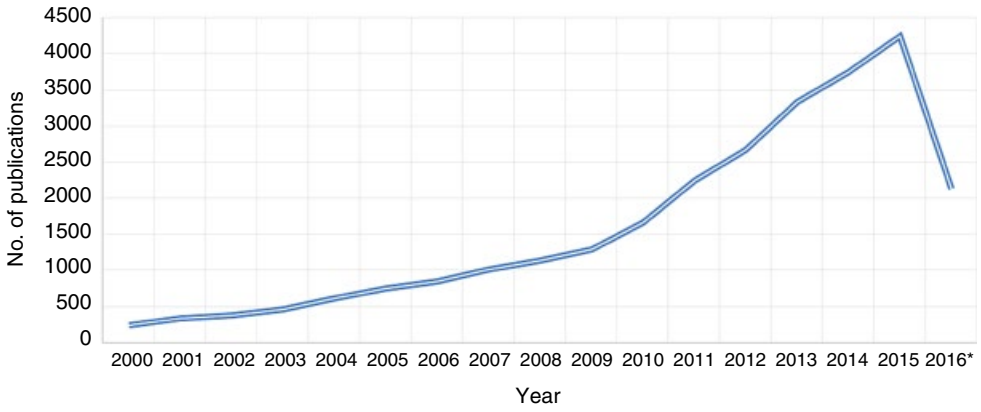


Figure 14.7 Growing interest in data mining techniques in scientific research. Information retrieved from PubMed based on "data mining OR machine learning" keywords. *search performed July 2016.

14.3.1 Genomics Data Mining

Computational methods of data analysis are considered the most advanced in the field of genomics. Next-generation sequencing technologies measuring DNA, RNA, and epigenetic patterns are the source of large genomics datasets, rich with information, but difficult to process and extract valuable information. There are numerous challenges accompanied with the analysis of such data. As high-throughput genomics experiments yield a huge number of candidate targets, the probability of false positive results is high. Relying solely on stringent statistical criteria might discriminate many real targets and introduce high false negative error. Thus, DM techniques need to account for this so-called multiple comparisons issue and minimize both types of errors. Moreover, high dimensionality and variability (noise) of genomics data further add to the complexity of the analysis [67].

A number of scientific publications report successful application of DM techniques in the area of genetics and genomics. ML techniques have been widely used for gene annotation, prediction of gene function, identification of regulatory elements, and so on [68]. Unsupervised learning methods of clustering as well as supervised learning algorithms are typically used to establish gene expression patterns in different study groups [69]. However, it has been pointed that many factors, both technical and biological, such as differences in analysis platforms, normalization methods, or gene expression correlation might be a source of discrepancy between identified prognostic markers [70]. Therefore, many efforts have been made aiming to define stable feature selection in the field of biomarker discovery [70,71].

Clustering of microarray data is commonly used as a method to efficiently reduce dimensionality of gene expression data. Figure 14.8 shows an example application of learning algorithm for building and evaluation of a model using scientific data. Both classes of clustering algorithms, that is, hierarchical and partitioning, have been applied in genomics data analysis [72]. Output of this analysis is a two-dimensional dendrogram, where genes with similar expression patterns are grouped together and connected by branches. These formed clusters might suggest co-expression, co-regulation, or existence of groups in samples or even detect outliers [73]. Importantly, interpretation of the results is often the most challenging part of clustering analysis. BN are another type of unsupervised learning methods that have been applied for gene expression analyses. Constructed network represents conditional dependencies between genes, which might be translated into possible interactions within cells and, thus, provide novel biological insights. For this reason, Bayes' nets attracted attention of scientists and were used as a discovery method in a number of genomics studies [74–76]. Lastly, classification methods are widely applied in microarray analyses to find gene expression patterns able to discriminate diseased from healthy people and predict disease stages or patient outcome.

14.3.2 Proteomics Data Mining

Similarly to genomics, a variety of DM techniques have been fruitfully applied in proteomics data analysis. DM application can provide information on individual proteins through protein annotations and prediction of protein structure, function, location, or possible interactions [77]. Moreover, proteomics DM enables gaining deeper understanding of protein networks in health and disease. Due to their high-throughput nature, mass spectrometry (MS)-based methods for characterization of protein content and expression in biological samples have largely benefited from DM techniques [78]. Primarily, mining can be performed on yet unprocessed information (mass peaks) to find distinctive patterns in data. Furthermore,

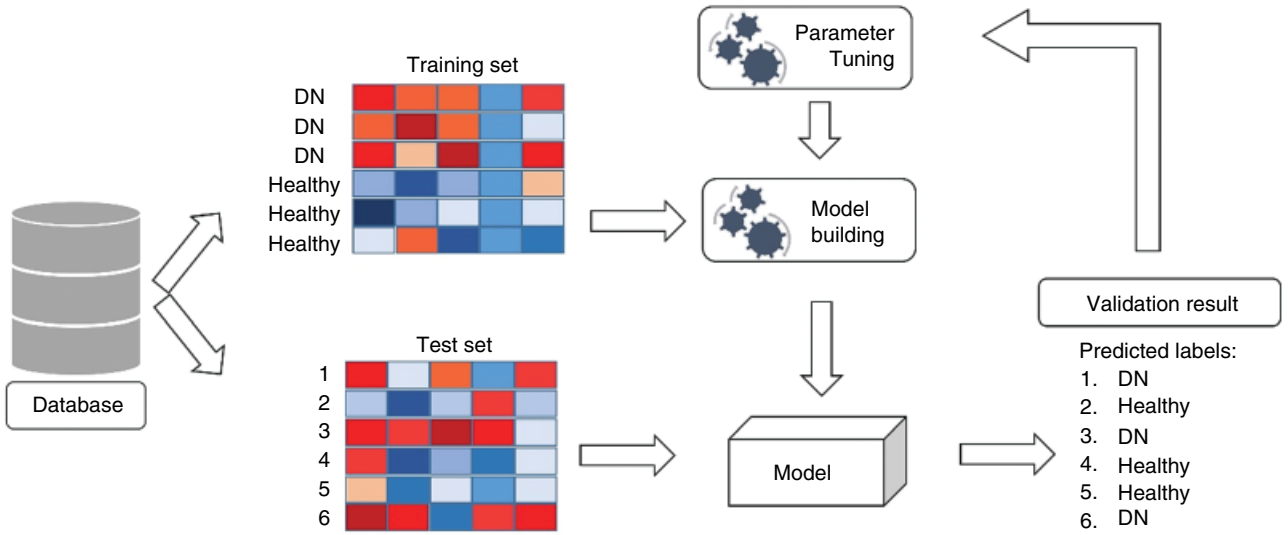


Figure 14.8 General example of machine learning application. A training set of labeled gene expression data is used as an input for the learning algorithm to build a model able to predict the label (diabetic nephropathy (DN) or healthy) for future gene expression data (unlabeled, test set). Based on the model, labels are assigned to the test set, and performance of the model is evaluated.

identified peptides and/or proteins can be used as an input for the learning algorithm, which can pinpoint the proteins discriminating patients from healthy controls and facilitate biomarker discovery and unraveling the biological mechanisms of different diseases [79].

ML algorithms are becoming a popular tool in the proteomics research. Supervised learning methods have been used for the development of classification tools, useful in diseases diagnosis and biomarker research. For example, in the work of Good et al., capillary electrophoresis coupled to MS (CE-MS) has been used to measure naturally occurring peptides in urine of CKD patients, healthy controls, and patients with different diseases. Using SVM algorithms, a polypeptide classifier (CKD 273) is able to diagnose CKD patients with 85.5% sensitivity and 100% specificity [80]. Along these lines, through SVM modeling, Jiang et al. have developed an interaction network-based proteomics classifier for the diagnosis of prostate cancer (PCa). Interestingly, one of the proteins comprising the classifier was also identified as an independent prognostic marker of PCa [81]. Other popular applications of learning algorithms in proteomics have been recently described in an excellent review by Kelchtermans et al. [82].

14.4 Bioinformatics Data Mining Tools

Proven benefits from DM application give way for the development of various informatics tools capable of performing mining tasks. There are many general DM systems that can be used for biological DM such as SAS Enterprise Miner, R, Matlab, SPSS, S-Plus, IBM Intelligent Miner, Microsoft Analysis Services, SGI MineSet, and Inxight VizServer. However, some biological DM tools such as GeneSpring, Spot Fire, VectorNTI, COMPASS, Statistics for Microarray Analysis, and Affymetrix Data Mining Tool have been developed. Also, a large number of biological DM tools are provided by the National Center for Biotechnology Information and by the European Bioinformatics Institute (www.ncbi.nlm.nih.gov).

14.5 Conclusions

In the era of emerging technologies and big data where enormous amounts of information are produced on a daily basis, there is a great need for solutions supporting efficient analysis and knowledge discovery. DM techniques offer promising solutions to deal with this task through the employment of sophisticated statistical algorithms combined with AI approaches and an extensive use of database management systems. The field of DM is rapidly evolving, providing constantly improving ways to mine data to detect hidden patterns, relationships, and interactions. The KDD is a step-by-step approach to guide analysts through the complex stages of data analysis with use of DM methodologies.

The biomedical research field has greatly benefited from DM as shown by the growing number of scientific publications where various mining techniques have been used. DM methodologies have proven successful in tackling many bioinformatics tasks and are

being continuously developed to enable effective data analysis. Therefore, given the ever-growing amount of biological data and its complexity, knowledge discovery approaches are fast becoming essential and are expected to play a significant role in the present and future research.

References

1 Silwattananusarn, T. & Tuamsuk, K. 2012. Data mining and its applications for knowledge management: literature review from 2007 to 2012. *Int J Data Min Knowl Manage Process*, 2, 13–24.

2 Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F. &Hua, L. 2012. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst*, 36, 2431–48.

3 Ramakrishnan N. & Grama A. Y. 2001. Mining scientific data. *Adv Comput*, 55, 119–69.

4 Holzinger, A., Dehmer, M. & Jurisica, I. 2014. Knowledge discovery and interactive data mining in bioinformatics—state-of-the-Art, future challenges and research directions. *BMC Bioinformatics*, 15(Suppl 6), I1.

5 Maimon, O. & Rokach, L. 2005. *Data Mining and Knowledge Discovery Handbook*. Springer, New York.

6 Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. 2000. CRISP-DM 1.0 Step-by-step data mining guide.

7 Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., et al. 2006. Machine learning in bioinformatics. *Brief Bioinformatics*, 7, 86–112.

8 Curbelo, R. J., Loza, E., De Yebenes, M. J. & Carmona, L. 2014. Databases and registers: useful tools for research, no studies. *Rheumatol Int*, 34, 447–52.

9 Connolly, T. M. & Begg, C. E. 2004. *DataBase Systems: A Practical Approach to Design, Implementation and Management* (4th Edition) (International Computer Science), Addison-Wesley Longman Publishing Co. Inc., Boston.

10 Stein, L. D. 2003. Integrating biological databases. *Nat Rev Genet*, 4, 337–45.

11 Zou, D., Ma, L., Yu, J. & Zhang, Z. 2015. Biological databases for human research. *Genomics Proteomics Bioinformatics*, 13, 55–63.

12 Papadopoulos, T., Krochmal, M., Cisek, K., Fernandes, M., Husi, H., Stevens, R., Bascands, J. L., et al. 2016. Omics databases on kidney disease: where they can be found and how to benefit from them. *Clin Kidney J*, 9, 343–52.

13 Ledolter, J. 2017. *Data Mining and Business Analytics with R*. John Wiley & Sons, Inc., Hoboken.

14 Biemann, C. 2007. *Unsupervised and Knowledge-free Natural Language Processing in the Structure Discovery Paradigm*. PhD thesis, University of Leipzig.

15 Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann Eugenics*, 7, 179–88.

16 Jovanovic, D., Gasic, B., Pavlovic, S. & Naumovic, R. 2013. Correlation of kidney size with kidney function and anthropometric parameters in healthy subjects and patients with chronic kidney diseases. *Ren Fail*, 35, 896–900.

17 Vaes, B., Beke, E., Truyers, C., Elli, S., Buntinx, F., Verbakel, J. Y., Goderis, G., et al. 2015. The correlation between blood pressure and kidney function decline in older people: a registry-based cohort study. *BMJ Open*, 5, e007571.

18 Prajapati, P., Singh, A. P. & Bendwal, S. 2013. Correlation between severity of chronic kidney disease and thyroid dysfunction. *J Indian Med Assoc*, 111, 514–6.

19 Cost, N. G., Noh, P. H., Devarajan, P., Ivancic, V., Reddy, P. P., Minevich, E., Bennett, M., et al. 2013. Urinary NGAL levels correlate with differential renal function in patients with ureteropelvic junction obstruction undergoing pyeloplasty. *J Urol*, 190, 1462–7.

20 Ayub, S., Khan, S., Ozair, U. & Zafar, M. N. 2014. Cystatin C levels in healthy kidney donors and its correlation with GFR by creatinine clearance. *J Pak Med Assoc*, 64, 286–90.

21 Suzuki, H., Tabata, T., Koizumi, H., Hohchi, N., Takeuchi, S., Kitamura, T., Fujino, Y., et al. 2014. Prediction of hearing outcomes by multiple regression analysis in patients with idiopathic sudden sensorineural hearing loss. *Ann Otol Rhinol Laryngol*, 123, 821–5.

22 Leffondre, K., Jager, K. J., Boucquemont, J., Stel, V. S. & Heinze, G. 2014. Representation of exposures in regression analysis and interpretation of regression coefficients: basic concepts and pitfalls. *Nephrol Dial Transplant*, 29, 1806–14.

23 Graham, U. & Ian, C. 2014. *A Dictionary of Statistics*. Oxford University Press, Oxford.

24 Parsa, A., Fuchsberger, C., Kottgen, A., O'Seaghdha, C. M., Pattaro, C., De Andrade, M., Chasman, D. I., et al. 2013. Common variants in Mendelian kidney disease genes and their association with renal function. *J Am Soc Nephrol*, 24, 2105–17.

25 Lu, J. L., Molnar, M. Z., Naseer, A., Mikkelsen, M. K., Kalantar-Zadeh, K. & Kovesdy, C. P. 2015. Association of age and BMI with kidney function and mortality: a cohort study. *Lancet Diabetes Endocrinol*, 3(9), 704–14.

26 Ginsberg, J. S., Zhan, M., Diamantidis, C. J., Woods, C., Chen, J. & Fink, J. C. 2014. Patient-reported and actionable safety events in CKD. *J Am Soc Nephrol*, 25, 1564–73.

27 Xu, R. & Wunsch, D. 2005. Survey of clustering algorithms. *IEEE Trans Neural Netw*, 16, 645–78.

28 Elith, J., Graham, C. H., Anderson, R. P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R. J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129–51.

29 Lin, N., Jiang, J., Guo, S. & Xiong, M. 2015. Functional principal component analysis and randomized sparse clustering algorithm for medical image analysis. *PLoS One*, 10, e0132945.

30 Keller, B. M., Nathan, D. L., Wang, Y., Zheng, Y., Gee, J. C., Conant, E. F. & Kontos, D. 2012. Estimation of breast percent density in raw and processed full field digital mammography images via adaptive fuzzy c-means clustering and support vector machine segmentation. *Med Phys*, 39, 4903–17.

31 Vivona, L., Cascio, D., Fauci, F. & Raso, G. 2014. Fuzzy technique for microcalcifications clustering in digital mammograms. *BMC Med Imaging*, 14, 23.

32 Zhang, Y., Padman, R. & Wasserman, L. 2014. On learning and visualizing practice-based clinical pathways for chronic kidney disease. *AMIA Annu Symp Proc*, 2014, 1980–9.

33 Mitchell, T. M. 1997. *Machine Learning*, McGraw-Hill, Inc., Maidenhead.

34 Bennett, K. P. 1994. Global Tree Optimization—A Non-Greedy Decision Tree Algorithm, Interface Foundation North America, Fairfax, VA.

35 International Transporter, C., Giacomini, K. M., Huang, S. M., Tweedie, D. J., Benet, L. Z., Brouwer, K. L., Chu, X., et al. 2010. Membrane transporters in drug development. *Nat Rev Drug Discov*, 9, 215–36.

36 Huang, G. M., Huang, K. Y., Lee, T. Y. & Weng, J. 2015. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. *BMC Bioinformatics*, 16 (Suppl 1), S5.

37 Friedman, N., Geiger, D. & Goldszmidt, M. 1997. Bayesian network classifiers. *Mach Learning*, 29, 131–63.

38 Larrañaga, P., Karshenas, H., Bielza, C. & Santana, R. 2013. A review on evolutionary algorithms in Bayesian network learning and inference tasks. *Inform Sci*, 233, 109–25.

39 Ben-Gal, I. 2008. *Bayesian Networks. Encyclopedia of Statistics in Quality and Reliability*. John Wiley & Sons, Ltd, Chichester.

40 Heckerman, D. 2008. A tutorial on learning with Bayesian networks. In: Holmes, D. & Jain, L. (eds.) *Innovations in Bayesian Networks*. Springer, Berlin/ Heidelberg.

41 Wang, J., Zuo, Y., Liu, L., Man, Y., Tadesse, M. G. & Ressom, H. W. 2014. Identification of functional modules by integration of multiple data sources using a Bayesian network classifier. *Circ Cardiovasc Genet*, 7, 206–17.

42 Singh, M. & Murthy, A. 2015. Prioritization of free-text clinical documents: a novel use of a bayesian classifier. *JMIR Med Informatics*, 3, e17.

43 Asadi, H., Dowling, R., Yan, B. & Mitchell, P. 2014. Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy. *PLoS One*, 9, e88225.

44 Chowdhury, D. R., Chatterjee, M. & Samanta, R. K. 2011. An artificial neural network model for neonatal disease diagnosis. *Int J Artif Intell Exp Syst*, 2, 96–106.

45 Huang, C.-J. & Liao, W.-C. 2004. Application of probabilistic neural networks to the class prediction of leukemia and embryonal tumor of central nervous system. *Neural Processing Lett*, 19, 211–26.

46 Jaafar, S. F. B. & Ali, D. M. Diabetes mellitus forecast using artificial neural network (ANN). Sensors and the International Conference on new Techniques in Pharmaceutical and Biomedical Research, 2005 Asian Conference on, September 5–7, 2005, pp. 135–9.

47 Borkowska, E. M., Kruk, A., Jedrzejczyk, A., Rozniecki, M., Jablonowski, Z., Traczyk, M., Constantinou, M., et al. 2014. Molecular subtyping of bladder cancer using Kohonen self-organizing maps. *Cancer Med*, 3, 1225–34.

48 Schilithz, A. O., Kale, P. L., Gama, S. G. & Nobre, F. F. 2014. Risk groups in children under six months of age using self-organizing maps. *Comput Methods Programs Biomed*, 115, 1–10.

49 Vanfleteren, L. E., Spruit, M. A., Groenen, M., Gaffron, S., Van Empel, V. P., Bruijnzeel, P. L., Rutten, E. P., et al. 2013. Clusters of comorbidities based on validated objective measurements and systemic inflammation in patients with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 187, 728–35.

50 Orru, G., Pettersson-Yeo, W., Marquand, A. F., Sartori, G. & Mechelli, A. 2012. Using Support Vector Machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. *Neurosci Biobehav Rev*, 36, 1140–52.

51 Molla, M., Waddell, M., Page, D. & Shavlik, J. 2004. Using machine learning to design and interpret gene-expression microarrays. *AI Mag*, 25, 23–44.

52 Argilés, À., Siwy, J., Duranton, F., Gayrard, N., Dakna, M., Lundin, U., Osaba, L., et al. 2013. CKD273, a new proteomics classifier assessing CKD and its prognosis. *PLoS One*, 8, e62837.

53 Marom, O., Nakhoul, F., Tisch, U., Shiban, A., Abassi, Z. & Haick, H. 2012. Gold nanoparticle sensors for detecting chronic kidney disease and disease progression. *Nanomedicine (Lond)*, 7, 639–50.

54 Harmston, N., Filsell, W. & Stumpf, M. P. 2010. What the papers say: text mining for genomics and systems biology. *Hum Genomics*, 5, 17–29.

55 Fluck, J. & Hofmann-Apitius, M. 2014. Text mining for systems biology. *Drug Discov Today*, 19, 140–4.

56 Zhu, F., Patumcharoenpol, P., Zhang, C., Yang, Y., Chan, J., Meechai, A., Vongsangnak, W., et al. 2013. Biomedical text mining and its applications in cancer research. *J Biomed Inform*, 46, 200–11.

57 Ananiadou, S., Pyysalo, S., Tsujii, J. & Kell, D. B. 2010. Event extraction for systems biology by text mining the literature. *Trends Biotechnol*, 28, 381–90.

58 Chatr-Aryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., et al. 2013. The BioGRID interaction database: 2013 update. *Nucleic Acids Res*, 41, D816–23.

59 Dweep, H., Sticht, C., Pandey, P. & Gretz, N. 2011. miRWalk—database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J Biomed Inform*, 44, 839–47.

60 Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., et al. 2013. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*, 41, D808–15.

61 Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Sohngen, C., et al. 2011. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res*, 39, D670–6.

62 Jensen, P. B., Jensen, L. J. & Brunak, S. 2012. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet*, 13, 395–405.

63 Nadkarni, G. N., Gottesman, O., Linneman, J. G., Chase, H., Berg, R. L., Farouk, S., Nadukuru, R., et al. 2014. Development and validation of an electronic phenotyping algorithm for chronic kidney disease. *AMIA Annu Symp Proc*, 2014, 907–16.

64 Souza, J., Matwin, S. & Japkowicz, N. 2002. Evaluating data mining models: a pattern language. *Proceedings of the 9th Conference on Pattern Language of Programs*, Illinois, USA.

65 Greene, C. S., Tan, J., Ung, M., Moore, J. H. & Cheng, C. 2014. Big data bioinformatics. *J Cell Physiol*, 229, 1896–900.

66 Zinovyev, A. 2015. Overcoming complexity of biological systems: from data analysis to mathematical modeling. *Math Model Nat Phenom*, 10, 186–205.

67 Lee, J.K., Williams, P.D. & Cheon, S. 2008. Data mining in genomics. *Clin Lab Med*, 28, 145–66.

68 Libbrecht, M. W. & Noble, W. S. 2015. Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16, 321–32.

69 Pyatnitskiy, M., Mazo, I., Shkrob, M., Schwartz, E. & Kotelnikova, E. 2014. Clustering gene expression regulators: new approach to disease subtyping. *PLoS One*, 9, e84955.

70 Cheng, W. C., Shu, W. Y., Li, C. Y., Tsai, M. L., Chang, C. W., Chen, C. R., Cheng, H. T., et al. 2012. Intra- and inter-individual variance of gene expression in clinical studies. *PLoS One*, 7, e38650.

71 He, Z. & Yu, W. 2010. Stable feature selection for biomarker discovery. *Comput Biol Chem*, 34, 215–25.

72 Pollard K. S., van der Laan M. J. 2005. Cluster analysis of genomic data. In: *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer-Verlag, New York, pp. 209–228.

73 Shannon, W., Culverhouse, R. & Duncan, J. 2003. Analyzing microarray data using cluster analysis. *Pharmacogenomics*, 4, 41–52.

74 Akutekwe, A. & Seker, H. 2014. A hybrid dynamic Bayesian network approach for modelling temporal associations of gene expressions for hypertension diagnosis. *Conf Proc IEEE Eng Med Biol Soc*, 2014, 804–7.

75 Lo, L. Y., Wong, M. L., Lee, K. H. & Leung, K. S. 2015. High-order dynamic Bayesian Network learning with hidden common causes for causal gene regulatory network. *BMC Bioinformatics*, 16, 395.

76 Tian, X. W. & Lim, J. S. 2015. Interactive naive Bayesian network: a new approach of constructing gene-gene interaction network for cancer classification. *Biomed Mater Eng*, 26 (Suppl 1), S1929–36.

77 Hooda, Y. & Kim, P. M. 2012. Computational structural analysis of protein interactions and networks. *Proteomics*, 12, 1697–705.

78 Elo, L. L. & Schwikowski, B. 2012. Mining proteomic data for biomedical research. *Wiley Interdiscip Rev Data Mining Knowl Discov*, 2, 1–13.

79 Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. & Bacardit, J. 2013. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *OMICS J Integrative Biol*, 17, 595–610.

80 Good, D. M., Zurbig, P., Argiles, A., Bauer, H. W., Behrens, G., Coon, J. J., Dakna, M., et al. 2010. Naturally occurring human urinary peptides for use in diagnosis of chronic kidney disease. *Mol Cell Proteomics*, 9, 2424–37.

81 Jiang, F. N., He, H. C., Zhang, Y. Q., Yang, D. L., Huang, J. H., Zhu, Y. X., Mo, R. J., et al. 2013. An integrative proteomics and interaction network-based classifier for prostate cancer diagnosis. *PLoS One*, 8, e63941.

82 Kelchtermans, P., Bittremieux, W., De Grave, K., Degroeve, S., Ramon, J., Laukens, K., Valkenborg, D., et al. 2014. Machine learning applications in proteomics research: how the past can boost the future. *Proteomics*, 14, 353–66.

15

-Omics and Clinical Data Integration

Gaia De Sanctis[1,2], Riccardo Colombo[1,3], Chiara Damiani[1,3], Elena Sacco[1,2], and Marco Vanoni[1,2]

[1] *SYSBIO, Centre of Systems Biology, Milan, Italy*
[2] *Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy*
[3] *Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy*

15.1 Introduction

The various components of a biological system do not act individually, but rather through complex hierarchical, coordinated, dynamical, and nonlinear interactions of a large number of components (e.g., proteins interacting with DNA, RNA, metabolites, and other proteins) that allow the functioning (or dysfunctioning) of the system itself [1]. Therefore, a biological function generates as an emergent property [2] of the system and is not ascribable or found in its single components, but only in their networking.

High-throughput technologies allow to collect genome-scale comprehensive molecular information—collectively referred to as -omics data—using an increasing number of sophisticated high-throughput technologies, including transcriptomics, proteomics (that includes the study of protein level and posttranslational modification data at the proteome scale), metabolomics, and interactomics. Indeed, our current understanding of biological functions is not limited by availability of vast amounts of data (big data), but rather by our ability to integrate and process them. Systems biology [3, 4] is the conceptual and operative approach needed to extract and integrate information from this huge amount of different -omics data. The systems biology approach systematically organizes, integrates, and rationalizes the different -omics data through statistical analysis, computer-aided modeling, and visualization. It requires different scientific competencies so to give them structure, improve our understanding of emergent properties and their design principles, and gain ability to predict the behavior of a system and to exploit it for applicative purposes (Figure 15.1).

Most common diseases affecting adults, including cardiovascular diseases, cancer, and diabetes, are multifactorial and derive by the interaction of several genetic and environmental factors concurring to phenotype and clinical manifestation [5–9]. As -omics data become available with ever-increasing accuracy and decreased cost, they can be used to guide the choice, design, and follow-up of effective therapeutic approaches, allowing to translate systems biology to medicine that aims at tackling the complexity of multifactorial diseases by means of systematic and integrated approaches for clinical purposes, that is, to allow a more efficient disease classification and identification of novel therapeutic targets. Post-genomics -omics-based systems medicine aims to transform diagnostic and therapeutic strategies being, in the next future, "personalized and predictive," namely, able to suggest the most potentially effective drug for any patient and to eventually foretell if and when a disease will occur and how it will develop [10–14].

Healthcare systems have nowadays to face considerable challenges connected with the highly variable clinical efficacy of current drugs as well to the huge costs associated with drug discovery, development, and clinical trials, inevitably causing economical suffering and high impacts on the financing of the sector. Indeed, a basic problem of the current disease classification system, based on phenotype determination, is that the same phenotype may derive from several disease mechanisms. Thus, a drug directed against one of those mechanisms would not be clinically effective in patients with different underlying mechanisms [15].

Let us take breast cancer as an example. During the last 30 years, the definition of a few biomarkers allowed to identify molecular breast cancer subgroups with different clinical characteristics, clinical courses, and sensitivities to existing therapies and allowed to design novel and more effective treatments for patients [16]. Patients with estrogen receptor-negative, human epidermal

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

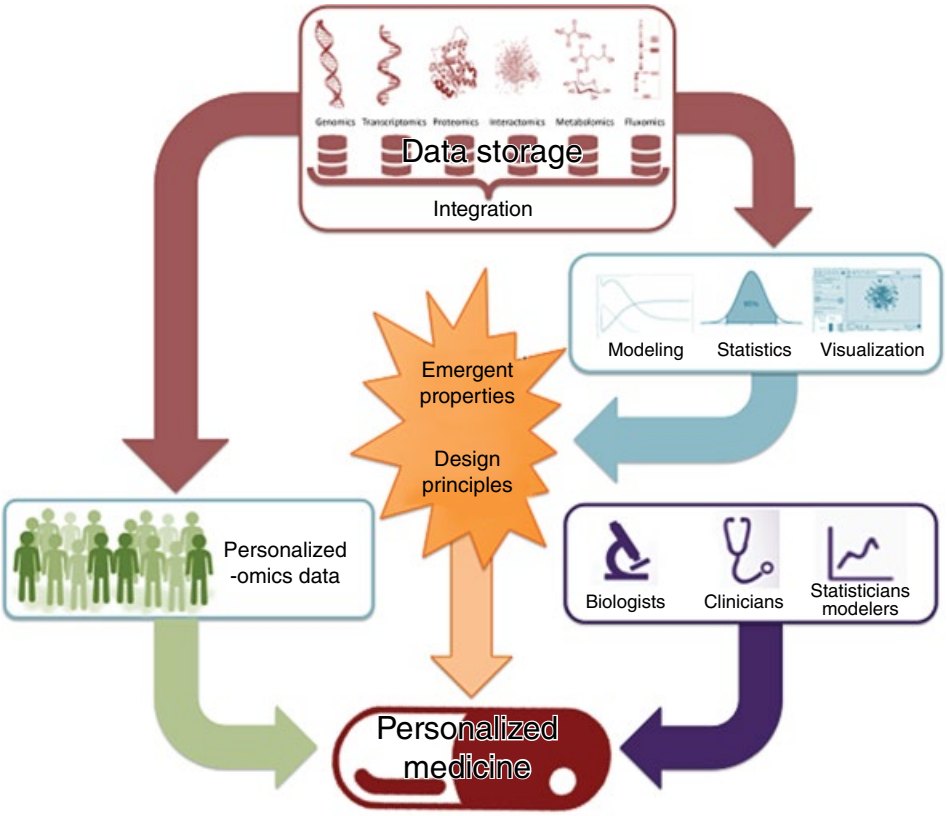


Figure 15.1 Overview of the systems biology approach toward the realization of personalized medicine. The many different -omics data currently available (reddish box) must be deposited in one or more database in order to systematically organize the information and to facilitate the data integration process. Integrated -omics data are analyzed (light-blue box) by means of statistical methods, computer-aided modeling, and visually represented in order to understand emergent properties and design principles of the biological system (orange cartoon). Personalized -omics data (green box), the knowledge of emergent properties/design principles, and different scientific competencies (purple box) will ultimately merge, allowing to develop personalized medical treatment of multifactorial diseases.

growth factor receptor 2-positive cancers are currently treated with the monoclonal antibody trastuzumab and have one of the more favorable prognoses of all breast cancer patients. However, trastuzumab is effective only in up to 50% of these patients, possibly because of various resistance mechanisms. As knowledge of the molecular events underlying the ability to respond to trastuzumab treatment increases, novel accurate predictive biomarkers—allowing to identify those patients who will respond to trastuzumab treatment—and/or novel drug targets will be identified. The increased resolution in the classification of these tumors will allow to develop new, highly targeted molecular therapeutics and at the same time to devise molecular diagnostic tools that will allow to implement a truly personalize medicine.

In this chapter we describe approaches to -omics integration that may uncover information hidden in each individual -omics. Integration of -omics data can be fully exploited if combined with modeling approaches, allowing to develop precision, personalized medicine of patients of multifactorial diseases, such as cancer.

15.2 Data Sources

-Omics technologies generate extremely large datasets and a quick web search can give a first picture of the huge variety of data sources publicly available. Assessing their relevance and quality may be a particularly hard task due to the heterogeneity of representations and notations.

Historically, the Human Genome Project has been the first -omics initiative related to human health. This pioneering study gave rise to a plethora of initiatives paving the way toward the current explosion of data generated by means of different -omics approaches. Focusing exclusively on data sources related to cancer disease, Table 15.1 describes some of the most relevant repositories of -omics data. Most of the available resources deal with genomics, transcriptomics, and proteomics data with some emphasis on cancer-related data. More recent initiatives tend to shed light on the heterogeneity of cancer in terms of genomic mutations and phenotypic differences among the different tumor subtypes. Table 15.1 highlights a substantial lack of

Table 15.1 Main data sources available for data integration.

Data source	Main features	Type of -omics data	Web page	References
Gene Expression Omnibus (GEO)	Repository of gene expression data from more than 2500 studies	Proteomics	http://www.ncbi.nlm.nih.gov/geo/	Edgar et al. [17], Barrett et al. [18]
Ensembl	Sequence data fed into a gene annotation system creating a set of predicted gene locations saved in a MySQL database for subsequent analysis and display	Genomics	http://www.ensembl.org	Hubbard et al. [19]
TRANSFAC database	Data on eukaryotic transcription factors and their miRNAs, binding sites, and regulated genes	Regulomics	http://www.biobase-international.com/product/transcription-factor-binding-sites	Wingender et al. [20]
1000 Genomes Project	Generic genetic variants whose frequencies are at least of 1% in the human population by NGS sequencing of genomes from many individuals. The raw and processed data associated with the 1000 resulting genomes are also stored and managed	Genomics	http://www.1000genomes.org/	Abecasis et al. [21], Abecasis et al. [22]
Encyclopedia of DNA Elements Project (ENCODE)	Integration-based approach aimed at the characterization, for a set of animal models/tissues/cell lines, of the profile of mRNA expression, histone marks and transcription factor binding profiling, DNA methylation, chromatin conformation, and location of active regulatory regions, among others	Genomics Transcriptomics Epigenomics Regulomics	http://genome.ucsc.edu/ENCODE/	Ecker et al. [23], ENCODE Project Consortium [24], Harrow et al. [25]
The Cancer Genome Atlas Project (TCGA)	Provides insights into the heterogeneity of different cancer subtypes by creating a map of molecular alterations for every type of cancer at multiple levels. For instance, the endometrial carcinoma has been characterized by mRNA, miRNA, protein, DNA methylation, copy number alterations, and somatic chromosomal aberrations	Transcriptomics Regulomics Proteomics Epigenomics Phenomics	http://cancergenome.nih.gov/	Weinstein et al. [26]
International Cancer Genome Consortium (ICGC)	Coordinates large-scale cancer genome studies in tumors from 50 cancer types/subtypes of main importance across the globe. More than 25 000 cancer genomes are studied at the genomics,	Genomics Epigenomics Transcriptomics	https://www.icgc.org/	Hudson et al. [27]
Cancer Genome Project (CGP)	Uses the human genome sequence and high-throughput mutation detection techniques to identify somatically acquired	Genomics Phenomics	http://www.sanger.ac.uk/research/projects/cancergenome/	Pleasance et al. [28]
Catalogue of Somatic Mutations in Cancer (COSMIC)	Contains data generated from the ICGC and TCGA studies, the Cancer Genome Project (CGP), and targeted sequencing of the NCI60 cell lines (a panel of 60 human cell lines) in known	Genomics Phenomics	http://cancer.sanger.ac.uk/cosmic	Bamford et al. [29]
Clinical Proteomic Tumor Analysis Consortium (CPTAC)	Has the goal of understanding the molecular basis of cancer through the application of proteomics technologies and workflows, systematically identifying proteins that derive from alterations in	Proteomics Genomics	https://proteomics.cancer.gov/programs/cptac	Ellis et al. [30], Zhang et al. [31]

Kyoto Encyclopedia of Genes and Genomes (KEGG)	KEGG contains representations of biological systems. It integrates genetic building blocks of genes and proteins, chemical building blocks of small molecules and reactions, and wiring diagrams of molecular interaction and reaction networks. Thus, KEGG databases are categorized into systems, genomic, chemical, and health information	Many, including genomics, proteomics, Interactomics, and metabolomics	http://www.kegg.jp	Kanehisa and Goto [32]
Multi-Omics Profiling Expression Database (MOPED)	Omics expression database that contains over five million protein and gene expression records. It links to various protein and pathway databases, including GeneCards, Panther, Entrez, UniProt, KEGG, SEED, and Reactome. Protein identifiers are integrated from GeneCards, GI, RefSeq, Locus Tag, UniProt, WormBase, and SGD	Mainly proteomics and transcriptomics	http://moped.proteinspire.org (accessible only with valid certificate)	Kolker et al. [33], Higdon et al. [34]
Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	Database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations, derived from genomic context, high-throughput experiments, co-expression, and previous knowledge. STRING, currently covering about 10 million proteins from 2031 organisms, quantitatively integrates interaction data from these sources for a large number of organisms and transfers information between these organisms where applicable	Interactomics	http://string-db.org/	Snel et al. [35]
Human Protein Atlas (HPA)	Database of information for almost all human protein-coding genes. Data are available on expression and localization of proteins based on both RNA and protein data	Proteomics	http://www.proteinatlas.org	Uhlen et al. [36]
Human Metabolome Database (HMDB)	Database containing 41 993 metabolite entries, 5 701 protein sequences are linked to these metabolite entries. Each entry has around 110 data fields with 2/3 of the information on chemical/clinical data and the rest regarding enzymatic or biochemical data	Metabolomics	http://www.hmdb.ca	Wishart et al. [37], Wishart et al. [38], Wishart et al. [39]
MINT	The database contains experimentally verified protein–protein interactions mined from the scientific literature through human	Interactomics Bibliomics	http://mint.bio.uniroma2.it/mint/Welcome.do	Licata et al. [40]
IntAct	Database of molecular interaction data, here every interaction is derived from literature or submitted directly by the user	Interactomics	http://www.ebi.ac.uk/intact/	Orchard et al. [41]
BioGRID	Repository with interaction data manually curated. It contains 56 300 publications for 1 060 041 protein and genetic interactions,	Interactomics Regulomics	http://thebiogrid.org	Oughtred et al. [42]

information on metabolomics, a technology having a great potential to impact clinical practice (e.g., biomarkers for diagnosis, monitoring, and definition of new therapeutic targets). In this context, a current challenge is the definition of metabolite resources with comprehensive spectral libraries, various integrative approaches, and serious considerations for clinical validation of the identified biomarkers [43].

15.3 Integration of Different Data Sources

The definition and the population of a database is by far the most effective way to represent and organize a wide range of data; however, biological databases are affected by the lack of uniformity in types and formats of data sources, mainly due to the lack of a unique standard. Databases of pathways are an example of this problem. For example, some of the 547 biological pathways reported in Pathguide (http://www.pathguide.org) are similar and redundant but are defined with different boundaries and components. This heterogeneity has to be taken into account when genome analysis methods based on pathways are applied (e.g., in Refs. [44–47]). Indeed, the same input data can generate different results if different databases are used for the analysis [48]. To overcome these issues, Cantor and colleagues proposed to use multiple databases for each analysis [49] in order to balance divergences among databases and/or to validate similar results obtained from different data sources. Gomez-Cabrero and colleagues, while reviewing data integration in the -omics era, advocate the need to create standards at earlier stages when novel data-type resources are developed [50].

One of the first definitions of database integration has been formulated in the context of the smoothening of redundancies between databases, pointing out the need to access different databases with overlapping content and to connect several of them, as if the user were interacting with one single information system. In general, since 1980s, two main approaches have been defined to efficiently integrate data coming from different sources:

1) **Data warehousing:** Data warehousing consists in the storage of all the data belonging to a certain category from different databases in only one large database, according to the process named ETL (extract, transform, and load) [51].
2) **Federated approaches:** In contrast to data warehouse, in federated databases data remains in the original data source. The integration consists in mapping data from each source on the federated database; thus, the end user can operate on a simple database. This data storage approach is relevant when the researcher needs updated information or must integrate a large amount of data deriving both from private and public databases [52].

The second approach is currently trending in the domain of life sciences. Proving this fact, one of the most common ways to connect data from several biological databases consists in implementing hypertext links between entries of different data sources. As such, link integration approaches represent a connection between web pages, whereas the actual integration method is then carried on by a user or by another application. As a matter of fact, this approach requires a significant amount of manual work in order to integrate data: scientific institutions that maintain biological databases have to face a time-consuming checking process to map the links between entries from distinct data sources. Therefore, given the high number of databases pertaining to the biological field, only links to the most used ones are typically set. A meaningful example is provided by one of the main search engines for health science databases: Entrez Global Query Cross-Database Search System (http://www.ncbi.nlm.nih.gov/sites/gquery), a tool that is able to retrieve information stored in several sources and regarding biomolecular sequences, structures, and literature references.

15.4 Integration of Different -Omics Data

It is becoming more and more evident that the integration of multi-omics layers is required for a deeper understanding of complex biological entities. To meet this goal, initial attempts of data integration reported in literature analyzed data from individual -omics separately. This phase was then followed by downstream actual integration of previous independent and parallel analyses outcomes. However, this method entails the loss of key emergent properties, which only become apparent by analyzing multi-omics dataset as a whole and not by studying the system as the sum of its parts [53]. The first step in the integration of multi-omics layers is the joining of information deriving from their pairs. In the following paragraphs, we will illustrate examples of pairwise integration between different -omics data.

15.4.1 Integrating Transcriptomics and Proteomics

Several studies in model organisms have shown that mRNA and protein expression profiles are often poorly correlated [54–56]. Proteins are generally more stable than mRNAs [57], so situations where a protein is still abundant in the near absence of the cognate mRNA

may arise. The opposite situation (i.e., low protein level while the corresponding transcript is high) may derive from poor translation of the mRNA. This happens either because the RNA itself is poorly translatable (due, e.g., to secondary structures that hamper translation [58, 59]) or because of interaction with other molecules, such as the trans-acting factors, RNA-binding proteins (RBPs), and small RNAs that bind to the mRNA and modify its translatability [60]. Among these natural antisense transcripts that regulate gene expression [61] are small (19–22 nucleotide) non-protein-coding RNA molecules (microRNAs or miRNAs). miRNAs downregulate expression of their target mRNAs through specific base pairing that results in decreased translation of the mRNA or leads to mRNA degradation [62]. Although a large number of human miRNAs are reported to be implicated in several developmental and adult disease states (e.g., cancer), many of their mRNA targets and their impact on phenotypes remain unknown [63]. Recently, due to the advance of high-throughput and low-cost experimental methods, there has been a huge development of computational methods based on sequence complementarity between the miRNA and the mRNAs [64].

The utility—and possibly the necessity—of integrating mRNA, miRNA, and protein expression in order to obtain a more comprehensive view of the system under study has been recently pointed out [65]. A possible approach to integration involves the analysis of individual -omics layers separately, whose results are then merged and compared. By way of example, Com and colleagues in a study of gentamicin nephrotoxicity report that transcriptomics and proteomics data were complementary and that their integration provided a more comprehensive picture of the putative nephrotoxicity mechanism of the antibiotic, consistent with histopathological evidence [66]. Although valuable, this approach misses the interconnection between the different -omics layers and may fail to uncover the system-level functional properties. By mapping transcriptomics and proteomics datasets on the protein interaction network and using chronic kidney disease as an example, Perco and colleagues show that such a joint analysis highlights pathways and processes characteristic for the phenotype under analysis that goes unnoticed when the two datasets are analyzed independently [67]. A similar, network-based methodology for integrative analysis of proteomics and transcriptomics data on psoriasis showed complementarities between two levels of cellular organization and allowed to identify common regulators—such as the most influential transcription factors and receptors—for two datasets [68]. Imielinski and colleagues identified subnetworks enriched in differentially expressed genes within networks built from proteins differentially expressed in estrogen receptor positive breast cancer tumors [69] from which a gene expression-based signature biomarker predictive of clinical relapse could be constructed.

Liu and colleagues [53] focused on the melanoma subset from NCI-60—a panel of 60 different human cancer cell lines from 9 different tissues. They quantified the additional information provided by their method compared with non-joint approaches and observed that integration and annotation in the analysis of different type of data changed the flow of information, with the joint analysis giving fully relevant molecular information only upon annotation of all mRNA, proteins, and miRNA. Particularly, compared with the separate analysis, the joint analysis better described melanogenesis, while the separate analyses failed to identify enrichment in melanin biosynthetic and metabolic processes, both related to the basal melanocyte physiology. A similar algorithm (iCluster) has been used to cross-correlate gene copy number and transcriptional profiling to discover potentially novel cancer breast and lung cancer subtypes by combining weak, consistent alteration patterns across subtypes [70]:

A final approach worth mentioning exploits the Bayesian framework to infer gene regulatory network form transcriptomics, whose accuracy is extended by combining prior knowledge [71] or protein–protein interaction (PPI) data [72].

The reader is referred to Haider and Pal [73] for a recent comprehensive review detailing other methods for integrating transcriptomics and proteomics networks.

15.4.2 Integrating Transcriptomics and Interactomics

Analysis of genome-wide expression profiles recently allowed to identify several disease markers (e.g., [74]), exploiting the link between perturbations of a particular phenotype and changes in mRNA levels. In this kind of analysis, each gene is scored for the ability of its expression pattern to discriminate between various classes of disease; subsequently, marker sets are selected based on attributed scores (signature-based approach). However, different marker sets for a specific disease are found among different studies (e.g., [75, 76]), likely because changes in expression of the few selected genes may be small compared with those of the downstream effectors, which may vary significantly among patients [77]. As such, a better strategy to identify markers would be to combine gene expression measurements over groups of genes that fall within common pathways [78–80]. Nevertheless, pathway-based analysis (to which gene-set enrichment analysis (GSEA) belongs) has the limitation that there is still no assignment of most human genes to a specific pathway.

A partial solution to these challenges lays in the use of PPI networks (the interactome) that provide a comprehensive map of functional interactions in the cell and allow the identification of subnetworks (composed by a group of proteins functionally linked to each other) that are significantly dysregulated in a disease of interest. In this regard, the development of a scoring scheme to assess the collective dysregulation of multiple interacting genes (mapped on the corresponding protein on the PPI network) and the development of efficient computational algorithms to search for subnetworks with significant scores represent the main methodological challenges. Commonly, the differential expression of each gene is first scored individually using a standard statistical test (e.g., *t*-test), and then subnetwork scores are computed as an aggregation of these individual differential expression scores. However, these methods provide limited systems level insights, since they assess the differential expression of functionally related genes individually and cannot capture patterns of coordinated dysregulation. An alternative strategy has been proposed in Ref. [81]. Here, authors illustrate a representation where genes having consistent expression patterns are mapped on PPI networks to form subnetworks that are significantly dysregulated in a disease of interest and may be conserved across multiple species [82]. Chuang and colleagues [83] applied a protein-network-based approach to identify markers of metastasis within gene expression profiles, with the aim of detecting genetic alterations and predicting the probability of metastasis in unknown samples. They show that the network-based method has many advantages compared with earlier analyses of differential expression:

1) The generated subnetworks provided models of the molecular mechanisms underlying metastasis.
2) Though analysis of differential expression usually does not allow detecting genes with known breast cancer mutations (such as KRAS, among others), these genes play a key role in the protein network by interconnecting many expression-responsive genes.
3) Subnetworks are remarkably more reproducible among different breast cancer cohorts than separate marker genes selected without network information.
4) Accuracy in prediction is higher with network-based classification, as demonstrated by selecting markers from one dataset and applying them to a second independent dataset.

A further evolution of the method, called interactome–transcriptome integration (ITI), consists in the integration of the analysis of several gene expression datasets (multi-dataset) to extract subnetworks that discriminate breast cancer distant metastasis [84]. The method showed increased performance on a vast collection of publicly available data and was validated on two independent breast cancer gene expression datasets [85, 86].

For a further dissertation on the essential role for the comprehension of biological systems of the integration between transcriptomics and interactomics (as well as with other categories of -omics data, such as genomics and proteomics), we refer the interested reader to a recent review [87] that summarizes strengths and weaknesses of the different approaches.

15.4.3 Integrating Transcriptomics and Metabolic Pathways

In order to better understand the role of differentially transcribed metabolic genes in the context of metabolic pathways, Patil and Nielsen [88] devised a technique of network enrichment whose goal is to identify the "reporter metabolites," that is, those spots in the metabolism where there is a crucial regulation to maintain homeostasis (i.e., a constant level of the metabolite) or to reset the concentration of the metabolite to a different level required for the correct functioning of the metabolic network. The first step of the procedure consists of mapping differential expression data on the corresponding enzymes of a genome-wide biochemical network (whose reconstruction process is described in the following section), adding a specification of the significance of differential gene expression. In this way, each metabolite node is scored based on the normalized transcriptional response of its neighboring enzymes. When dealing with differential data, the normalized transcriptional response is calculated as size-independent aggregated *Z*-scores of the neighboring enzymes. The scoring used to identify reporter metabolites is a test for the null hypothesis hereafter formulated: "Neighbor enzymes of a metabolite in the metabolic graph show the observed normalized transcriptional response by chance." Metabolites with the highest score are defined as reporter metabolites, that is, those metabolites around which transcriptional changes occur.

All in all, advantages of performing a multidimensional -omics analysis to obtain more information from human high throughput data instead of analyzing a single -omics data type can be summarized as follows:

1) Integration of multiple data types is a strategy to prevent information loss due to the fact that information on a biological entity (gene, protein, transcript, etc.) can suggest to refine other -omics analyses in order to fill information gaps or to correct wrong data associations.
2) Different data sources providing information on the same gene or pathway are less likely to produce "false positives."

3) Examination of different levels of regulation by means of an integrated approach is a promising way to unravel the functioning and fine regulation of the biological system under examination.

Readers interested in mathematical aspects of methods for multi-omics data integration may refer to a recent review [89]. The availability of dedicated servers for the analysis of multi-omics datasets, including transcriptomics, miRNomics, proteomics, and genomics may help to spot similarities and differences between the enrichments obtained from different -omics and widen the use of integrative multi-omics analyses [90].

As highlighted in "Introduction," many common diseases such as cancer, diabetes, and cardiomyopathy should be considered as network diseases. If the complexity of the network is not taken into account, we may fail in identifying a potential drug having high efficacy and low toxicity [91–94]. One of the main reasons of such a poor predictive power is that the exploitation of individual -omics platform does not provide enough information to link drug response with personalized -omics profile. Indeed, a stronger integration of different -omics platforms could validate data and help in clarifying the connections between -omics, as well as accelerate multi-target drug discovery [95]. Cellular subsystems have been defined by ontologies, such as Gene Ontology (GO). It has been proposed that such a hierarchical structure may guide the organization of -omics data. Interpretation of this "ontotype" through logical rules generated by machine learning techniques allowed predictions of the growth properties of over 2000 yeast strains carrying inactivation of two genes and could pave the way for interpretation of the phenotypic properties of complex diseases [96]. According to similar reasoning, an initiative to define the hallmark networks of cancer has been recently launched [97].

As we will see later, a further step, modeling, may be required to fully extract information hidden in -omics data structure according to mechanistic principles and generate experimentally testable predictions.

15.5 Visualization of Integrated -Omics Data

Consistently with the need for -omics data integration approaches, a strong need for tools able to represent them in an effective way has emerged among scientists and clinicians belonging to different communities. To satisfy this need, visual representations of -omics data have been extensively used to give an immediate representation of the complexity beyond the systems, to sum up relevant information [98], and to help to formulate hypotheses on represented systems.

The recourse to visualization strategies has been motivated by the fact that the human brain has a remarkable capability to process visual information in order to identify patterns (e.g., biochemical pathways) and relevant topological features (e.g., the presence of highly connected nodes called "hubs") [99]. The "visual complexity" of these representations ranges over various orders of magnitude spanning from the description of a small functional units (signal transduction and metabolic pathways, interaction pool of a protein), to the representation, at whole cell/tissue/organism level, of the interactions involving different -omics data.

The development of high-throughput -omics technologies has imposed a change of paradigm for the definition of these representations, shifting from the manual curation and refinement to fully (or partially) automatized procedures exploiting sophisticated software. Even if recent efforts have produced remarkable results (see Ref. [100] for an extensive review and Table 15.2 for a non exhaustive list of relevant and widely used tools for data visualization), in this domain some challenges are still open.

A first challenge is related to the scalability of the methods. Scalability issues are particularly relevant in network representations, an obvious and traditional way to visualize data and their relationships (generally nodes represent entities and edges represent relationships). This type of representation is intuitive and powerful for simple systems, but has also some scalability limitations: when the system size and complexity increases, also the "visual complexity" increases, and since most of the software make use of standard visualization packages, the most common layout of the network is often a very uninformative "hairball" (Figure 15.2a) [98].

The wide usage of this primitive layout is mainly due not only to the lack of knowledge on the inner organization of the network (e.g., cellular localization of elements, molecular functions, structure of protein complexes, etc.) but also to the difficulty of representing the system in a way expected by the user (e.g., the arrangement of metabolic or signaling pathways using an immediately recognizable shape). To move from the uninformative "hairball layout" to a more meaningful representation, several algorithms have been devised to visually organize the network, on the basis of given criteria, such as node degree distribution, geometrical representations (e.g., circles, grids), directionality of the process (hierarchical representations), and physical simulations, modifying both the spatial layout and placing information (i.e., -omics data) on network elements (color/dimension/shape of nodes and edges). In Figure 15.2, we provide an example of useful -omics

Table 15.2 Visualization tools focused on interaction networks.

Visualization tool	Main features	Functions and compatibility	Advantages	Web page	References
Arena 3D	Standalone free application Allows visualizing biological multilayer networks in 3D			http://www.arena3d.org/	Pavlopoulos et al. [101]
BioLayout Express3D	Layout, visualization, and clustering of large-scale networks in both 3D and 2D. Supports both unweighted and weighted graphs. Uses a graphic render, so that the size of networks that can be processed is limited	Highly interactive: it is possible to switch between 2D and 3D representations, zoom in/out, rotate or move the network. Markov Clustering algorithm is incorporated and data are automatically separated in distinct groups. Compatible with Cytoscape	Offers different analytical approaches to microarray data analysis	http://www.biolayout.org	Freeman et al. [102]
CellDesigner	Structured diagram editor for drawing gene-regulatory and biochemical networks	Visual representation of biochemical semantics, direct integration with SBML ODE Solver and Copasi, and linkage to SBW-powered simulator modules	Intuitive user interface helps to draw a diagram with the standard SBGN notation	http://www.celldesigner.org	Funahashi et al. [103]
Cytoscape	Stand-alone Java application. Provides 2D representations of large-scale networks (up to hundredth thousands of nodes and edges). Supports directed, undirected, and weighted graphs and has powerful visual styles	Highly interactive: possible zoom in and out and browsing of the network; organization of multiple networks and possibility to compare them; allows to select subsets of nodes/interactions and search for active subnetworks/pathway modules; incorporates statistical analysis of the network. Compatible with other tools. Allows to import mRNA expression profiles, gene functional annotations from GO, and KEGG	Visualization of molecular interaction networks and their integration with gene expression profiles and other data. Allows the manipulation and comparison of multiple networks. Many plug-ins are available for more specialized analysis	http://www.cytoscape.org	Shannon et al. [104]
E-cell 3D	Software platform to model, simulate, and analyze complex, heterogeneous, and multi-scale biochemical reaction systems	E-Cell 3D exploits the advanced graphics APIs of MacOS X; however this is the only supported operative system. 3D networks can be navigated using Nintendo Wii remote controller	Models stored in Systems Biology Markup Language (SBML) XML file can be directly converted to E-Cell 3D	http://ecell3d.iab.keio.ac.jp/index.html	Tomita et al. [105]
iPath	Interactive Pathways Explorer (iPath) is a web-based tool for the visualization, analysis, and customization of the various pathways maps	KEGG-based overview maps	Extensive map customization and data mapping capabilities. All maps in iPath can be easily converted to various bitmap and vector graphical formats for easy inclusion in documents or further processing	http://pathways.embl.de	Yamada et al. [106]
MapMan	A user-driven tool that displays large datasets onto diagrams of metabolic pathways or other processes	Based on Java and hence cross platform		http://mapman.gabipd.org/	Thimm et al. [107]

Medusa	Open-source Java application. Provides 2D representation of networks up to a few hundred nodes and edges. Uses nondirected, multi-edge connections, allowing the simultaneous representation of more than one connection between two bioentities	Highly interactive: allows the selection and analysis of subsets of nodes. A text search can be applied to find nodes. Medusa has its own text file format not compatible with other visualization tools or integrated with other data sources	Shows multi-edge connections, each line representing different concepts of information. It is optimized for PPI data as taken from STRING	https://sites.google.com/site/medusa3visualization/	Hooper and Bork [108]
Ondex	Stand-alone freely available open-source application. Provides 2D representations of directed, undirected, and weighted networks. Handles large-scale networks of hundred thousands of nodes and edges and supports bidirectional connections. Different types of data are separating in different disks–circles interconnected with each other	Various filters allow to selectively add or remove connected nodes from the display. A tree-like subgraph can be extracted from a given node and the most important nodes at any level can be determined. A filter is available to import microarray expression level data. Data may be imported through many databases, among which are TRANSFAC, Gene Ontology, and KEGG	Ability to combine heterogeneous data types into one network. Suitable for text mining, sequence, and data integration analysis	http://www.ondex.org/	Koehler et al. [109], Kohler et al. [110], Skusa et al. [111]
Osprey	Stand-alone application running under a wide range of platforms. Provides 2D representations of directed, undirected, and weighted networks. Not efficient for large-scale network analysis but provides various layout options and ways to arrange nodes in different geometric distributions	Provides several features for functional assessment and comparative analysis of different networks together with network and connectivity filters and dataset superimposing. Also allows to cluster genes by GO processes. Data can be loaded either by using different text formats or by connecting directly to several databases	Various filtering capabilities render Osprey a powerful tool for network manipulation. The key feature is the ability to incorporate new interactions into an already existing network	http://tinyurl.com/osprey1/	Breitkreutz et al. [112]
Pajek	Stand-alone Windows application. Offers 2D and pseudo3D representations and supports single, directed, and weighted graphs. Suitable for large-scale networks with thousands or millions of nodes and vertices. Great variety of layout options. Separates data into layers, allowing the display of hierarchical relationships. Can handle dynamic graphs and reveal how networks change over time	Highly interactive, many clustering methods. Allows decomposition of a large network into several smaller networks and detection of clusters in them It has its own input file format, not compatible with commonly used formats; not connected with any biological data sources	Variety of layout algorithms facilitating exploration and pattern identification within networks	http://pajek.imfm.si/	Batagelj and Mrvar [113]
PathVisio	Pathway analysis and drawing software to draw edit and analyze biological pathways. Experimental data can be easily visualized on pathways and relevant pathways that are over-represented in a dataset can be easily found	Provides a basic set of features for pathway drawing, analysis, and visualization. Additional features are available as plug-ins	Plug-ins extended functionalities and can also be customized for an advanced use	http://www.pathvisio.org	Kutmon et al. [114]

(*Continued*)

Table 15.2 (Continued)

Visualization tool	Main features	Functions and compatibility	Advantages	Web page	References
Pathway Tools	A tool to guide the user through creation, editing, querying, visualization, and analysis of Pathway Genome Databases	Wide diffusion in different research communities	Pathway Tools -omics viewers allow -omics datasets to be graphically painted onto three system-level diagrams: a diagram of the full metabolic network of the organism, a diagram of the full regulatory network of the organism, and a diagram of the full genome of the organism It can also depict data from multi-omics data types simultaneously, such as mixing gene-expression and metabolomics data in one diagram	http://bioinformatics.ai.sri.com	Karp et al. [115]
PIVOT	Java application free for academics. Projects in 2D and uses single non directed lines to show relationships between bioentities. No limits in the size of data presented	Allows the expansion of the network, to highlight dense areas of the map, to visualize a subarea of a big network. Many features help to navigate and interpret the interaction map and to connect remote proteins to the displayed map through graph-theory algorithms. Configured to work with proteins from human, yeast, *Drosophila*, and mouse links to external web information pages	Best suited for visualizing PPIs and identifying relationships between them	http://acgt.cs.tau.ac.il/pivot/	Orlev et al. [116]
ProMeTra	An open-source framework that provides visualization methods for multi-omics datasets	The integration of genomics and transcriptomics datasets originating from different services	Format SVG is used to facilitate the visualization of the results of complex functional genomics experiments	http://fusion.cebitec.uni-bielefeld.de	Neuweger et al. [117]
Tulip	Stand-alone free application. Allows generic visualization of extremely large networks and supports 3D visualization			http://tulip.labri.fr/TulipDrupal/	Auber [118]
VANTED	Stand-alone free application. Supports combined visualization of abundance data, networks, and pathways			https://immersive-analytics.infotech.monash.edu/vanted/	Junker et al. [119]

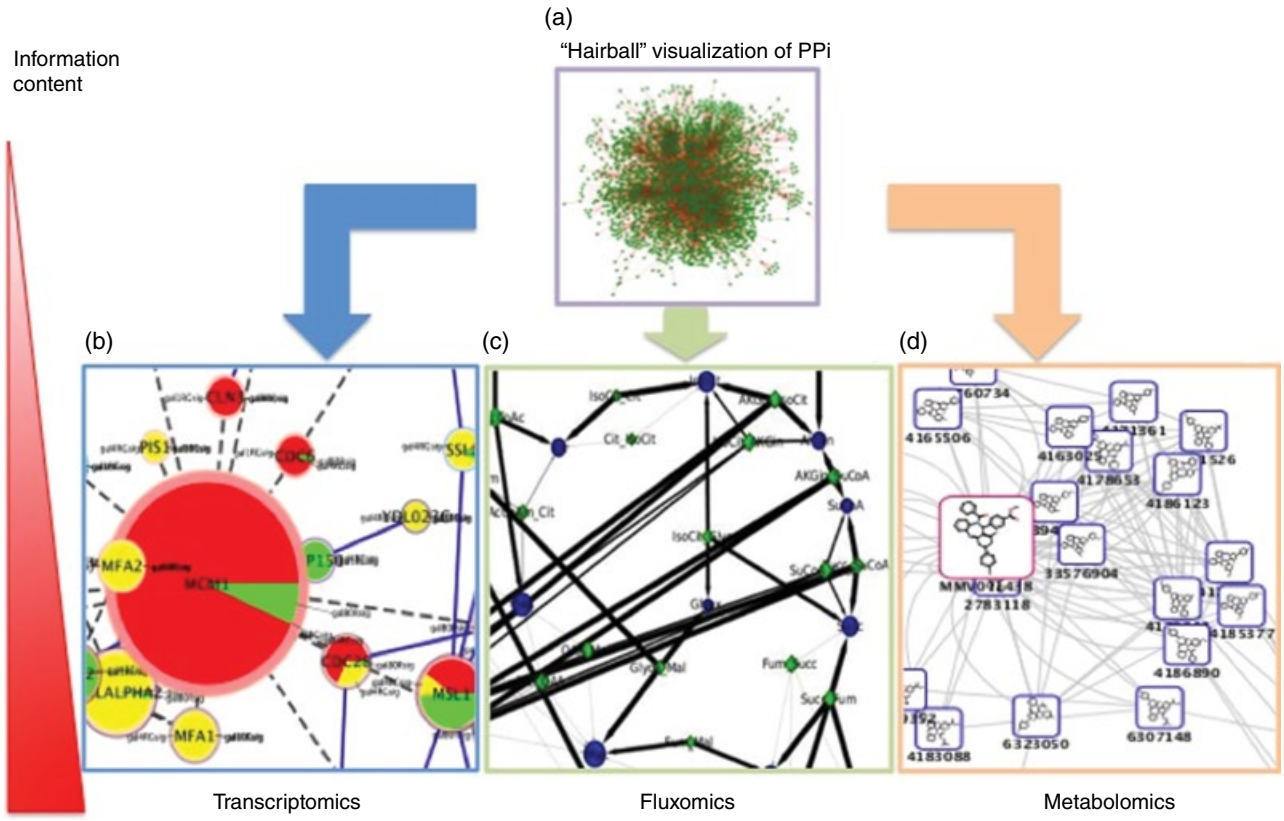


Figure 15.2 From uninformative to meaningful -omics data visualization with Cytoscape. Adding data on the network in a rational way improves the information content of the representation. On the top part, an example of hairball layout obtained using data from high-confidence protein–protein interactions [120] (a). On the lower level, examples of -omics data mapped on a network modifying its elements (nodes and edges color/dimension, layout) exploiting Cytoscape apps (Node Chart Plug-in for transcriptomics (b), CyFluxViz for fluxomics, (c) and chemViz2 for metabolomics (d)).

data visualization: in box b, transcriptomics data have been mapped on nodes (genes) using a color code for expression level (red for upregulation, green for downregulation, yellow for no change), while slice size in the pie chart is proportional to the number of experiments where the gene has the same expression pattern (up-/downregulation or no change). In the same graph, node size is proportional to the node degree (i.e., nodes having a large number of connections—hubs—have a larger size). Edges in box b indicate interactions between proteins encoded by genes represented in nodes, dashed edges indicating a low confidence interaction. In box c, the layout of a metabolic network has been manually defined accordingly to a commonly used representation (in the panel a portion of the tricarboxylic acid (TCA) cycle). Metabolites are represented with blue nodes having sizes proportional to the node degree, while reaction nodes are marked with green diamond nodes. Edge thickness is proportional to the value of the flux through a given reaction. In box d, another metabolic network has been represented using a default layout. However metabolite nodes have been represented using boxes inside of which structural formulas are shown; the abundance of every metabolite can be represented here coloring the border of the box accordingly to a color gradient.

A promising way to face visual complexity emerging from -omics size networks is represented by clustering approaches (e.g., MCODE) that are integrated with network visualization tools and used, for example, to predict higher-order protein complexes from the interaction data. Network clustering is a new kind of clustering method that is performed using correlation networks, in which each node is a gene and each edge indicates co-expression of two genes under a given experimental condition (Figure 15.3). Available tools include BioLayout Express 3D and Cytoscape.

A challenge is connected to the retrieval of desired information and to the network navigation for the exploration of the "surroundings" of a given element, an activity that could generate insights to direct the investigation of the system.

Another challenge can be identified in the enrichment of the visualization by adding further information (e.g.,

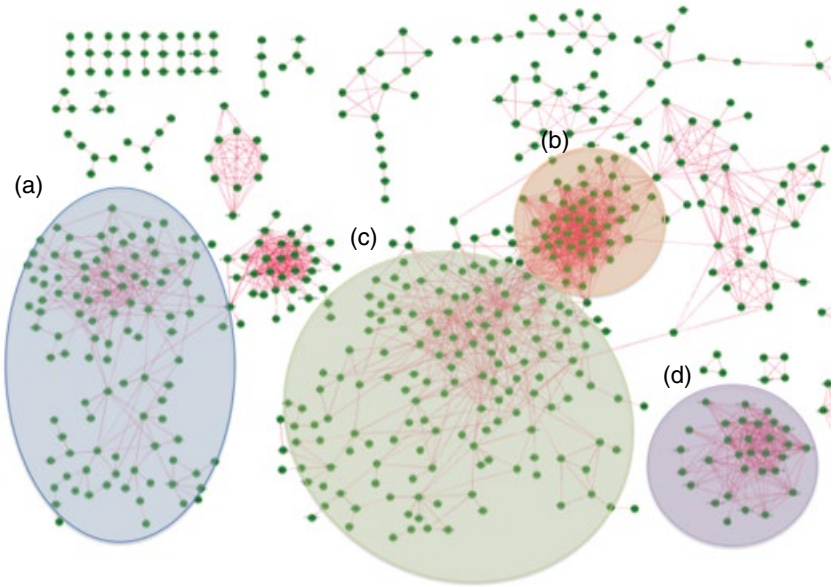


Figure 15.3 Network showing a clustering of co-expression data using Cytoscape. *Source:* Dataset from Prieto et al. [121]. In the network, genes are mapped as nodes, while edges represent the co-expression relation. Identified clusters (isolated groups of nodes highlighted with ellipses) represent different cellular functions, for example, A, mitochondrial metabolism; B, ribosome; C, nuclear related metabolism; and D, immune response.

attributes from external sources and database) while maintaining a good readability of the relevant information. In this context, when network enrichment is used to find pathways or networks where genes are significantly over-represented, a valid aid for the interpretation of the results of the analysis is the superimposition on the reference map of the metabolite concentrations or significance levels toward a certain metric by means of dedicated tools, such as MapMan, Pathway Tools Omics Viewer (Figure 15.4), and ProMeTra.

Lastly, future perspectives on -omics visualization through networks representations include (i) the exploration of three-dimensional layouts, that is, multiple networks (representing each one a homogeneous type of data) linked among them to provide a more complete understanding of the system (e.g., BioLayout Express 3D); (ii) combinations of both three-dimensional layouts and temporal descriptions (e.g., E-Cell 3D); and (iii) layouts that mix aspects of classic and three-dimensional representation (e.g., Arena3D).

Besides network representations, visualization of -omics data can be performed through complementary approaches that aim at aggregating information and reduce visual complexity. In particular in the context of transcriptomics (expression profiles), many tools implement scatter plots combined with dimensionality reduction, profile plots, heat maps, dendrograms, and clustering.

An interesting cloud-based, community-driven resource (GenomeSpace, http://www.genomespace.org) has been recently presented [122]. Through the implementation of workflows, GenomeSpace aims to make the use of integrative analysis accessible to non-programmers.

15.6 Integration of -Omics Data into Models

The statistical and machine learning approaches to data integration illustrated earlier provide a first attempt to identify the biochemical pathways perturbed in different patients and statistical correlations with drug responses. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity, but fails to deliver a system-level understanding of the molecular mechanisms behind the emergence of different phenotypes [123]. On the other hand, network biology approaches that apply topological and graph theory concepts to biological networks statistically inferred from -omics data or reconstructed according to *a priori* knowledge are important for understanding the structure and properties of the integrated cellular network and its modular structure, with the ultimate aim of understanding its organizational principles [124]. Enrichment of network modules, which integrate -omics data with known or predefined molecular scaffolds, allows the identification of the portions of the network that are most active under a given condition [125]. Nevertheless, by relying on a static conceptualization of the network while ignoring its integrated dynamics in state space, also these methods fail to provide a mechanistic understanding of the disease, which would be desirable to reliably predict individual drug response.

The mechanistic understanding of a system requires the integration of these data under mathematical and relational models that can describe dynamically the relationships between their components. Two main computational

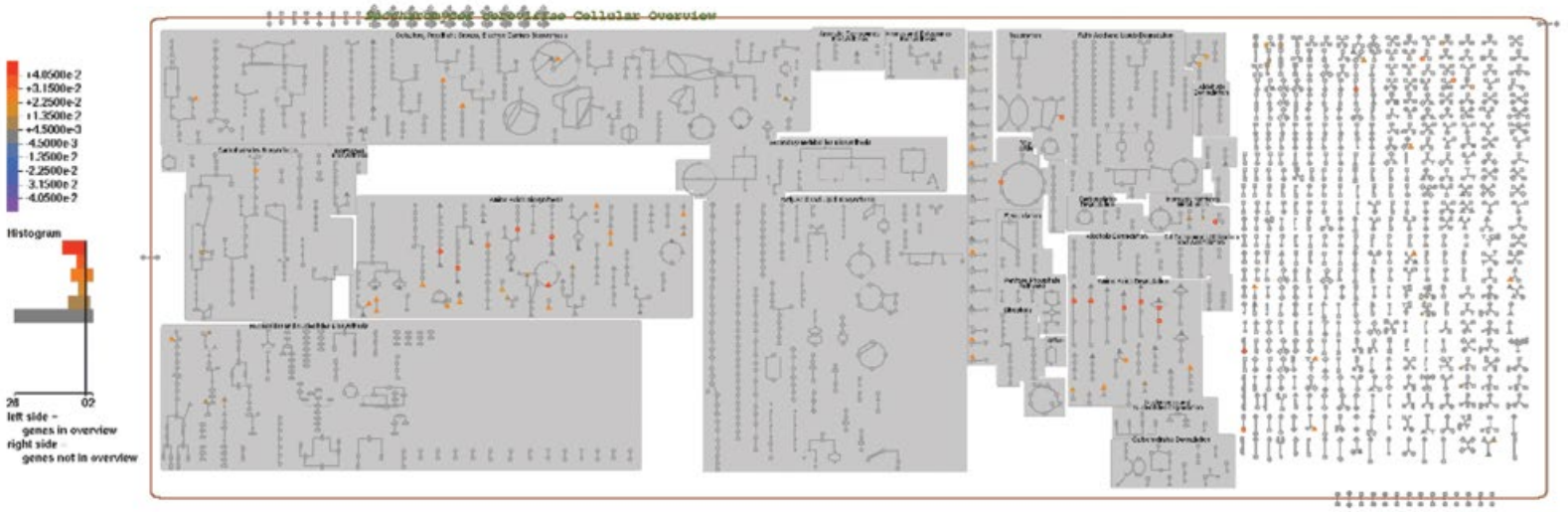


Figure 15.4 A cellular overview of the metabolic network of yeast generated with Pathway Tools. In this visualization, metabolic reactions are subdivided in pathways (grey boxes), exchange reactions are placed across the border of a rectangle representing the cellular membrane, while reactions not assigned to a specific pathway are grouped on the right side. The metabolic map has been enriched with data (colored nodes) on reporter metabolites (see section "Integrating Transcriptomics and Metabolic Pathways" for a reference on reporter metabolites), the color of the node is set accordingly to the confidence value for the reporter metabolite identification (top left color bar), the distribution of nodes having a given confidence value is shown on a histogram in the bottom left corner.

frameworks allow to predict the phenotype that emerges from a given biological network structure: kinetic modeling and constraint-based modeling.

Kinetic modeling allows estimating the evolution in time of the concentration of each network component in a reacting system (such as metabolites, transcripts, or proteins). The transition from one state of the network to the following one is determined by the interaction with the other network components and by rate law equations. The traditional way of modeling the time evolution of the molecular populations in a reacting system is to use ordinary differential equations (ODEs). However, when more appropriate, an approach that considers stochastic fluctuations can be applied.

An example of data integration into kinetic modeling is provided by the kinetic model of glycolysis in yeast [126] and *Plasmodium falciparum* [127]. Each glycolytic enzyme was kinetically characterized and the parameters of kinetic equations (e.g., Michaelis–Menten) were chosen to best fit the experimental kinetic data; the resulting rate laws incorporated into the model. The difficulty in obtaining kinetic parameters and their appropriateness for *in vivo* situations makes it difficult if not impossible to scale up traditional kinetic models to large (genome-wide) networks. A recent paper [128] integrated metabolic profiling data obtained from the plasma of different patients into a whole-cell, metabolic kinetic model of a red blood cell (RBC) (that includes 55 transport and 87 intracellular reactions). The models allowed to identify individuals at risk for a drug side effect and protective genetic variations, proving the feasibility and usefulness of "personalized" kinetic models, whose use will accelerate discoveries in characterizing individual metabolic variation. Still, routine integration of -omics data into kinetic modeling remains a problematic task that is awaiting improved methodologies.

On the contrary, constraint-based modeling is a framework well suited for metabolic network modeling and multi-omics data integration, which is capable of providing a deeper understanding of metabolic functions than data alone [129]. Constraint-based modeling relies on the idea of excluding phenotypes that do not abide by the imposed constraints, iteratively restricting the space of possible phenotypes until getting the most plausible one(s) [130]. Fundamental assumption of this kind of techniques is a pseudo-steady state for internal metabolites concentrations. As compared with kinetic modeling, constraint-based modeling has the substantial advantage of not requiring any knowledge on kinetic parameters governing reaction rates. Recent developments of constraint-based models account for gene expression reconstructions that use approximate stoichiometric relationships between the level of enzymes and their cognate catalyzed fluxes to compute feasible, optimal, and spatially resolved states describing the cellular composition at the molecular level [131].

15.6.1 Multi-Omics Data Integration into Genome-Scale Constraint-Based Models

The starting point for multi-omics data integration, within the constraint-based framework, is the description of the entire metabolism of a given organism as a network. This goal is attainable, thanks to the increased access to genome sequencing and annotation techniques. Moving from functional gene annotation, a metabolic reaction can be associated with each metabolic gene, that is, the reaction catalyzed by the corresponding enzyme. Once the identified reactions are grouped by metabolite, a genome-wide metabolic network is obtained. According to this paradigm, several genome-wide reconstructions are today available for different organisms, from microorganisms to human. An example is provided by Recon2 [132] and HMR [133], which encompass virtually all the reactions that in principle can occur in human metabolism and are therefore considered as *generic* reconstructions. Genome-wide generic reconstructions can be customized on specific cell types or tissues (or even patients) by exploiting several kinds of -omics data and appropriate algorithms (for a review see Ref. [134]). The so-obtained *specific* network represents the subnetwork that is known to be active in a given cell, according to its transcriptome, proteome, metabolome, and fluxome [135].

The family of -omics data that can most naturally be incorporated into genome-wide networks is fluxomics data. Constraint-based models allow indeed to specify the boundaries for the flux allowed for a given reaction. Constraints on nutrient intake and secretion fluxes (exchange reactions in the constraint-based terminology) are determinants in reducing the space of possible phenotypes.

The incorporation of transcriptome data to further constrain the flux distribution solution space is less straightforward. The main approaches are (i) the switch approach (e.g., GIMME and iMAT), using on/off reaction fluxes based on threshold expression levels, and (ii) the valve approach (e.g., E-Flux and PROM) to regulate reaction fluxes according to relative gene/protein expressions [136]. See Ref. [137] for a recent review.

Paradoxically, metabolite profiling may be the kind of -omics data that can most difficultly be integrated into genome-wide models (as reviewed in Ref. [138]). However, several algorithms, such as the INIT and tINIT [135, 139], have been proposed as an effective strategy to extract the portions of a generic GW model that is active in a given tissue or cell type, according to heterogeneous biological evidence, including metabolome. Based on

proteome, or on transcriptome when the former is not available, INIT assigns weights to the reactions in the HMR according to their different levels of evidence in the specific tissue or cell type. A unitary weight is also assigned to demand reactions (reactions that remove metabolites from the network) according to detected metabolites to impose the capability to accumulate a set of metabolites. An optimization process is then performed with the aim of maximizing as much as possible the reactions fluxes with a high weight (since the corresponding enzymes have a high expression level) while minimizing the others. Reactions that carry flux in the obtained optimal flux distribution are assigned to the tissue- or cell-specific model.

A more complex approach for integrating quantitative proteomics and metabolomics data with genome-scale metabolic network models, called integrative -omics-metabolic analysis (IOMA), was also proposed [140], which requires a mechanistic model of reaction rates. To evaluate the predictive performance of IOMA, the authors applied it to predict metabolic flux for RBC for which a detailed kinetic model is available. Remarkably, they demonstrated the advantages in the use of both proteomics and metabolomics to infer metabolic flux, as compared with inputting only one of the sources.

Once an active network is obtained according to the different integration algorithms, flux balance analysis (FBA) is then typically applied to determine the flux distribution that maximizes or minimize a specified objective.

15.7 Data Integration and Human Health

The integration of different -omics data, without the aid of computational models, has allowed identifying biomarkers of different human diseases. We focus here on the next step: how integration of data into models may improve system-level understanding of human diseases and, in perspective, may help in defining novel drug targets and better therapeutic regimens.

15.7.1 Applications to Metabolic Diseases

Genome-wide metabolic networks find their natural application in the study of metabolic diseases. In the simplest case, inborn error of metabolism (IEM) can be easily simulated by "deleting" the reaction catalyzed by the enzyme coded by the defecting gene. Metabolic biomarkers can then be predicted by monitoring the change in their feasible exchange flux [141]. Indeed, Recon 2 predicted 54 reported biomarkers for 49 different IEMs, with an accuracy of 77% [132]. However, metabolic network modeling has also been successfully applied to the investigation of more complex metabolic diseases, such as diabetes. As an example, the integration of transcriptome data and metabolic pathways, through pathways enrichment analysis, has supported the identification of reporter metabolites that allow to distinguish nonalcoholic fatty liver disease from healthy patients [142].

Varemo and colleagues [143] elucidated metabolic alterations in skeletal myocytes associated with type 2 diabetes at a system level, by generating cell-type-specific RNA-sequencing (RNA-seq) data for human myocytes and studying the correlation of this data with proteome data for myocytes from the Human Protein Atlas. Then, the authors constructed a comprehensive myocyte genome-wide model using these data and mapped transcriptional changes related to type 2 diabetes on the myocyte genome-wide model. An extensive transcriptional regulation in type 2 diabetes emerged, particularly around pyruvate oxidation, branched-chain amino acid catabolism and tetrahydrofolate metabolism, connected through the downregulated dihydrolipoamide dehydrogenase.

Jozefczuk and colleagues [144] analyzed network features of hepatic steatosis, another common metabolic disease. The authors generated gene-set enrichment and over-representation analysis through the pathway database integration system ConsensusPathDB. Network analysis of expression data of steatosis samples versus control revealed several pathways and functional modules of the disease, on which a first model prototype of steatosis related processes was developed. The prototype model included a minimal network, comprising a regulatory network (based on the transcription factor SREBF1) linked to a metabolic network of glycerolipid and fatty acid biosynthesis (including the downstream transcriptional targets of SREBF1). As the glutathione pathway was among the pathways enriched in steatosis versus control, the authors mapped mRNA expression data to a kinetic model of the glutathione synthesis pathway, focusing on a subset of complete pathways, rather than all genes of the genome. Then, Jozefczuk and co-authors extended this approach to other pathways important for liver regulation and functioning, such as fatty acid biosynthesis, fatty acid metabolism, bile acid pathway, gluconeogenesis, urea cycle, glycolysis, TCA cycle, and glyoxylate shunt. An object-oriented, comprehensive, multi-pathway, and multi-tissue *in silico* platform to investigate hepatic metabolism and its associated deregulations has been constructed. The SteatoNet model's ability to effectively describe biological behavior has been proven by its ability to identify metabolic flux

alterations previously identified experimentally in liver patients and animal models [145].

15.7.2 Applications to Cancer Research

Besides metabolic diseases, modeling of metabolic networks finds a large application in cancer research, where alterations in metabolism have been identified as a major hallmark of cancer [5, 134, 146]. FBA—typically exploited to predict physiologically relevant growth rates or the rate of metabolite production as a function of the underlying biochemical networks [147, 148]—is particularly useful to investigate the metabolic reprogramming performed by cancer cells [149]. FBA allows to identify, given a specified nutrient availability, the distribution of metabolic flux across the various pathways that maximize growth. In fact, enhanced growth indistinctly characterizes cancer cells and can be regarded as their "purpose." With this aim, the Human Metabolic Atlas offers a collection of tissue-specific reconstructions for both health and tumor tissues, obtained with the INIT algorithm, starting from the generic human reconstruction HMR and from -omics data in public databases such as the Human Protein Atlas [36].

The Human Metabolic Atlas also includes functional personalized GEMs for six hepatocellular carcinoma (HCC) patients [139]. Agren et al. [139] identified strong differences among the six HCC patients and simulated the effect of potential antimetabolites, by blocking the reactions that the corresponding metabolite engages in. They identified potential antimetabolites with antiproliferative or cytotoxic effect against HCC tumors for all six patients. Among these potential antimetabolites, they experimentally evaluated the effect of an L-carnitine analog on HepG2 cell proliferation, confirming their genome-scale modeling predictions.

In 2015, Asgari and colleagues [150] used the human metabolic model Recon1 as a scaffold to reconstruct tissue-specific models with the E-Flux method, which maps gene expression data into a genome-wide model by constraining the maximum possible flux through the reactions. Then, through FBA, the authors computed the reaction fluxes between normal and corresponding cancer cells in their subsystems. They found that the distribution of increased and decreased metabolic fluxes was unrelated to the significantly up- and downregulated metabolic genes of the associated cancer. Thus, they demonstrated that expression pattern of all metabolic genes (and not just significant up- and downregulated ones) plays a key role in metabolic rewiring of cancer cells. Consistently, rather than differential expression of specific genes, 7 subsystems (out of 13 common to all considered cancer cells) appear to be responsible for the Warburg effect: glutamine metabolism, nucleotides, glycolysis, oxidative phosphorylation, pentose phosphate pathway, TCA cycle, and pyruvate metabolism. Therefore, the Warburg effect appears to be a consequence of metabolic adaptation.

GEMs indeed represent a valuable tool to investigate the rationale behind metabolic events associated with cancer like the Warburg effect [151]. In this regard, Shlomi and colleagues inserted a solvent capacity constraint to the genome-scale human metabolic reconstruction Recon1 [152] to show how aerobic glycolysis emerges as the best strategy for growth. Along similar lines, [153] used a reduced flux balance model of ATP production constrained by the glucose uptake capacity and by the solvent capacity of cell's cytoplasm to demonstrate that the Warburg effect is a favorable catabolic state for rapidly proliferating cells with high glucose uptake capacity.

When studying metabolic plasticity and the ability of cells to adapt to changing environmental conditions, "core models" may be a valuable alternative to GEMs by allowing to highlight the more relevant properties of the network [154]. Di Filippo and colleagues extracted and manually curated, from the corresponding GEMs in the Human Metabolic Atlas, specific constraint-based core models for liver, breast, and lung tumors. A core model reconstructed starting from the original general human metabolic network was used as a reference. The three tumor models showed common metabolic properties reported in different kind of tumors: downregulation of respiratory chain, enhanced glycolytic flux, and stimulated utilization of glutamine via reductive carboxylation. Metabolic flux distribution among the three tumors was significantly different. Reactions that were present in the reference model, but absent in the tumors models, were isolated. Their insertion into the cancer models resulted in a less cancerous phenotype, and vice versa, their deletion from the generic models lead to a more cancerous phenotype. A group of reactions in particular was identified to be critically responsible for the reversion of tumor models toward less cancerous phenotypes [155]. The group includes the transport of phosphates from cytosol and mitochondrion, whose role for the correct functioning of the respiratory chain was indeed demonstrated in literature [156]. The metabolic advantages provided by particular metabolic events as compared with alternative phenotypes may also be investigated by comparing ensembles of flux distributions consistent with alternative strategies [157].

The workflow of constraint-based data integration approaches to cancer is schematically illustrated in Figure 15.5.

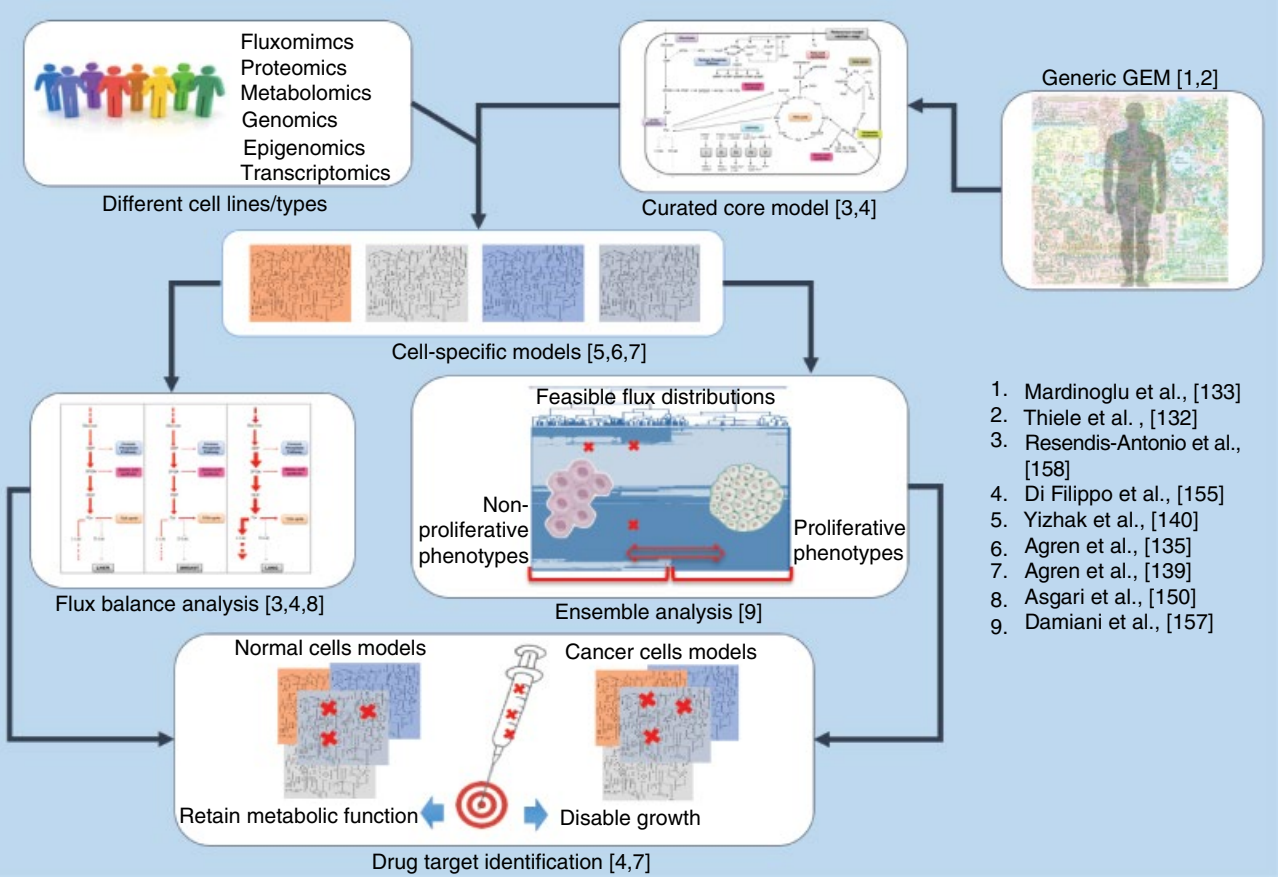


Figure 15.5 Workflow of constraint-based data integration approaches to cancer. Workflow from the extraction and curation of a model from the generic human metabolic map (top-right box), to its customization according to -omics data (following the arrows downstream of top boxes), to its flux balance analysis to estimate fluxes and other approaches (such as the ensemble approach) to explore the space of possible phenotypes, and to the simulation of possible drug targets (bottom box). For each process some references are reported as an example.

15.8 Conclusions

Integration of different -omics technologies allows to better extract hidden information in each dataset, allowing an unprecedented precision in the definition of the molecular phenotype of patient-derived samples. Correlation of these high-resolution molecular phenotypes to the clinical outcome provides precious indications for the development of novel stratification procedures to be used in the choice of the more appropriate therapeutic regimen. Integration of (multi)-omics data into mathematical models of diseased networks—notably metabolic networks—allows *ex post* examination of patients data collections. These personalized models provided the proof of principle of their ability to identify fragility points and to design appropriate personalized therapeutic regimens. As technical improvements and reductions in cost make it easier and easier to collect -omics data and more powerful and efficient computational methods are devised, it will be possible to apply this workflow in real time and use it as a guide in the design of a patient's personalized therapy, enabling the customization of medical care to the specific phenotype of each patient rather than providing a single, conventional treatment.

References

1 Langley, S. R., Dwyer, J., Drozdov, I., Yin, X. & Mayr, M. 2013. Proteomics: from single molecules to biological pathways. *Cardiovascular research*, 97, 612–22.

2 Bhalla, U. S. & Iyengar, R. 1999. Emergent properties of networks of biological signaling pathways. *Science*, 283, 381–7.

3 Alberghina, L. & Westerhoff, H. V. 2005. Systems Biology: did we know it all along? in *Systems Biology: Definitions and Perspectives*. Topics in Current Genetics, Springer-Verlag, Berlin/Heidelberg, 13, pp. 3–9.

4 Kitano, H. 2002. Systems biology: a brief overview. *Science*, 295, 1662–4.

5 Alberghina, L., Gaglio, D., Moresco, R. M., Gilardi, M. C., Messa, C. & Vanoni, M. 2014. A systems biology road map for the discovery of drugs targeting cancer cell metabolism. *Current pharmaceutical design*, 20, 2648–66.

6 Hornberg, J. J., Bruggeman, F. J., Westerhoff, H. V. & Lankelma, J. 2006. Cancer: a Systems Biology disease. *Bio Systems*, 83, 81–90.

7 Kitano, H. 2004. Cancer as a robust system: implications for anticancer therapy. *Nature reviews. Cancer*, 4, 227–35.

8 Kitano, H. 2007. The theory of biological robustness and its implication in cancer. *Ernst Schering Research Foundation workshop*, (61), 69–88.

9 Wruck, W., Kashofer, K., Rehman, S., Daskalaki, A., Berg, D., Gralka, E., Jozefczuk, J., Drews, K., Pandey, V., Regenbrecht, C., Wierling, C., Turano, P., Korf, U., Zatloukal, K., Lehrach, H., Westerhoff, H. V. & Adjaye, J. 2015. Multi-omic profiles of human non-alcoholic fatty liver disease tissue highlight heterogenic phenotypes. *Scientific data*, 2, 150068.

10 Auffray, C. & Hood, L. 2012. Editorial: systems biology and personalized medicine—the future is now. *Biotechnology journal*, 7, 938–9.

11 Hood, L., Balling, R. & Auffray, C. 2012. Revolutionizing medicine in the 21st century through systems approaches. *Biotechnology journal*, 7, 992–1001.

12 Hood, L. & Tian, Q. 2012. Systems approaches to biology and disease enable translational systems medicine. *Genomics, proteomics & bioinformatics*, 10, 181–5.

13 Tanaka, H. 2010. Omics-based medicine and systems pathology. A new perspective for personalized and predictive medicine. *Methods of information in medicine*, 49, 173–85.

14 Tian, Q., Price, N. D. & Hood, L. 2012. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. *Journal of internal medicine*, 271, 111–21.

15 Gustafsson, M., Nestor, C. E., Zhang, H., Barabasi, A. L., Baranzini, S., Brunak, S., Chung, K. F., Federoff, H. J., Gavin, A. C., Meehan, R. R., Picotti, P., Pujana, M. A., Rajewsky, N., Smith, K. G., Sterk, P. J., Villoslada, P. & Benson, M. 2014. Modules, networks and systems medicine for understanding disease and aiding diagnosis. *Genome medicine*, 6, 82.

16 Hortobagyi, G. N. 2012. Toward individualized breast cancer therapy: translating biological concepts to the bedside. *The oncologist*, 17, 577–84.

17 Edgar, R., Domrachev, M. & Lash, A. E. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30, 207–10.

18 Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S. & Soboleva, A. 2013. NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*, 41, D991–5.

19 Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I. & Clamp, M. 2002. The Ensembl genome database project. *Nucleic acids research*, 30, 38–41.

20 Wingender, E., Dietze, P., KARAS, H. & Knuppel, R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic acids research*, 24, 238–41.

21 Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. & Mcvean, G. A. 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467, 1061–73.

22 Abecasis, G. R., Auton, A., Brooks, L. D., Depristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. & Mcvean, G. A. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491, 56–65.

23 Ecker, J. R., Bickmore, W. A., Barroso, I., Pritchard, J. K., Gilad, Y. & Segal, E. 2012. Genomics: ENCODE explained. *Nature*, 489, 52–5.

24 Encode Project Consortium 2011. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9, e1001046.

25 Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B. L., Barrell, D., Zadissa, A., Searle, S., Barnes, I., Bignell, A., Boychenko, V., Hunt, T., Kay, M., Mukherjee, G., Rajan, J., Despacio-Reyes, G., Saunders, G., Steward, C., Harte, R., Lin, M., Howald, C., Tanzer, A., Derrien, T., Chrast, J., Walters, N., Balasubramanian, S., Pei, B., Tress, M., Rodriguez, J. M., Ezkurdia, I., Van Baren, J., Brent, M., Haussler, D., Kellis, M., Valencia, A., Reymond, A., Gerstein, M., Guigo, R. & Hubbard, T. J. 2012. GENCODE: the reference human genome annotation for The ENCODE project. *Genome research*, 22, 1760–74.

26 Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C. & Stuart, J. M. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45, 1113–20.

27 Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., Bernabe, R. R., Bhan, M. K., Calvo, F., Eerola, I., Gerhard, D. S., Guttmacher, A., Guyer, M., Hemsley, F. M., Jennings, J. L., Kerr, D., Klatt, P., Kolar, P., Kusada, J., Lane, D. P., Laplace, F., Youyong, L., Nettekoven, G., Ozenberger, B., Peterson, J., Rao, T. S., Remacle, J., Schafer, A. J., Shibata, T., Stratton, M. R., Vockley, J. G., Watanabe, K., Yang, H., Yuen, M. M., Knoppers, B. M., Bobrow, M., Cambon-Thomsen, A., Dressler, L. G., Dyke, S. O., Joly, Y., KATO, K., Kennedy, K. L., Nicolas, P., Parker, M. J., Rial-Sebbag, E., Romeo-Casabona, C. M., Shaw, K. M., Wallace, S., Wiesner, G. L., Zeps, N., Lichter, P., Biankin, A. V., Chabannon, C., Chin, L., Clement, B., De Alava, E., Degos, F., Ferguson, M. L., Geary, P., Hayes, D. N., Johns, A. L., Kasprzyk, A., Nakagawa, H., Penny, R., Piris, M. A., Sarin, R., Scarpa, A., Van De Vijver, M., Futreal, P. A., Aburatani, H., Bayes, M., Botwell, D. D., Campbell, P. J., Estivill, X., Grimmond, S. M., Gut, I., Hirst, M., Lopez-Otin, C., Majumder, P., Marra, M., Mcpherson, J. D., Ning, Z., Puente, X. S., Ruan, Y., Stunnenberg, H. G., Swerdlow, H., Velculescu, V. E., Wilson, R. K., Xue, H. H., Yang, L., Spellman, P. T., Bader, G. D., Boutros, P. C., Flicek, P., Getz, G., Guigo, R., Guo, G., Haussler, D., Heath, S., Hubbard, T. J., Jiang, T., et al. 2010. International network of cancer genome projects. *Nature*, 464, 993–8.

28 Pleasance, E. D., Cheetham, R. K., Stephens, P. J., Mcbride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M. L., Ordonez, G. R., Bignell, G. R., Ye, K., Alipaz, J., Bauer, M. J., Beare, D., Butler, A., Carter, R. J., Chen, L., Cox, A. J., Edkins, S., Kokko-Gonzales, P. I., Gormley, N. A., Grocock, R. J., Haudenschild, C. D., Hims, M. M., James, T., Jia, M., Kingsbury, Z., Leroy, C., Marshall, J., Menzies, A., Mudie, L. J., Ning, Z., Royce, T., Schulz-Trieglaff, O. B., Spiridou, A., Stebbings, L. A., Szajkowski, L., Teague, J., Williamson, D., Chin, L., Ross, M. T., Campbell, P. J., Bentley, D. R., Futreal, P. A. & Stratton, M. R. 2010. A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463, 191–6.

29 Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P. A., Stratton, M. R. & Wooster, R. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer*, 91, 355–8.

30 Ellis, M. J., Gillette, M., Carr, S. A., Paulovich, A. G., Smith, R. D., Rodland, K. K., Townsend, R. R., Kinsinger, C., Mesri, M., Rodriguez, H. & Liebler, D. C. 2013. Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer discovery*, 3, 1108–12.

31 Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M. C., Zimmerman, L. J., Shaddox, K. F., Kim, S., Davies, S. R., Wang, S., Wang, P., Kinsinger, C. R., Rivers, R. C., Rodriguez, H., Townsend, R. R., Ellis, M. J., Carr, S. A., Tabb, D. L., Coffey, R. J., Slebos, R. J. & Liebler, D. C. 2014. Proteogenomic characterization of human colon and rectal cancer. *Nature*, 513, 382–7.

32 Kanehisa, M. & Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28, 27–30.

33 Kolker, E., Higdon, R., Haynes, W., Welch, D., Broomall, W., Lancet, D., Stanberry, L. & Kolker, N. 2012. MOPED: Model Organism Protein Expression Database. *Nucleic acids research*, 40, D1093–9.

34 Higdon, R., Stewart, E., Stanberry, L., Haynes, W., Choiniere, J., Montague, E., Anderson, N., Yandl, G., Janko, I., Broomall, W., Fishilevich, S., Lancet, D., Kolker, N. & Kolker, E. 2014. MOPED enables discoveries through consistently processed proteomics data. *Journal of proteome research*, 13, 107–13.

35 Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. 2000. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic acids research*, 28, 3442–4.

36 Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Bjorling, L. & Ponten, F. 2010. Towards a knowledge-based Human Protein Atlas. *Nature biotechnology*, 28, 1248–50.

37 Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., Fung, C., Nikolai, L., Lewis, M., Coutouly, M. A., Forsythe, I., Tang, P., Shrivastava, S., Jeroncic, K., Stothard, P., Amegbey, G., Block, D., Hau, D. D., Wagner, J., Miniaci, J., Clements, M., Gebremedhin, M., Guo, N., Zhang, Y., Duggan, G. E., Macinnis, G. D., Weljie, A. M., Dowlatabadi, R., Bamforth, F., Clive, D., Greiner, R., Li, L., Marrie, T., Sykes, B. D., Vogel, H. J. & Querengesser, L. 2007. HMDB: the Human Metabolome Database. *Nucleic acids research*, 35, D521–6.

38 Wishart, D. S., Knox, C., GUO, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., LI, L., Vogel, H. J. & Forsythe, I. 2009. HMDB: a knowledgebase for the human metabolome. *Nucleic acids research*, 37, D603–10.

39 Wishart, D. S., Jewison, T., Guo, A. C., Wilson, M., Knox, C., Liu, Y., Djoumbou, Y., Mandal, R., Aziat, F., Dong, E., Bouatra, S., Sinelnikov, I., Arndt, D., Xia, J., Liu, P., Yallou, F., Bjorndahl, T., Perez-Pineiro, R., Eisner, R., Allen, F., Neveu, V., Greiner, R. & Scalbert, A. 2013. HMDB 3.0—the Human Metabolome Database in 2013. *Nucleic acids research*, 41, D801–7.

40 Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L. & Cesareni, G. 2012. MINT, the molecular interaction database: 2012 update. *Nucleic acids research*, 40, D857–61.

41 Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R. C., Meldal, B., Melidoni, A. N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., Van Roey, K., Cesareni, G. & Hermjakob, H. 2014. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42, D358–63.

42 Oughtred, R., Chatr-Aryamontri, A., Breitkreutz, B. J., Chang, C. S., Rust, J. M., Theesfeld, C. L., Heinicke, S., Breitkreutz, A., Chen, D., Hirschman, J., Kolas, N., Livstone, M. S., Nixon, J., O'Donnell, L., Ramage, L., Winter, A., Reguly, T., Sellam, A., Stark, C., Boucher, L., Dolinski, K. & Tyers, M. 2016. Use of the BioGRID database for analysis of yeast protein and genetic interactions. *Cold Spring Harbor protocols*, 2016(1), 86–94.

43 More, T., Roychoudhury, S., Gollapalli, K., Patel, S. K., Gowda, H., Chaudhury, K. & Rapole, S. 2015. Metabolomics and its integration with systems biology: PSI 2014 conference panel discussion report. *Journal of proteomics*, 127, 73–9.

44 Menashe, I., Maeder, D., Garcia-Closas, M., Figueroa, J. D., Bhattacharjee, S., Rotunno, M., Kraft, P., Hunter, D. J., Chanock, S. J., Rosenberg, P. S. & Chatterjee, N. 2010. Pathway analysis of breast cancer genome-wide association study highlights three pathways and one canonical signaling cascade. *Cancer research*, 70, 4453–9.

45 Sloan, C. D., Shen, L., West, J. D., Wishart, H. A., Flashman, L. A., Rabin, L. A., Santulli, R. B., Guerin, S. J., Rhodes, C. H., Tsongalis, G. J., Mcallister, T. W., Ahles, T. A., Lee, S. L., Moore, J. H. & Saykin, A. J. 2010. Genetic pathway-based hierarchical clustering analysis of older adults with cognitive complaints and amnestic mild cognitive impairment using clinical and neuroimaging phenotypes. *American journal of medical genetics. Part B, Neuropsychiatric genetics: the official publication of the International Society of Psychiatric Genetics*, 153B, 1060–9.

46 Swaminathan, S., Shen, L., Risacher, S. L., Yoder, K. K., West, J. D., Kim, S., Nho, K., Foroud, T., Inlow, M., Potkin, S. G., Huentelman, M. J., Craig, D. W., Jagust, W. J., Koeppe, R. A., Mathis, C. A., Jack, C. R., JR., Weiner, M. W. & Saykin, A. J. 2012. Amyloid pathway-based candidate gene analysis of [(11)C]PiB-PET in the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. *Brain imaging and behavior*, 6, 1–15.

47 Zhang, M., Liang, L., Xu, M., Qureshi, A. A. & Han, J. 2011. Pathway analysis for genome-wide association study of basal cell carcinoma of the skin. *PloS one*, 6, e22760.

48 Elbers, C. C., Van Eijk, K. R., Franke, L., Mulder, F., Van Der Schouw, Y. T., Wijmenga, C. & Onland-Moret, N. C. 2009. Using genome-wide pathway analysis to unravel the etiology of complex diseases. *Genetic epidemiology*, 33, 419–31.

49 Cantor, R. M., Lange, K. & Sinsheimer, J. S. 2010. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *American journal of human genetics*, 86, 6–22.

50 Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. & Tegner, J. 2014. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8 Suppl 2, I1.

51 Ponniah, P. 2004. *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*. John Wiley & Sons, Inc., Hoboken.

52 Sheth, A. P. & Larson, J. A. 1990. Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys (CSUR)*, 22.3, 183–236.

53 Liu, Y., Devescovi, V., Chen, S. & Nardini, C. 2013. Multilevel omic data integration in cancer cell lines: advanced annotation and emergent properties. *BMC systems biology*, 7, 14.

54 Ghazalpour, A., Bennett, B., Petyuk, V. A., Orozco, L., Hagopian, R., Mungrue, I. N., Farber, C. R., Sinsheimer, J., Kang, H. M., Furlotte, N., Park, C. C., Wen, P. Z., Brewer, H., Weitz, K., Camp, D. G., 2ND, Pan, C., Yordanova, R., Neuhaus, I., Tilford, C., Siemers, N., Gargalovic, P., Eskin, E., Kirchgessner, T., Smith, D. J., Smith, R. D. & Lusis, A. J. 2011. Comparative analysis of proteome and transcriptome variation in mouse. *PLoS genetics*, 7, e1001393.

55 Pascal, L. E., True, L. D., Campbell, D. S., Deutsch, E. W., Risk, M., Coleman, I. M., Eichner, L. J., Nelson, P. S. & Liu, A. Y. 2008. Correlation of mRNA and protein levels: cell type-specific gene expression of cluster designation antigens in the prostate. *BMC genomics*, 9, 246.

56 Yeung, E. S. 2011. Genome-wide correlation between mRNA and protein in a single cell. *Angewandte Chemie*, 50, 583–5.

57 Schwanhausser, B., Busse, D., LI, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. 2011. Global quantification of mammalian gene expression control. *Nature*, 473, 337–42.

58 Hinnebusch, A. G. 2014. The scanning mechanism of eukaryotic translation initiation. *Annual review of biochemistry*, 83, 779–812.

59 Stefanovic, B. 2013. RNA protein interactions governing expression of the most abundant protein in human body, type I collagen. *Wiley interdisciplinary reviews. RNA*, 4, 535–45.

60 Szostak, E. & Gebauer, F. 2013. Translational control by 3'-UTR-binding proteins. *Briefings in functional genomics*, 12, 58–65.

61 Nishizawa, M., Okumura, T., Ikeya, Y. & Kimura, T. 2012. Regulation of inducible gene expression by natural antisense transcripts. *Frontiers in bioscience*, 17, 938–58.

62 Ambros, V. 2004. The functions of animal microRNAs. *Nature*, 431, 350–5.

63 Nam, S., LI, M., Choi, K., Balch, C., Kim, S. & Nephew, K. P. 2009. MicroRNA and mRNA integrated analysis (MMIA): a web tool for examining biological functions of microRNA expression. *Nucleic acids research*, 37, W356–62.

64 Muniategui, A., Pey, J., Planes, F. J. & Rubio, A. 2013. Joint analysis of miRNA and mRNA expression data. *Briefings in bioinformatics*, 14, 263–78.

65 Tebaldi, T., Re, A., Viero, G., Pegoretti, I., Passerini, A., Blanzieri, E. & Quattrone, A. 2012. Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. *BMC genomics*, 13, 220.

66 Com, E., Boitier, E., Marchandeau, J. P., Brandenburg, A., Schroeder, S., Hoffmann, D., Mally, A. & Gautier, J. C. 2012. Integrated transcriptomic and proteomic evaluation of gentamicin nephrotoxicity in rats. *Toxicology and applied pharmacology*, 258, 124–33.

67 Perco, P., Muhlberger, I., Mayer, G., Oberbauer, R., Lukas, A. & Mayer, B. 2010. Linking transcriptomic and proteomic data on the level of protein interaction networks. *Electrophoresis*, 31, 1780–9.

68 Piruzian, E., Bruskin, S., Ishkin, A., Abdeev, R., Moshkovskii, S., Melnik, S., Nikolsky, Y. & Nikolskaya, T. 2010. Integrated network analysis of transcriptomic and proteomic data in psoriasis. *BMC systems biology*, 4, 41.

69 Imielinski, M., Cha, S., Rejtar, T., Richardson, E. A., Karger, B. L. & Sgroi, D. C. 2012. Integrated proteomic, transcriptomic, and biological network analysis of breast carcinoma reveals molecular features of tumorigenesis and clinical relapse. *Molecular & cellular proteomics: MCP*, 11 (6), M111.014910.

70 Shen, R., Olshen, A. B. & Ladanyi, M. 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25, 2906–12.

71 Zhang, Y., Deng, Z., Jiang, H. & Jia, P. 2007. Inferring gene regulatory networks from multiple data sources via a dynamic bayesian network with structural EM. *Data integration in the life sciences*, 4544, 204–14.

72 Nariai, N., Kim, S., Imoto, S. & Miyano, S. 2004. Using protein-protein interactions for refining gene networks estimated from microarray data by Bayesian networks. *Pacific symposium on biocomputing*, 336–47.

73 Haider, S. & Pal, R. 2013. Integrated analysis of transcriptomic and proteomic data. *Current genomics*, 14, 91–110.

74 Ramaswamy, S., Ross, K. N., Lander, E. S. & Golub, T. R. 2003. A molecular signature of metastasis in primary solid tumors. *Nature genetics*, 33, 49–54.

75 Van't Veer, L. J., Dai, H., Van De Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. & Friend, S. H. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415, 530–6.

76 Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., Yang, F., Talantov, D., Timmermans, M., Meijer-Van Gelder, M. E., YU, J., Jatkoe, T., Berns, E. M., Atkins, D. & Foekens, J. A. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365, 671–9.

77 Ein-Dor, L., Kela, I., Getz, G., Givol, D. & Domany, E. 2005. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, 21, 171–8.

78 Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the national academy of sciences of the United States of America*, 102, 15545–50.

79 Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S. & Park, P. J. 2005. Discovering statistically significant pathways in expression profiling studies. *Proceedings of the national academy of sciences of the United States of America*, 102, 13544–9.

80 Wei, Z. & Li, H. 2007. A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23, 1537–44.

81 Chen, J. & Yuan, B. 2006. Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, 22, 2283–90.

82 Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., Mccuine, S., Uetz, P., Sittler, T., Karp, R. M. & Ideker, T. 2005. Conserved patterns of protein interaction in multiple species. *Proceedings of the national academy of sciences of the United States of America*, 102, 1974–9.

83 Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D. & Ideker, T. 2007. Network-based classification of breast cancer metastasis. *Molecular systems biology*, 3, 140.

84 Garcia, M., Millat-Carus, R., Bertucci, F., Finetti, P., Birnbaum, D. & Bidaut, G. 2012. Interactome-transcriptome integration for predicting distant metastasis in breast cancer. *Bioinformatics*, 28, 672–8.

85 Desmedt, C., Haibe-Kains, B., Wirapati, P., Buyse, M., Larsimont, D., Bontempi, G., Delorenzi, M., Piccart, M. & Sotiriou, C. 2008. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 14, 5158–65.

86 Van De Vijver, M. J., He, Y. D., Van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., Van Der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H. & Bernards, R. 2002. A gene-expression signature as a predictor of survival in breast cancer. *The New England journal of medicine*, 347, 1999–2009.

87 Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I. & Stagljar, I. 2015. Fundamentals of protein interaction network mapping. *Molecular systems biology*, 11, 848.

88 Patil, K. R. & Nielsen, J. 2005. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proceedings of the national academy of sciences of the United States of America*, 102, 2685–9.

89 Bersanelli, M., Mosca, E., Remondini, D., Giampieri, E., Sala, C., Castellani, G. & Milanesi, L. 2016. Methods for the integration of multi-omics data: mathematical aspects. *BMC bioinformatics*, 17 (Suppl 2), 15.

90 Stockel, D., Kehl, T., Trampert, P., Schneider, L., Backes, C., Ludwig, N., Gerasch, A., Kaufmann, M., Gessler, M., Graf, N., Meese, E., Keller, A. & Lenhof, H. P. 2016. Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, 32(10), 1502–8.

91 Kitano, H. 2007. A robustness-based approach to systems-oriented drug design. *Nature reviews. Drug discovery*, 6, 202–10.

92 Ogilvie, L. A., Wierling, C., Kessler, T., Lehrach, H. & Lange, B. M. 2015. Predictive modeling of drug treatment in the area of personalized medicine. *Cancer informatics*, 14, 95–103.

93 Westerhoff, H. V. 2015. Network-based pharmacology through systems biology. *Drug discovery today. Technologies*, 15, 15–6.

94 Wierling, C., Kessler, T., Ogilvie, L. A., Lange, B. M., Yaspo, M. L. & Lehrach, H. 2015. Network and systems biology: essential steps in virtualising drug discovery and development. *Drug discovery today. Technologies*, 15, 33–40.

95 Leung, E. L., Cao, Z. W., Jiang, Z. H., Zhou, H. & LIU, L. 2013. Network-based drug discovery by integrating systems biology and computational technologies. *Briefings in bioinformatics*, 14, 491–505.

96 Yu, M. K., Kramer, M., Dutkowski, J., Srivas, R., Licon, K., Kreisberg, J., NG, C. T., Krogan, N., Sharan, R. & Ideker, T. 2016. Translation of genotype to phenotype by a hierarchy of cell subsystems. *Cell systems*, 2, 77–88.

97 Krogan, N. J., Lippman, S., Agard, D. A., Ashworth, A. & Ideker, T. 2015. The cancer cell map initiative: defining the hallmark networks of cancer. *Molecular cell*, 58, 690–8.

98 Suderman, M. & Hallett, M. 2007. Tools for visually exploring biological networks. *Bioinformatics*, 23, 2651–9.

99 Bucci, E. M., Natale, M. & Poli, A. 2011. Protein networks: generation, structural analysis and exploitation. In Yang, N.-S. (Ed), *Systems and Computational Biology: Molecular and Cellular Experimental Systems*. INTECH, Shanghai.

100 Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D. & Gavin, A. C. 2010. Visualization of omics data for systems biology. *Nature methods*, 7, S56–68.

101 Pavlopoulos, G. A., O'Donoghue, S. I., Satagopam, V. P., Soldatos, T. G., Pafilis, E. & Schneider, R. 2008. Arena3D: visualization of biological networks in 3D. *BMC systems biology*, 2, 104.

102 Freeman, T. C., Goldovsky, L., Brosch, M., Van Dongen, S., Maziere, P., Grocock, R. J., Freilich, S., Thornton, J. & Enright, A. J. 2007. Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS computational biology*, 3, 2032–42.

103 Funahashi, A., Matsuoka, Y., Jouraku, A., Morohashi, M., Kikuchi, N. & Kitano, H. 2008. CellDesigner 3.5: a versatile modeling tool for biochemical networks. *Proceedings of the IEEE*, 96, 1254–65.

104 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, 13, 2498–504.

105 Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., Saito, K., Tanida, S., Yugi, K., Venter, J. C. & Hutchison, C. A., 3rd. 1999. E-CELL: software environment for whole-cell simulation. *Bioinformatics*, 15, 72–84.

106 Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. & Bork, P. 2011. iPath2.0: interactive pathway explorer. *Nucleic acids research*, 39, W412–5.

107 Thimm, O., Blasing, O., Gibon, Y., Nagel, A., Meyer, S., Kruger, P., Selbig, J., Muller, L. A., Rhee, S. Y. & Stitt, M. 2004. MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *The plant journal: for cell and molecular biology*, 37, 914–39.

108 Hooper, S. D. & Bork, P. 2005. Medusa: a simple tool for interaction graph analysis. *Bioinformatics*, 21, 4432–3.

109 Koehler, J., Rawlings, C., Verrier, P., Mitchell, R., Skusa, A., Ruegg, A. & Philippi, S. 2005. Linking experimental results, biological networks and sequence analysis methods using Ontologies and Generalised Data Structures. *In silico biology*, 5, 33–44.

110 Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P. & Philippi, S. 2006. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics*, 22, 1383–90.

111 Skusa, A., Ruegg, A. & Kohler, J. 2005. Extraction of biological interaction networks from scientific literature. *Briefings in bioinformatics*, 6, 263–76.

112 Breitkreutz, B. J., Stark, C. & Tyers, M. 2003. Osprey: a network visualization system. *Genome biology*, 4, R22.

113 Batagelj, V. & Mrvar, A. 1998. Pajek—program for large network analysis. *Connections*, 21, 47–57.

114 Kutmon, M., Van Iersel, M. P., Bohler, A., Kelder, T., Nunes, N., Pico, A. R. & Evelo, C. T. 2015. PathVisio 3: an extendable pathway analysis toolbox. *PLoS computational biology*, 11, e1004085.

115 Karp, P. D., Paley, S. M., Krummenacker, M., Latendresse, M., Dale, J. M., Lee, T. J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., Altman, T., Paulsen, I., Keseler, I. M. & Caspi, R. 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings in bioinformatics*, 11, 40–79.

116 Orlev, N., Shamir, R. & Shiloh, Y. 2004. PIVOT: protein interactions visualization tool. *Bioinformatics*, 20, 424–5.

117 Neuweger, H., Persicke, M., Albaum, S. P., Bekel, T., Dondrup, M., Huser, A. T., Winnebald, J., Schneider, J., Kalinowski, J. & Goesmann, A. 2009. Visualizing post genomics data-sets on customized pathway maps by ProMeTra-aeration-dependent gene expression and metabolism of Corynebacterium glutamicum as an example. *BMC systems biology*, 3, 82.

118 Auber, D. 2004. Tulip—a huge graph visualization framework. In Jünger, M. & Mutzel, P. (Eds.) *Graph Drawing Software*. Springer-Verlag, Berlin/Heidelberg, pp. 105–26.

119 Junker, B. H., Klukas, C. & Schreiber, F. 2006. VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC bioinformatics*, 7, 109.

120 Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. & Bork, P. 2002. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417, 399–403.

121 Prieto, C., Risueno, A., Fontanillo, C. & De Las Rivas, J. 2008. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. *PloS one*, 3, e3911.

122 Qu, K., Garamszegi, S., Wu, F., Thorvaldsdottir, H., Liefeld, T., Ocana, M., Borges-Rivera, D., Pochet, N., Robinson, J. T., Demchak, B., Hull, T., Ben-Artzi, G., Blankenberg, D., Barber, G. P., Lee, B. T., Kuhn, R. M., Nekrutenko, A., Segal, E., Ideker, T., Reich, M., Regev, A., Chang, H. Y. & Mesirov, J. P. 2016. Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *Nature methods*, 13, 245–7.

123 Chang, M. T., Asthana, S., Gao, S. P., Lee, B. H., Chapman, J. S., Kandoth, C., Gao, J., Socci, N. D., Solit, D. B., Olshen, A. B., Schultz, N. & Taylor, B. S. 2016. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology*, 34, 155–63.

124 Barabasi, A. L., Gulbahce, N. & Loscalzo, J. 2011. Network medicine: a network-based approach to human disease. *Nature reviews. Genetics*, 12, 56–68.

125 International Cancer Genome Consortium, M. C. A. P. A. W. G. 2015. Pathway and network analysis of cancer genomes. *Nature methods*, 12, 615–21.

126 Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., Van Der Weijden, C. C., Schepper, M., Walsh, M. C., Bakker, B. M., Van Dam, K., Westerhoff, H. V. & Snoep, J. L. 2000. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *European journal of biochemistry/FEBS*, 267, 5313–29.

127 Penkler, G., Du Toit, F., Adams, W., Rautenbach, M., Palm, D. C., Van Niekerk, D. D. & Snoep, J. L. 2015. Construction and validation of a detailed kinetic model of glycolysis in *Plasmodium falciparum*. *The FEBS journal*, 282, 1481–511.

128 Bordbar, A., Mccloskey, D., Zielinski, D. C., Sonnenschein, N., Jamshidi, N. & Palsson, B. O. 2015. Personalized whole-cell kinetic models of metabolism for discovery in genomics and pharmacodynamics. *Cell systems*, 1, 283–92.

129 Hyduke, D. R., Lewis, N. E. & Palsson, B. O. 2013. Analysis of omics data with genome-scale models of metabolism. *Molecular bioSystems*, 9, 167–74.

130 Bordbar, A., Monk, J. M., King, Z. A. & Palsson, B. O. 2014. Constraint-based models predict metabolic and associated cellular functions. *Nature reviews. Genetics*, 15, 107–20.

131 O'Brien, E. J. & Palsson, B. O. 2015. Computing the functional proteome: recent progress and future prospects for genome-scale models. *Current opinion in biotechnology*, 34, 125–34.

132 Thiele, I., Swainston, N., Fleming, R. M., Hoppe, A., Sahoo, S., Aurich, M. K., Haraldsdottir, H., Mo, M. L., Rolfsson, O., Stobbe, M. D., Thorleifsson, S. G., Agren, R., Bolling, C., Bordel, S., Chavali, A. K., Dobson, P., Dunn, W. B., Endler, L., Hala, D., Hucka, M., Hull, D., Jameson, D., Jamshidi, N., Jonsson, J. J., Juty, N., Keating, S., Nookaew, I., Le Novere, N., Malys, N., Mazein, A., Papin, J. A., Price, N. D., Selkov, E., SR., Sigurdsson, M. I., Simeonidis, E., Sonnenschein, N., Smallbone, K., Sorokin, A., Van Beek, J. H., Weichart, D., Goryanin, I., Nielsen, J., Westerhoff, H. V., Kell, D. B., Mendes, P. & Palsson, B. O. 2013. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31, 419–25.

133 Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Nookaew, I., Jacobson, P., Walley, A. J., Froguel, P., Carlsson, L. M., Uhlen, M. & Nielsen, J. 2013. Integration of clinical data with a genome-scale metabolic model of the human adipocyte. *Molecular systems biology*, 9, 649.

134 Yizhak, K., Chaneton, B., Gottlieb, E. & Ruppin, E. 2015. Modeling cancer metabolism on a genome scale. *Molecular systems biology*, 11, 817.

135 Agren, R., Bordel, S., Mardinoglu, A., Pornputtapong, N., Nookaew, I. & Nielsen, J. 2012. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS computational biology*, 8, e1002518.

136 Saha, R., Chowdhury, A. & Maranas, C. D. 2014. Recent advances in the reconstruction of metabolic models and integration of omics data. *Current opinion in biotechnology*, 29, 39–45.

137 Blazier, A. S. & Papin, J. A. 2012. Integration of expression data in genome-scale metabolic network reconstructions. *Frontiers in physiology*, 3, 299.

138 Topfer, N., Kleessen, S. & Nikoloski, Z. 2015. Integration of metabolomics data into metabolic networks. *Frontiers in plant science*, 6, 49.

139 Agren, R., Mardinoglu, A., Asplund, A., Kampf, C., Uhlen, M. & Nielsen, J. 2014. Identification of anticancer drugs for hepatocellular carcinoma through personalized genome-scale metabolic modeling. *Molecular systems biology*, 10, 721.

140 Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. & Shlomi, T. 2010. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26, i255–60.

141 Shlomi, T., Cabili, M. N. & Ruppin, E. 2009. Predicting metabolic biomarkers of human inborn errors of metabolism. *Molecular systems biology*, 5, 263.

142 Mardinoglu, A., Agren, R., Kampf, C., Asplund, A., Uhlen, M. & Nielsen, J. 2014. Genome-scale metabolic modelling of hepatocytes reveals serine deficiency in patients with non-alcoholic fatty liver disease. *Nature communications*, 5, 3083.

143 Varemo, L., Scheele, C., Broholm, C., Mardinoglu, A., Kampf, C., Asplund, A., Nookaew, I., Uhlen, M., Pedersen, B. K. & Nielsen, J. 2015. Proteome- and transcriptome-driven reconstruction of the human myocyte metabolic network and its use for identification of markers for diabetes. *Cell reports*, 11, 921–33.

144 Jozefczuk, J., Kashofer, K., Ummanni, R., Henjes, F., Rehman, S., Geenen, S., Wruck, W., Regenbrecht, C., Daskalaki, A., Wierling, C., Turano, P., Bertini, I., Korf, U., Zatloukal, K., Westerhoff, H. V., Lehrach, H. & Adjaye, J. 2012. A systems biology approach to deciphering the etiology of steatosis employing patient-derived dermal fibroblasts and iPS cells. *Frontiers in physiology*, 3, 339.

145 Naik, A., Rozman, D. & Belic, A. 2014. SteatoNet: the first integrated human metabolic model with multi-layered regulation to investigate liver-associated pathologies. *PLoS computational biology*, 10, e1003993.

146 Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of cancer: the next generation. *Cell*, 144, 646–74.

147 Lewis, N. E., Nagarajan, H. & Palsson, B. O. 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature reviews. Microbiology*, 10, 291–305.

148 O'Brien, E. J., Monk, J. M. & Palsson, B. O. 2015. Using genome-scale models to predict biological capabilities. *Cell*, 161, 971–87.

149 Jerby, L. & Ruppin, E. 2012. Predicting drug targets and biomarkers of cancer via genome-scale metabolic modeling. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 18, 5572–84.

150 Asgari, Y., Zabihinpour, Z., Salehzadeh-Yazdi, A., Schreiber, F. & Masoudi-Nejad, A. 2015. Alterations in cancer cell metabolism: the Warburg effect and metabolic adaptation. *Genomics*, 105, 275–81.

151 Shlomi, T., Benyamini, T., Gottlieb, E., Sharan, R. & Ruppin, E. 2011. Genome-scale metabolic modeling elucidates the role of proliferative adaptation in causing the Warburg effect. *PLoS computational biology*, 7, e1002018.

152 Mo, M. L., Jamshidi, N. & Palsson, B. O. 2007. A genome-scale, constraint-based approach to systems biology of human metabolism. *Molecular bioSystems*, 3, 598–603.

153 Vazquez, A. 2010. Optimal cytoplasmatic density and flux balance model under macromolecular crowding effects. *Journal of theoretical biology*, 264, 356–9.

154 Cazzaniga, P., Damiani, C., Besozzi, D., Colombo, R., Nobile, M. S., Gaglio, D., Pescini, D., Molinari, S., Mauri, G., Alberghina, L. & Vanoni, M. 2014. Computational strategies for a system-level understanding of metabolism. *Metabolites*, 4, 1034–87.

155 Di Filippo, M., Colombo, R., Damiani, C., Pescini, D., Gaglio, D., Vanoni, M., Alberghina, L. & Mauri, G. 2016. Zooming-in on cancer metabolic rewiring with tissue specific constraint-based models. *Computational biology and chemistry*, 62, 60–9.

156 Palmieri, F. 2004. The mitochondrial transporter family (SLC25): physiological and pathological implications. *Pflugers Archiv: European journal of physiology*, 447, 689–709.

157 Damiani, C., Pescini, D., Colombo, R., Molinari, S., Alberghina, L., Vanoni, M. & Mauri, G. 2014. An ensemble evolutionary constraint-based approach to understand the emergence of metabolic phenotypes. *Natural computing*, 13, 321–331.

158 Resendis-Antonio, O., Checa, A. & Encarnación, S. 2010. Modeling core metabolism in cancer cells: surveying the topology underlying the Warburg effect. *PLoS one*, 5(8), e12383.

16

Generation of Molecular Models and Pathways

Amel Bekkar, Julien Dorier, Isaac Crespo, Anne Niknejad, Alan Bridge, and Ioannis Xenarios

Vital-IT, SIB Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland

16.1 Introduction

In this chapter we will discuss methods to construct and analyze comprehensive and predictive models of biological systems. Such models include all of the elements and regulatory structures necessary to model the dynamics of the normal system. They also allow us to elucidate how perturbations in one or more regulatory interactions will alter those dynamics. Many disease processes such as cancer or diabetes arise from alterations in regulatory network components that transcend the classical definitions of signaling pathways, and modeling allows us to study their behavior as part of a complete system.

Over the last few decades, a wide range of modeling frameworks has been developed. These combine prior knowledge from traditional hypothesis-driven experiments (in which the function of individual gene products is carefully investigated using orthogonal approaches) or existing databases with -omics-level signatures of biological systems derived from high-throughput technologies. Decades of research on how individual regulatory interactions assemble into signaling pathways provides a rich source of information for the assembly of more complex predictive systems-level models—the "bottom-up" approach to model construction. This complements the -omics-driven top-down approach that also includes the application of computational techniques to analyze (and in some cases reduce) complex high-throughput data in order to facilitate the process of model construction (Figure 16.1).

Modeling regulatory networks involves two main steps. The first is to build the primary structure of the network, while the second is to choose and apply the appropriate mathematical approach to simulate its dynamical behavior. In the following section we will discuss the first step, that of network construction. Networks may be constructed using experimental data (such as high-throughput -omics data), existing published knowledge, or a combination of the two. Tools that can infer a regulatory network from high-throughput data [1–3] include REVEAL (that uses Boolean modeling [4]), BMA [5] and ScanBMA [6] (that use probabilistic modeling to generate Bayesian networks), and ARACNE [2, 7], CLR [8], and C3NET [9] (that use mutual information to assemble correlation networks). Here we will focus on strategies that take advantage of existing published knowledge to construct regulatory networks for complex biological systems. We call these networks prior knowledge networks (PKNs).

16.2 PKN Construction Through Expert Biocuration

The scientific literature is the most useful source of information when starting to build a PKN. From the existing knowledge and reported experimental evidence, one can assemble a directed graph that includes the relevant components of the system of interest as well as the relationships between them (inhibition or activation). The most popular research tools are the classical search engines such as *PubMed* (*NCBI*) and *Google Scholar*, but several text mining tools were developed over the last decade to scan the literature such as *iHOP* [10], an online text mining service that provides a gene-guided network to access PubMed abstracts. The text mining literature contains numerous softwares and platforms (for a comprehensive review see Ref. [11]).

Pathway and reaction databases are other precious curated sources of knowledge. KEGG Pathway Database provides manually drawn maps of interactions and reactions. Reactome [12] is also a manually curated database of human pathways and interactions. Other databases are more disease specific such as Atlas of Cancer Signaling Networks (ACSN) [13] where molecular

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

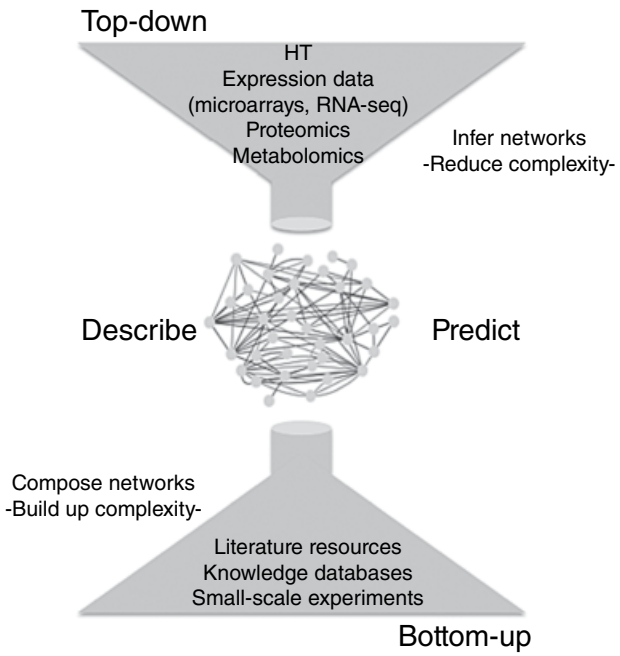


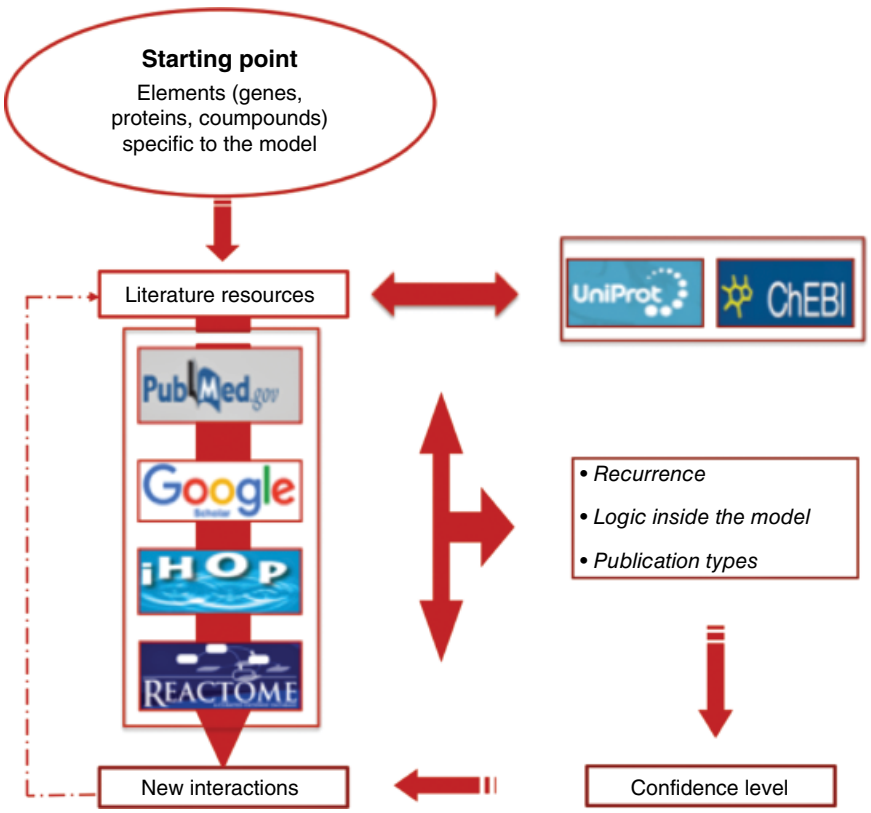
Figure 16.1 Modeling regulatory network at the intersection of the top-down and the bottom-up methodologies to describe biological systems and predict their behavior/outcome.

mechanisms involved in cancer are collected and represented in the form of five interconnected maps, each covering signaling processes involved in apoptosis, cell survival, cell motility, cell cycle, and DNA repair. IntAct [14], which is part of the IMEx consortium (www.imex.org), also provides curated molecular interaction database. All these resources provide tooling to search and overlay experimentally determined measurement from gene to protein expression.

A number of commercial resources are also available to help with the reconstruction and overlay of PKNs, for example, Ingenuity Pathway Analysis, Ariadne, and MetaCore, that provide several services such as manually curated database of signaling pathways, text mining, and knowledge extraction and a number of tools to analyze and manipulate networks. These services are not open source or freely accessible, rendering their access costly.

These PKN resources presented in Figure 16.2 are good starting points for network modeling since they offer a global view of the system. However, they all have the drawback that they are not contextualized for specific biological cases since they merge information from different experimental, tissue, and cellular contexts. Most of time these PKNs need to be refined and trained to experimental data before being used for *in silico* simulation experiments and analysis. This issue is addressed in Section 16.3.1.5.

Figure 16.2 Curation workflow for prior knowledge network (PKN) construction and the different steps necessary to build a contextualized PKN.



16.3 Modeling and Simulating the Dynamical Behavior of Networks

The choice of modeling approach will depend on the complexity of the biological question (or knowledge network) as well as the quantity and type of data available. We will discuss here several modeling approaches that range from the quantitative modeling using ordinary differential equations (ODEs) to qualitative and logical modeling. Quantitative modeling provides detailed information about the dynamical behavior of a biological system but requires experimental data on kinetic parameters that are rarely available. Quantitative modeling is therefore generally applied to the study of small and well-characterized systems. Qualitative modeling approximates quantitative behaviors using a limited number of defined states and does not require experimental data on kinetic parameters. This feature makes qualitative modeling a powerful means to study and predict the behavior of large networks for which detailed kinetic data is not available. An overview of the modeling workflow is presented in Figure 16.3.

16.3.1 Logic Models

16.3.1.1 Boolean Networks

Logical modeling of biological systems was pioneered by Kauffman [15, 16]. Among logic-based methods, the simplicity of Boolean models makes them an attractive means to describe large networks. A Boolean network consists of a set of nodes representing genes, proteins, regulatory RNAs, small molecules, and other components that are relevant to the biological process of interest. The node states are binary, where 1 represents an active node and 0 an inactive node—and where activity could be the result of protein production or an activating modification such as phosphorylation. Interactions between nodes are represented by edges that denote the influence of one node over another—either activation or inhibition. The state of a node (0 or 1, corresponding to the logical values FALSE or TRUE) is determined by the state of its input nodes. Interaction among inputs is captured by Boolean functions as combinations of elementary AND, OR, and NOT gates that generate logic rules for target activation/inactivation. The identity of gates is determined based on prior knowledge and experimental observations.

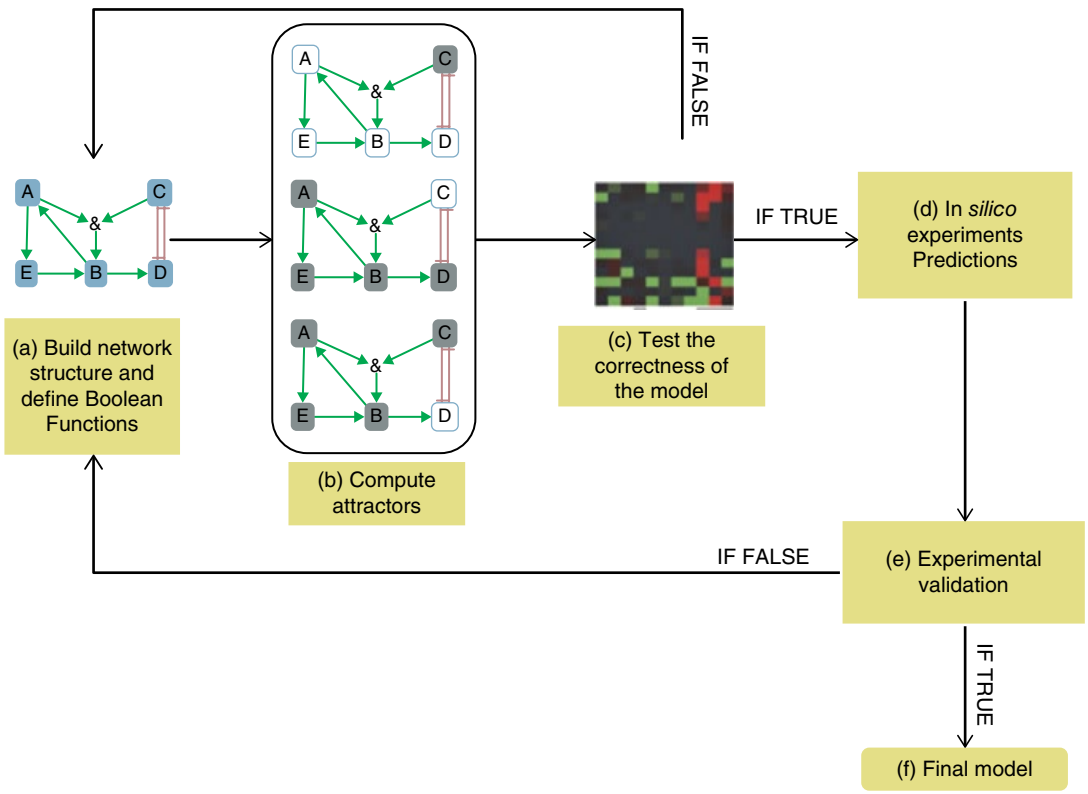


Figure 16.3 Modeling workflow. Network construction and iterative optimization against experimental datasets. (a) Toy example of a Boolean network with five nodes A, B, C, D, and E and one AND logic gate; arrow edges are activations and T edges are inhibitions. (b) Network attractors: grey nodes are active (1) and white nodes are inactive (0). (c) Test the correctness of the model compared with experimental data: expression data, proteomic data, or other data relevant for the biological question. (d) *In silico* simulations and experimental prediction. (e) Experimental validation of simulation outcomes. (f) Final model.

For example, a node A is active if one of its activators B or C is active, and its inhibitor D is inactive; the logical function of A would be A = (B or C) AND NOT D.

Simulation of network dynamics involves (i) updating the nodes based on their inputs (Boolean function) and (ii) searching for the long-term behavior of the network where the node states stabilize, also called attractors.

An updating scheme is needed when simulating dynamics using Boolean models [17]. In the synchronous updating scheme, we make the assumption that all events in the system have similar timescales and all genes change their state simultaneously. However, in the asynchronous model, one node state is updated at a time. There are deterministic and stochastic asynchronous schemes. In stochastic schemes, the node that is updated in the next time step $t+1$ is chosen randomly, whereas in deterministic asynchronous schemes, the nodes are updated according to a predetermined order. The requirement of asynchronous models for high CPU time prevents their use for analysis of large and highly connected networks; hence there is a need for the synchronous models as alternative.

A snapshot of the state of all the nodes in the network at a time t is called the state of the network. Attractors are the states where the system stabilizes. They represent the stable behavior of the system. Simulation allows one to identify these attractors. Simple attractors feature only a single state, while complex attractors feature multiple states among which the system oscillates. Positive functional feedback loops (sequences of edges by which nodes positively influence their own activation) are a feature of systems with multiple stable states where functional negative feedback loops generate cyclic attractors [18].

Identifying network's attractors is biologically relevant since attractors may be associated with cellular phenotypes, assuming the network has been well designed. Attractor analysis allows one to compare the activation level of network components with prior knowledge and experimental data. An attractor must be able to reproduce prior experimental observations; if it fails the network structure and Boolean functions should be checked. These comparisons are used to *test the correctness of the model*. Data used to test the correctness of the model must be different than those used to construct the model. Knockout experiments could be used as training sets to optimize the network. Several iterations of curation and optimization may be necessary to obtain a model that best reproduces known experimental behavior. One example of iteration process is the recent study of Flobak et al. on drug combination simulations to test drug synergy in gastric cancer cells [19]. The authors removed iteratively from the PKN components not targeted by drugs to obtain a network sufficiently small to allow asynchronous simulations while consistent with the whole network behavior, enabling thus the analysis of all single and pair inhibitions. Their simulation predicted synergetic inhibitory action of 5 combinations from a total of 21, of which 4 of these predictions were experimentally confirmed, thereby demonstrating the benefits of such an approach.

Simulation of the Boolean network behavior and the identification of attractors can be performed using a number of tools. *BoolSim*/*genYsis* [17] provides a set of algorithms for synchronous and asynchronous node update to compute the attractors, as well as functions to perform gene perturbation experiments. It has been used, inter alia, for simulating *in silico* perturbation experiments that enabled the identification of IL-11 as novel pluripotency-associated factor capable of sustaining self-renewal in human pluripotent stem cells [20]. *GINsim* [21] and *SQUAD* [22] are other simulation tools designed for qualitative modeling. Unlike *BoolSim* that is a command line tool, GINsim—used, inter alia, to compute the steady states of the previously cited model of drug synergy in gastric cancer cells [19]—and *SQUAD* provide user-friendly graphical interfaces, which is more convenient especially for non-computational scientists. Moreover, *SQUAD* enables to create a continuous dynamical system and localizes its steady states that are located near the steady states of the discrete system; it has been successfully used to reproduce the behavior of the regulatory network implicated in T-helper cell differentiation.

Logical modeling is becoming nowadays a popular modeling framework generating the need of format and tool standardization. The *Consortium for Logical Models and Tools* (CoLoMoTo) [23] is an international open community that brings together modelers, curators, and developers of methods and tools. It aims at the definition of standards for model representation and interchange and the establishment of criteria for the comparison of methods, models, and tools. The colomoto.org website contains the methods, formats, and software relevant for logical modeling.

The Boolean approach provides a way to narrow down possible experimental combinations that need to be tested. It has been successfully used to study several biological systems. Chasapi et al. [24] describe a model of the septation initiation network in *Schizosaccharomyces pombe* that includes 54 nodes and 124 edges. Their model was able not only to reproduce known experimental outputs such as the septation blockage by *cdc11* and *cdc16* deletion but also to make *in silico* the counterintuitive double-mutant phenotypic prediction that Sid4p mutant cells would septate if they express Cdc7p in high levels. This prediction has been validated *in vivo*, demonstrating the power of qualitative modeling in hypothesis generation and prediction of experimental outcomes.

Another example of successful collaboration between Boolean modeling and experimental approaches is the recent study by Guex and Crespo et al. Researchers combined prior knowledge with a network inference strategy that consists of using experimental measurements and *in silico* predictions to construct and refine the network that controls TIE-2-expressing monocytes (TEM) pro-angiogenic function. This network was then used to identify, through simulations, combined treatments that would alter TEM pro-angiogenic activity. All the minimal perturbations that bring the TEM from highly pro-angiogenic phenotype to the weak pro-angiogenic one and vice versa were tested, and four of the five predicted combined treatments were validated experimentally. The treatments proved to be highly efficient and may constitute a novel valid therapeutic strategy in breast cancer [25].

16.3.1.2 Probabilistic Boolean Networks (PBN)

In order to encode Boolean functions, it is necessary to understand how different inputs that are the regulators of a given target (protein, gene, etc.) operate to influence its activity/state so that appropriate logic connectives (AND, OR, NOT) can be used to encode the Boolean function. However, the lack of experimental evidence raises the need to express uncertainty in the regulatory logic and overcome the deterministic inflexibility of Boolean networks. In probabilistic Boolean networks (PBN), each node can have several Boolean functions and each function can have a probability based on prior data. A regulation function is then randomly selected at each time step for each node according to its probability [26, 27]. The result is a sequence of states that constitutes a Markov chain with fixed transition probabilities. Shmulevich et al. [26] used a subnetwork designed from human glioma gene expression data to demonstrate their proposed method for the steady-state analysis. Their study relies on the use of Monte Carlo methods to analyze the corresponding Markov chains. Thus, they used a subnetwork of 15 genes constructed from human glioma gene expression dataset and then analyzed the joint steady-state probabilities of combinations of two genes: Tie-2 and NF-κB, Tie-2 and TGFB3, and TGFB3 and NF-κB. Their results seem to be consistent with what is known about these genes in tumorigenesis. Indeed they show that TGFB3 and NF-κB are directly linked, which is very likely since TGFB3 is a homolog of TGFB1, already known to have a direct regulatory relationship with NF-κB.

16.3.1.3 Multiple Value Modeling

In some cases such as when the effect of a component on its target depends on more than two levels of activation, it may be necessary to consider further levels of activation of node states rather than the simple binary Boolean values (0/1). To take into account such cases, several methods and tools were proposed to extend the Boolean modeling by allowing a range of intermediate values for node states. Tools such as BoolSim [17, 28] and GINsim [21] designed at first for Boolean modeling were extended to take over multi-valued nodes in Boolean networks. However, in some cases there is a need to encode more complex rules than those allowed by the conventional Boolean framework. To emphasize this point, we discuss in the next section the fuzzy logic (FL)-based modeling.

16.3.1.4 Fuzzy Logic-Based Modeling

FL modeling extends simple Boolean modeling by allowing a range of intermediate values for node states rather than the simple binary values. The intermediate values in FL models can be descriptive classes such as "low," "medium," and "high," with logic operators connecting these descriptive inputs to a specific output, offering an interesting intermediate alternative between Boolean models (BM) and ODEs. Inputs are mapped to outputs thanks to list of if-then statements called rules where the "if" part is the antecedent and the "then" part the consequent.

In FL the truth of any statement is a matter of degree. Before the rules can be evaluated, inputs must be fuzzified: membership functions are used to assign values of inputs to a descriptive input class. They define how each input is mapped to a membership value called *degree of membership* that ranges from 0 (no membership) to 1 (full membership). If the antecedent has more than one input, fuzzy operators are applied. As illustrated in Table 16.1, to resolve the statements the min() function can be used to resolve the AND, the max() to resolve the OR, and the additive complement for the NOT. Indeed if we keep the fuzzy values at their extremes 0 and 1, the standard Boolean logic applies, and we can easily demonstrate that the min(), max(), and the additive complement preserve the results of AND, OR, and NOT truth tables that remain unchanged. In the context of FL modeling, we can then consider other intermediate values.

FL rules need not to be unique in contrast to Boolean logic where only one rule can be used and a weight can

Table 16.1 Example of the use of membership functions.

A	B	A and B	min(A,B)	A or B	max(A,B)	NOT A	1 – A
0	0	0	0	0	0	1	1
0	1	0	0	1	1		
1	0	0	0	1	1	0	0
1	1	1	1	1	1		

be applied to each FL rule to specify whether it should be used or not. The final step aims to assign one value to the output. It is called *defuzzification*; the results of multiple rules are combined and resolved to determine the output value. Thanks to this flexibility of FL gates, intermediate levels of activity and complex processing functions can be modeled.

Aldridge et al. used this method to model signaling network induced by tumor necrosis factor (TNF), epidermal growth factor (EGF), and insulin in human colon carcinoma cells [29]. They describe step by step the assembly of the signaling network and the application of the FL framework to incorporate qualitative and noisy data and produce relevant quantitative predictions. They used as starting point a PKN diagram, and then logic gates and their associated membership functions were generated from the cellular responses to the cytokines treatments. Their model simulations recapitulated most features of their data, generated new insights regarding mitogen-activated protein kinase 2 (MK2) and extracellular signal-regulated kinase (ERK) pathways cross-talk, and uncovered unexpected inhibition of the inhibitor of nuclear factor kappa-B kinase (IKK) by EGF treatment. These observations show once again the ability and usefulness of network modeling to generate testable biological predictions that cannot be obvious from simple inspection of the data.

Subsequently a new approach to FL modeling named constrained fuzzy logic (cFL) was introduced by Morris et al., enabling the formal training of a PKN to experimental data and resulting in a quantitative network model [30]. Their method was implemented in the *CellNetOptimizer* software. It comprises three main stages: First, the PKN is converted into a cFL model. Second the model is trained to experimental data. Third, trained models are refined and reduced. This method is limited by some issues such as the excessive CPU time requirement that was addressed by a nonlinear programming (NLP) optimization formulation approach [31].

Another recent study combined PKN and FL modeling to address the question of downregulation of hepatic detoxification during inflammation [32]. The response of primary human hepatocytes to IL-6 stimulation was investigated. The approaches used were chemical perturbation experiments, single inhibitions of STAT3, PI3K, and MAPK and then combinatorial inhibitions upon IL-6 stimulation to train the model. The R library CNORfuzzy [33] was used for the analysis. The resulting model suggested a central role of RXRα receptor as link between inflammatory signaling and drug-metabolizing enzymes and transporters. This prediction has been experimentally validated by siRNA-mediated RXRα gene silencing.

16.3.1.5 Contextualization of PKNs Using Experimental Data

PKNs are a good starting point when modeling gene/signaling regulatory networks. Most PKNs are derived by combining the results from many different experimental systems and are not specific to a given biological context such as a tissue or cell type. These may have distinct variants of "textbook" signaling pathways, and these may also be altered in disease states. PKNs are also biased in their composition as individual components may have been studied to varying degrees.

The detection of truly active signaling topologies based on comprehensive experimental data is then necessary to accurately model specific biological systems. To more accurately model specific biological systems, starting PKNs may be trained or contextualized using experimental data from that system. Some of the methods for this contextualization process are described as follows:

SigNetTrainer is based on an integer linear programming formulation to encode constrains on the qualitative behavior of the nodes. It proposes a set of algorithms to detect and remove inconsistencies between measurements and PKN topology [34]. It has been applied by Michailidou et al. to improve the mechanistic knowledge of human hepatocellular carcinoma chemoprevention in order to provide a strategy to assess the preclinical candidates [35]. The PKN has been obtained by merging pathways from databases such as KEGG and Ingenuity into a signaling network and then contextualized to proteomic data.

CellNOptR is another software for building logic models by training PKNs to experimental data. It converts the PKN to a Boolean network and identifies the optimal sub-model that fits the multiple perturbations experimental data [33]. It has been used, inter alia, by Vega et al. to investigate the mechanism governing budding yeast cell signaling pathway interactions [36]. Models of high osmolarity glycerol (HOG) and the mating pheromone response pathways were trained to phosphoproteomic time course data corresponding to the stimulation with NaCl, pheromone, and both stimuli.

PRUNET is a recently developed software where a single experiment comparing two cellular phenotypes is sufficient for contextualization. Besides the PKN, it takes as input a list of up- and downregulated genes resulting from the comparison between two stable cellular phenotypes and returns a contextualized network [37]. The authors demonstrated the applicability of their method using four previously published models as examples: epithelial–mesenchymal transition [38], T-helper lymphocyte differentiation [39], induction of pluripotent stem cells [40], and differentiation of human embryonic stem cells into cardiomyocytes [41].

They added to these models consistent and inconsistent interactions and then used training sets from the same publications for contextualization. The resulting networks were then used to predict expression values and simulate perturbations to evaluate whether contextualized networks reproduce expected response.

16.3.1.6 Ordinary Differential Equations

Quantitative models need real-valued parameters over a continuous timescale. A detailed model of regulation can be described by ODEs. ODEs were among the first approaches used by researchers to model cell physiology in the 1970s [42]. They provide detailed information about the dynamic of the network, but the fact that they require high-quality kinetic data makes them applicable to only few systems.

These equations describe the instantaneous change of molecular concentrations of each component as a function of the level of its regulators. These changes are expressed as reaction-rate equations that have the mathematical form

$$\frac{dxi}{dt} = fi(x), \quad 1 \le i \le n$$

There are n equations; x is the vector concentration of proteins, RNAs, or metabolites and fi a usually nonlinear function. These equations can be extended to take into account concentrations (u) of input elements such as externally supplied nutrients:

$$\frac{dxi}{dt} = fi\left(x, u\right), \quad 1 \le i \le n$$

Solving these rate equations depends on f. In general, these equations are difficult to solve analytically when $fi(x)$ are nonlinear. One way to work around this issue is to have resort to numerical simulations. The approximate values of $x_1, x_2, \ldots, x_n$ are calculated for consecutive time points $t_1, t_2, \ldots, t_n$ to approximate the exact solution of the equation. Several tools have been developed to create such models and simulate their behaviors such as SCAMP, DBSolve, MIST, and GEPASI. SMBL software guide lists (http://sbml.org/SBML_Software_Guide/SBML_Software_Summary), inter alia, ODE-based simulators and analysis tools.

Alternatively one can simplify the model. The Hill and Michaelis–Menten equations are the most frequent simplification used to model biochemical phenomena of small systems [43]. *Valenime et al*. [44] used the Hill functions to model a dynamic model of the gene regulatory network involved in *Arabidopsis* flowering time. Their network integrated eight genes that represent the core of the network responsible for flowering time regulation and for which the available experimental data provide a clear description of their mutual interactions. Model parameters were estimated from gene expression time courses of the selected eight genes in wild-type and various genetic backgrounds. A total of 35 parameters in 6 equations were estimated per gene. The model was validated by simulating changes in expression levels in mutants and comparing these predictions with independent expression data as well as comparing predicted and experimental flowering times for several double mutants.

ODE approaches are well suited for modeling cell cycles; one example is the *Caulobacter crescentus* model that describes the molecular mechanism for control of the cell division cycle [45]. In this study authors constructed around the three master regulators of cell division cycle: CtrA, GcrA, and DnaA, a mathematical model of the temporal dynamics of the regulatory elements. Parameter values for rate equations (rate constants, binding constants, and thresholds) were determined from experimental data. The model successfully reproduced the behavior of wild-type, mimicked correctly the phenotypes of many mutants, and predicted phenotypes of novel mutants.

A second example is the budding yeast cell-cycle regulatory network model. This topic was studied from various sides by Tyson et al. [46–48]. They use mathematical modeling to study the different aspects of yeast cell cycle such as the START and FINISH events of mitosis [47]. Their models, most of the time, accurately describe the behavior of wild-type cells such as growth, division, and regulation and are supported by the phenotypes of hundreds of mutant strains, giving new insights on different aspects of yeast cell-cycle control and regulation.

16.3.1.7 Piecewise Linear Differential Equations

In most situations the required detailed information for ODE modeling on reaction mechanism and the large number of parameters needed for the equations are not available. Indeed, determining such parameters implies the need of large and detailed experimental data, hence the need to create more qualitative models. Piecewise linear differential equations (PLDE) have favorable mathematical properties that facilitate the analysis.

PLDE have been at first proposed by Glass and Kauffman in 1973 [16] and were then extensively studied from a theoretical point of view. However, because of the difficulty of applying them for large networks, they have been rarely used to model real biological systems [49]. Softwares developed for qualitative modeling such as *Genetic Network Analyzer* (GNA) [50] or *BooleanNet* [51] allow to study such models and to evaluate their steady states. GNA has been used to build a model of nutritional stress response in *Escherichia coli* [52]. The network included six genes that play a key role in the

response of the cell to carbon starvation, and then based on experimental literature, a PLDE model has been constructed. GNA was then used to simulate the transition from exponential to stationary phase and the reentry into exponential phase. The model was finally validated by comparing the simulation results with experimental data.

16.3.1.8 Constraint-Based Modeling

Constraint-based modeling has been widely used to model large-scale metabolic pathways unmanageable with kinetic modeling approaches. It aims to model the dynamics of metabolism at cellular level by using constraints that limit cell behavior such as mass balance, energy balance, and flux limitations to differentiate between those network states that are achievable by the system from those that are not. One of the most applied methods is flux balance analysis (FBA) in which the stoichiometry of the underlying biochemical network constrains the steady states that may be reached by the system [53, 54]. It is worth mentioning here the MetaNetX.org platform that provides a suite of tools for accessing, analyzing, and manipulating metabolic networks, including FBA [55].

The first step of a FBA is to define the system: as reported in the PKN reconstruction section, metabolic pathways can be built from prior knowledge by assembling metabolites and metabolic reactions. Several databases provide organism-specific metabolic pathways. MetaNetX.org also provides access to hundreds of genome-scale metabolic networks and pathways [55].

Following the construction of an organism-specific network of metabolic reactions, these are converted into matrix form: **S** is the stoichiometric matrix where each row represents a metabolite and each column represents a reaction and **V** is the matrix of fluxes. Figure 16.4 shows **S** and **V** matrices for a toy model composed by three metabolites.

At steady state, the flux through each reaction is given by $\mathbf{S}\cdot\mathbf{V}=\mathbf{0}$, it is the null space of **S**. As large models contain more reactions than metabolites, there is more than one possible solution to these equations; however a bounded solution space is identified, and additional constraints on the basis of experimental measurements in a cell can be used to determine the steady state. An additional constraint could be the maximization of biomass production that is a widely used objective function as this represents a reasonable metabolic goal under most growth conditions (see following text). This (or another) objective function is then used to solve the equations. Several computational tools have been developed to solve these equations using linear programming such as the COBRA Toolbox [56], OptFlux [57], and FAME [58].

In order to study the effects of environmental or genetic perturbations on cellular metabolism, several methods have been developed to integrate measurements of changes at the transcript, protein, and metabolite level under the FBA framework. Machado and Herrgard evaluated several published methods for integrating transcriptomics data to genome-scale metabolic models [59] but found that none performed better than the conventional FBA. This indicates that further improvements are necessary for methods designed for a predictive purpose.

Other methods were developed for the production of contextual metabolic models using experimental data: mCADRE [60] is a method developed by Wang et al. to infer context-specific networks based on gene expression data and metabolic network topology. This method has been used to reconstruct 126 human tissues and cell type-specific metabolic models.

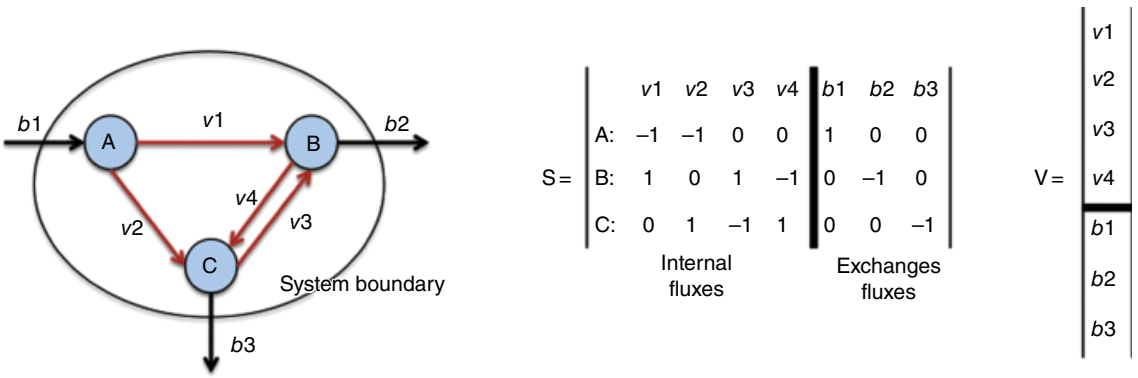


Figure 16.4 Example of metabolic network with three metabolites—A, B, and C. The system contains three reactions that are internal fluxes with one reversible reaction (*v*3, *v*4) and three exchanges fluxes (*b*1, *b*2, and *b*3). **S** is the stoichiometric matrix, **V** the matrix of fluxes, at steady state $\mathbf{S}\cdot\mathbf{V}=0$, A: $-v1-v2+b1=0$, B: $v1+v3-v4-b2=0$, C: $v2-v3+v4-b3=0$.

INIT [61] is another algorithm developed to integrate several types of data to generate metabolic networks that are active in a certain context. It has been applied to reconstruct genome-scale metabolic networks for 69 human cell types and 16 cancer cell types. The program used the Human Protein Atlas as source to assess the presence/absence of enzymes in the different cell types, tissue-specific gene expression data, and metabolomics data as additional constraints.

Protein levels should provide a more accurate snapshot of metabolism than transcript levels. Montezano et al. propose a new approach for defining the FBA objective function from proteomics data for the bacterium *Mycobacterium tuberculosis* exposed to mefloquine [62]. FBA requires the specification of an objective function representing the goal of the cell; a commonly used goal is biomass synthesis. However, biomass maximization is not expected to be the main metabolic goal of the cell when exposed to antibiotics. The authors propose a method to determine the objective function based on the level of proteins in the cases where the organism is under drug-induced stress. They calculate from protein levels in the sample the coefficients needed. Their method is summarized as follows: after a normalization step, they defined two cases; the simplest case is when each enzyme catalyzes one reaction. In this case the coefficient c of each reaction is the normalized quantitative value of the enzyme obtained with the proteomics experiment. A more complex situation is when one reaction is catalyzed by a combined action of multiple enzymes. In this case a Boolean expression that describes the combined action of all enzymes is solved in order to obtain the final value of c.

16.3.1.9 Hybrid Models

Mathematical models based on PKNs have guided our understanding of many systems at several levels ranging from gene regulatory network to metabolic models, providing insights into the qualitative behavior of the system or in a more quantitative evaluation of its components. However each approach has advantages and limits, captures different aspects of biology, and consequently yields only a part of the global picture. By combining diverse methods, one can move toward more global and realistic models that better describe the behavior of living organisms at the molecular level with a greater level of detail.

An integrative method that combines different modeling procedures was used to model the whole-cell life cycle of *Mycoplasma genitalium* [63, 64]. Twenty-eight sub-models representing gene networks, signaling, and metabolism describing functional processes were modeled separately. These sub-models described different areas of cell biology form transport and metabolism to host interaction encompassing DNA replication and maintenance, RNA/protein synthesis, and maturation and cytokinesis. All sub-models were then unified by linking their common inputs and outputs.

Next, the authors verified that the model reproduces the training data. To do that, they simulated 128 wild-type cells in culture environment predicting cell mass and growth rate and also molecular properties as count, localization, and activity of molecules. Their model calculations were consistent with many observed cellular features. Then independent datasets that were not used to construct the model were employed for validation.

Subsequently, the authors used this model to perform different *in silico* predictions: they observed previously undetected cellular behaviors including *in vivo* rates of protein–DNA association and an inverse relationship between the duration of DNA replication initiation and the rate of DNA replication. They also evaluated the global distribution of energy in the cell and performed all

Table 16.2 Tools cited in the text by task.

Task	Tools
PKN construction	PubMed
	Google Scholar
	iHOP
	Reactome
	KEGG Pathway
	IntAct
	ACSN
	Ingenuity Pathway Analysis
	MetaCore
Boolean modeling	BoolSim
	GINsim
	SQUAD
Constrained fuzzy logic	CellNetOptimizer
PKN contextualization	SigNetTrainer
	CellNOptR
	PRUNET
ODEs	SCAMP
	DBSolve
	MIST
	GEPASI
PLDE	Genetic Network Analyzer
	BooleanNet
FBA	MetaNetX
	COBRA Toolbox
	OptFlux
	FAME

possible single-gene *in silico* perturbations to evaluate essential genes required for cellular life. The model agreed with previously observed genes essentiality in 79% of the cases.

The authors demonstrated the feasibility of such hybrid models, and the fact that these models have succeeded in recapitulating a broad set of experimental data and provided insight into several biological processes supports the idea that such an approach should be helpful to tackle the complexity of biological systems.

16.4 Conclusions

Several modeling frameworks were introduced in this chapter (Table 16.2) with some examples of applications. Nevertheless the cited methods do not constitute an exhaustive list of all existing methodologies; we could almost say that each model is unique since the adopted strategy depends on the biological questions, the available data, the network size, and the technical limitations.

Frequently these models miss several significant pieces and constitute only an approximation to reality; however they demonstrated through numerous studies the last decades that they are useful on helping us to better understand the systems for which they are formulated and consequently bring the knowledge closer to what is actually happening in living cells and organisms.

The development and improvement of modeling methods together with the advancements in the high-throughput technologies as single-cell measurements will certainly drive more detailed and more accurate modeling.

References

1 De Smet R, Marchal K. Advantages and limitations of current network inference methods. *Nat Rev Microbiol*. 2010;8(10):717–729.

2 He F, Balling R, Zeng A-P. Reverse engineering and verification of gene networks: principles, assumptions, and limitations of present methods and future perspectives. *J Biotechnol*. 2009;144(3):190–203.

3 Linde J, Schulze S, Henkel SG, Guthke R. Data- and knowledge-based modeling of gene regulatory networks: an update. *Excli J*. 2015:346–378.

4 Eduati F, Corradin A, Di Camillo B, Toffolo G. A Boolean approach to linear prediction for signaling network modeling. *PLoS One*. 2010;5(9):e12789.

5 Yeung KY, Dombek KM, Lo K, et al. Construction of regulatory networks using expression time-series data of a genotyped population. *Proc Natl Acad Sci U S A*. 2011;108(48):19436–19441.

6 Young W, Raftery AE, Yeung K. Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Syst Biol*. 2014;8(1):47.

7 Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. *Nat Genet*. 2005;37(4):382–390.

8 Faith JJ, Hayete B, Thaden JT, et al. Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*. 2007;5(1):e8.

9 Altay G, Emmert-Streib F. Structural influence of gene networks on their inference: analysis of C3NET. *Biol Direct*. 2011;6(1):31.

10 Hoffmann R, Valencia A. A gene network for navigating the literature. *Nat Genet*. 2004;36(7):664.

11 Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*. 2006;7(2):119–129.

12 Croft D, Mundo AF, Haw R, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2014;42(Database issue):D472–D477.

13 Kuperstein I, Bonnet E, Nguyen H-A, et al. Atlas of cancer signalling network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*. 2015;4:e160.

14 Orchard S, Ammari M, Aranda B, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res*. 2014;42(Database issue):D358–D363.

15 Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. *J Theor Biol*. 1969;22(3):437–467.

16 Glass L, Kauffman SA. The logical analysis of continuous, non-linear biochemical control networks. *J Theor Biol*. 1973;39(1):103–129.

17 Garg A, Di Cara A, Xenarios I, Mendoza L, De Micheli G. Synchronous versus asynchronous modeling of gene regulatory networks. *Bioinformatics*. 2008;24(17):1917–1925.

18 Comet J-P, Noual M, Richard A, et al. On circuit functionality in Boolean networks. *Bull Math Biol*. 2013;75(6):906–919.

19 Flobak Å, Baudot A, Remy E, et al. Discovery of drug synergies in gastric cancer cells predicted by logical modeling. *PLoS Comput Biol*. 2015;11(8):e1004426.

20 Peterson H, Abu Dawud R, Garg A, et al. Qualitative modeling identifies IL-11 as a novel regulator in

maintaining self-renewal in human pluripotent stem cells. *Front Physiol.* 2013;4:303.
21 Chaouiya C, Naldi A, Thieffry D. Logical modelling of gene regulatory networks with GINsim. *Methods Mol Biol.* 2012;804:463–479.
22 Di Cara A, Garg A, De Micheli G, Xenarios I, Mendoza L. Dynamic simulation of regulatory networks using SQUAD. *BMC Bioinformatics*. 2007;8(1):462.
23 Naldi A, Monteiro PT, Mussel C, et al. Cooperative development of logical modelling standards and tools with CoLoMoTo. *Bioinformatics*. 2015;31(7):1154–1159.
24 Chasapi A, Wachowicz P, Niknejad A, et al. An extended, Boolean model of the septation initiation network in *S. pombe* provides insights into its regulation. *PLoS One*. 2015;10(8):e0134214.
25 Guex N, Crespo I, Bron S, et al. Angiogenic activity of breast cancer patients' monocytes reverted by combined use of systems modeling and experimental approaches. *PLoS Comput Biol*. 2015;11(3):e1004050.
26 Shmulevich I, Gluhovsky I, Hashimoto RF, Dougherty ER, Zhang W. Steady-state analysis of genetic regulatory networks modelled by probabilistic Boolean networks. *Comp Funct Genomics*. 2003;4(6):601–608.
27 Shmulevich I, Dougherty ER, Kim S, Zhang W. Probabilistic Boolean networks: a rule-based uncertainty model for gene regulatory networks. *Bioinformatics*. 2002;18(2):261–274.
28 Garg A, Mendoza L, Xenarios I, DeMicheli G. Modeling of multiple valued gene regulatory networks. *Conf Proc IEEE Eng Med Biol Soc*. 2007;2007:1398–1404.
29 Aldridge BB, Saez-Rodriguez J, Muhlich JL, Sorger PK, Lauffenburger DA. Fuzzy logic analysis of kinase pathway crosstalk in TNF/EGF/insulin-induced signaling. *PLoS Comput Biol*. 2009;5(4):e1000340.
30 Morris MK, Saez-Rodriguez J, Clarke DC, Sorger PK, Lauffenburger DA. Training signaling pathway maps to biochemical data with constrained fuzzy logic: quantitative analysis of liver cell responses to inflammatory stimuli. *PLoS Comput Biol*. 2011;7(3):e1001099.
31 Mitsos A, Melas IN, Morris MK, Saez-Rodriguez J, Lauffenburger DA, Alexopoulos LG. Non linear programming (NLP) formulation for quantitative modeling of protein signal transduction pathways. *PLoS One*. 2012;7(11):e50085.
32 Keller R, Klein M, Thomas M, et al. Coordinating role of RXRα in downregulating hepatic detoxification during inflammation revealed by fuzzy-logic modeling. *PLoS Comput Biol*. 2016;12(1):e1004431.
33 Terfve C, Cokelaer T, Henriques D, et al. CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms. *BMC Syst Biol.* 2012;6:133.
34 Melas IN, Samaga R, Alexopoulos LG, Klamt S. Detecting and removing inconsistencies between experimental data and signaling network topologies using integer linear programming on interaction graphs. *PLoS Comput Biol*. 2013;9(9):e1003204.
35 Michailidou M, Melas IN, Messinis DE, et al. Network-based analysis of nutraceuticals in human hepatocellular carcinomas reveals mechanisms of chemopreventive action. *CPT Pharmacometrics Syst Pharmacol*. 2015;4(6):350–361.
36 Vaga S, Bernardo-Faura M, Cokelaer T, et al. Phosphoproteomic analyses reveal novel cross-modulation mechanisms between two signaling pathways in yeast. *Mol Syst Biol*. 2014;10(12):767.
37 Rodriguez A, Crespo I, Androsova G, Del Sol A. Discrete logic modelling optimization to contextualize prior knowledge networks using PRUNET. *PLoS One*. 2015;10(6):e0127216.
38 Moes M, Le Béchec A, Crespo I, et al. A novel network integrating a miRNA-203/SNAI1 feedback loop which regulates epithelial to mesenchymal transition. *PLoS One*. 2012;7(4):e35440.
39 Mendoza L, Pardo F. A robust model to describe the differentiation of T-helper cells. *Theory Biosci*. 2010;129(4):283–293.
40 Chang R, Shoemaker R, Wang W. Systematic search for recipes to generate induced pluripotent stem cells. *PLoS Comput Biol*. 2011;7(12):e1002300.
41 Gu Y, Liu G-H, Plongthongkum N, et al. Global DNA methylation and transcriptional analyses of human ESC-derived cardiomyocytes. *Protein Cell*. 2014;5(1):59–68.
42 Shu J, Shuler ML. A mathematical model for the growth of a single cell of *E. coli* on a glucose/glutamine/ammonium medium. *Biotechnol Bioeng*. 1989;33(9):1117–1126.
43 Chen WW, Niepel M, Sorger PK. Classic and contemporary approaches to modeling biochemical reactions. *Genes Dev*. 2010;24(17):1861–1875.
44 Leal Valentim F, Mourik Sv, Posé D, et al. A quantitative and dynamic model of the *Arabidopsis* flowering time gene regulatory network. *PLoS One*. 2015;10(2):e0116973.
45 Li S, Brazhnik P, Sobral B, Tyson JJ. A quantitative study of the division cycle of *Caulobacter crescentus* stalked cells. *PLoS Comput Biol*. 2008;4(1):e9.
46 Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ. Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell*. 2004;15(8):3841–3862.
47 Kraikivski P, Chen KC, Laomettachit T, Murali TM, Tyson JJ. From START to FINISH: computational analysis of cell cycle control in budding yeast. *NPJ Syst Biol Appl.* 2015;1:15016.

48 Hancioglu B, Tyson JJ. A mathematical model of mitotic exit in budding yeast: the role of Polo kinase. *PLoS One*. 2012;7(2):e30810.

49 De Jong H, Gouzé J-L, Hernandez C, Page M, Sari T, Geiselmann J. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol*. 2004;66(2):301–340.

50 de Jong H, Geiselmann J, Hernandez C, Page M. Genetic network analyzer: qualitative simulation of genetic regulatory networks. *Bioinformatics*. 2003;19(3):336–344.

51 Albert I, Thakar J, Li S, Zhang R, Albert R. Boolean network simulations for life scientists. *Source Code Biol Med.* 2008;3:16.

52 Batt G, Ropers D, de Jong H, et al. Validation of qualitative models of genetic regulatory networks by model checking: analysis of the nutritional stress response in *Escherichia coli*. *Bioinformatics.* 2005;21 (Suppl 1):i19–i28.

53 Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol*. 2010;28(3):245–248.

54 Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol*. 2003;14(5): 491–496.

55 Ganter M, Bernard T, Moretti S, Stelling J, Pagni M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics*. 2013;29(6):815–816.

56 Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nat Protoc*. 2007;2(3):727–738.

57 Rocha I, Maia P, Evangelista P, et al. OptFlux: an open-source software platform for in silico metabolic engineering. *BMC Syst Biol.* 2010;4:45.

58 Boele J, Olivier BG, Teusink B. FAME, the flux analysis and modeling environment. *BMC Syst Biol.* 2012;6:8.

59 Machado D, Herrgård M. Systematic evaluation of methods for integration of transcriptomic data into constraint-based models of metabolism. *PLoS Comput Biol*. 2014;10(4):e1003580.

60 Wang Y, Eddy JA, Price ND. Reconstruction of genome-scale metabolic models for 126 human tissues using mCADRE. *BMC Syst Biol.* 2012;6:153.

61 Agren R, Bordel S, Mardinoglu A, Pornputtapong N, Nookaew I, Nielsen J. Reconstruction of genome-scale active metabolic networks for 69 human cell types and 16 cancer types using INIT. *PLoS Comput Biol.* 2012;8(5):e1002518.

62 Montezano D, Meek L, Gupta R, Bermudez LE, Bermudez JCM. Flux balance analysis with objective function defined by proteomics data-metabolism of *Mycobacterium tuberculosis* exposed to Mefloquine. *PLoS One*. 2015;10(7):e0134014.

63 Karr JR, Sanghvi JC, Macklin DN, et al. A whole-cell computational model predicts phenotype from genotype. *Cell*. 2012;150(2):389–401.

64 Macklin DN, Ruggero NA, Covert MW. The future of whole-cell modeling. *Curr Opin Biotechnol.* 2014;28:111–115.

17

Database Creation and Utility

Magdalena Krochmal[1], Katryna Cisek[2], and Holger Husi[3,4]

[1] *Proteomics Laboratory, Biomedical Research Foundation Academy of Athens, Athens, Greece*
[2] *Mosaiques Diagnostics GmbH, Hannover, Germany*
[3] *Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK*
[4] *Department of Diabetes and Cardiovascular Science, Centre for Health Science, University of the Highlands and Islands, Inverness, UK*

17.1 Introduction

The exponential growth and availability of data is a prominent characteristic of our era. In fact, the amount of information produced every day is bigger than ever, and an efficient way to store, access, and process it has become a significant challenge. That is why database systems (DBS) became an essential tool in data management life cycle and are commonly used in all domains where information is highly valued. The emergence of database technology has stimulated the development and progress of many businesses and applications.

Scientific research is one of the fields where databases are used in order to organize information such as scientific publications, molecules, biological pathways, patient records, experimental results, and so on [1]. Databases serve as the foundation for considerable progress in scientific disciplines ranging from computing to biology [2]. With the advent of high-throughput technologies that allow for simultaneous examination of thousands of molecules (genes, proteins, metabolites), adoption of DBS in order to support the process of extraction and analysis became inevitable due to the size and intricacy of the data [3]. In the field of bioinformatics, where computer science and biology meet, knowledge is often inferred from analysis of enormous amounts of complex data, and databases are necessary. The combination of computational and experimental systems biology approaches allows for the elucidation of complex biological mechanisms and offers efficient process modeling tools [4].

This chapter introduces basic concepts of DBS and data integration. Initially, technical aspects of DBS and data models are presented. In particular, relational databases (RDB) are discussed. Subsequently, biological databases and their application in research with special emphasis on -omics data are reviewed.

17.2 Database Systems

17.2.1 Introduction to Databases

The formal definition of a database states that it is a shared collection and description of logically related data, designed to meet the information needs of an organization [5]. With the exponential growth of information volume, databases became essential for data storage and particularly data analysis and reporting. Regardless of the field of application, whether it is for business or research purposes, some general definitions related to DBS are universal, and their knowledge is essential for good understanding of database utility. Therefore, this section presents an overview of the most important technical concepts of DBS and database design.

17.2.2 Data Life Cycle and Objectives of Database Systems

Data life cycle is a term describing the flow of the data in the entire data management process. In scientific research, the purpose of data life cycle management is to support scientific discovery, reliability, and reproducibility through collection of high-quality data that can be effectively managed and reused [6]. Data life cycle starts with the project plan and the definition of the scope of a project. Next, data is collected from various sources and curated to meet the purpose and design criteria. A good study design ensures the availability of data for discovery and reuse in the future. Subsequently, the data and their description (metadata) are stored in a repository from where they can be accessed, analyzed, and interpreted in the discovery step. Lastly, data undergo various quality controls, cataloguing, and classification and are placed into storage. The data life cycle is depicted in Figure 17.1.

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

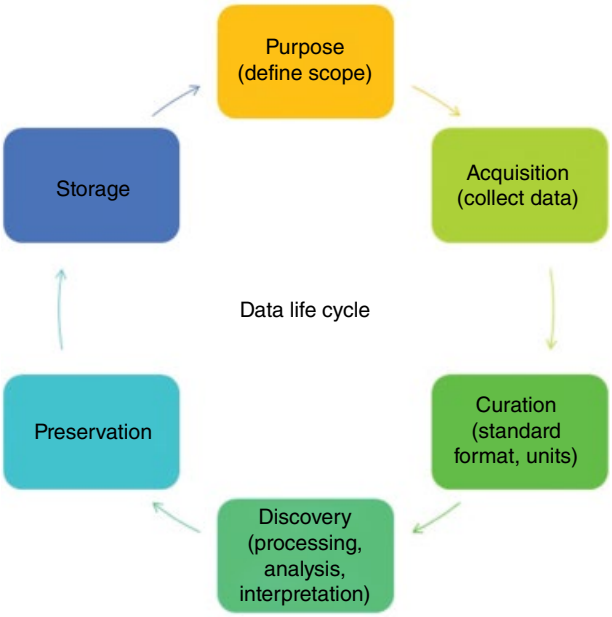


Figure 17.1 Data life cycle. Concept supporting the management of the data and the flow of an information system's data throughout its life cycle: from creation, collection, and discovery process until storage or disposal.

The database is a component of a bigger structure, called DBS, that consists of elements for database management, related hardware, software, and end users. A simplified representation of DBS is illustrated in Figure 17.2. Crucial element in DBS structure, apart from the database itself, is the database management system (DBMS)—a computer software designed for the management of the database. It serves as an interface between end users or software applications and the database, allowing for data querying and manipulation and simultaneously ensuring data accessibility, consistency, and security [7]. With relation to database usage, three main roles can be distinguished, that is, database administrator, application programmers, and end users [5]. The task assigned to database administrator is database design and maintenance such as server and applications upgrade, structure modifications, backup and recovery, performance optimization, and monitoring of system security. Application programmers develop and implement interfaces enabling end users to access and query the database resources according to their requirements. The objective of the DBMS is to ensure that the data is correctly handled at each point of data life cycle. Main goals include ensuring [7]:

- Data availability—Accessibility of data to users
- Data integrity—Maintaining and assuring data accuracy (formats, units, etc.)

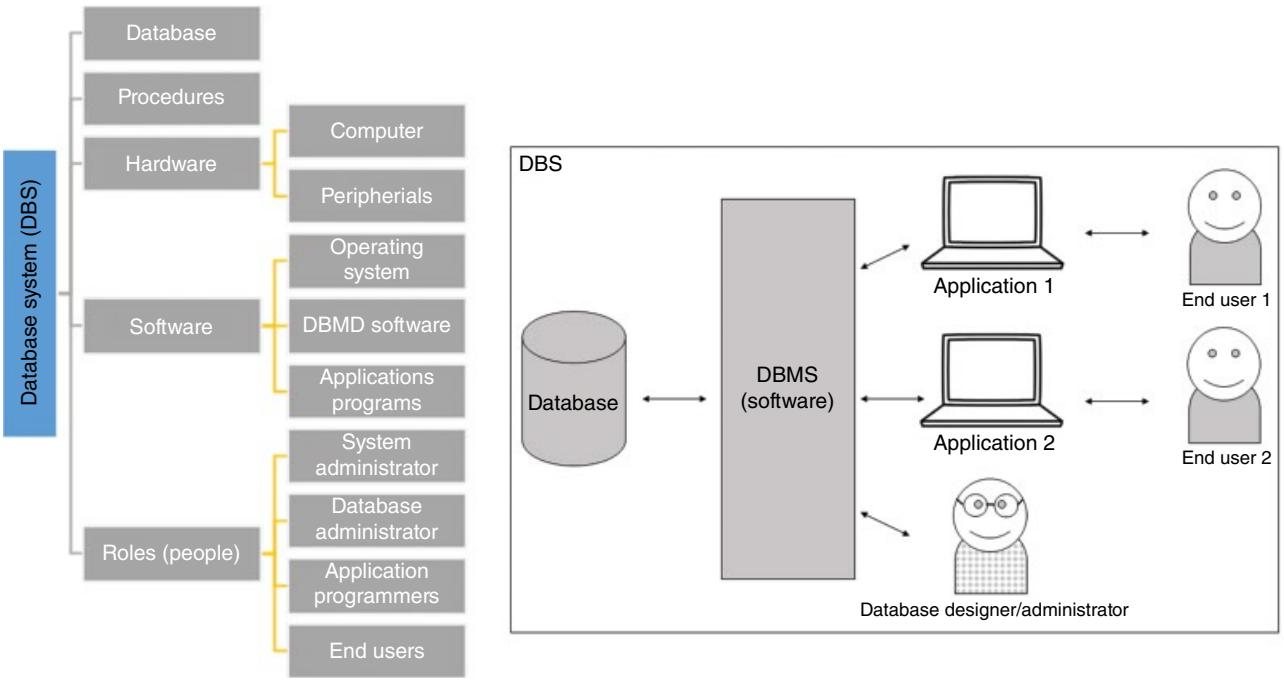


Figure 17.2 Simplified representation of database system (DBS) and components. End users communicate with the software systems/applications, which, in turn, communicate (through the programming interface) with the DBMS. The DBMS communicates with the operating system to store data in and/or extract data from the database.

- Data consistency—Minimizing number of redundancies and duplicates
- Data security—Introducing various levels of security for users, password protection, and so on

17.2.3 Advantages and Limitations

Data management using databases and DBMS has significant advantages. File-based repositories, where the same information might be stored in many files, require a lot of storage space and contain redundant information. Database design aims at data integration, where the problem of redundancy is minimized and controlled by DBMS. Moreover, lack of redundancy reduces the risk of inconsistencies in the data. If any change occurs, there is no need to update each file separately, but the update can be performed only once, resulting in the new value being available to all users at the same time. Importantly, DBMSs ensure data integrity—consistency and validity of information through use of constraints that specify formats that are allowed to be introduced in the database. Another advantage of a centralized database approach is that additional information can be extracted by integrating data. New dependencies can be derived from integration of multiple sources, thus bringing an added value compared with analysis of each source independently. Additionally, DBMSs enable efficient sharing of the data, as they are stored in one central repository. Moreover, DBMSs provide protection of the database from unauthorized users (through roles, passwords, or logs). Data security is especially important for confidential data. Lastly, an undeniable benefit of databases is cost reduction through integration of multiple sources into one repository, which is of value especially to large organizations, where maintenance of one central system might have a lower total cost than several small systems running independently [7].

Apart from the advantages of DBMS, some drawbacks have to be pointed out. The process of designing the database, its maintenance, and usage is quite complex. To take full advantage of the system, it is vital for all people involved in the process (from designers to end users) to understand its structure and functionality. Failing to properly design and use all available tools can result in unforeseen problems and limited performance. Additionally, adoption of DBMS is costly in memory size required for the software to work efficiently.

17.2.4 Database Design Models

In order to better understand the evolution of database design, it is of importance to understand the hierarchy of data. The basic building blocks of the database are "raw" facts (*data*) that have none or little meaning when not organized in a logical manner. What defines the "raw" information is the value of its *field* (e.g., a field might define a telephone number, price, or date). Many related *fields* consist a *record*, which is a set of connected values that describe a particular item (e.g., a record of an employee will consist of fields such as name, birthday, salary, position, etc.). A collection of all (e.g., a file of all students enrolled at the university) is called a *file*. Multiple files can be integrated into a database, which is at the top of the data hierarchy tree. Figure 17.3 depicts the data hierarchy levels.

In the course of years, approaches of database design have evolved from simple file-based system toward more advanced and elegant solutions. The three most common implementation database models are hierarchical, network, and relational models (Figure 17.4). Hierarchical and network databases have been substituted by set-oriented RDB, which are now transforming into object-oriented and multimedia DBS [5]. Moreover, non-RDB models emerged to address the requirements of new applications and handle the explosion in the volume and variety of data in recent years. In the history of database evolution, the following approaches have been adopted:

1) File-based system
   The very first attempt used in the past to store data in an organized manner resulted in databases in the form of separate files in folders [8]. This system performed adequately for storage and retrieval of small datasets, but had many limitations, as it required a significant effort to extract, process, or cross-reference information, due to the fact that data might be stored in separate files and formats. Major limitations included data duplication (which is undesirable due to excessive use of storage space, loss of data integrity, and consistency), data dependence (difficult to change once defined structure), incompatible file formats (structure dependent on programming language used for the development), and reliance on the developer for reporting (fixed queries, difficult to adjust reports to users' needs). Additionally, lack of data security, integrity, and possibility to access the data simultaneously by several users led to the development of alternative database solutions.
2) Hierarchical data structure
   Hierarchical databases follow a treelike data structure. Records are linked from top to bottom following strict hierarchy rules. It is an example of one-to-many (1 : M) relationship, where each parent can have many children, but each child has only one parent. Hierarchical databases are suitable for simple data representation. Among the advantages, data security,

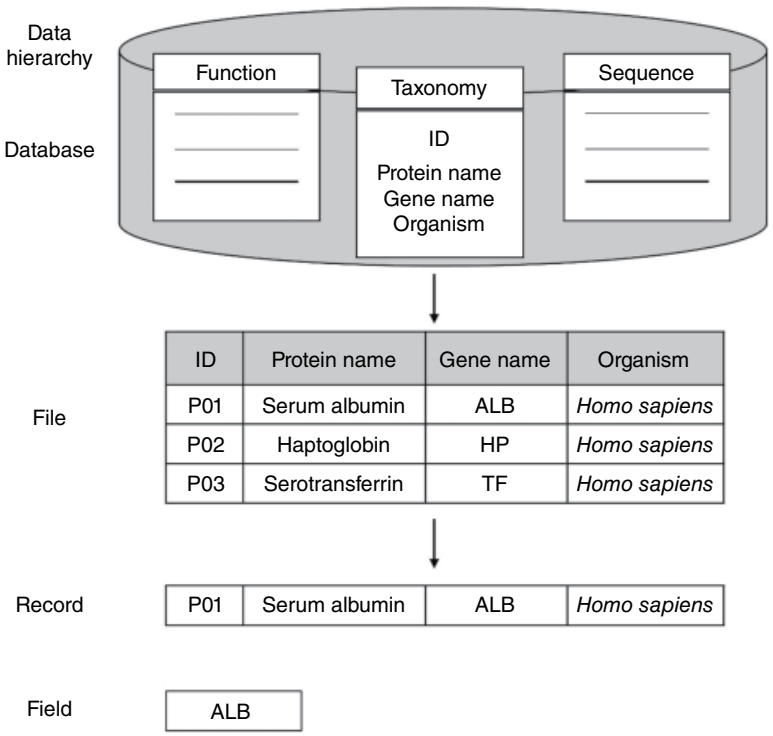


Figure 17.3 Hierarchy of data. Database consists of integrated files, which are a collection of related records. Each record is a collection of related fields, where single facts or attributes are stored.

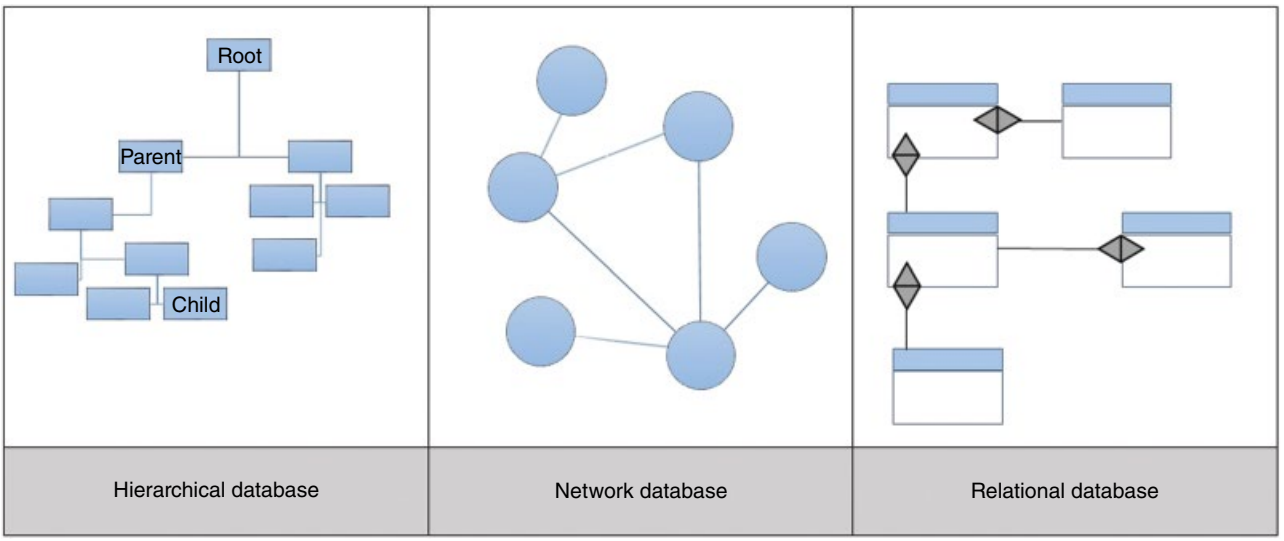


Figure 17.4 Schematic representation of hierarchical, network, and relational database models.

independence, and integrity can be pointed out. Unfortunately, this type of database is quite complex and inflexible in management and requires the knowledge of its physical structure. Implementation of relationships other than one-to-many is not possible, and development of applications is often time consuming and complicated.

3) Network data structure
Network databases are similar to hierarchical model but provide more flexibility by means of the design of existing relationships. Each relationship (called a *set*) is a composition of at least two entities—an *owner* (parent) and a *member* (child) record, where many-to-many (M:M) relationships are allowed. In comparison with hierarchical models, network systems allow the modeling of more natural relationships between entities.

4) Relational data structure
The relational model following the well-established relational database management system (RDBMS)

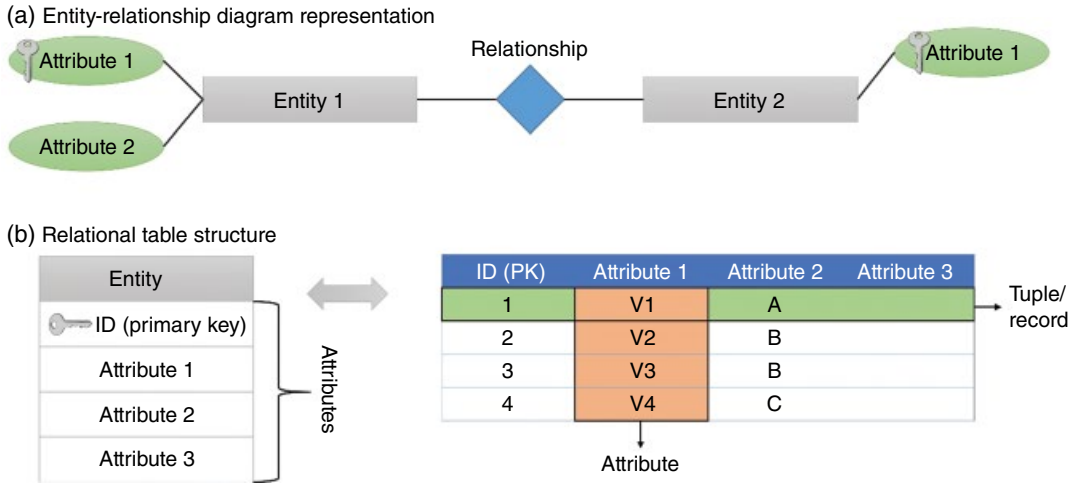


Figure 17.5 Basic concepts of relational model. (a) Database entities (tables) are matched based on data in key columns, forming a relationship. (b) Relational model organizes data into tables (relations) of columns and rows, with a unique identifier of each row (primary key).

principles is currently the most widely used in database design. The concept of RDB was introduced by Edgar E. Codd in 1970. The strength of this approach lies in solid fundaments founded on mathematical principles—theory of sets and linear algebra. Basic concepts of RDB (illustrated in Figure 17.5) [7] are as follows:

- Data is organized in tables, where each table represents an object, concept, or thing (called an *entity*).
- Each entity contains a set of properties that describe it (*attributes*).
- Dependencies between entities are represented by *relationships*.
- Relationships can take three logical forms: one-to-one, one-to-many, and many-to-many.
- *Relation* is a set of entities and their relationships.
- *Tuple*/*record* is one row in a table.
- An attribute or combination of attributes that uniquely identify a tuple is called a *primary key*.
- Structured Query Language (SQL) is the programming language to access and manipulate data in the database.

Importantly, in RDB, data is only stored once, assuring data consistency, easy management (updates, deletions), and efficient storage. Additionally, data stored and organized in tables is much more comprehensive and easy to understand. Scalability is another advantage of RDB: new data can be easily added without the need to modify existing records. In addition, the usage of relational algebra in the database queries ensures that there is no ambiguity, which may be a problem in network-type databases.

Drawbacks that can be pointed out include the complexity of an RDB and difficulty of creation regarding definition of entities, relationships, and constraints. Furthermore, usage of such database requires the knowledge of its structure and relationships, which in some cases can be intricate. Nevertheless, it is certainly worth putting a significant effort in the design phase. Well-designed database limits the number of errors in the data and significantly decreases the amount of information that has to be inserted, because of redundancy control. Additionally, the capabilities of DBMS systems provide great performance and tools to manage data effectively.

5) Non-relational (distributed) structure

The concept of non-relational (NoSQL) databases has emerged quite recently, as a response to ever-growing volumes of data and the need of more flexible ways of data management and storage, as compared with RDB (described earlier). The term NoSQL was first used in 1998 for an RDB that omitted the use of SQL and was ultimately assigned to all databases following non-relational structure [9]. Therefore, NoSQL is used as an umbrella term for collection of concepts about data storage and manipulation in all databases that do not follow the popular, rigid principles of the relational model. NoSQL databases feature flexibility, scalability, and better performance, compromising transactional integrity and efficient data querying (no support for joins and order by operations) [10]. Non-RDB structure types include (Figure 17.6):

- Key-value stores (hash tables)—Database structure designed for storage of associative arrays in a form of couples (key, value). As in RDB, a key is a unique identifier of assigned value, whereas the value can be either a structured or completely unstructured element (i.e., binary large object (BLOB), e.g., image, audio). Implementation of the key-value model is

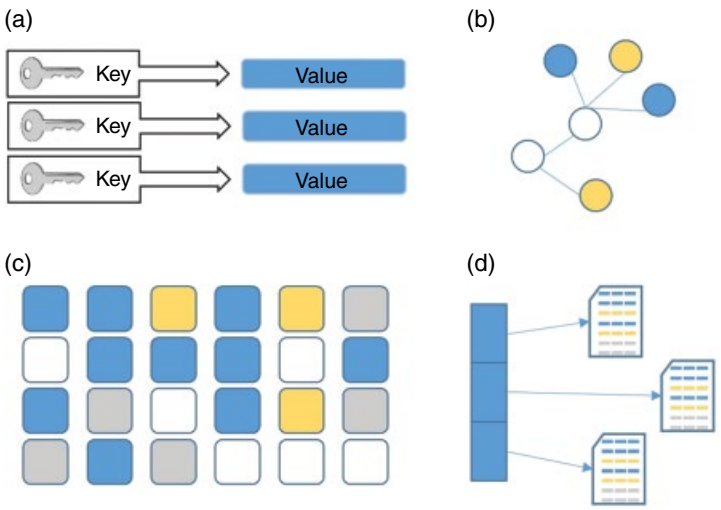


Figure 17.6 Non-SQL database types. (a) Key-value stores—key is a unique attribute of the content (value). (b) Graph databases—storage of interconnected data represented by nodes and edges. (c) Column family stores—key-value pair, where the key is mapped to a value that is a set of columns. (d) Document database—data is stored in the form of multiple documents (e.g., JSON, XML), which contain more complex grouping of key-value pairs.

considered to be the simplest among non-RDB types. Great scalability and performance are among the main advantages of key-value stores; however queries and update operations on parts of the value are often inefficient.

Examples: Redis (www.redis.io), Project Voldemort (www.project-voldemort.com)

- Column family stores—NoSQL object containing columns of related data. Similarly to key-value stores, unique identifier (key) points to value, this is distributed among a set of columns. The columns are arranged by column family and given the schema-less structure; each row can contain a different number of columns. Therefore column family stores are excellent for storing and processing large amounts of data, distributed among many machines. These databases are highly scalable and offer high query performance.

  Examples: Google Bigtable (https://cloud.google.com/bigtable/)
- Graph databases—Databases use graph structures and graph theory—study of graphs to model pairwise relations between objects. A graph is a set of nodes (which represent data entities, e.g., employees, accounts), which are connected by edges (depicting relationships between nodes). Graph databases, with use of relationships, are designed to allow simple and rapid retrieval of complex hierarchical structures that are difficult to model in relational systems.

  Examples: Neo4J (www.neo4j.com), InfiniteGraph (www.objectivity.com/products/infinitegraph/)
- Document databases—In document databases data is stored in documents, and typically each document contains all information about the record and its associated data to avoid splitting the document into smaller key/value pairs. The semi-structured documents are usually stored in formats like JSON (JavaScript Object Notation) or XML (Extensible Markup Language). Document databases are considered an advanced version of key-value stores where nested values associated with each key are allowed.

  Example: MongoDB (www.mongodb.com), Apache CouchDB (couchdb.apache.org)

17.2.5 Development Life Cycle

An important aspect of software engineering the development plan, which assures that the whole process from design to implementation is performed correctly. The development plan is a collection of small steps, each focusing on a separate aspect of the process. Phases of database development are presented in Figure 17.7. Depicted schema can be applied to any database model, that is, hierarchical, network, or relational.

The design process is initiated by defining the requirements of data collection and analysis. The mission statement and aims of the DBS are defined. This step helps describe precisely the purpose of the development and plan the following steps of the life cycle. Consecutively, database design starts and consists of three phases: conceptual design, logical design, and physical design. Conceptual (semantic) data model is derived from thorough data analysis. The aim of this phase is to gather information about the data in order to design an optimal database structure that fits the requirements of users. Semantic model provides information of formal data structure and existing constraints, irrespective of design approach, system, and software. In the logical database design phase, conceptual design is used to define which data model will be applied. Logical schema, the output of

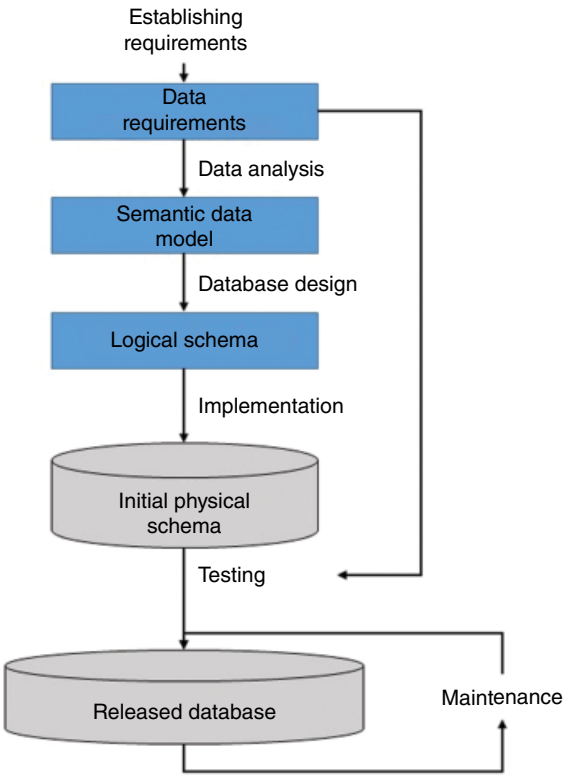


Figure 17.7 Database design process.

this step, describes all tables, primary and foreign keys, attributes, and constraints enforced by requirements. The most representative implementation of semantic model has to be chosen, taking into consideration various design criteria such as control of duplication, flexibility of change, efficiency, and usability. Based on the final logical schema, the physical database can be created within the respective DBMS. The subsequent implementation steps vary with regard to the type of particular DBMS used, but in general involve specification of storage schema, security measures, and user roles. Once the database is created (all defined tables and constraints are implemented), data can be populated. This process can either be initiated by transfer of existing data or the use of applications for the database. Before the database can be released to public use, it needs testing in order to identify any errors and assure that it is compliant with the requirements. The final version should be under constant monitoring and maintenance process to ensure best performance [11].

17.2.6 Database Transactions, Structured Query Language (SQL)

A unit of work performed on a database is called a transaction. Transaction refers to a sequence of actions performed in logical order and represents any change in the database. The main purpose of transactions is to assure that any action performed on data is safe and will remain consistent in case of system failure. Additionally, transactions isolate programs accessing the database to avoid errors caused by concurrent processes. Transactions are described by four main properties (acronym ACID) that assure their proper execution:

- **A**tomicity—Ensures that a single unit of work is either completed successfully or all changes are aborted and rolled back to the initial state
- **C**onsistency—Ensures that any transaction will bring database from one state to another with respect to all rules and constraints stated in the definition
- **I**solation—Controls the execution of concurrent transactions, assuring that the effects of an incomplete transaction might not affect another transaction
- **D**urability—Ensures that once a transaction has been committed, changes are stored permanently (even after database failure, power loss, etc.)

SQL is a standard programming language for accessing and manipulating databases. In RDB it is used to retrieve and update data in tables. Slightly different types of SQL versions exist, depending on the DBMS vendor. Most common types include:

- PL/SQL—Procedural Language/Structured Query Language by Oracle Corporation
- TSQL—Transact-SQL (T-SQL) by Microsoft and Sybase
- PL/pgSQL—Procedural Language/PostgreSQL by the PostgreSQL

In general, all versions have to be American National Standards Institute (ANSI) standard compliant and support major keywords such as SELECT, UPDATE, DELETE, INSERT, ALTER, and others. SQL commands can be divided into three functional groups: Data Definition Language (DDL), Data Manipulation Language (DML), and Data Control Language (DCL), respectively. DDL statements concern database schemas, creation of tables, and descriptions. DML keywords are the most common statements for data retrieval, deletion, or update. Lastly, DCL commands are used to set rights and permissions to database users.

17.2.7 Data Analysis and Visualization

The ultimate goal of development of the database, apart from the need of a central data repository, is to extract useful information from the data. Databases come with a number of powerful tools that support data analysis and visualization of the results, helping to gain valuable knowledge from big volumes of data.

In the business vocabulary, the set of techniques aiming at transformation of raw data into meaningful information is called business intelligence (BI). Analysis of the data may include simple SQL queries and statistical analysis, as well as complex multidimensional analysis and data mining (Knowledge Discovery in Databases (KDD)). To support the process there are multiple software tools available (open source and commercial), with the most common example of Excel spreadsheets. Moreover, for more sophisticated analysis, there is a possibility to use different programming languages to perform analyses and calculations adjusted to the needs and requirements of the study. Database exploration and discovery is also possible through visual presentation of data in the form of graphs, charts, tables, and so on [12]. Visual model outcome is often much more informative for human's perception [13] and can provide ad hoc conclusions about the data quality or model performance.

Importantly, data analysis for scientific applications is far more complex as each time different research questions have to be answered; thus there is no universal approach to tackle all the problems. Implementation of good scientific software is hard and requires in-depth knowledge of the data and reasoning of the processes. There are different programming environments (often open source) that come with a number of libraries and packages for data manipulation and processing, some offering also visualization capabilities. Among the most commonly used are the commercially available Matlab and SAS and open-source Octave, R, Python (with NumPy/SciPy libraries) [14]. A non-exhaustive list of available programming environments presented in Table 17.1.

17.3 Biological Databases

Growing number of biological databases is the natural consequence of the advent of high-throughput technologies and significant decrease of cost of such experiments. Biological databases not only serve as a data storage also, more importantly, help with data organization improving data retrieval, visualization, and analysis. Moreover, data repositories facilitate data sharing and exchange, which is crucial for scientific research [2]. Lastly, this form of storage allows for integration of information from various sources in an automated manner, giving an opportunity to model complex biological processes and mine for intrinsic knowledge.

Databases in research cover diverse scientific topics and might be classified based on many criteria. With regard to data coverage, comprehensive and specialized databases can be distinguished. Comprehensive resources contain data of general interest, whereas specialized databases focus only on specific type of data (e.g., specific disease) or organism. Additionally, biological repositories can be divided with regard to level of biocuration, that is, primary databases containing raw data and secondary, containing processed information. Finally, databases can be characterized according to the type of biological information stored. Most general categories are DNA, RNA, protein, expression, pathway, disease, nomenclature, literature and standard, and ontology databases [2]. Examples of databases following this classification are presented in Table 17.2.

Table 17.1 List of available data analysis software for scientific applications.

	Software	Link	Description
Commercial	Matlab	www.mathworks.com/products/matlab/	Technical computing language and interactive environment for algorithm development, data visualization, data analysis, and simulation
	SAS	www.sas.com	Software suite for advanced analytics, multivariate analyses, business intelligence, data management, and predictive analytics
	SPSS	www.ibm.com/software/analytics/spss/	Software package used for statistical analysis, developed by IBM
Open source	R	www.r-project.org	Programming language and software environment for statistical computing, data mining, and visualization
	Octave	www.gnu.org/software/octave/	High-level interpreted language, primarily intended for numerical computations, data manipulation, and visualization
	Python	www.python.org	SciPy and NumPy libraries—software for mathematics, science, and engineering

Table 17.2 Types of biological databases.

Type of database	Example
DNA	GenBank [15], GeneCards [16]
RNA	RNAcentral [17], miRBase [18]
Protein	UniProt [19], Protein Data Bank (PDB) [20]
Expression	Human Protein Atlas [21], TiGER [22]
Pathway	Reactome [23], KEGG [24]
Disease	MalaCards [25], CKDdb [26], peptiCKDdb [27], Nephroseq [28]
Literature	PubMed [29]
Standard and ontologies	Gene Ontology [30], HGNC [31]

17.3.1 Development Life Cycle

Development of scientific databases is not straightforward as design has to be adjusted to the type of biological data. Thus, the design process requires not only experience in database design but also biological knowledge and well-defined needs and expectations that the resource has to meet.

17.3.1.1 Data Extraction

Semantic searching is an integral part of data retrieval, collection, and curation. Especially in the field of biology and specifically in omics studies, data originate from different sources and in different formats and nomenclature is not standardized or harmonized. These drawbacks make it critical to search for data by contextual meaning rather than individual terms in order to extract the maximum of relevant information, regardless of the nomenclature. To this effect semantic searching includes different query options, from general to focused, as well as searching by concepts, terms, synonyms, and any variations on these themes.

In the same way, efficient data compilation requires proper infrastructure for effective data storage and management. Semantic Web (SW) technologies fulfill this requirement with a standardized framework and furthermore, allow widespread public open access on the World Wide Web. SW technologies employ a standard data modeling language called Resource Description Framework (RDF), its own query language SPARQL Protocol and RDF Query Language (SPARQL), as well as its own schema language to represent knowledge and define concepts called Web Ontology Language (OWL). Semantic knowledge bases (KB) offer integrated solutions where the data is organized, harmonized, and interlinked and easy to use with the aforementioned framework for formulating new hypotheses [32].

17.3.1.2 Semantic Tools for -Omics

Expert Protein Analysis System (ExPASy) is a resource portal from the Swiss Institute of Bioinformatics (SIB), who also developed and maintained the UniProtKB/Swiss-Prot database. The ExPASy portal compiles many systems biology and -omics resources and tools, including Basic Local Alignment Search Tool (BLAST) for nucleic acid and protein sequence searching and Mascot for protein identification from mass spectrometry data, as well as the Systems Biology Research Tool (SBRT), an integrated and easy-to-use open-source application program interface (API) capable of supporting various plug-ins. In contrast to the ExPASy portal, which is merely a common entry point for resources developed and maintained by many different groups, SBRT is an integrated platform capable of performing over 35 different methods or functions for analyzing stoichiometric networks (e.g., identifying reaction pathways) across topics such as graph theory, algebra, geometry, and statistics, with additional features available via process plug-ins. This integrated platform can perform multiple computational processes controlled via a text-based input file or command line and also interface with other external packages, such as Mathematica, R, GLPK, Xerces, and Metatool [33].

R statistical computing open-source software implements Bioconductor, a platform of over 900 packages for the bioinformatic handling and statistical analysis of high-throughput genomic data, including DNA, RNA, chromatin immunoprecipitation, three-dimensional architecture of genomes using Hi-C, methylome and ribosome profiling, as well as tools for microarray, proteomic, metabolomic, flow cytometry, and quantitative imaging. The highly integrative environment of Bioconductor within the R statistical software enables the user to create complex workflows with multiple inputs and data types, which can then be submitted as workflow vignettes to the Bioconductor user community, or alternatively the user can adapt an already publicly available Bioconductor workflow. A popular workflow is the "RNA-seq workflow: gene-level exploratory analysis and differential expression," which includes data input matrix preparation, exploratory analysis and visualization, differential expression analysis, plotting results, annotating and exporting results, removing hidden batch effects, time course experiments, and session information for the whole workflow. Additional packages, such as SGSeq, which is used for prediction, quantification, and visualization of alternative transcript events from RNA-seq data, can be adapted into the existing workflow for customization [34].

Bioconductor packages can also directly link with publicly available data sources or databases, such as the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO), a public repository of

microarray data. GEOquery (https://www.bioconductor.org/packages/release/bioc/html/GEOquery.html) is a tool that links GEO and Bioconductor and allows the user to search for and download a particular GEO dataset and convert it to Bioconductor compatible format. Another tool, GEOsubmission (https://www.bioconductor.org/packages/release/bioc/html/GEOsubmission.html), prepares microarray data for submission to GEO. In order to do a more complex data search by accessing the GEO metadata information for samples, platforms, and datasets, the users can utilize package GEOmetadb [35] or GeoSearch [36], which expands the search terms to find all gene names and their aliases and outputs a summary of search results with common biology keywords. Lastly, geoR (http://www.leg.ufpr.br/geoR/) incorporates a set of functions for geostatistical analysis in various categories including data preparation, descriptive analysis, empirical variogram (data relation/correction with Euclidean distance), variogram results/plots/lines, nonparametric variogram fitting, profile likelihood, kriging, simulation of Gaussian random fields, and other functions including traditional, likelihood-based, and Bayesian inference/prediction for Gaussian and transformed Gaussian models.

Cytoscape (www.cytoscape.org) is an open-source integrated platform for analysis and visualization of complex network data by using various apps. Information from publications is integrated using Agilent Literature Search (http://apps.cytoscape.org/apps/agilentliteratureresearch), which mines literature using semantic querying (multiple terms, aliases, concept lexicons, e.g., species) and user-selected search engines. The molecule associations collected by this app can then be used to create interaction networks, or the literature results can be mapped to a de novo network to provide further support in the context of a biological process or disease. MiMIplugin (http://apps.cytoscape.org/apps/mimiplugin) collects and merges data from protein interaction databases including Michigan Molecular Interactions (MiMI), BIND, DIP, HPRD, RefSeq, Swiss-Prot, IPI, and CCSB-HI1 along with molecular interactions and their attributes for network generation. The app is also integrated with other NCBI tools for literature information, document summary creation, and pathway matching. Another network app, ConsensusPathDBplugin (http://apps.cytoscape.org/apps/consensuspathdbplugin), attempts to resolve inconsistencies between different databases at the level of complex protein–protein, genetic, metabolic, signaling, gene regulatory, and drug–target interactions and biochemical pathways. In addition to generating non-redundant and seamless consensus pathways from 32 public resources relative to *Homo sapiens*. The tool is also available for data from other species, such as yeast and mouse.

17.3.2 Existing Biological Repositories

17.3.2.1 Information Sources for -Omics

GeneCards (http://www.genecards.org/) is an integrated database of human genes that includes automatically mined genomics, proteomics, and transcriptomics information, as well as orthologies, disease relationships, single nucleotide polymorphisms (SNPs), gene expression, gene function, and service links for ordering assays and antibodies. The GeneCards database, originated in 1997, is being developed and maintained by the Crown Human Genome Center at the Weizmann Institute of Science. As of 2015 it contains a total of 152 704 genes, out of which 21 965 are protein coding, 16 329 pseudogenes, 1 754 genetic loci, 134 gene clusters, and 5 473 uncategorized entries. Moreover, GeneCards is affiliated with other databases such as MalaCards (a human disease database), LifeMap (a discovery database), PathCards (a pathway unification database), and GeneLoc (a genome locator database). Analysis tools integrated into GeneCards include GeneAnalytics (a gene set analyzer), VarElect (NGS phenotyper), GeneALaCart (a GeneCards batch query tool), and GenesLikeMe (a related genes finder). The GeneCards Suite of interlinked databases and analysis tools is one of the most comprehensive sources relative to genetics.

The NCBI (http://www.ncbi.nlm.nih.gov/) is a portal to a plethora of biomedical and genomics resources including literature, health, genomes, genes, proteins, and chemicals and encompasses widely used -omics databases, such as GEO and Online Mendelian Inheritance in Man (OMIM). GEO compiles microarray, next-generation sequencing, and other forms of high-throughput functional genomics data from the research community in compliance with grants or journals for open access to the data. The data stored includes raw data, processed data, and descriptive metadata, all indexed, cross-linked, and searchable. About 90% of the datasets comprise gene expression studies focusing on disease, development, evolution, immunity, ecology, toxicology, and metabolism. The rest of the datasets in GEO are functional genomics and epigenomics studies elucidating genome methylation, chromatin structure, genome copy number variations, and genome–protein interactions. Data explorers, analyzers, and visualizers are also available as extra features for GEO data discovery. Although OMIM is curated by McKusick–Nathans Institute of Genetic Medicine, the Johns Hopkins University School of Medicine, it is considered a phenotypic companion to the Human Genome Project, a National Institutes of Health (NIH) initiative like NCBI. OMIM is a catalog of human genes and genetic disorders, based on published literature, and therefore cross-indexed with PubMed. The records include information on clinical features, inheritance,

population genetics, heterogeneity, genotype/phenotype correlations, cloning, gene structure, gene function, mapping, and so on, attempting to elucidate relationships between genetic variation and phenotypic expression.

The ExPASy gateway comprehensive searches can be performed in various scientific databases maintained and curated by SIB, such as Eukaryotic Promoter Database (EPD), miROrtho (the catalog of animal microRNA genes), MyHits (protein sequences and motifs), PROSITE (protein families and domains), Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (protein interaction networks for systems biology), SWISS-2DPAGE (2D gel database), SWISS-MODEL Repository (three-dimensional protein structure models), and UniProtKB/Swiss-Prot (a curated protein sequence database providing a high level of up-to-date annotation). UniProtKB is one of the most comprehensive resources for protein information, encompassing data from the literature as well as verified computational analysis. For each sequence in UniProtKB/Swiss-Prot, gene and species information is harmonized, and discrepancies between sequences are annotated for alternative splicing, natural variation, incorrect initiation sites, incorrect exon boundaries, frameshifts, and so on. Computational predictions, including posttranslational modifications, transmembrane domains and topology, signal peptides, domain identification, and protein family classification are manually evaluated. Automatically annotated entries are part of UniProtKB/TrEMBL and await manual curation. UniRef, UniParc, Proteomes, and Supporting data including literature citations, taxonomy, subcellular locations, cross-ref databases, diseases, and keywords are fully integrated in UniProtKB.

17.3.2.2 Renal Information Sources for -Omics

In nephrology, several databases have been developed to collect information for computational modeling, including repositories focused on one -omics type, such as peptiCKDdb and Nephroseq, and multi-omics resources, such as the Chronic Kidney Disease database (CKDdb), the Kidney and Urinary Pathway Knowledge Base (KUPKB), and GeneKid.

The peptiCKDdb (www.peptiCKDdb.com) is a repository of mined peptidomics and proteomics datasets originating from scientific literature related to chronic kidney disease (CKD). It can serve as a knowledge base for scientists seeking confirmation of their findings, as well as a source of data for integrative analysis supporting biomarker research in the field of renal pathology. This resource currently stores data from a total of 119 publications. The main features include user-friendly interface for fast and easy browsing of the records, multiparametric search engine, results visualization, and data export features.

Nephroseq (www.nephroseq.org), as the name suggests, is a data mining engine of pre-analyzed clinical and molecular transcriptomics datasets of kidney disease and its comorbidities from human and mouse studies. Currently, there are two datasets from renal gene expression experiments constituting the analytical base of the resource. In addition to being a database with gene expression profiles for molecules of interest, this resource also integrates Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, predicted microRNA targets, and Human Protein Reference Database (HPRD) Interaction Sets and allows for co-expression, outlier, heterogeneity, and concept analysis that enables meta-analysis of gene expression trends.

The CKDdb (www.padb.chem.gla.ac.uk/ckddb/) stores microRNA, genomics, peptidomics, proteomics, and metabolomics information relevant to CKD, collected from over 300 studies in the literature and integrated into the Pan-omics Analysis DataBase (PADB) using gene and protein clusters (CluSO) and mapping of orthologous genes (OMAP) between species. This resource integrates highly diverse omics data across various species in one platform and allows for a systematic evaluation of CKD-relevant pathways using a systems biology approach.

The KUPKB (www.kupkb.org) compiles mRNA, miRNA, metabolite, and protein datasets from literature as well as GEO relevant to kidney pathology and physiology; this information is implemented using SW technologies to standardize content published and shared on the Internet. Moreover, KUPKB is linked to additional resources, such as NCBI gene, UniProt, HomoloGene, and KEGG, allowing complex queries to return all the relevant linked information, across species and including biological pathways. Additionally, KUPKB is equipped with the network visualization tool, KUPNetViz, which allows for integration of experimental data from KUPKB (such as renal locations and diseases) and external resources (Gene Ontology, KEGG, miRNA data), providing a comprehensive and informative network view.

Lastly, GeneKid, a pipeline created for the SysKid consortium project, which aims to develop new diagnostics and treatments for CKD, focuses on harmonizing heterogeneous omics data by using the genes' annotation network ("symbolization") to build a unified omics network. The challenge of this approach is assigning all nonunique gene identifiers to one correct Human Genome Organisation (HUGO) symbol, made difficult by nomenclature variability between laboratories, especially for linking genes with cellular metabolites; to improve this linkage, symbolization is augmented using the Human Metabolome Database (HMDB) and DrugBank database.

17.3.3 Application in Research

17.3.3.1 Data Mining on Large Multi-Omics Datasets

Omics data encompasses more than just genomics, transcriptomics, proteomics, peptidomics, and metabolomics; there are many other fields, such as epigenomics and phenomics, that focus on gene regulation and environmental interactions, respectively. The integration of data from these fields will allow the molecular characterization of diseases and ultimately lead to personalized medicine [37]. However, data mining large omics datasets is more than just a computational issue, because the real challenge is gaining biological insight from large-scale high-throughput experiments. This is because each omics does not function alone; genes, transcripts, proteins, metabolites, and so on are all part of an intricate network of interactions, and this interactome is dynamic in nature and responds to environmental stimuli [38]. Therefore, even though each of these approaches can generate vast amounts of data, each omics is merely a component of the whole system and must be integrated in order to achieve a global perspective, especially when attempting to decipher mechanisms, leading to diseases.

Disease etiologies must be deciphered in comparison with a healthy interactome. Thus, specific biological processes, pathways, and interactions leading to disease can be identified, and possible therapeutic interventions can be inferred [39]. Omics data integration starts by compiling all the data by mining various databases, such as scientific literature, knowledge databases, and sequence and structural databases, as well as assigning functional, regulatory, dynamic, or metadata content. This is due to the complexity of the various omics data types, which prevents the collection, deposition, and linkage of all omics data in a common database, or even in databases sharing a common architecture. Standardization and universal linkage of all types of omics data into a single interlinked data structure is one of the main challenges of systems biology [40].

17.3.3.2 Multi-Omics Tools for Researchers

The discovery of dysregulated processes or pathways in disease is the key to successful prediction of novel drug targets and treatments. In the last two decades, there has been an explosion of *in silico* tools for computational analysis and pathway integration of omics data [41], both commercially available, including Ingenuity Pathway Analysis (IPA) (http://www.ingenuity.com/) and MetaCore (http://lsresearch.thomsonreuters.com/), and freeware, such as Cytoscape (http://www.cytoscape.org/) and InCroMAP (http://www.ra.cs.uni-tuebingen.de/software/InCroMAP/), as well as online web tools such as 3omics (http://3omics.cmdm.tw/) and IMPaLA (http://impala.molgen.mpg.de/). Although each tool may implement different methodological approaches (e.g., Cytoscape is made up of various plug-ins implemented in a common platform, while 3omics is a web-accessible tool) and algorithms (e.g., each Cytoscape plug-in has its own independent algorithm, while 3omics uses a text mining algorithm to merge literature data). However, the common characteristics of all integration tools are that they are based on or linked to an underlying database that stores information about known cellular and signaling and biochemical pathways [42, 43].

While commercially available tools such as IPA and MetaCore have been developed with their own proprietary databases, most freeware, such as Cytoscape and InCroMAP, utilize public databases in their underlying framework. The most known pathway databases include KEGG (http://www.genome.jp/kegg/), WikiPathways (http://www.wikipathways.org/), and Reactome (http://www.reactome.org/). Disease-related pathway information from all of these databases can be accessed using the Cytoscape platform's various applications. CyKEGGParser (http://apps.cytoscape.org/apps/cykeggparser) is an application that not only accesses and visualizes KEGG pathway information but is also capable of merging, correcting, and editing pathways (tuning of protein–protein interactions within pathway maps). The WikiPathways (http://apps.cytoscape.org/apps/wikipathways) application allows for importing, visualizing, querying, and merging of pathways, along with generating customized networks. The ReactomeFIPlugIn (http://apps.cytoscape.org/apps/reactomefiplugin) application contains functions for pathway enrichment and functional relationships for a set of genes, constructs customized networks based on complex queries, and performs automated or manual annotations. InCroMAP is a stand-alone software capable of performing analysis on mRNA, miRNA, DNA methylation, and protein modifications; however it only supports KEGG pathway information. 3omics is a web-accessible data mining tool for integrating transcriptomics, proteomics, and metabolomics and can extrapolate missing omics data; it also only implements KEGG pathway data. However, IMPaLA, also an online tool, lacks graphical visualization but incorporates information from many databases, making it a very comprehensive tool.

17.3.3.3 Limitations of Multi-Omics Tools

While tools such as IMPaLA aim to consolidate many different knowledge sources and databases into a comprehensive model, a lot of redundancies and inconsistencies within the information are generated. This is due to the lack of standardization and harmonization of the

vast amount of information available for different omics and biological pathways [44]. Likewise, WikiPathways is a consensus database with manual community curation of pathways collected from literature as well as other primary source databases, such as KEGG and Reactome. Although these different databases collect the same information, they do not share a standard nomenclature and pathway classification system to avoid redundancies. For example, common names are by no means unique; the KEGG "ECM-receptor interaction" (hsa04512) receives 163 hits in WikiPathways. However, this is the result of a simple pathway name query, where the words "receptor" and "interaction" are part of many other pathway names. A query containing only "ECM" yields 7 hits; however, none of them is the "ECM-receptor interaction" pathway from KEGG. Such inconsistencies highlight the necessity for standard comprehensive naming conventions; however, no such standard exists for biochemical pathways. This is due to the fact that as primary databases were created independently, developers followed their own naming conventions, frequently using the names of ligands, receptors, main targets, and so on for the nomenclature. Frequently, the same pathways identified in the literature received different common names introducing ambiguity and challenges in curating the pathway information. While Gene Ontology (GO) terms were suggested as a solution to conflicting common names, they suffer from the same problem; KEGG and Reactome integrate their own ontologies, and standardization is a crucial topic among many data curation communities [45].

17.3.3.4 Future Outlook for Multi-Omics

Computational power and capabilities are growing at a geometric rate; new tools are being developed, and databases are evolving new architectures and data structures. Starting with the basic concept of a node representing a one-dimensional data point and a pathway linking the nodes in a two-dimensional representation of the data points and their relationships, it follows that networks of interlinked pathways representing dynamic cellular functions must then be implemented as multidimensional data structures. Whatever the underlying structure, flexible data formats and compatibility can ensure access to information and accessibility for different platforms. However, the development of a novel and robust unified standard for annotation and nomenclature is the key for successful data integration [46, 47].

17.4 Conclusions

Development of databases has proven to be beneficial in all fields, from business to science, where data, and most importantly knowledge derived from it, is important. There are numerous approaches to database design; nevertheless, all of them aim at the same goal—to provide a high level of data organization in order to facilitate storage, management, and analysis of the data.

Databases hold a great potential for research purposes and in the era of high-throughput platforms are an essential tool to support the discovery process. Data repositories serve the research community by providing easy access to information with the possibility to answer a specific biological question through use of queries. Importantly, the collection of data in a searchable form greatly facilitates the process of literature mining, which can be very time-consuming and requires detailed analysis of the manuscripts or supplements. Moreover, if a high level of organization and standardization is achieved, then data can be easily mined or integrated, in search for additional information, which is not initially obvious. The number of biological databases is still growing, and new databases are being developed in order to make the best use of existing data, offering the possibility for new discoveries in the field of research and bioinformatics. Numerous discoveries are due to the use of databases. Thus, databases are essential for current and future scientific research.

References

1 Curbelo, R. J., Loza, E., De Yebenes, M. J. & Carmona, L. 2014. Databases and registers: useful tools for research, no studies. *Rheumatol Int*, 34, 447–452.

2 Zou, D., Ma, L., Yu, J. & Zhang, Z. 2015. Biological databases for human research. *Genomics Proteomics Bioinformatics*, 13, 55–63.

3 Stein, L. 2013. *Creating Databases for Biological Information: An Introduction*. Curr Protocols in Bioinformatics, Chapter 9, Unit 9.1. John Wiley & Sons, Inc., New York.

4 Wynn, M. L., Consul, N., Merajver, S. D. & Schnell, S. 2012. Logic-based models in systems biology: a predictive and parameter-free network analysis method. *Integr Biol (Camb)*, 4, 1323–1337.

5 Connolly, T. M. & Begg, C. E. 2005. *Database Systems: A Practical Approach to Design, Implementation, and Management*. Addison-Wesley, Harlow/New York.

6 Lenhardt, W. C., Ahalt, S., Blanton, B., Christopherson, L. & Idaszak, R. 2014. Data management lifecycle and

software lifecycle management in the context of conducting science. *J Open Res Softw*, 2(1), e15.
7 Foster, E. C. & Godbole, S. V. 2010. *Database Systems: A Pragmatic Approach*, Xlibris Corporation, Bloomington.
8 Stein, L. 2002. *Creating Databases for Biological Information: An Introduction*. Current Protocols in Bioinformatics. John Wiley & Sons, Inc., New York.
9 Strauch, C. 2011. NoSQL databases. *Lecture Selected Topics on Software-Technology Ultra-Large Scale Sites*. Stuttgart Media University, Stuttgart.
10 Pokorny, J. 2013. NoSQL databases: a step to database scalability in web environment. *Int J Web Inf Syst*, 9, 69–82.
11 Watt, A. 2015. *Database Design*, 2nd Edition, BCcampus, Victoria.
12 Keim, D.A. & Kriegel, H.-P. 1994. Using visualization to support data mining of large existing databases, *Proceedings of the IEEE Visualization '93 Workshop*, San Jose, CA, in: Lecture Notes in Computer Science. Springer, Vol. 871, pp. 210–229.
13 Kopanakis, I. & Theodoulidis, B. 2003. Visual data mining modeling techniques for the visualization of mining outcomes. *J Vis Lang Comput*, 14, 543–589.
14 Janert, P. K. 2010. *Data Analysis with Open Source Tools*, O'Reilly Media, Inc., Sebastopol.
15 Benson, D. A., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. 2014. GenBank. *Nucleic Acids Res*, 42, D32–D37.
16 Safran, M., Chalifa-Caspi, V., Shmueli, O., Olender, T., Lapidot, M., Rosen, N., Shmoish, M., Peter, Y., Glusman, G., Feldmesser, E., Adato, A., Peter, I., Khen, M., Atarot, T., Groner, Y. & Lancet, D. 2003. Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*, 31, 142–146.
17 RNAcentral Consortium. 2015. RNAcentral: an international database of ncRNA sequences. *Nucleic Acids Res*, 43, D123–D129.
18 Kozomara, A. & Griffiths-Jones, S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*, 42, D68–D73.
19 Bateman, A., Martin, M.J. & O'Donovan, C. 2015. UniProt: a hub for protein information. *Nucleic Acids Res*, 43, D204–D212.
20 Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M. & Bourne, P. E. 2011. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*, 39, D392–D401.
21 Ponten, F., Schwenk, J. M., Asplund, A. & Edqvist, P. H. 2011. The Human Protein Atlas as a proteomic resource for biomarker discovery. *J Intern Med*, 270, 428–446.
22 Liu, X., Yu, X., Zack, D. J., Zhu, H. & Qian, J. 2008. TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, 9, 271.
23 Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., Caudy, M., Garapati, P., Gillespie, M., Kamdar, M. R., Jassal, B., Jupe, S., Matthews, L., May, B., Palatnik, S., Rothfels, K., Shamovsky, V., Song, H., Williams, M., Birney, E., Hermjakob, H., Stein, L. & D'Eustachio, P. 2014. The reactome pathway knowledgebase. *Nucleic Acids Res*, 42, D472–D477.
24 Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S. & Kanehisa, M. 2008. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res*, 36, W423–W426.
25 Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Stein, T. I., Bahir, I., Belinky, F., Morrey, C. P., Safran, M. & Lancet, D. 2013. MalaCards: an integrated compendium for diseases and their annotation. *Database (Oxford)*, 2013, bat018.
26 Fernandes, M. & Husi, H. 2015. FP222 the chronic kidney disease database (CKDdb). *Nephrol Dial Transplant*, 30, iii141.
27 Krochmal, M., Fernandes, M., Filip, S., Pontillo, C., Husi, H., Zoidakis, J., Mischak, H., Vlahou, A. & Jankowski, J. 2016. PeptiCKDdb-peptide- and protein-centric database for the investigation of genesis and progression of chronic kidney disease. *Database (Oxford)*, 2016. doi:10.1093/database/baw128.
28 Nephroseq Research Edition. University of Michigan, Ann Arbor. https://www.nephroseq.org/resource/login.html (accessed October 16, 2017).
29 The Europe PMC Consortium. 2015. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res*, 43, D1042–D1048.
30 Blake, J.A., Christie, K.R. & Dolan, M.E. 2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res*, 43, D1049–D1056.
31 Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. 2015. Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res*, 43, D1079–D1085.
32 Antezana, E., Blonde, W., Egana, M., Rutherford, A., Stevens, R., De Baets, B., Mironov, V. & Kuiper, M. 2009. BioGateway: a semantic systems biology tool for the life sciences. *BMC Bioinformatics*, 10, S11.
33 Wright, J. & Wagner, A. 2008. The Systems Biology Research Tool: evolvable open-source software. *BMC Syst Biol*, 2, 55.
34 Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., Bravo, H. C., Davis, S., Gatto, L., Girke, T., Gottardo, R., Hahne, F., Hansen, K. D., Irizarry, R. A., Lawrence, M., Love, M. I., Macdonald, J., Obenchain, V., Oles, A. K., Pages, H., Reyes, A., Shannon, P., Smyth, G. K., Tenenbaum, D.,

Waldron, L. & Morgan, M. 2015. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*, 12, 115–121.

35 Zhu, Y., Davis, S., Stephens, R., Meltzer, P. S. & Chen, Y. 2008. GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, 24, 2798–2800.

36 Ji, Z. & Ji, H. 2015. GEOsearch: GEOsearch. R package version 1.0.1. Available at https://www.bioconductor.org/packages/release/bioc/html/GEOsearch.html (accessed August 22, 2017).

37 Zhang, X., Kuivenhoven, J. A. & Groen, A. K. 2015. Forward individualized medicine from personal genomes to interactomes. *Front Physiol*, 6, 364.

38 Yugi, K., Kubota, H., Hatano, A. & Kuroda, S. 2016. Trans-omics: how to reconstruct biochemical networks across multiple "omic" layers. *Trends Biotechnol*, 34(4), 276–290.

39 Pesce, F., Pathan, S. & Schena, F. P. 2013. From -omics to personalized medicine in nephrology: integration is the key. *Nephrol Dial Transplant*, 28, 24–28.

40 Chavan, S. S., Shaughnessy, J. D., J. R. & Edmondson, R. D. 2011. Overview of biological database mapping services for interoperation between different "omics" datasets. *Hum Genomics*, 5, 703–708.

41 Kim, T. Y., Kim, H. U. & Lee, S. Y. 2010. Data integration and analysis of biological networks. *Curr Opin Biotechnol*, 21, 78–84.

42 Kuo, T. C., Tian, T. F. & Tseng, Y. J. 2013. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. *BMC Syst Biol*, 7, 64.

43 Su, G., Morris, J. H., Demchak, B. & Bader, G. D. 2014. Biological network exploration with cytoscape 3. *Curr Protoc Bioinformatics*, 47, 8.13.1–8.13.24.

44 Harel, A., Dalah, I., Pietrokovski, S., Safran, M. & Lancet, D. 2011. Omics data management and annotation. *Methods Mol Biol*, 719, 71–96.

45 Chowdhury, S. & Sarkar, R. R. 2015. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database (Oxford)*, 2015. doi: 10.1093/database/bau126.

46 Alyass, A., Turcotte, M. & Meyre, D. 2015. From big data analysis to personalized medicine for all: challenges and opportunities. *BMC Med Genomics*, 8, 33.

47 Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. & Tegner, J. 2014. Data integration in the era of omics: current and future challenges. *BMC Syst Biol*, 8 (Suppl 2), I1.

Part III

Test Cases CKD and Bladder Carcinoma

18

Kidney Function, CKD Causes, and Histological Classification

Franco Ferrario[1], Fabio Pagni[1,2], Maddalena Bolognesi[1], Elena Ajello[1], Vincenzo L'Imperio[1,2], Cristina Masella[3], and Giovambattista Capasso[3]

[1] *Nephropathology Center, University of Milano-Bicocca, San Gerardo Hospital, Monza, Italy*
[2] *Department of Medicine and Surgery, Pathology, University of Milano-Bicocca, San Gerardo Hospital, Monza, Italy*
[3] *Nephrology and Dialysis Unit, Second University of Naples, Policlinico Nuovo Napoli, Naples, Italy*

18.1 Introduction

Kidneys are the setting of different biologic processes such as the fine regulation of fluids and electrolytes balance, the metabolism of toxins and drugs with tubular secretion of both active and inactive components, the synthesis of active hormones like Vitamin D by means of 1-α hydroxylation of 25-OH Vitamin D and the degradation of part of circulating insulin at proximal tubules [1]. However, to assess renal function, clinicians inevitably need to evaluate the glomerular filtration rate (GFR) [2], expressed in ml/minute. GFR represents the physiological process by measuring blood flow through the semipermeable barrier of glomerular tufts that generate an ultrafiltrate. Its composition changes during the passage through the tubular segments and flows via the ureters to the bladder before being eliminated in the form of urine. Several factors influence the value of GFR. The maintenance of a stable systemic hemodynamic status guarantees an appropriate renal plasma flow and GFR. However, at the glomerular level, sophisticated pathways of vasoactive and hormonal molecules regulate single-nephron hemodynamic status acting on the afferent and efferent arteriolar vascular tone and protecting glomeruli against sudden and/or periodic changes of circulating plasma volume [3–5]. The final goal is to maintain a stable GFR. Starling forces drive the ultrafiltrate formation, and are essentially represented by the balance between oncotic and hydrostatic pressures at the two sides of the semi-permeable barrier. Kidneys filter 180 l of blood per day through a total of two millions of glomeruli which compose a very wide vascular surface area (1 m^2) and receive a significant amount of blood as 20% of the cardiac output is destined to kidneys. A great number of variables as physical activity, pregnancy, and also pathological conditions like chronic heart failure, body fluid loss or overload, and chronic kidney disease (CKD), may act both on systemic and local hemodynamic balance. The effect of these factors may be reflected by changes of GFR. Furthermore, glomerular filtration is influenced by age, sex, ethnicity drugs, protein dietary intake, and its value needs to be correlated to body surface area in order to reduce interindividual variability. Healthy people have GFR higher than 100 ml/min and the average normal value is 120–130 ml/min [6].

18.2 The Evaluation of Glomerular Filtration Rate

The evaluation of GFR is at the moment the best clinical tool to measure renal function in the clinical setting. It is essential to reveal the presence of chronic kidney insufficiency and to establish the severity of the disease, as the current guidelines define CKD as the presence of abnormalities of kidney structure or function (GFR < 60 ml/min/1.73 m^2) [7]. GFR estimation is also necessary to monitor the decline of renal function over time, to perform CKD prognosis, and to assess cardiovascular risk. Decisions about drug dosage and the eligibility for radiological examinations with contrast agents are also dependent on the degree of renal function. GFR can be measured indirectly (mGFR) by determining the clearance of exogenous markers administered at the moment of the evaluation; alternatively it can be estimated (eGFR) from serum levels of endogenous filtration markers (mainly creatinine). The development of techniques and equations to determine measured and estimated GFR was initiated at the first decades of the last century. Alving and Smith were among the pioneers who introduced the

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

use of exogenous substances to assess renal function [8–10] and their work led to the identification of inulin, a 5200 Da inert uncharged fructose polymer, as the ideal molecule. Since then, the evaluation of urinary and plasma clearance of inulin has become the gold standard to measure GFR indirectly. Other exogenous tracers such as iothalamate and iohexol have been tested [11–13] and have shown a good safety profile in patients during clearance measurements, with a performance comparable to that of inulin. For iohexol, urinary clearance is unnecessary because the kinetic of the tracer is excellently represented by plasmatic clearance [14–16]. However, in routine clinical practice, clearance measurements require an inpatient setting and are not always easy to apply and comfortable for patients because repeated blood sampling and, for some markers, timed urine collection are necessary. For these reasons, in 1970s, clinicians set up the basis of the first equations for estimating GFR using endogenous markers [17]. In principle, the plasmatic concentration of the ideal analyte in a steady state condition should correlate with the reciprocal of the GFR level, with no need for sequential blood and urine testing and with no interference from other excretion pathways [2]. However, serum levels of endogenous filtration marker are influenced by different variables independent from GFR: synthesis (from cells or dietary intake), extrarenal (intestinal or biliary metabolism) and renal elimination, tubular reabsorption and secretion [18, 19]. All these processes cannot be measured directly; thus, the equations to estimate GFR use clinical measurable parameters to substitute unmeasurable variables.

Commonly used endogenous markers include low-molecular-weight metabolites like urea and creatinine, and low-molecular-weight proteins such as cystatin C, β2-microglobulin, and β-trace protein. Creatinine is the most commonly used among markers; it is a breakdown product of muscle creatine phosphate freely filtered by glomeruli but also secreted by renal proximal tubules [20]. Several non-GFR determinants affect serum creatinine [18]. Age and sex influence the amount of total body muscle mass and, consequently, the level of creatinine turnover [21]. Several drugs antagonize molecular mechanisms of creatinine tubular secretion (such as trimethoprim and cimetidine) producing a decline of creatinine clearance and increase plasmatic concentration of the metabolite. This event does not truly affect the real GFR even if the final result is the rise of serum creatinine [20, 22]. Physical phenotype, prevalence of lean or fat body mass, belonging to different ethnicity (Asian, Black, Hispanic), body integrity, or amputation must be taken into account in the estimation of GFR based on creatinine concentration. Finally, concomitant diseases such as cancer, lymphoproliferative disorders, heart disease, and also malnutrition and neuromuscular diseases influence creatinine turnover [21].

Independent from the selected marker, it is fundamental to consider that the performance of the eGFR equations to reflect the ideal-measured GFR declines considerably in non-steady-state conditions [18]. When acute reduction of GFR occurs, as in acute kidney injury, there is a lag time before eGFR alterations manifest and adequately reflect the magnitude of real GFR decline. This happens because the accumulation of the endogenous marker after an abrupt fall in renal excretion, to an extent that overcomes extrarenal metabolism, takes several hours. This is also applicable when GFR rises: an increase in eGFR reflecting the GFR increase will be apparent after some time, when a new steady state is established. Cockroft and Gault formula that measures the creatinine clearance form serum and clinical parameters that represent the first consistent attempt to define eGFR [17]. However, the formula has limitations as creatinine clearance overestimates GFR because of the tubular secretion of creatinine. Moreover, there is no adjustment for clinical parameters in the equation (body surface area and height are not included) [23].The first significant step for establishing a widely accepted eGFR equation was made in 1999 when the Modification of Diet in Renal Disease (MDRD) Study equation was developed [24, 25]. The MDRD Study was a randomized trial dealing with the effect of low-protein dietary intake and the reduction of blood pressure on progression of CKD. In a second phase, the MDRD formula has been updated with the introduction of standardized serum creatinine, the isotope dilution mass spectrometry (IDMS), according to the guidelines of the US task force of the National Kidney Disease Education Program (NKDEP), in order to abolish inaccuracies derived from interlaboratories variability [26]. Thus the original formula was modified in 2006 as a 4-variable equation with creatinine measurement standardized to IDMS. The 2006 MDRD equation shows greater precision and correspondence to mGFR values compared to previous eGFR estimation methods [27]. However, the correspondence with mGFR declines at less-compromised stages of renal function as it tends to underestimate the GFR in this subset of population. The Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) research group has developed the CKD-EPI estimation equation based on creatinine, validated in populations from different studies. Among clinical and biochemical parameters evaluated, two diverse coefficients were assigned according to the level of serum creatinine. The CKD-EPI equation produces a lesser bias for eGFR greater than 60 ml/min and has allowed the reclassification of the CKD stage for a consistent proportion of patients of the study populations [28, 29].

The CKD-EPI group has also investigated potential advantages of cystatin C over creatinine as filtration marker for eGFR. It is a low-molecular-weight serum protein freely filtered by the glomerulus and reabsorbed by tubular epithelium. Comparative analysis, however, has demonstrated similar accuracy between equations using the two different markers [30]. Better performance with higher accuracy was obtained from the combined formula CKD-EPI based on standardized serum creatinine and cystatin C assays (2012 CKD-EPI$_{cr\text{-}cys}$), also in the elderly and among different sex and ethnic groups [31, 32]. Higher accuracy of the combined equation is derived mostly from greater precision.

Current guidelines recommend the use of a GFR-estimating equation based on creatinine for initial evaluation of renal function. When a more accurate definition is necessary (e.g., in case of titration of nephrotoxic agents), it is preferable to use eGFR based on serum cystatin or to perform mGFR with serum or urinary clearance of reference markers (as inulin, iohexol) if possible. It is always recommended to use the updated estimating equations: the 2009 CKD-EPI for creatinine and the 2012 CKD-EPI for cystatin and creatinine–cystatin equations [7]. If eGFR$_{cr}$ in adults is 45–59 ml/min/1.73 m^2 without any other sign of CKD, it is necessary to apply eGFR$_{cr\text{-}cys}$ if there is need for diagnosis confirmation. In case of eGFR$_{cr\text{-}cys}$ < 60 ml/min CKD is confirmed.

18.3 Causes of CKD

Many diseases may affect kidney function, both systemic and primary renal disorders. When the pathogenetic cause has limited duration the result can be a transient kidney insufficiency that recovers when the disease resolves, and it can take days or weeks. Otherwise, persistent disorders may permanently compromise renal function producing CKD, whose prevalence varies according to geographic areas. Hypertension and diabetes are the most frequent causes of CKD in western countries mainly because of diet and life style characterized by elevated intake of salt and hydrogenated fats and very scant physical activity. However, to better characterize the causes of CKD, we can distinguish primary renal disorder from systemic diseases with secondary renal involvement.

- Primary renal disorders:

  Glomerulonephritis: the inflammation of the filtering unit of kidney, the glomerulus, generally due to an immunological cause. Predisposing genetic loci may play a role in the development of these disorders especially when a second causative event occurs, such as viral or bacterial systemic infection (hepatitis B or C, syphilis, TBC). In other cases, glomerulonephritis can be the result of a primary renal autoimmune disease with loss of immune tolerance toward molecules commonly recognized as "self." Glomerulonephritis usually manifests with urinary alterations as hematuria and proteinuria, and sometimes with hypertension and initial decline of GFR. However, chronic glomerular inflammation may alter substantially the structural integrity of the vascular tuft and produce a permanent impairment of renal function, potentially progressive over several years.

  Genetic disorders: Among all causes, genetic disorders account for 15–20% of all CKD diagnoses, mostly presenting with early onset impairment of renal function. In the wide group of genetic disorders, congenital abnormalities of kidney and urinary tract (CAKUT syndromes), steroid resistant nephrotic syndromes, and ciliopathies are thought to be responsible for more than 70% of cases of CKD with genetic origin [33, 34]. Ciliopathies embrace many of known cystic disorders: autosomal dominant polycystic kidney disease type 1 and 2, autosomal recessive polycystic kidney disease, nephronophthisis type 1–9, medullary cystic disease, and Bardet–Biedl syndrome. The pathogenesis of CAKUT syndromes lies on an abnormal nephrogenesis process during embryological life and may present with different phenotypes: kidney agenesis, renal hypodysplasia, and ureteropelvic junction obstruction. Among glomerular diseases, causative mutations frequently involve structural glomerular proteins such as nephrin, podocin, α-actinin-4, or some of collagen subtypes which compose the glomerular basal membrane as in the Alport syndrome (*COL4A4* and *COL4A5* genes). Dent syndrome is determined by a mutation of a chloride channel (CLCN5) expressed on the intracellular vesicles which mediate molecules trafficking to the cell surface. Dent syndrome causes a generalized dysfunction of the proximal tubules with urinary wasting of substances normally reabsorbed and it is often associated with severe nephrocalcinosis which may severely compromise renal function and lead to CKD.

  Renal stone disease: Nephrolithiasis is a common renal disorder whose prevalence has increased worldwide during the last 20 years across sex, age, and race, principally because of metabolic factors (eating and drinking habits) [35]. Stone formers may experience recurrent acute episodes of nephrolithiasis, especially in the presence of predisposing factors such as abnormalities of the urinary tract, infections (struvite calculi), hereditary disorders (as polycystic kidney disease), hyperoxaluria, and

cystinuria. In some of these cases, the coexistence of diffuse nephrocalcinosis, the presence of massive stone deposition in the urinary tract, and the recurrence of obstructive episodes may damage the renal parenchyma permanently and determine a certain degree of renal insufficiency.

Chronic kidney infections: Chronic pyelonephritis is often associated with abnormalities of the urinary system (such as vesicoureteral reflux, neurogenic bladder) that allow the urine to flow back toward kidneys; rarely it is caused by systemic infections or infective processes of the intra-abdominal organs, that secondarily involve kidneys. Whatever the cause, chronic pyelonephritis may determine progressive loss of renal function with recurrent episodes of hyperpyrexia.

- Secondary renal disorders

Cardiovascular diseases: *Hypertension* is one of the most frequent causes of CKD, together with diabetes, in western countries. Atherosclerosis, smoking, dyslipidemia and high salt intake are recognized as the common predisposing factors. The genetic background also plays an important role, even though not all the genetic loci that contribute to the development of the disease have been identified. High blood pressure alters the integrity of the arterial wall producing stiffening and thickening of blood vessels. These alterations also involve renal circulation and the final effect is a diffuse nephroangiosclerosis and the impairment of the finely tuned glomerular circulation. Over time, these functional alterations lead to structural alterations represented by the disruption of the glomerular filtering barriers. *Diabetes* also markedly contributes to vascular dysfunction because of alteration of carbohydrates and lipid metabolism. In this setting, circulating and locally generated products of advanced glycosylation improperly react with the normal proteic component of the arterial wall. This event inevitably predisposes to endothelial dysfunction and inflammation which compromises vascular integrity. *Chronic heart failure* is another leading cause of CKD, especially in the elderly [36]. Chronic renal hypoperfusion manifests because of depressed cardiac ejection fraction and reduced effective circulating volume that are the pathogenetic mechanisms at the basis of renal insufficiency. The *renal artery stenosis*, especially when bilateral, is responsible for secondary hypertension resistant to multidrug therapy. In the youth it is frequently caused by fibromuscular dysplasia while in the elderly it is generally due to severe atherosclerosis. The effect is a chronic consistent reduction of renal plasma flow that produces an ischemic injury with overactivation of compensatory neurohormonal mechanisms (RAAS system). Over time, the severe hypertension and the ischemic persistent damage result in chronic kidney insufficiency.

Autoimmune disorders: The autoimmune diseases are the setting of the systemic inappropriate reaction of the immunological system against molecules normally recognized as "self." The result is a chronic and clinically meaningful, uncontrolled inflammatory response which may involve every organ and tissue [37, 38]. The phenotypic presentation may vary largely and in case of renal manifestations of the disease, it is usual to observe alterations of urinary sediment with proteinuria and/or hematuria. A certain degree of GFR decline is frequently observed already at the onset of the renal involvement, and in some cases renal function can be permanently impaired with patients manifesting CKD. Rheumatoid arthritis, cryoglobulinemia, LES, thyroiditis, and vasculitis are some of the autoimmune disorders that frequently present with renal involvement. In the kidney, the immune-mediated damage can be determined by a direct cytotoxic effect or the generation of immune complexes (IC) composed of an immunoglobulin and an antigen. In the latter case, the IC can aggregate "in situ" in the glomeruli or in the systemic circulation, and thereafter settle in the kidney.

Tubule-interstitial nephropathies: This heterogeneous group of diseases embraces different causative agents such as infections; toxic agents like antibiotics, salicylates, chemotherapy drugs, lymphoproliferative diseases, nephrocalcinosis, gout, heavy metals; and chemical substances like paraquat, herbicides [39]. The tubules and the peritubular interstitium constitute an essential functional unit, and the damage of this sophisticated system translates into the loss of all related reabsorbitive and metabolic pathways, together with the disruption of the mechanisms that guarantee the ability to modulate urine tonicity. The renal impairment can be mediated by diverse pathogenic mechanisms: an exaggerated oxidative stress, the activation of a concomitant immunological response aimed to remove the toxic agent, and a direct interference with biological systems devoted to the synthesis of energetic molecules (ATP). All these processes may have deleterious effects on renal hemodynamics and on the integrity of the glomerular barrier. According to the duration of exposure to causative agents and to the extent of renal damage, this group of tubule-interstitial nephropathies may result in chronic kidney insufficiency.

Amyloidosis: It is a severe disease characterized by the systemic extracellular accumulation of a proteic, low-molecular–weight, insoluble substance, the amyloid that cannot be cleared by the scavenger cells of the immune system nor metabolized by any enzymatic pathway. Thus, the amyloid accumulates and can be found in every organ and system. Several subtypes of amyloidosis are known, according to the abnormal protein, for example, the amyloidosis AA typical of chronic inflammatory diseases (rheumatoid arthritis, bowel inflammatory disorders), amyloidosis AL associated with overproduction of altered immunoglobulins in lymphoproliferative diseases, and amyloidosis ATTR caused by accumulation of transthyretin. In the renal amyloidosis, the mesangial space is completely filled with the protein fibril deposits that are also found on the wall of arterioles and glomerular capillaries [40]. These protein fibrils appear bright green to a polarized light microscope, after Red Congo staining. Both the glomerular hemodynamic and the integrity of filtering barrier are inevitably compromised and the final event is the organ failure.

For these reasons, a complete clinical–pathological assessment of these patients is crucial to determine their prognosis and to employ the proper therapeutic regimen, also to reduce the incidence of ESRD and renal failure.

18.3.1 Histological Classification of CKD

Renal biopsy is an "invasive" procedure, with possible risk of complications. Thus, its evaluation should be performed in centers with recognized nephropathological expertise and supported by all methodologies (LM, IF, EM, Immunohistochemistry), essential for a correct diagnosis. The value of performing a renal biopsy in a patient with suspected glomerular disease is twofold. First, of course, it allows establishing a diagnosis of a specific disease or category of disease, although in some cases this may be largely apparent from the patient's clinical history and serologic data. This frequently happens in case of systemic conditions, such as systemic lupus erythematosus (SLE) that often involve the kidney. The second, and perhaps more useful for the clinician, is to provide prognostic information regarding the likely clinical course of the patient, and the likelihood of improving this prognosis with therapeutic intervention, most notably immunosuppressive therapy. For many glomerular diseases, although their identification on renal biopsy is generally straightforward, their histologic appearance on biopsy, much like their clinical presentation, can vary greatly. This histologic appearance represents a snapshot of prior and ongoing events within the kidney, and has been shown in studies performed over several decades to be correlated to some extent with both the clinical presentation and evolution, including response to therapeutic intervention and likelihood of progression to end-stage renal disease (ESRD) [41–45]. It is for this reason that morphologic classification systems of several different glomerular diseases have been proposed, particularly those with the greatest degree of morphologic heterogeneity, including lupus nephritis (LN) [46–50], IgA nephropathy (IgAN) [51–55], and ANCA-associated vasculitis [56–58]. However, the value of these classification systems remains to be definitively established. This value is dependent on a number of different parameters. First, the validity of any disease classification system is based on the criteria of reproducibility (measured by the k value) and precision (which is the s.d. of variation), which ensure that a classification is widely applicable by pathologists around the world, with acceptably low intraobserver and interobserver variation, and necessarily implicates a level of simplicity that allows the classification to be employed within the context of routine clinical practice. Second and most important is the ability of the classification to provide prognostic information regarding the likelihood of disease progression, above and beyond the available clinical data at the time of biopsy and during follow-up, and/or to provide information useful in identifying those patients who are likely to respond to certain therapeutic interventions. Third, morphologic classification is a dynamic process, because periodic discoveries are made that add to our knowledge of etiology or pathogenesis, identify new markers of prognosis and/or therapeutic responsiveness, and identify new therapies. Thus, a widely applicable histologic classification system should truly be a working classification capable of undergoing modification in response to new knowledge without a significant loss of precision or ease of utilization. An example of the latter is the Banff working classification for renal allograft pathology, which over the past two decades has undergone key modifications in response to generation of new knowledge, such as the increased recognition of the importance of antibody-mediated rejection. Clinical Renal Syndromes are relatively few:

- Asymptomatic proteinuria
- Asymptomatic hematuria
- Nephrotic syndrome
- Nephritic syndrome
- Rapidly progressive renal failure
- Acute renal failure
- Chronic renal failure

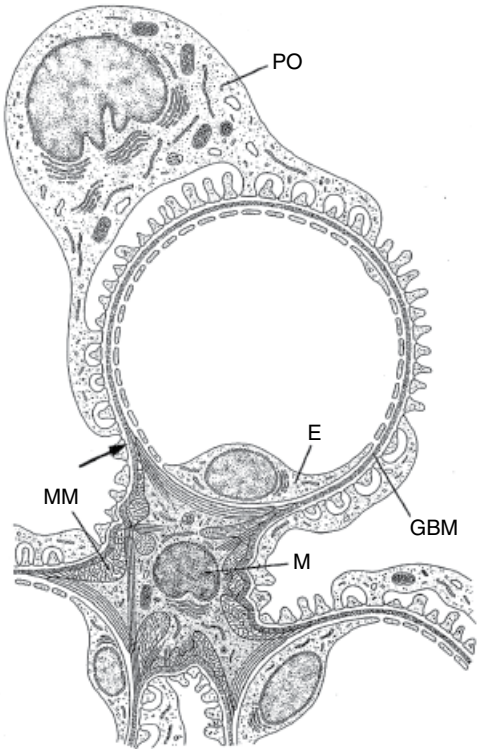


Figure 18.1 In this cartoon, the schematic structure of a normal glomerulus can be appreciated. The capillary stalk is composed of mesangial cells (M) from which the branching of capillary lumen starts. The capillary wall is composed, from the inside to the outside, by fenestrated endothelial cells (E), the glomerular basement membrane (GBM), and from the foot processes of podocytes (PO).

Also, glomerular cells are relatively few, as depicted in the cartoon in Figure 18.1:

- Mesangial
- Endothelial
- Visceral epithelial (Podocytes)
- Parietal epithelial

Conversely the number of totally different types of nephritis is extremely large and every year new entities are described.

Disease that typically cause the nephrotic syndrome

Focal segmental glomerulosclerosis (all variants)
Idiopathic membranous glomerulopathy
Minimal change glomerulopathy
Diabetic glomerulosclerosis
Type I membranoproliferative glomerulonephritis
Idiopathic mesangioproliferative glomerulonephritis
Amyloidosis
C1q nephropathy
Fibrillary glomerulonephritis
Monoclonal immunoglobulin deposition disease
Type II membranoproliferative glomerulonephritis
Pre-eclampsia/eclampsia
Immunotactoid glomerulopathy
C3 nephropathy
Collagenofibrotic glomerulopathy

Disease that typically cause hematuria and nephritis

LN
IgA nephropathy
Idiopathic IC proliferative glomerulonephritis
Pauci-immune ANCA-associated vasculitis
Postinfectious acute diffuse proliferative glomerulonephritis
Thin basement-membrane lesion
Antiglomerular basement membrane antibody glomerulonephritis
Alport disease

Diseases other than glomerulonephritis that typically cause acute renal failure

Thrombotic microangiopathy (all types)
Acute tubulointerstitial nephritis
Acute interstitial nephritis (IgG4 related)
Acute tubular necrosis
Atheroembolization
Light chain cast nephropathy
Cortical necrosis

Diseases other than those already listed that typically manifest as chronic renal failure

Arterionephrosclerosis
Chronic sclerosing glomerulonephritis
ESRD not otherwise specified
Chronic tubulointerstitial nephritis
Miscellaneous other diseases
Adequate tissue with nonspecific abnormalities
No pathologic lesion identified

Moreover, the lesions associated to a single nephritis can be extremely variable in both characteristics and intensity, requiring a proper classification. Some examples are reported in Tables 18.1–18.3 and Figure 18.2.

For all these reasons, a correct diagnostic evaluation of renal biopsy needs nephropathological expertise supported by all methodologies. To stress this point,

Table 18.1 International Society of Nephrology/Renal Pathology Society (ISN/RPS) 2003 classification of lupus nephritis.

Class I Minimal mesangial lupus nephritis
Class II Mesangial proliferative lupus nephritis
Class III Focal lupus nephritis
Class IV Diffuse lupus nephritis
Class V Membranous lupus nephritis

Table 18.2 Classification scheme for ANCA-associated glomerulonephritis.

Class	Inclusion criteria[a]
Focal	≥50% normal glomeruli
Crescentic	≥50% glomeruli with cellular crescents
Mixed	<50% normal, <50% crescentic, <50% globally sclerotic glomeruli
Sclerotic	≥50% globally sclerotic glomeruli

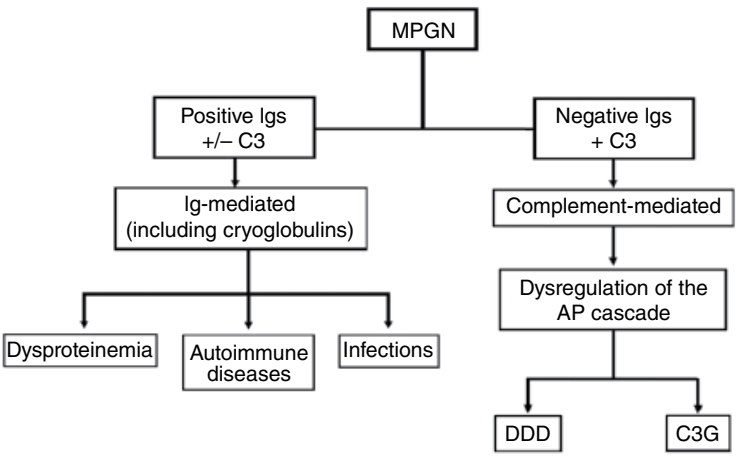
[a] Pauci-immune staining pattern on immunofluorescence microscopy (IM) and ≥1 glomerulus with necrotizing or crescentic glomerulonephritis on light microscopy (LM) are required for inclusion in all four classes.

Table 18.3 Glomerular classification of DN.

Class	Description	Inclusion criteria
I	Mild or nonspecific LM changes and EM-proven GBM thickening	Biopsy does not meet any of the criteria mentioned below for class II, III, or IV
		GBM > 395 nm in female and >430 nm in male individuals of 9 years of age and older[a]
IIa	Mild mesangial expansion	Biopsy does not meet criteria for class III or IV
		Mild mesangial expansion in >25% of the observed mesangium
IIb	Severe mesangial expansion	Biopsy does not meet criteria for class III or IV
		Severe mesangial expansion in >25% of the observed mesangium
III	Nodular sclerosis (Kimmelstiel–Wilson lesion)	Biopsy does not meet criteria for class IV
		At least one convincing Kimmelstiel–Wilson lesion
IV	Advanced diabetic glomerulosclerosis	Global glomerular sclerosis in >50% of glomeruli
		Lesion from classes I through III

LM, light microscopy.
[a] On the basis of direct measurement of GBM width by EM, these individual cutoff levels may be considered indicative when other GBM measurements are used.

Figure 18.2 MPGN—a simple classification.



some striking examples of the main histological lesions are presented in Figures 18.3–18.12, in which four different diseases are depicted.

All these examples confirm that the evaluation of renal biopsy should be performed in centers with recognized nephropathological expertise and supported

Endocapillary hypercellularity

Definition: Increase of mesangial cells, endothelial cells, infiltrating leukocytes causing narrowing of the glomerular capillary lumina.

Figure 18.3 Definition of endocapillary hypercellularity.

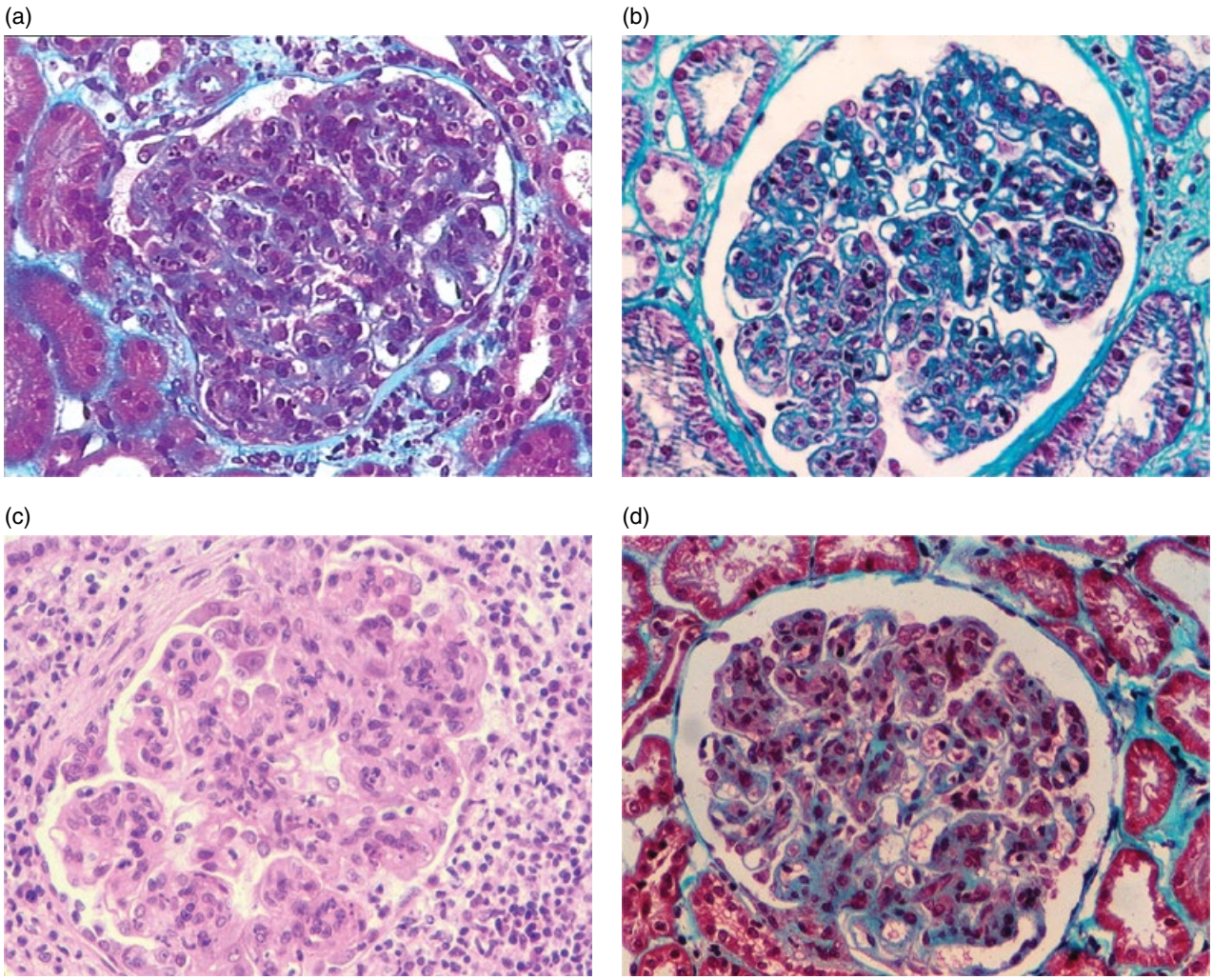


Figure 18.4 In figure a particular pattern of glomerular injury is represented, known as "endocapillary hypercellularity," as the main manifestation of four different glomerular diseases. In Figure 18.3 there is the definition of endocapillary hypercellularity. Figure 18.4 shows the light microscopy images of four glomeruli with a similar endocapillary hypercellularity but affected by different pathology ((a), (b), (d) stained with trichrome stain, (c) stained with Hematoxylin and Eosin, ×20).

by various methodologies (LM, IF, EM, Immunohistochemistry), essential for a correct diagnosis. Moreover, it is necessary to establish a strict clinico-pathological correlation.

18.4 Assessment of Disease Progression and Response to Therapy for the Individual: Interval Renal Biopsy

Repeating the renal biopsy in individual patients is not practiced widely. Although disease classifications are useful at a population level, it is harder to demonstrate that they are of value in providing clinically meaningful prognostic information for the individual. Given that serum creatinine is a poor marker of renal function [59–61], this is best assessed by interval biopsy to monitor both disease progression and response to therapy. This is most practiced in LN and in renal transplantation, although it is assuming a wider role in other diseases such as ANCA vasculitis.

18.5 Recent Advances: Pathology at the Molecular Level

In recent years, molecular biology research has moved on from studying single genes, their transcripts (messenger RNA (mRNA)), or proteins to study groups of molecules within a given domain in parallel with microarray technology [62]. In addition, there has been a growing

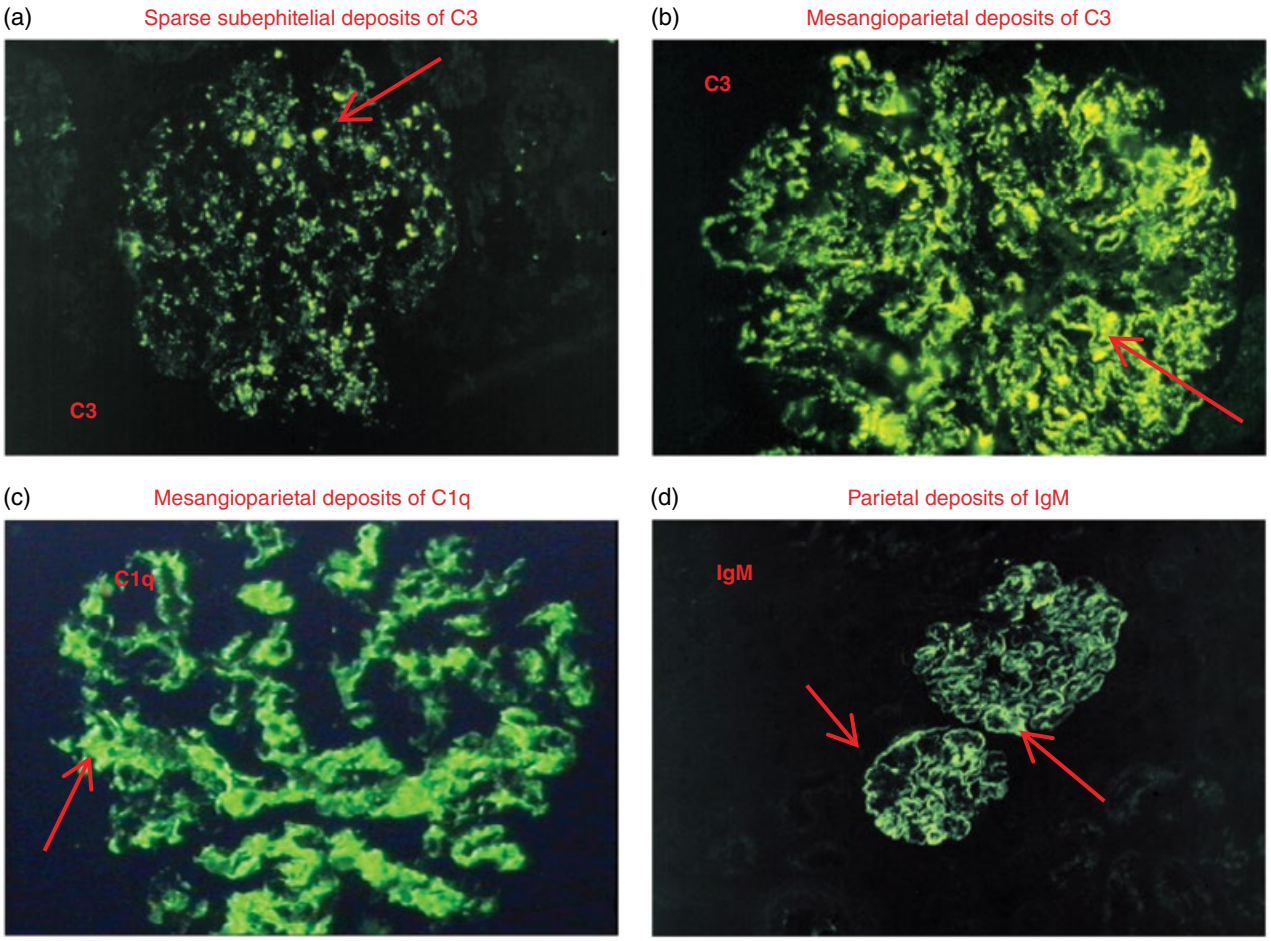


Figure 18.5 In this figure a particular pattern of glomerular injury is represented, known as "endocapillary hypercellularity," as the main manifestation of four different glomerular diseases. The same glomeruli was stained with immunofluorescence technique for different antisera. In (a) there are sparse subepithelial deposits of C3, in (b) mesangioparietal deposits of C3, in (c) mesangioparietal deposits of C1q, and in (d) parietal deposits of IgM.

interest in the role of microRNAs in kidney homeostasis and disease [63]. MicroRNAs are endogenous, short noncoding lengths of RNA that control the expression of many genes. The microRNAs may be detected by a number of techniques including microarray technology and quantitative PCR. The hope is that such "omics" approaches will serve as a molecular microscope focused on new ways of examining renal biopsy tissue and help elucidate disease mechanisms and identify novel biomarkers that will aid diagnosis, prognosis, and treatment [64, 65]. Microarray techniques can be preferred over conventional pathology because of their ability to identify and quantify thousands of transcripts at once and to measure early and rapid changes in disease processes before the resulting pathological lesions are detectable. However, microarrays cannot give information on anatomical relationships and the source of a particular transcript. Nevertheless, it is likely that these molecular markers together with histology will aid in the diagnostic precision of the kidney biopsy. The employment of different molecular techniques in this field are necessary for the development of new noninvasive biomarkers—such as urine proteomics chips—that need to be validated against the biopsy as a gold standard alongside the clinical data [66]. Despite the advances that have been made over the past few years within the field of molecular characterization of the renal biopsy, there have been few clinically relevant correlates. Along with the identification of soluble urokinase plasminogen activator receptor, the enquiry into the molecular pathogenesis of FSGS has become more feasible with techniques such as laser capture microdissection (LCM), which allows the investigation of the glomerular gene expression profiles of patients with primary FSGS using a microarray of mRNA isolated from formalin-fixed renal biopsies [67]. LCM and mass spectrometry has also shown itself to be a

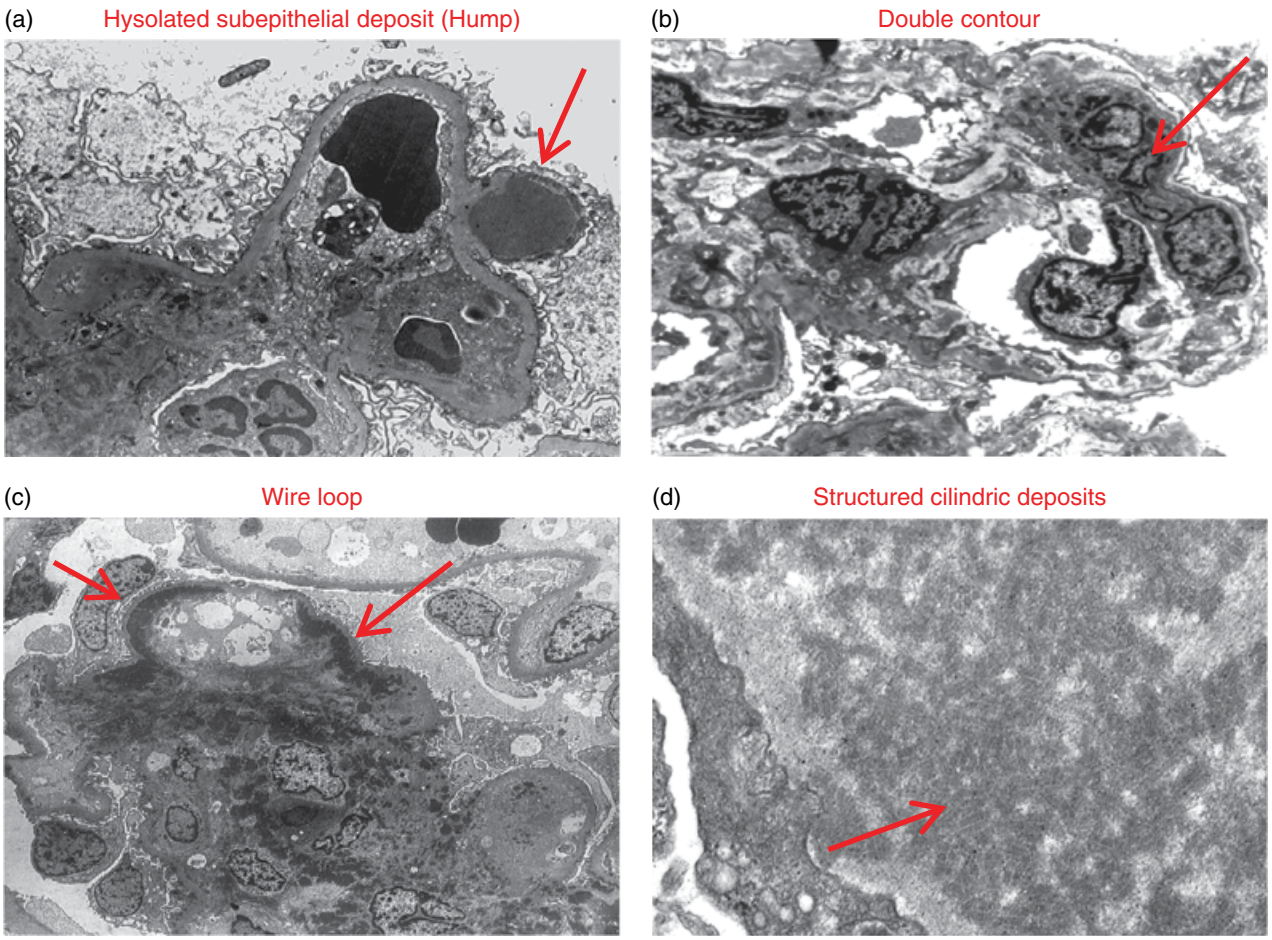


Figure 18.6 In this figure a particular pattern of glomerular injury is represented, known as "endocapillary hypercellularity," as the main manifestation of four different glomerular diseases. It depicts the electron microscopy feature of each case. In (a) the presence of subepithelial deposits, in (b) the double contour aspect of basement membrane, in (c) the so-called wire loops, and (d) reveal the formation of structured cilindric deposits.

Final diagnosis

- A: Acute postinfectious glomerulonephritis
- B: Primary membranoproliferative glomerulonephritis
- C: Lupus nephritis (Class IV-G diffuse global)
- D: Cryoglobulinemic glomerulonephritis

Figure 18.7 In this figure a particular pattern of glomerular injury is represented, known as "endocapillary hypercellularity," as the main manifestation of four different glomerular diseases. The diagnosis for each case is reported.

Nodular patterns

Definition: Marked and diffuse mesangial matrix enlargement with nodular appearance with no or few mesangial cells proliferation

Figure 18.8 In this figure a particular pattern of glomerular injury is represented, called "nodular," as the main manifestation of four different glomerular diseases. There is the definition of nodular pattern.

valuable proteomic tool, allowing the characterization of rarer forms of renal amyloidosis from renal biopsy [68] These include those associated with deposition of fibrinogen a-chain, apolipoprotein A-1 and A-IV, transthyretin, and gelsolin. Such precise phenotyping should allow a better genetic counseling and disease-specific treatments to be implemented. LCM and mass spectrometry is also useful in determining the type of immunoglobulins and complement factors in IC and complement-mediated glomerulonephritis, respectively [69]. Finally, newer stains for IgG subtypes have allowed the identification of novel monoclonal forms of proliferative GN [70] and IgG4-associated autoimmune interstitial nephritis [71].

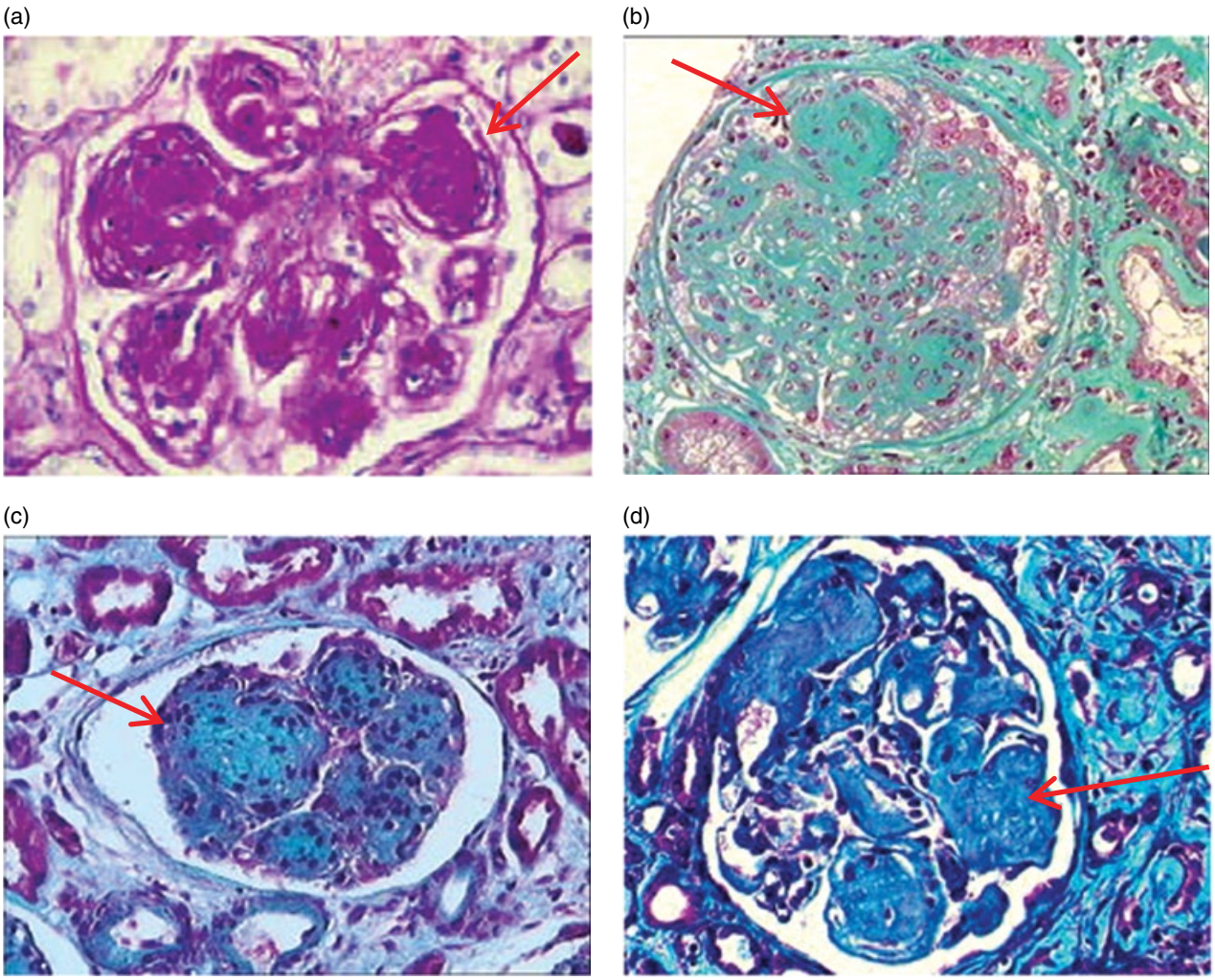


Figure 18.9 In this figure a particular pattern of glomerular injury is represented, called "nodular," as the main manifestation of four different glomerular diseases. The light microscopy images of four glomeruli with a similar nodular pattern but affected by different pathology ((a) stained with PAS; (b), (c), and (d) stained with trichrome stain ×20).

18.6 Digital Pathology

The interpretation of images of tissues and cells at a resolution higher than the naked human eye is the core of pathology. For a long time the microscope has been the only available instrumentation to this aim, over centuries providing live images at increasing resolution through ever-improving optics [72]. During the last decades, significant technical advances were implemented in optical pathology [73, 74], such as the introduction of digital cameras producing still images and microscope-mounted video cameras that allow live examination of slides (dynamic images). These still or dynamic images can be transferred by the means of network connections to remote sites to be assessed by another pathologist, a process commonly called telepathology [75, 76]. This has found applications such as teleconsultation and frozen section diagnosis for the intraoperative diagnosis. Approximately a decade ago, further improvements of these techniques resulted in the creation of digital slide scanners [77]. These slide scanners produce whole slide images (WSI, also called digital or virtual slides) that combine the advantages of images from live cameras (whole slide access) and digital cameras (high resolution). WSI are explored using an image viewer, which enables the examination of digital slides in a manner comparable to the use of a conventional microscope in three aspects. First, WSI can be explored at different magnifications, with the additional advantage of in-between magnifications, if provided by the viewer software. Secondly, navigation of the slides in each direction

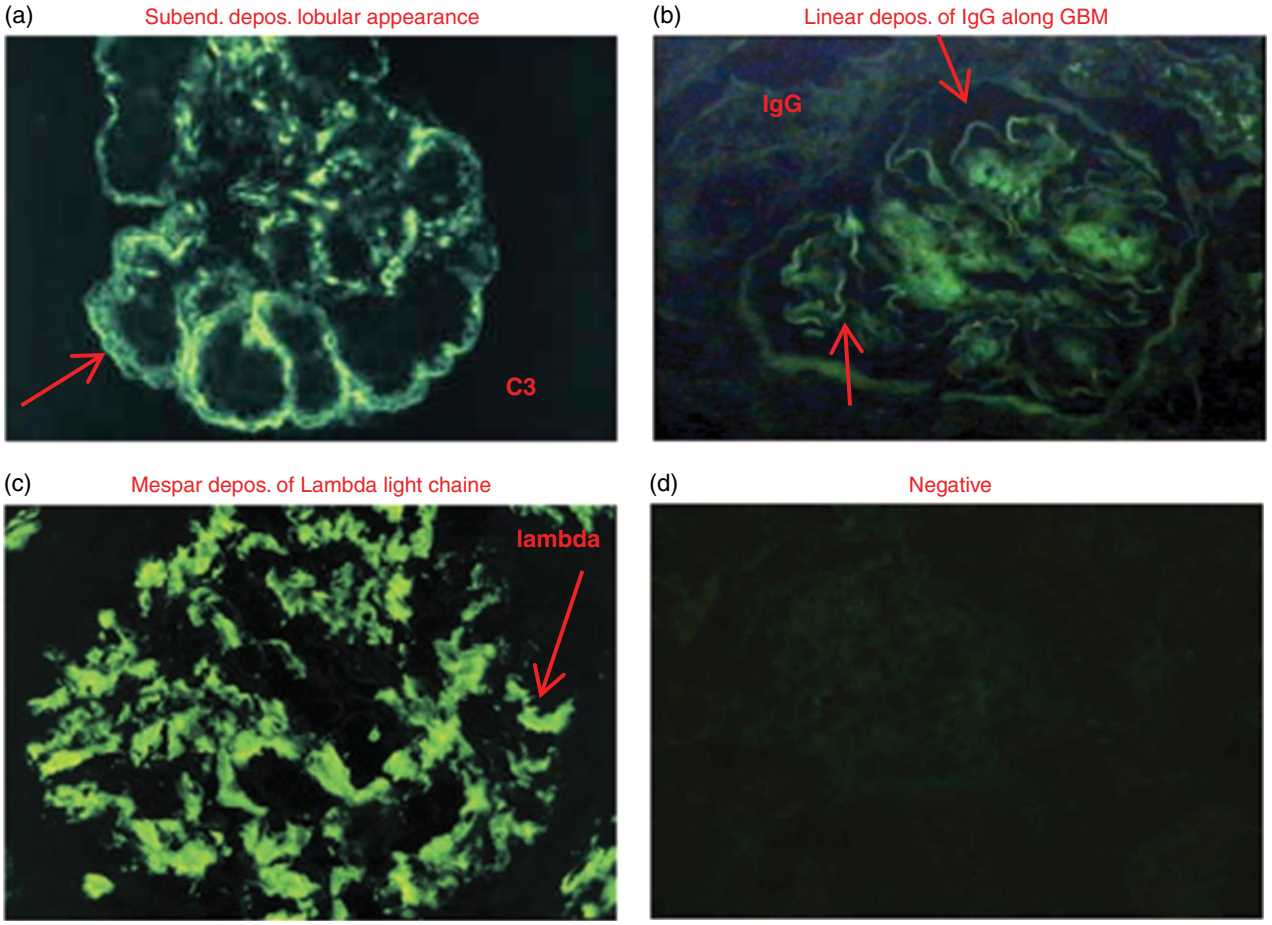


Figure 18.10 In this figure a particular pattern of glomerular injury is represented, called "nodular," as the main manifestation of four different glomerular diseases. The same glomeruli is stained with immunofluorescence technique for different antisera. In (a) there are continuous subendothelial deposits with lobular appearance of C3, in (b) linear deposits of IgG along the glomerular basement membrane (GBM), in (c) mesangioparietal deposits of lambda light chain, and in (d) the negativity to any antiserum tested.

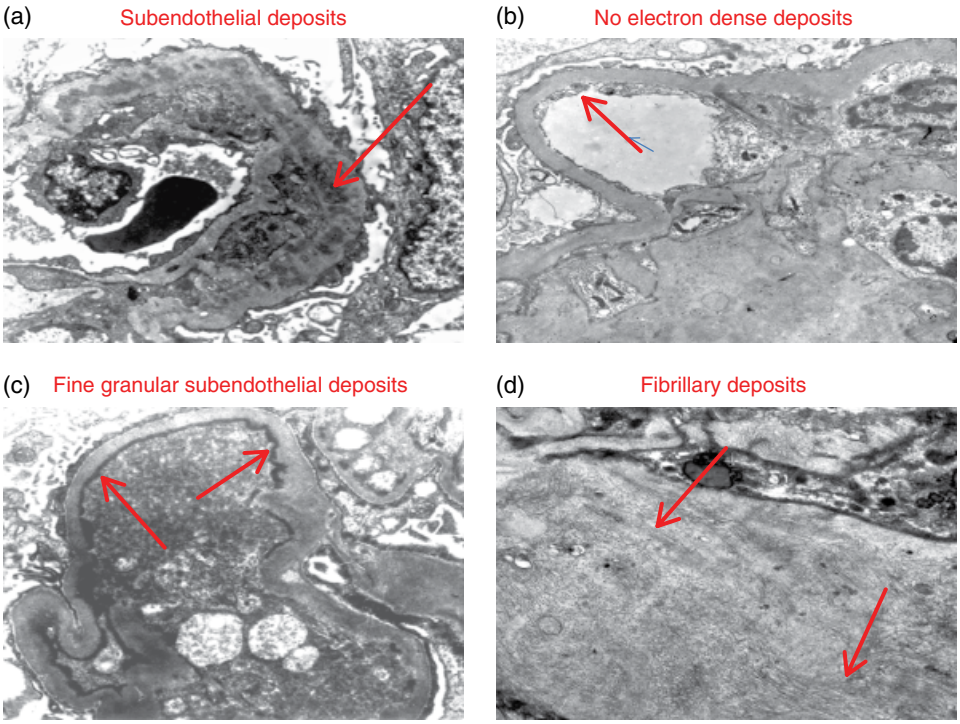


Figure 18.11 In this figure a particular pattern of glomerular injury is represented, called "nodular," as the main manifestation of four different glomerular diseases. Depicts the electron microscopy feature of each case. In (a) the presence of subendothelial deposits, in (b) the absence of immunodeposits in the GBM, in (c) the finely granular subendothelial deposits, and (d) reveal the formation of fibrillary deposits.

Final diagnosis

- A: Primary membranoproliferative glomerulonephritis
- B: Diabetic nodular glomerulosclerosis
- C: Light chain deposition disease (lambda)
- D: Amyloidosis AA

Figure 18.12 In this figure a particular pattern of glomerular injury is represented, called "nodular," as the main manifestation of four different glomerular diseases. The diagnosis for each case is reported.

is possible. Thirdly, some scanners allow scanning more than one focus plane, thereby even allowing focusing up and down [78–80]. Furthermore, WSI have several advantages over conventional slides: Image viewers are able to show an overview image together with the higher-power view, resulting in better orientation within the slide when viewing at higher magnification and more easy navigation to other regions of interest. Image viewers can display several slides side by side, so the examiner can compare structural details between slides or different stains of the same tissue area. WSI can be made available instantaneously to multiple examiners from all over the world through the internet without the need for a microscope. Focusing is carried out during scanning, requiring less user interaction. The quality of WSI is constant over time; it can be used directly for automated image analysis and morphometry and it can also be integrated within the electronic patient records, together with other images. Figure 18.13 shows a screenshot of a WSI as it is seen with an image viewer.

18.7 Conclusions

A renal biopsy is a relatively safe procedure with a well-defined risk profile allowing patients and clinicians to have a complete view of the issue and decide whether performing one or not. This is particularly important, since it represents an irreplaceable part of the diagnostic process, providing also prognostic and mechanistic insights. It is a rich source of information and it is likely that it will deliver specific molecular and cellular patterns of disease that will enable targeted therapy in the future. Renal biopsies also facilitate "bedside-to-bench" research that further defines the mechanisms and pathogenesis of progressive renal injury, with the potential of new therapies.

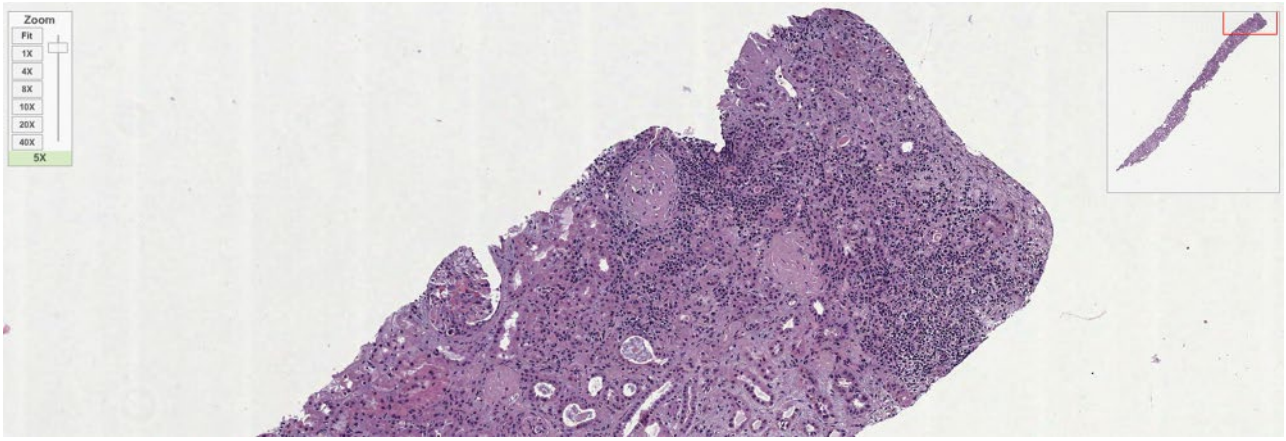


Figure 18.13 An example of the WSI viewer.

References

1 Sahay M., Kalra S., Bandgar T. Renal endocrinology: the new frontier. *Indian J Endocrinol Metab*. 2012 Mar–Apr;16(2):154–5.
2 Stevens L.A., Coresh J., Greene T. et al. Assessing kidney function—measured and estimated glomerular filtration rate. *N Engl J Med.* 2006 Jun 8;354(23): 2473–83.
3 Hall J.E., Guyton A.C. *Guyton and Hall textbook of medical physiology*. 12th edition. Philadelphia, PA: Saunders/Elsevier; 2011. p. 303–22.
4 Kastner P.R., Hall J.E., Guyton A.C. Control of glomerular filtration rate: role of intrarenally formed angiotensin II. *Am J Physiol.* 1984 Jun;246(6 Pt 2):F897–906.

5 Kassirer J.P. Clinical evaluation of kidney function—glomerular function. *N Engl J Med.* 1971 Aug 12;285(7):385–9.

6 Wesson L. *Physiology of the human kidney*. New York: Grune & Stratton; 1969.

7 Chapter 1: Definition and classification of CKD. *Kidney Int Suppl (2011).* 2013 Jan;3(1):19–62.

8 Miller B.F., Alving A.S., Rubin A.S. The renal excretion of inulin at low plasma concentration of this compound, and its relationship to the glomerular filtration rate in normal, nephritic and hypertensive individuals. *J Clin Investig.* 1940;19(1):89–94.

9 Smith H. Comparative physiology of the kidney. In: Smith H, ed. *The kidney: structure and function in health and disease*. New York: Oxford University Press, 1951: p. 520–74.

10 Vurek G.G., Pegram S.E. Fluorometric method for the determination of nanogram quantities of inulin. *Anal Biochem*. 1966;16:409–15.

11 Capasso G., Unwin R.J., Pica A. et al. Iothalamate measured by capillary electrophoresis is a suitable alternative to radiolabeled inulin in renal micropuncture. *Kidney Int.* 2002 Sep;62(3):1068–74.

12 Israelit A.H., Long D.L., White M.G. et al. Measurement of glomerular filtration rate utilizing a single subcutaneous injection of 125I-iothalamate. *Kidney Int*. 1973 Nov;4(5):346–9.

13 Isaka Y., Fujiwara Y., Yamamoto S. et al. Modified plasma clearance technique using nonradioactive iothalamate for measuring GFR. *Kidney Int.* 1992 Oct;42(4):1006–11.

14 Gaspari F., Perico N., Ruggenenti P. et al. Plasma clearance of nonradioactive iohexol as a measure of glomerular filtration rate. *J Am Soc Nephrol.* 1995 Aug;6(2):257–63.

15 Gaspari F., Perico N., Remuzzi G. Application of newer clearance techniques for the determination of glomerular filtration rate. *Curr Opin Nephrol Hypertens.* 1998 Nov;7(6):675–80.

16 Gaspari F., Perico N., Remuzzi G. Measurement of glomerular filtration rate. *Kidney Int Suppl*. 1997 Dec;63:S151–4.

17 Cockcroft D.W., Gault M.H. Prediction of creatinine clearance from serum creatinine. *Nephron.* 1976;16(1):31–41.

18 Stevens L.A., Levey A.S. Measured GFR as a confirmatory test for estimated GFR. *J Am Soc Nephrol.* 2009 Nov;20(11):2305–13.

19 Stevens L.A., Zhang Y., Schmid C.H. Evaluating the performance of equations for estimating glomerular filtration rate. *J Nephrol.* 2008 Nov-Dec; 21(6):797–807.

20 Levey A.S. Measurement of renal function in chronic renal disease. *Kidney Int.* 1990 Jul;38(1):167–84.

21 Jafar T.H., Islam M, Jessani S. et al. Level and determinants of kidney function in a South Asian population in Pakistan. *Am J Kidney Dis.* 2011 Nov;58(5):764–72.

22 Berglund F., Killander J., Pompeius R. Effect of trimethoprim-sulfamethoxazole on the renal excretion of creatinine in man. *J Urol.* 1975 Dec;114(6):802–8.

23 Stevens L.A., Levey A.S. Frequently asked questions about GFR estimates. New York: National Kidney Foundation https://www.kidney.org/content/frequently-asked-questions-about-gfr-estimates. Accessed October 16, 2017.

24 Levey A.S., Bosch J.P., Lewis J.B. et al. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. *Ann Intern Med.* 1999 Mar 16;130(6):461–70.

25 Levey A.S., Greene T., Kusek J. et al. A simplified equation to predict glomerular filtration rate from serum creatinine. *J Am Soc Nephrol.* 2000;11:155A.

26 Levey A.S., Coresh J., Greene T. et al. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Ann Intern Med.* 2006 Aug 15;145(4):247–54.

27 Stevens L.A., Coresh J., Feldman H.I. et al. Evaluation of the modification of diet in renal disease study equation in a large diverse population. *J Am Soc Nephrol.* 2007 Oct;18(10):2749–57.

28 Levey A.S., Stevens L.A., Schmid C.H. et al. A new equation to estimate glomerular filtration rate. *Ann Intern Med.* 2009 May 5;150(9):604–12.

29 Stevens L.A., Nolin T.D., Richardson M.M. et al. Comparison of drug dosing recommendations based on measured GFR and kidney function estimating equations. *Am J Kidney Dis.* 2009 Jul;54(1):33–42.

30 Stevens L.A., Coresh J., Schmid C.H. et al. Estimating GFR using serum cystatin C alone and in combination with serum creatinine: a pooled analysis of 3,418 individuals with CKD. *Am J Kidney Dis.* 2008 Mar;51(3):395–406.

31 Inker L.A., Schmid C.H., Tighiouart H. et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. *N Engl J Med.* 2012 Jul 5;367(1):20–9.

32 Schaeffner E.S., Ebert N., Delanaye P. et al. Two novel equations to estimate kidney function in persons aged 70 years or older. *Ann Intern Med.* 2012 Oct 2;157(7):471–81.

33 Vivante A., Hildebrandt F. Exploring the genetic basis of early-onset chronic kidney disease. *Nat Rev Nephrol.* 2016 Mar;12(3):133–46.

34 Hildebrandt F. Genetic kidney diseases. *Lancet.* 2010 Apr 10;375(9722):1287–95.

35 Romero V., Akpinar H., Assimos D.G. Kidney stones: a global picture of prevalence, incidence, and associated risk factors. *Rev Urol.* 2010 Spring;12(2–3):e86–96.

36 Smilde, T.D., Hillege, H.L., Navis, G. et al. Impaired renal function in patients with ischemic and nonischemic chronic heart failure: association with neurohormonal activation and survival. *Am Heart J.* 2004;148:165–72.

37 Quigg R. J. Glomerular injury induced by antibody and complement. *Semin Nephrol.* 1991;11:259–67.

38 Heyman B., Wiersma E.J., Kinoshita T. In vivo inhibition of the antibody response by a complement receptor-specific monoclonal antibody. *J Exp Med.* 1990;172:665–8.

39 Ghane S.F., Assadi F. Drug-induced renal disorders. *J Renal Inj Prev.* 2015 Sep 1;4(3):57–60.

40 Dember L.M. Amyloidosis-associated kidney disease. *J Am Soc Nephrol.* 2006 Dec;17(12):3458–71.

41 National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: evaluation, classification and stratification. *Am J Kidney Dis.* 2002;39:1–266.

42 Coresh J., Selvin E., Stevens L.A. et al. Prevalence of chronic kidney disease in the United States. *JAMA.* 2007;298:2038–47.

43 Hallan S.I., Coresh J., Astor B.C. et al. International comparison of the relationship of chronic kidney disease prevalence and ESRD risk. *J Am Soc Nephrol.* 2006;17:2275–84.

44 Levey A.S., de Jong P.E., Coresh J. et al. The definition, classification, and prognosis of chronic kidney disease: a KDIGO Controversies Conference report. *Kidney Int.* 2011;80:17–28.

45 Collins A.J., Foley R.N., Herzog C. et al. United States Renal Data System, USRDS 2012 Annual Data Report: Atlas of Chronic Kidney Disease and End-Stage Renal Disease in the United States. National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases: Bethesda, MD; 2011.

46 Weening J.J., D'Agati V.D., Schwartz M.M. et al. The classification of lupus nephritis revisited. *Kidney Int.* 2004;65:521–30.

47 Markowitz G.S., D'Agati V.D. Classification of lupus nephritis. *Curr Opin Nephrol Hypertens.* 2009;18:220–5.

48 Yokoyama H., Wada T., Hara A. et al. The outcome and a new ISN/RPS 2003 classification of lupus nephritis in Japanese. *Kidney Int.* 2004;66:2382–8.

49 Najafi C.C., Korbet S.M., Lewis E.J. et al. Significance of histologic patterns of glomerular injury upon long-term prognosis in severe lupus glomerulonephritis. *Kidney Int.* 2001;59:2156–63.

50 aHill G.S., Delahousse M., Nochy D. et al. Class IV-S versus class IV-G lupus nephritis: clinical and morphologic differences suggesting different pathogenesis. *Kidney Int.* 2005;68:2288–97.bYu F., Tan Y., Wu L.H. et al. Class IV-G and IV-S lupus nephritis in Chinese patients: a large cohort study from a single center. *Lupus.* 2009 Oct;18(12):1073–81.

51 Haas M. Histologic subclassification of IgA nephropathy: a clinicopathologic study of 244 cases. *Am J Kidney Dis.* 1997;29:829–42.

52 Radford M.G., Donadio J.V., Bergstralh E.J. et al. Predicting renal outcome in IgA nephropathy. *J Am Soc Nephrol.* 1997;8:199–207.

53 D'Amico G. Natural history of idiopathic IgA nephropathy: role of clinical and histological prognostic factors. *Am J Kidney Dis.* 2000;36:227–37.

54 Cattran D.C., Coppo R., Cook H.T. et al. The Oxford classification of IgA nephropathy: rationale, clinicopathological correlations, and classification. *Kidney Int.* 2009;76:534–45.

55 Roberts I.S., Cook H.T., Troyanov S. et al. The Oxford Classification of IgA nephropathy: pathology definitions, correlations, and reproducibility. *Kidney Int.* 2009;76: 546–56.

56 Berden A.E., Ferrario F., Hagen E.C. et al. Histopathologic classification of ANCA-associated glomerulonephritis. *J Am Soc Nephrol.* 2010;21: 1628–36.

57 de Lind van Wijngaarden R.A., Hauer H.A., Wolterbeek R. et al. Clinical and histologic determinants of renal outcome in ANCA-associated vasculitis: a prospective analysis of 100 patients with severe renal involvement. *J Am Soc Nephrol.* 2006;17: 2264–74.

58 Chang D.Y., Wu L.H., Liu G. et al. Re-evaluation of the histopathologic classification of ANCA-associated glomerulonephritis: a study of 121 patients in a single center. *Nephrol Dial Transplant.* 2012;27: 2343–9.

59 Pagni F., Galimberti S., Goffredo P. et al. The value of repeat biopsy in the management of lupus nephritis: an international multicentre study in a large cohort of patients. *Nephrol Dial Transplant.* 2013;28:3014–23.

60 Daleboudt G.M., Bajema I.M., Goemaere N.N. et al. The clinical relevance of a repeat biopsy in lupus nephritis flares. *Nephrol Dial Transplant.* 2009;24: 3712–7.

61 Lu J., Tam L.S., Lai F.M. et al. Repeat renal biopsy in lupus nephritis: a change in histological pattern is common. *Am J Nephrol.* 2011;34:220–5.

62 Ewis A.A., Zhelev Z., Bakalova R. et al. A history of microarrays in biomedicine. *Exp Rev Mol Diag.* 2005;5:315–28.

63 Chandrasekaran K., Karolina D.S., Sepramaniam S. et al. Role of microRNAs in kidney homeostasis and disease. *Kidney Int.* 2012;81:617–27.

64 Williams W.W., Taheri D., Tolkoff-Rubin N. et al. Clinical role of the renal transplant biopsy. *Nat Rev Nephrol.* 2012;8: 110–21.

65 Halloran P.F., de Freitas D.G., Einecke G. et al. An integrated view of molecular changes, histopathology and outcomes in kidney transplants. *Am J Transplant*. 2010;10: 2223–30.

66 Waikar S.S., Betensky R.A., Emerson S.C. et al. Imperfect gold standards for kidney injury biomarker evaluation. *J Am Soc Nephrol*. 2012;23:13–21.

67 Hodgin J.B., Borczuk A.C., Nasr S.H. et al. A molecular profile of focal segmental glomerulosclerosis from formalin-fixed, paraffin-embedded tissue. *Am J Pathol*. 2010;177:1674–86.

68 Sethi S., Vrana J.A., Theis J.D. et al. Laser microdissection and mass spectrometry-based proteomics aids the diagnosis and typing of renal amyloidosis. *Kidney Int*. 2012;82:226–34.

69 Sethi S., Vrana J.A., Theis J.D. et al. Mass spectrometry based proteomics in the diagnosis of kidney disease. *Curr Opin Nephrol Hypertens*. 2013;22(3):273–80.

70 Nasr S.H., Satoskar A., Markowitz G.S. et al. Proliferative glomerulonephritis with monoclonal IgG deposits. *J Am Soc Nephrol*. 2009;20:2055–64.

71 Raissian Y., Nasr S.H., Larsen C.P. et al. Diagnosis of IgG4-related tubulointerstitial nephritis. *J Am Soc Nephrol*. 2011;22: 1343–52.

72 Huisman A., Looijen A., van den Brink S.M., van Diest P.J. Creation of a fully digital pathology slide archive by high-volume tissue slide scanning. *Hum Pathol.* 2010;41:751–7.

73 Weinstein R.S.. Innovations in medical imaging and virtual microscopy. *Hum Pathol.* 2005;36:317–9.

74 Teodorovic I., Isabelle M., Carbone A. et al. TuBaFrost 6: virtual microscopy in virtual tumour banking. *Eur J Cancer*. 2006;42:3110–6.

75 Weinstein R.S. Prospects for telepathology. *Hum Pathol.* 1986;17:433–4.

76 Baak J.P., van Diest P.J., Meijer G.A. Experience with a dynamic inexpensive video-conferencing system for frozen section telepathology. *Anal Cell Pathol.* 2000;21:169–75.

77 Glatz-Krieger K., Spornitz U., Spatz A., Mihatsch M.J., Glatz D. Factors to keep in mind when introducing virtual microscopy. *Virchows Arch.* 2006;448:248–55.

78 Schrader T., Niepage S., Leuthold T. et al. The diagnostic path, a useful visualisation tool in virtual microscopy. *Diagn Pathol.* 2006;1:40.

79 Romer D.J., Yearsley K.H., Ayers L.W. Using a modified standard microscope to generate virtual slides. *Anat Rec B New Anat.* 2003;272:91–7.

80 Gongora J.H., Barcelo H.A. Telepathology and continuous education: important tools for pathologists of developing countries. *Diagn Pathol.* 2008;3(Suppl. 1):S24.

19

CKD: Diagnostic and Other Clinical Needs

Alberto Ortiz

Laboratory of Nephrology, IIS-Fundacion Jimenez Diaz, School of Medicine, UAM, Madrid, Spain
REDinREN, Madrid, Spain
Pathology, IIS-Fundacion Jimenez Diaz, School of Medicine, UAM, Madrid, Spain
IRSIN, Madrid, Spain

19.1 The Evolving Concept of Chronic Kidney Disease

Chronic kidney disease (CKD) is now defined, according to the widely implemented Kidney Disease: Improving Global Outcomes (KDIGO) guideline, as abnormalities of kidney structure or function, present for more than 3 months, with implications for health [1]. The guideline further provides a list of specific criteria that allows the diagnosis of CKD when present for more than 3 months. Importantly, just one criterion is required to diagnose CKD. These criteria include markers of kidney damage, such as pathological albuminuria (>30 mg/g creatinine), urine sediment abnormalities, electrolyte and other abnormalities due to tubular disorders, abnormalities detected by histology, and structural abnormalities detected by imaging or history of kidney transplantation. Alternatively, CKD may be diagnosed solely based on the presence of a decreased estimated glomerular filtration rate (eGFR <60 ml/min/1.73 m^2), although physicians should be very well aware that eGFR is estimating muscle mass from age and sex data and thus, may be misleading when patients have higher or lower muscle mass than expected by these two parameters. KDIGO goes on to recommend the categorization of CKD according to cause, GFR, and albuminuria, which may be better remembered by using the "CGA" acronym. Thus, we should emphasize that every CKD patient should be diagnosed with a cause for CKD, as discussed below. In addition, a G category for GFR and an A category for albuminuria should be provided as follows: Albuminuria A1, urinary albumin creatinine ratio (UACR) <30 mg/g; A2, UACR 30–300 mg/g; A3, UACR >300 mg/g. GFR: G1, eGFR >90 ml/min/1.73 m^2; G2, eGFR 89–60 ml/min/1.73 m^2; G3a, eGFR 59–45 ml/min/1.73 m^2; G3b, eGFR 44–30 ml/min/1.73 m^2; G4, eGFR 29–15 ml/min/1.73 m^2; G5, eGFR <15 ml/min/1.73 m^2. Albuminuria categories A2–A3 or GFR categories G3a–G5 are, by themselves, equivalent to a diagnosis of CKD. A key inference of this categorization is that albuminuria should be assessed whenever CKD is suspected. This is not the case now for many physicians outside nephrology and thus, CKD is underdiagnosed [2, 3]. Underdiagnoses of early stages of CKD, characterized by pathological albuminuria and preserved global kidney function (to the extent that eGFR provides a measure of this), deprives the patient of undergoing evaluation for a cause that can be treated early and also of the possibility to start early prevention of CKD complications, including cardiovascular mortality and CKD progression [4–7].

The current concept of CKD, which does not require a decreased kidney function, encompasses a wider population than prior definition of kidney disease and is an advance over the concept of chronic kidney insufficiency held in the eighties and early nineties (Figure 19.1). Back then, patients were only referred to the nephrologist when they were young and serum creatinine was above normal limits, meaning than GFR was below 50–60 ml/min or if the patient presented symptoms of kidney disease. Both are late manifestations of kidney injury. The situation was even worse in the sixties, when nephrologists treated "uremic syndrome," that is, very late manifestations of kidney failure.

In our view, the concept of CKD is still evolving. The recent realization that systemic inflammation leads to decreased kidney expression of klotho, a kidney-generated hormone with antiaging, vascular protective, and nephroprotective properties, suggests that kidney injury starts much earlier than previously thought [8–12].

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

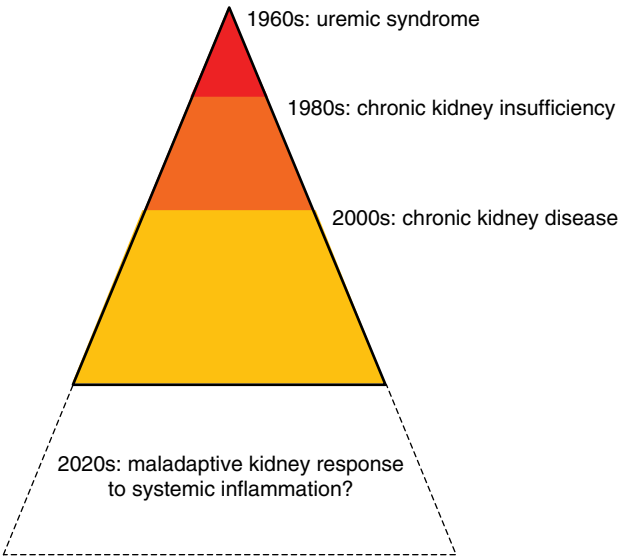


Figure 19.1 The evolving concept of chronic kidney disease (CKD).

In this regard, klotho deficiency may underlie the connection between pathological albuminuria and kidney disease progression or cardiovascular mortality when renal function (assessed by GFR) is still preserved [13].

19.2 A Growing Epidemic

The most important risk factor for CKD is old age. While the prevalence of CKD in the general adult population hovers around 10% [2], it is much higher in the elderly (Figure 19.2). Thus, around two-thirds of persons aged 80 or above have CKD [3]. In this regard, humankind may be divided into two groups: those that have CKD and those that will have CKD, if they live long enough. Moreover, there is a discrepancy between the prevalence of known CKD among the elderly and the real prevalence of CKD (what has been termed "occult" CKD). For some authors this is no major issue, since they believe that CKD in the elderly is part of the "physiological" decrease in GFR with aging. However, it is difficult to consider a "physiological" decrease in GFR that is associated with more than doubling of the risk of all-cause death (Figure 19.3) [14]. Taking into account that at that age the baseline risk of death is already around 60/1000 patient years, doubling the risk is associated with a much greater increase in absolute death rate than greater fold-changes in death risk at younger ages. This is linked with another thought: CKD in the elderly does not occur suddenly; it is frequently a slowly progressive process. Thus, screening programs at younger ages (we would suggest at age 50 and every few years thereafter) may identify patients at earlier stages of the disease, thus allowing early intervention. Moreover, there is plenty of room for improvement in the care of elderly patients with CKD (the margin for improvement in mortality is around 80 deaths/1000 patient years). Thus, this growing segment of the population should be the focus of intensive research and pilot kidney health programs. By contrast, current guidelines suggest that the bulk of elderly patients with CKD should be managed by primary care physicians, according to current state-of-the-art Medicine [15]. This attitude will only perpetuate the high mortality of elderly CKD patients. Lack of referral to Nephrology will deprive them of an in-depth study in search of a treatable etiology and of the latest management strategies for CKD, while depriving the nephrologists of a population base for further research into the contributors to the high mortality, in order to develop novel therapeutic approaches.

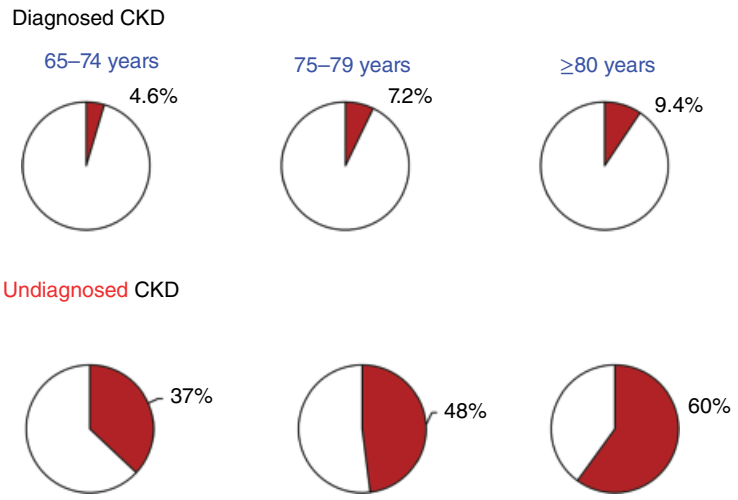


Figure 19.2 CKD: an aging-associated disease (*Source:* Drawn using data from Ref. [3]).

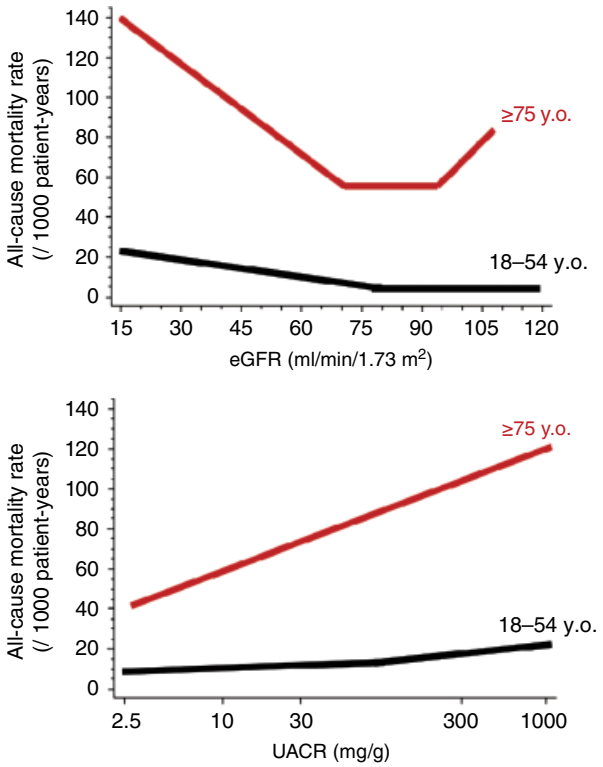


Figure 19.3 CKD: an aging-associated disease that increases the risk of death. Risk for individuals in the 18–54 years age range depicted in black and for individuals 75 years of age or older in red. (*Source:* Drawn using data from Ref. [14]).

19.3 Increasing Mortality from Chronic Kidney Disease

The existence of a higher mortality cannot be dissociated from the concept of CKD, since the eGFR and albuminuria thresholds that define CKD were chosen based on the risk for CKD progression to end-stage kidney disease (ESRD), chronic kidney failure (CKF) in KDIGO nomenclature; and also based on the risk for all-cause and cardiovascular mortality [1]. This increased mortality and CKD progression risk is what the expression "with implications for health" in the KDIGO definition of CKD refers to. In this regard, recent epidemiological data on the global impact of CKD on mortality are worrisome and demand action.

On January 15, 2015, The Lancet published the most recent iteration of the Global Burden of Disease study (GBD2013) [16]. Successive iterations of the study are aimed at updating data and at improving the quality of the data. GDB2013 data show that in terms of median percentage increase in rank 1990–2013 for years of life lost (YLL), CKD was the fastest growing nontransmissible cause of death, growing faster than other well-known and well-publicized causes of death such as diabetes and several heart diseases (Figure 19.4). Only HIV infection, a transmissible form of death, grew faster in this period.

The prior GBD2010 report found that CKD was among the top three fastest growing main causes of death worldwide: the absolute number of deaths from CKD increased by 82% from 1990 to 2010 [17]. Age-standardized death rates from CKD increased by 15% in this period, while rates for most diseases fell, including other nontransmissible diseases such as major vascular diseases, chronic pulmonary disorders, most forms of cancer, and liver cirrhosis [17]. This means that for most nontransmissible causes of death, the global increment in the number of deaths is due to ageing of the population, while for CKD both ageing of the population and increased age-adjusted deaths were contributing. It is possible that reducing death rates from other diseases allows persons to live long enough to develop CKD. In this regard, despite

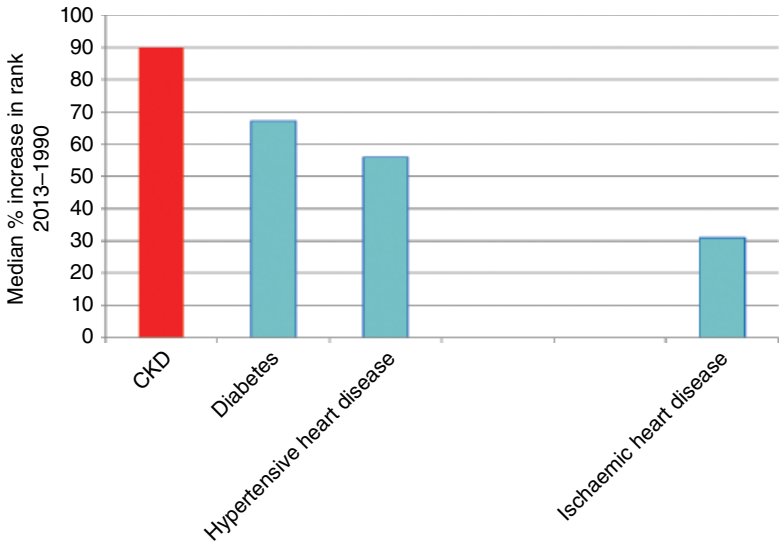


Figure 19.4 CKD is the nontransmissible cause of death that increased the most in the past 20 years according to the global burden of disease 2013 (GBD2013) study [16]. For comparison other key causes of death from chronic disease are shown.

general advances in patient care, mortality of CKF patients remains 10–100-fold higher than in the age-matched general population [18]. The higher mortality of CKF patients cannot be pinpointed to a single cause and is evident for both cardiovascular and noncardiovascular causes. A key concept is that both causes of death (cardiovascular and noncardiovascular) are heterogeneous. Thus, CKF is associated with an increased risk of dying from very different cardiovascular causes, such as atherosclerosis-related cardiovascular disease, arrhythmia, and pulmonary embolism and the pathogenesis and therapy of these entities differs. In some cases, Nephrology has lagged behind other specialties. Hemodialysis is now the most frequent cause of catheter-related bacteremia in the United States, after new standards of care have dramatically decreased catheter-related bacteremia in the intensive care unit [19].

Another issue is that, while much emphasis has been placed on the relative increase in risk of death (fold-increase over age-matched controls), the demographic characteristics of the dialysis population with a predominance of elderly individuals, make the aged primary victims in terms of excess absolute number of deaths over the general population. This concept should be emphasized, since observational studies suggest that a nihilistic approach to therapy in elderly patients might contribute to the increased death rate. That is, lack of therapeutic intervention of very diverse types (from anti-diabetic therapy to beta-blockers or phosphate binders) is associated with increased risk of death in CKF [18]. However, clinical trials are needed to prove a cause–effect relationship.

By far the greatest contributor to CKF mortality worldwide is lack of access to renal replacement therapy. There are not enough resources worldwide to provide dialysis or transplantation to 3 200 000 persons in need every year: there are only resources to provide new renal replacement therapy to around 440 000 persons every year [20]. Limited access to renal replacement therapy is also a problem in Europe, not only in Asia, Africa, and Latin America [21–23]. Inequalities between European citizens in access to renal transplantation are notorious. In addition, the incidence and prevalence of renal replacement therapy in Greece, Belgium (French- or Dutch-speaking), or Portugal is higher than in other European countries and almost double compared to neighboring Netherlands or Spain. In lower-income European countries, a low renal replacement therapy incidence may represent lack of access to needed healthcare (e.g., Montenegro 26 pmp, eightfold lower than in Belgium). However, differences between high-income countries remain unexplained. Given that the working age population peaked worldwide in 2012, it is unlikely that a shrinking working population will develop the resources to treat the ever-growing aging population at highest risk of CKF. Emphasis should be made on early diagnosis, mainly early etiologic diagnosis, and early therapy of CKD, preferably etiology-based therapy, to prevent progression, as well as in developing new concepts for renal replacement therapy that may be more widely applied, perhaps by means of tissue engineering.

19.4 The Issue of Cause and Etiologic Therapy

There is a dearth of etiology-based therapy in CKD as well as of tools for the noninvasive assignment of etiology. Moreover, nonspecific therapy for CKD has seen little innovation in recent decades [24].

Therapy for diabetic kidney disease (DKD) is still based on drugs first shown to decrease proteinuria in this condition in the mid-eighties [25]. Multiple clinical trials have failed to identify novel therapeutic approaches to be used in daily clinical practice as an add-on to renin–angiotensin system blockade to improve outcomes [26, 27]. Today, DKD remains a clinical diagnosis and renal biopsies are rarely performed. Disease heterogeneity may have contributed to negative trials and newer, noninvasive biomarkers of diagnosis and disease stage are sorely needed.

Hypertensive kidney disease is the second-most frequent cause of CKF in the United States and in at least some European countries [24]. However, many now doubt that hypertensive kidney disease even exists [28]. In the United States, hypertensive kidney disease is diagnosed mainly in African Americans. However, recent data point to a genetic basis for CKD in African Americans: a genetic variant in the *ApoL1* gene that confers resistance to infection by *Trypanosoma brucei* [29, 30]. These genetic variants confer sensitivity to DKD and HIV nephropathy. It is highly likely that APOL1-related kidney injury has been misdiagnosed as hypertensive nephropathy for decades. The therapeutic implications of this finding are potentially huge, since if hypertension is a consequence, rather than the cause of kidney disease, the therapeutic approach should directly address kidney injury rather than blood pressure. There is similar confusion regarding the prevalence of hypertensive nephropathy in Europe [31]. The adjusted incident rates of RRT per million population for hypertensive nephropathy ranged from 38.8 in Iceland to 4.2 in Finland and 4.2 in Scotland. In European countries that reported the existence of nonhypertensive renal vascular disease as a cause of RRT, the ratio of hypertensive to nonhypertensive vascular disease as a cause of RRT ranged from

0.5 in Croatia and Austria (reflecting a predominance of nonhypertensive vascular causes) to 32 in Norway (reflecting a preponderance of hypertensive vascular causes of CKF). Are these differences real? Or given the absence of specific biomarkers and histological features, should the term hypertensive nephropathy be interpreted as an educated way of referring to CKD of unknown origin? In this regard, a pathogenic therapeutic approach cannot be prescribed if the cause of the nephropathy is unclear or even explicitly unknown, which is the case for 2–60% of RRT patients in different European countries.

Glomerulonephritis, interstitial nephropathies, and cystic diseases are also major causes of CKF. Current understanding of glomerular disease is still not enough to provide specific pathogenic therapy. Glomerulonephritides are still treated with either nonspecific proteinuria lowering medication or nonspecific immune suppressants. In this regard, glomerulonephritides are still classified based on morphological criteria that date back to the seventies. This most probably represents an oversimplification that throws into the same morphological basket different conditions in terms of pathogenesis, severity, and progression potential that may require very different therapeutic approaches. A molecular classification, similar to that used for some malignancies, may allow a better assessment for pathogenesis, staging, and therapy selection for personalized medicine.

There is also a lack of understanding of the etiology and pathogenesis of most primary chronic tubulointerstitial nephropathies that contribute to the scarcity of biomarkers and specific diagnostic criteria and therapy.

It is yet unknown whether general management of CKD may result in improved outcomes for autosomal dominant polycystic kidney disease (ADPKD) in terms of preservation of renal function [32]. The recent approval of tolvaptan for the treatment of ADPKD in Japan and in Europe may provide the stimulus for further research into the phenotypic heterogeneity of ADPKD and lead to the identification of noninvasive biomarkers that guide the indication and follow-up of therapy [33].

19.5 Unmet Medical Needs: Biomarkers and Therapy

The major unmet medical needs in Nephrology and suggested actions have been recently reviewed [24]. Development of novel diagnostic, risk stratification, and individualization tools to personalize therapeutic approaches is sorely needed.

Accurate, sensitive, specific, and noninvasive diagnostics tests should allow the identification of the etiology of CKD both early in the course of the disease and when patients seek medical care at a later stage. The creation of a cohort of young adults without prior known kidney disease with sequential clinical and analytical follow-up and sequential biobanking may provide the materials that allow the identification of early markers of kidney disease.

A molecular or pathophysiological classification of kidney diseases and specifically glomerulonephritis is needed to complement or even replace the current morphological or clinicopathological ones. This should identify specific molecular signatures and targets and predict progression and response to therapy in order to guide the indication and monitoring of specific therapeutic approaches. Systems biology approaches applied to blood, urine, or kidney tissue in humans or experimental systems hold promise for such classification or identification of components of the classification [34–39].

Imaging techniques should advance in order to allow repetitive, noninvasive monitoring of kidney inflammation and fibrosis and assessment of diverse kidney functions in a dynamic manner that allows characterization of active versus chronic lesions and progressing versus stable chronic lesions.

Novel preventive and therapeutic approaches should be developed based on etiologic and pathophysiological insights. Nephrology needs novel therapeutic approaches for unmet needs and these can only be developed by a precise and detailed understanding of pathophysiological events, the evaluation of preclinical models relevant to the human situation, and the improved design of lower-cost clinical trials.

The importance of representative animal models cannot be overemphasized [40–43]. In the absence of adequate animal models, we are left with testing novel potential therapeutic approaches directly in humans, exposing those pioneers to unknown risks. Conversely, inadequate animal models lead to equivocal hypothesis, destined to fail when tested in clinical trials.

The bedside-to-bench interaction should intervene at every step of clinical development. As an example, enzyme replacement therapy (ERT) for Fabry disease was expected to solve the enzymatic defect and prevent progression of this proteinuric nephropathy. However, ERT failed to prevent progression once nephropathy was present [44], raising the need to better understand the molecular mediators of tissue injury. This led to the identification of lyso-Gb3, a glycolipid that accumulates in Fabry disease but is not normalized by ERT, as a promoter of fibrosis, a hallmark of the disease, in podocytes [45–47]. This knowledge led to the characterization of inhibitors of lyso-Gb3 actions on podocytes, including

vitamin D receptor activators and inhibitors of TGFBeta1 and Notch signaling [45–47].

Eventually, these advances should be validated in clinical trials. Clinical trial design optimization is required in Nephrology, following a series of big misses. Furthermore, the number of randomized controlled trials published in Nephrology from 1966 to 2002 was 50–90% lower than in every other medical specialty [48]. Still in 2014, a systematic review of Clinicaltrials.gov disclosed that only <3% of trials were classified as Nephrology [49].

A major issue in Nephrology RCTs is the definition of surrogate end-points that allow the design of adequately powered trials that are feasible in terms of sample size and length of follow-up. The emphasis on hard end-points is problematic for many kidney diseases with a natural history measured in decades: advanced cases must be enrolled in order to get sufficient events, but advanced cases may be less responsive to therapy, especially to etiologic therapy, given that a sizable amount of renal mass may have already been lost. In this regard, the qualification of novel biomarkers that may be used as surrogate end-points should also be facilitated [38].

19.6 Conclusions

CKD is a highly prevalent and growing problem which leads to both CKF requiring renal replacement therapy and to premature mortality. The GBD study has observed that CKD is among the fastest growing causes of premature death worldwide. This is not surprising, given the chronic underfunding of both the research effort and population-level screening programs and interventions. As a consequence, we lack tools for early noninvasive diagnosis of CKD, for predicting progression in nonalbuminuric CKD and for disease staging. This, associated to insufficient knowledge of pathophysiology and scarcity of RCTs, has led to very few therapeutic innovations in recent decades. Only major funding efforts by private and public investors, together with a global rethinking of the strategy to tackle the CKD epidemic and to provide novel modes of renal function replacement, may eventually control the CKD epidemic in the same way that progress is being made to prevent and treat cardiovascular disease and cancer. In this regard, disadvantaged populations are particularly prone to CKD [50, 51].

Acknowledgments

Grant support: ISCIII and FEDER funds PI13/00047, PIE13/00051, EUTOX, Sociedad Española de Nefrologia, ISCIII-RETIC REDinREN RD012/0021, Comunidad de Madrid CIFRA S2010/BMD-2378. Salary support: Programa Intensificación Actividad Investigadora (ISCIII/Agencia Laín-Entralgo/CM) to AO.
Disclosure: No competing interests.

References

1 Kidney Disease: Improving Global Outcomes (KDIGO) CKD Work Group. KDIGO 2012 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney Int Suppl.* 2013;3:1–150.
2 Otero A, de Francisco A, Gayoso P, Garcia F. Prevalence of chronic renal disease in Spain: results of the EPIRCE study. *Nefrologia.* 2010;30(1):78–86.
3 Stevens LA, Li S, Wang C, et al. Prevalence of CKD and comorbid illness in elderly patients in the United States: results from the Kidney Early Evaluation Program (KEEP). *Am J Kidney Dis.* 2010;55(3 Suppl 2):S23–S33.
4 Kidney Disease: Improving Global Outcomes (KDIGO) CKD–MBD Work Group. KDIGO clinical practice guideline for the diagnosis, evaluation, prevention, and treatment of chronic kidney disease–mineral and bone disorder (CKD–MBD). *Kidney Int.* 2009;76(Suppl 113):S1–S130.
5 Ortiz A, Sanchez-Niño MD. The demise of calcium-based phosphate binders. *Lancet.* 2013;382:1232–1234.
6 Gallegos-Villalobos A, Portolés J, Ortiz A. Application of the new cholesterol guidelines. *N Engl J Med.* 2014;371:77–78.
7 Tonelli M, Wanner C. Kidney disease: Improving Global Outcomes Lipid Guideline Development Work Group Members. Lipid management in chronic kidney disease: synopsis of the kidney disease: Improving Global Outcomes 2013 clinical practice guideline. *Ann Intern Med.* 2014;160:182.
8 Moreno JA, Izquierdo MC, Sanchez-Niño MD, et al. NFκB-dependent regulation of Klotho expression by the inflammatory cytokines TWEAK and TNFα in kidney cells. *J Am Soc Nephrol.* 2011;22:1315–1325.
9 Sanchez-Niño MD, Sanz AB, Ortiz A. Klotho to treat kidney fibrosis. *J Am Soc Nephrol.* 2013 Apr;24(5):687–689.
10 Izquierdo MC, Perez-Gomez MV, Sanchez-Niño MD, et al. Klotho, phosphate and inflammation/ageing in chronic kidney disease. *Nephrol Dial Transplant.* 2012 Dec;27 (Suppl 4):iv6–iv10.

11 Kurosu H, Yamamoto M, Clark JD, et al. Suppression of aging in mice by the hormone Klotho. *Science*. 2005 Sep 16;309(5742):1829–1833.

12 Kuro-o M, Matsumura Y, Aizawa H, et al. Mutation of the mouse klotho gene leads to a syndrome resembling ageing. *Nature*. 1997 Nov 6;390(6655):45–51.

13 Ortiz A, Fernandez-Fernandez B. Humble kidneys predict mighty heart troubles. *Lancet Diabetes Endocrinol*. 2015 Jul;3(7):489–491.

14 Hallan SI, Matsushita K, Sang Y, et al. Age and association of kidney measures with mortality and end-stage renal disease. *JAMA*. 2012 Dec 12;308(22):2349–2360.

15 Martínez-Castelao A, Górriz JL, Segura-de la Morena J, et al. Consensus document for the detection and management of chronic kidney disease. *Nefrologia*. 2014;34(2):243–262.

16 GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age -sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*. 2015 Jan 10;385(9963):117–171.

17 Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012;380:2095–2128.

18 Ortiz A, Covic A, Fliser D, et al. Epidemiology, contributors to, and clinical trials of mortality risk in chronic kidney failure. *Lancet*. 2014;383:1831–1843.

19 James MT, Conley J, Tonelli M, Manns BJ, MacRae J, Hemmelgarn BR. Meta-analysis: antibiotics for prophylaxis against hemodialysis catheter-related infections. *Ann Intern Med*. 2008;148:596–605.

20 Anand S, Bitton A, Gaziano T. The gap between estimated incidence of end-stage renal disease and use of therapy. *PLoS One*. 2013;8(8):e72860.

21 Gonzalez-Espinoza L, Ortiz A. 2012 ERA-EDTA registry annual report: cautious optimism on outcomes, concern about persistent inequalities and data black-outs. *Clin Kidney J*. 2015 Jun;8(3):243–247.

22 Pippias M, Stel VS, Abad Diez JM et al. Renal replacement therapy in Europe: a summary of the 2012 ERA-EDTA Registry Annual Report. *Clin Kidney J*. 2015;8(3):248–261.

23 Rosa-Diez G, Gonzalez-Bedat M, Pecoits-Filho R, et al. Renal replacement therapy in Latin American end-stage renal disease. *Clin Kidney J*. 2014 Aug;7(4):431–436.

24 Ortiz A. Translational nephrology: what translational research is and a bird's-eye view on translational research in nephrology. *Clin Kidney J*. 2015 Feb;8(1):14–22.

25 Taguma Y, Kitamoto Y, Futaki G, et al. Effect of captopril on heavy proteinuria in azotemic diabetics. *N Engl J Med*. 1985;313:1617–1620.

26 Fernandez-Fernandez B, Ortiz A, Gomez-Guerrero C, Egido J. Therapeutic approaches to diabetic nephropathy-beyond the RAS. *Nat Rev Nephrol*. 2014;10:325–346.

27 Perez-Gomez MV, Sanchez-Niño MD, Sanz AB, et al. Horizon 2020 in diabetic kidney disease: the clinical trial pipeline for add-on therapies on top of renin angiotensin system blockade. *J Clin Med*. 2015;4:1325–1347.

28 Freedman BI, Sedor JR. Hypertension-associated kidney disease: perhaps no more. *J Am Soc Nephrol*. 2008 Nov;19(11):2047–2051.

29 Genovese G, Friedman DJ, Ross MD, et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*. 2010;329:841–845.

30 Lipkowitz MS, Freedman BI, Langefeld CD, et al. Apolipoprotein L1 gene variants associate with hypertension-attributed nephropathy and the rate of kidney function decline in African Americans. *Kidney Int*. 2013;83:114–120.

31 ERA-EDTA Registry. Annual report 2012. http://www.era-edta-reg.org/files/annualreports/pdf/AnnRep2012.pdf; accessed August 23, 2017.

32 Rodriguez-Osorio L, Perez-Gomez VM, Ortiz A. Decreasing incidence of renal replacement therapy over time at the critical 50-59-year age range suggests a role for nephroprotective therapy in ADPKD. *Kidney Int*. 2015 Jul;88(1):194.

33 Torres VE, Chapman AB, Devuyst O, et al. Tolvaptan in patients with autosomal dominant polycystic kidney disease. *N Engl J Med*. 2012 Dec 20;367(25):2407–2418.

34 Schanstra JP, Zürbig P, Alkhalaf A, et al. Diagnosis and prediction of CKD progression by assessment of urinary peptides. *J Am Soc Nephrol*. 2015 Aug;26(8):1999–2010.

35 Posada-Ayala M, Zubiri I, Martin-Lorenzo M, et al. Identification of a urine metabolomics signature in patients with advanced-stage chronic kidney disease. *Kidney Int*. 2014 Jan;85(1):103–111.

36 Benito-Martin A, Ucero AC, Zubiri I, et al. Osteoprotegerin in exosome-like vesicles from human cultured tubular cells and urine. *PLoS One*. 2013 Aug 23;8(8):e72387.

37 Zubiri I, Posada-Ayala M, Sanz-Maroto A, et al. Diabetic nephropathy induces changes in the proteome of human urinary exosomes as revealed by label-free comparative analysis. *J Proteomics*. 2013 Nov 7;96C:92–102.

38 Mischak H, Ioannidis JP, Argiles A, et al. Implementation of proteomic biomarkers: making it work. *Eur J Clin Invest*. 2012;42:1027–1036.

39 Siwy J, Schanstra JP, Argiles A, et al. Multicentre prospective validation of a urinary peptidome-based classifier for the diagnosis of type 2 diabetic nephropathy. *Nephrol Dial Transplant.* 2014;29:1563–1570.

40 Ramos AM, González-Guerrero C, Sanz A, et al. Designing drugs that combat kidney damage. *Expert Opin Drug Discov.* 2015 Apr;3:1–16.

41 Ortiz A, Sanchez-Niño MD, Izquierdo MC, et al. Translational value of animal models of kidney failure. *Eur J Pharmacol.* 2015 Jul 15;759:205–220.

42 Díaz-García JD, Gallegos-Villalobos A, Gonzalez-Espinoza L, Sanchez-Niño MD, Villarrubia J, Ortiz A. Deferasirox nephrotoxicity-the knowns and unknowns. *Nat Rev Nephrol*. 2014 Oct;10(10):574–586.

43 Sanz AB; Sanchez-Niño MD, Martín-Cleary C, Ortiz A, Ramos AM. Progress in the development of animal models of acute kidney injury and its impact on drug discovery. *Expert Opin Drug Discovery.* 2013;8:879–895.

44 Warnock DG, Ortiz A, Mauer M, et al. Renal outcomes of agalsidase beta treatment for Fabry disease: role of proteinuria and timing of treatment initiation. *Nephrol Dial Transplant*. 2012;27:1042–1049.

45 Sanchez-Niño MD, Carpio D, Sanz AB, Ruiz-Ortega M, Mezzano S, Ortiz A. Lyso-Gb3 activates Notch1 in human podocytes. *Hum Mol Genet*. 2015 Oct 15;24(20):5720–5732.

46 Sanchez-Niño MD, Sanz AB, Carrasco S, et al. Globotriaosylsphingosine actions on human glomerular podocytes: implications for Fabry nephropathy. *Nephrol Dial Transplant.* 2011 Jun;26(6):1797–1802.

47 Weidemann F, Sanchez-Niño MD, Politei J, et al. Fibrosis: a key feature of Fabry disease with potential therapeutic implications. *Orphanet J Rare Dis.* 2013;8:116.

48 Strippoli GF, Craig JC, Schena FP. The number, quality, and coverage of randomized controlled trials in nephrology. *J Am Soc Nephrol.* 2004;15:411–419.

49 Inrig JK, Califf RM, Tasneem A, et al. The landscape of clinical trials in nephrology: a systematic review of Clinicaltrials.gov. *Am J Kidney Dis*. 2014;63:771–780.

50 Martín-Cleary C, Ortiz A. CKD hotspots around the world: where, why and what the lessons are. A CKJ review series. *Clin Kidney J.* 2014;7:519–523.

51 Garcia-Garcia G, Jha V, on behalf of the World Kidney Day Steering Committee. CKD in disadvantaged populations. *Clin Kidney J.* 2015 Feb;8(1):3–6.

20

Molecular Model for CKD

Marco Fernandes[1], *Katryna Cisek*[2], *and Holger Husi*[1,3]

[1] *Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK*
[2] *Mosaiques Diagnostics GmbH, Hannover, Germany*
[3] *Department of Diabetes and Cardiovascular Science, Centre for Health Science, University of the Highlands and Islands, Inverness, UK*

20.1 Introduction

Chronic kidney disease (CKD) is widely recognized as having a multifactorial origin with both acquired and inherited risk factors, and poses a major health and financial burden in modern societies. Taking into account the last two decades, the number of CKD-related deaths has risen by 82.3%, which is the third largest increase among the top 25 leading causes of death worldwide [1]. Furthermore, the health costs associated with CKD morbidity as the treatment of end-stage renal disease (ESRD) in many developed countries can easily ascend to 3% of their internal healthcare budget [2].

Understanding the molecular basis underlying onset and progression of multifactorial diseases such as CKD requires an interdisciplinary combination of existing fundamental knowledge with new data obtained from several omics platforms. Thereby, we can foresee that Systems Biology approaches are required in order to tackle the biological complexity that characterizes kidney diseases.

20.2 Data-Driven Approaches and Multiomics Data Integration

Systems biology requires comprehensive data at all molecular levels. Hence using a data-driven approach based on differential molecular expression profiles a disorder and/or disorder progression stages could be a promising approach to find potential body fluid-accessible biomarkers and therapeutic targets. Nevertheless, harmonizing single-level omics data obtained from different studies and mining the increasing volume of data across databases in an efficient way is still a big challenge [3].

With the advent of high-throughput technologies, their application in the biomedical field was a foreseen logical step. However, until recently, integration of multiomic data was not a common approach. Therefore, we briefly present a potential workflow (Figure 20.1) to handle multilevel 'omic data that could lead to new research questions and/or clinical applications.

The application of high-throughput omic platforms leads to the generation of large datasets (e.g., spot intensities, spectral data) that are commonly stored in local databases. Then, the data undergo several steps of cleaning, filtering, normalization, reduction, and other preprocessing steps that are dependent of their intrinsic properties. Thereafter, data are mapped to several (ideally) external database repositories in order to ensure future reuse and as well to allow the use of a wide range of analysis tools that normally require different molecular identifiers. This molecular information can be then integrated in a further stage by means of a meta-analysis or by cross-normalization of data from different acquisition platforms. A data integration approach can be used in order to develop models than can have a topological origin, that is, connectivity between nodes in a tridimensional network. An alternative is to recreate the cell environment and dynamics by describing the interactions on a qualitative and quantitative manner and add the underlying data for connectivity; for example, protein–protein interactions (PPIs), molecular co-occurrence, ontologies, enzymatic reactions for metabolites, and so on. Another approach relies merely on a statistical assessment, which tries to deconvolute the representation of features and variables by mathematical methods and thus requires input of biological context after the analysis. The state-of-art model would consider all of these three components simultaneously: network topology, molecular

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

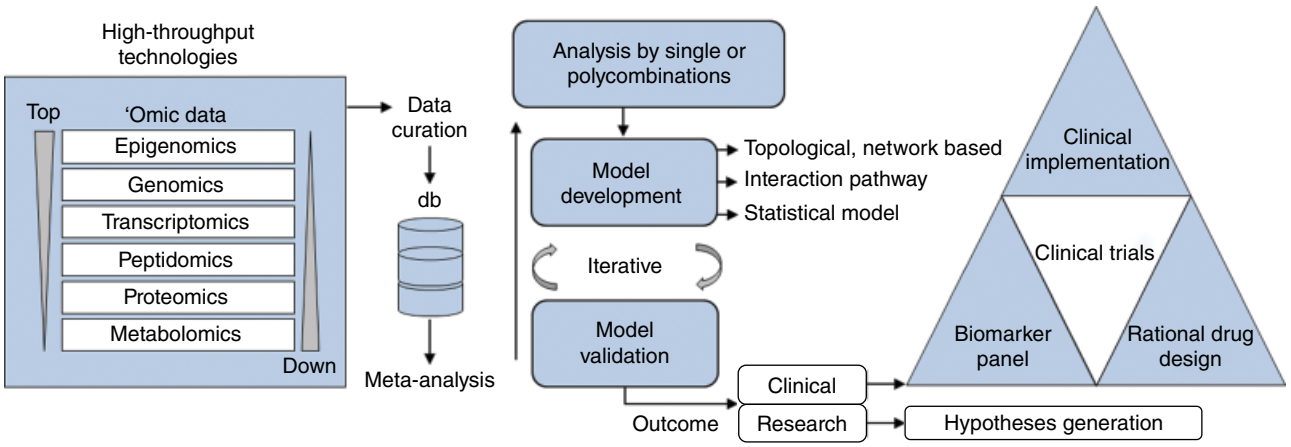


Figure 20.1 Purposed workflow for a data-driven approach in biomarker discovery, starting from data acquisition till clinical implementation. db, database.

Table 20.1 Model development in several disease/conditions and omics platforms.

Disease/condition	Organism	'omic platform	Type of model	Data source	Model validation	Reference
AKI	Mouse	Proteomics	de novo pathway	Generated plus literature	Immunohistochemistry	[5]
Diabetic nephropathy	Human	Multiomics	Protein interaction network	omics studies, patent text and clinical trials		[6]
Muscle Invasive Bladder Cancer	Human	Transcriptomics	Protein interaction network	gene2pubmed		[7]
CKD	Human	—	Logic programming	Clinical data		[8]
DM-associated vascular disease	Mouse	Proteomics	de novo pathway	Generated	Immunohistochemistry	[9]

interactions, and statistical criteria in order to provide the most robust representation of changes associated with a disease state. Moreover, every new model needs to be corroborated, either by adding new data to the previous model and then test how it responds, or additionally with validation of the features that support the model (experimental validation: immunohistochemistry, qRT-PCR, luciferase reporter assay, ELISA, etc.). This is an iterative process leading to the creation of a model and involves several cycles of adding new data and testing its validity again. The model can then serve as a clinical tool for patient stratification or it can be applied in research for generating new hypotheses [4].

In this chapter, we are going to focus on and describe the application of standalone and web-based tools used in pathway(s) visualization and the development of disease models (Table 20.1) based on data acquired by several omics technologies. Finally, we are going to present CKD as a case study in order to illustrate the development of a disease model in a stepwise way, from single models to fully integrative and combinatorial models.

20.2.1 Database Resources

The ongoing impressive growth of the amount of scientific data, caused by rapid technological improvements and common usage of high-throughput technologies, created the need for effective data handling and analysis. Together with increased capabilities of generating, collecting, storing, and managing information, data mining is becoming an indispensable tool in research. The employment of database systems into the research workflow, resulted in easier information retrieval and management, and hence empowered the development of more comprehensive data analysis techniques that enhances knowledge extraction. This is particularly important in life sciences, given the high complexity and multidimensionality of biological data and the growing trend toward integrative analysis and modeling [10].

In nephrology, several databases (Table 20.2) have been developed to collect information for computational modeling, including the Chronic Kidney Disease database (CKDdb), the Kidney and Urinary Pathway Knowledge Base (KUPKB), GeneKid, and Nephroseq (formerly known

Table 20.2 Databases within the scope of CKD or of general scope that can be filtered by conditions and 'omics platforms based on expression profiles.

Name	Organism	'omic platform	Data source	Description	URL
CKD-related					
CKDdb	Multispecies	Multiomics	Publications	Clustered, differential expression	www.padb.org/ckddb
KUPKB	Multispecies	Multiomics	Publications	Differential expression	www.kupkb.org
Nephroseq	Multispecies	Transcriptomics	GEO	Differential expression	www.nephroseq.org
PeptiCKDdb	Human	Proteomics	Publications	Differential expression	www.peptickddb.com
RGED	Human	Transcriptomics	GEO	Differential expression	http://rged.wall-eva.net
PKDB	Human	SNP analysis	Publications	Gene variants for ADPKD	http://pkdb.pkdcure.org
General scope					
GEO profiles	Multispecies	Gene expression	User submission	Differential expression	www.ncbi.nlm.nih.gov/geoprofiles
EBI expression atlas	Multispecies	Transcriptomics	User submission	Differential and baseline expression	www.ebi.ac.uk/gxa
MOPED	Multispecies	Transcriptomics and proteomics	User submission and GEO	Differential expression	www.proteinspire.org/MOPED
UPdb	Human	Proteomic fingerprint	Own data and literature	Peak profiling	www.padb.org/updb/updb.html
LSSR—Large Scale Screening Resource	Multispecies	Proteomics	Publications	Differential expression	www.padb.org/lssr
PRIDE	Multispecies	Proteomics	User submission	Differential expression	www.ebi.ac.uk/pride/archive

as Nephromine). CKDdb stores microRNA, genomics, peptidomics, proteomics, and metabolomics information relevant to CKD, collected from over 300 studies in the literature and integrated into the Pan-omics Analysis DataBase (PADB) using gene and protein clusters (CluSO) and mapping of orthologous genes (OMAP) between species. This resource integrates highly diverse omics data across various species in one platform and allows for a systematic evaluation of CKD-relevant pathways using a systems biology approach. KUPKB compiles mRNA, miRNA, metabolite and protein datasets from literature, as well as Gene Expression Omnibus (GEO) relevant to kidney pathology and physiology; this information is implemented using Semantic Web technologies, a protocol to standardize content published and shared on the Internet. Moreover, KUPKB is linked to additional resources, such as NCBI gene, UniProt, Homologene, and Kyoto Encyclopedia of Genes and Genomes (KEGG), allowing complex queries to return all the relevant linked information, across species and including biological pathways. GeneKid, a pipeline created for the SysKid consortium project, which aims to develop new diagnostics and treatments for CKD, focuses on harmonizing heterogeneous omics data by using the genes' annotation network ("symbolization") to build a unified omics network. The challenge of this approach is assigning all nonunique gene identifiers to one correct HUGO gene nomenclature committee (HGNC) symbol, due to nomenclature variability between laboratories, especially for linking genes with cellular metabolites. In order to improve this linkage, metabolite identities are retrieved using the Human Metabolome Database (HMDB) and DrugBank database. Lastly, Nephroseq is a data-mining engine of preanalyzed clinical and molecular transcriptomics datasets of kidney disease and its co-morbidities from human and mouse studies. In addition to being a database with gene expression profiles for molecules of interest, this resource also integrates KEGG pathways, predicted microRNA targets, Human Protein Reference Database (HPRD) interaction sets, and allows for co-expression, outlier detection, and concept analysis that permits meta-analysis of gene expression trends. But while these data sources provide public access to various omics datasets and bioinformatics mining tools, none fully integrates the whole landscape of omics disciplines into a comprehensive, dynamic, and visual model of cellular biochemistry. Genecards (http://www.genecards.org/) is an integrated database of human genes that includes automatically mined genomic, proteomic, and transcriptomic information, in

addition to gene orthologues, disease–disease relationships, single nucleotide polymorphisms (SNPs), gene expression, gene function, and providing web links for ordering assays and antibodies. The GeneCards database, created in 1997, is being updated and managed by the Crown Human Genome Center at the Weizmann Institute of Science in Israel. Entrez Gene (http://www.ncbi.nih.gov/entrez/) is the National Center for Biotechnology Information (NCBI) database for genetic data, including nomenclature, map location, gene products, and their attributes; markers; phenotypes; and links to citations, sequences, variation details, maps, expression, homologs, protein domains, and external databases. This database houses the genomes that have been completely sequenced or that are of interest to the research community and are scheduled for intense sequence analysis. The content of Entrez Gene represents the result of the effort of both biocuration and automated data integration from NCBI's Reference Sequence project (RefSeq), from collaborating model organism databases, and from many other databases available from NCBI. OMIM is the acronym for the Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM). This database is a catalogue of human genes and genetic disorders. The National Center for Biotechnology Information (NCBI) has developed the online interface for this database, which is administered by Dr Victor A. McKusick and his colleagues at Johns Hopkins. The database contains textual information and references, connections to the bibliographic medical database MEDLINE and to sequence records hosted in the NCBI Entrez system, and links to additional related resources at NCBI and elsewhere. Disease Genes Database (http://www.proteinlounge.com/epath3d/eprotein-overview.asp) is a repository of disease-causing genes and their respective diseases. The database is linked to the Protein Lounge Pathway Database and Protein Database with additional information on each particular gene of interest. The content of the database includes disease—gene searches, gene sequence information, gene protein products, gene family information, pathway information for the genes, gene signal transduction, and related article information. Disease Centered Central Mutation Databases is a consortium of disease-causing genetic databases, such as the Asthma genetic database, the repertory of Familial Mediterranean Fever (FMF) and Hereditary Inflammatory Disorders Mutations database, and the Keio Mutation Databases using Mutation View for the analysis of eye, heart, ear, brain, and cancer disease-causing genes (http://www.hgvs.org/disease-centered-central-mutation-databases).

20.2.2 Software Tools and Solutions

The number of Bioinformatics tools associated with omics and disease analysis is vast; thus here we only present the ones that are more user friendly and generate end results allowing a meaningful biological interpretation (Table 20.3).

Table 20.3 Overall view of standalone and web-based tools to assist in pathway(s) visualization and in the development of disease models.

Name	Type	License	(+) aspects	(–) aspects	URL
Cytoscape	Network generation	Free	Hundreds of applications available	Cytoscape-version dependent apps	http://www.cytoscape.org
PathVisio	Pathway construction	Free	Statistics associated with pathways	Unable to handle metabolites and genes/proteins simultaneously	www.pathvisio.org
KEGG mapper	Pathway mapping	Free	Simultaneously search and coloring features	Without statistical output; Errors on IDs matching	www.genome.jp/kegg/mapper.html
VisANT	Integrative visual data-mining of pathways	Free	Visualization of dynamic flux processes	Mainly focused on metabolic networks	http://visant.bu.edu
Ingenuity IPA	Integration of multiomics	Commercial	User-friendly, no special skills required	Lack of reproducibility. No trace of the underlying databases used	www.ingenuity.com/products/ipa
MetaCore	Data-mining and pathway analysis	Commercial	User-friendly, no special skills required	No public details of how it works	https://portal.genego.com
3omics	Integration of human transcriptomic, proteomic, and metabolomic data	Free	Literature-based imputation of missing molecular elements	Phenotypic analysis missing	http://3omics.cmdm.tw

Many modern high-throughput technologies lead to the generation of large-scale and complex datasets, including PPIs (e.g., Yeast two-hybrid, Y2H, screening), protein–DNA interactions (Chromatin Immunoprecipitation, ChIP, Assays), kinase–substrate interactions (e.g., protein microarrays) [11], qualitative and quantitative genetic interactions (e.g., RNA interference, RNAi, high-throughput screening) [12], gene co-expression, and so on [13] The "Big Data" problem can be solved by the development of Bioinformatics tools that are able to handle these large datasets in order to reduce their complexity to a level that enables rational interpretation and could provide new biological insights. Cytoscape is an open-source software tool running in the Java programming environment and is capable to assemble large-scale datasets into biological networks, in such a way that allows simultaneous integration, visualization, and analysis of interacting molecular components with associated expression data into a combined conceptual framework [14]. The Cytoscape platform offers several applications (301 applications available on November 19, 2016) for a diverse array of uses and analysis types. A summary description and assessment of some of these applications can be additionally seen in a review by Saito and collaborators [13].

20.2.2.1 Gene Ontology (GO) and Pathway-Term Enrichment

To overcome the challenge of providing biological meaning and context when dealing with, for example, large-scale gene data from high-throughput experiments, many ontology sources exist in order to capture biological information in a meaningful way. The Gene Ontology (GO) consortium [15] aims to capture the increasing knowledge on gene function in a controlled vocabulary applicable to a wide range of organisms. Even though it is designated as GO, it represents genes and gene product attributes on matters of their associated biological processes (BP), cellular components (CC), and molecular functions (MF). Between the terms there is a hierarchical relationship (parent–child). Due to the complexity of hierarchy structure, the terms can be in several different levels. The specificity of the terms varies along the tree: from very general terms (in first levels of GO) to very specific. For a complete view on the studied process, several ontology sources should be consulted in order to integrate their complementary information. The amount of information associated with each source, each individual gene, is overwhelming and renders the analysis of the relationship between genes and between terms very difficult to represent and elucidate. Also, for closely related terms, a high degree of redundancy of their associated genes exists. In order to perform an improved GO analysis, the ClueGO application [16] for Cytoscape was developed and allows visualization and integration of nonredundant biological terms for large clusters of genes in a functionally grouped network. Alongside with the analysis of a single gene set list, that is, cluster, ClueGO is able to perform comparison of several clusters, illustrating the specificity and also the common aspects of their functionality by allowing comparison between datasets. From the ontology sources used, the terms are selected by different filter criteria. Related terms which share similar associated genes can be combined in order to reduce redundancy [16].

ClueGO is used for the integration and visualization of GO and pathway terms sourced from KEGG [17], WikiPathways [18], and Reactome [19]. The resultant ClueGO network is established based on kappa statistics (chance-corrected measures of agreement), in which the kappa coefficient (ranging 0–1) shows the concordance on how any given gene and/or gene product pairs share similar annotations of terms [16, 20]. The visualization of the generated network is driven by the selection of the kappa coefficient, in which a high coefficient leads to the connections only among closely related terms with significant overlap in associated gene products. On the other hand, a lower coefficient will allow visualizing the connections between less related terms. There are several analysis parameters (GO terms/pathway selection, levels of GO tree interval, kappa score network connectivity, among other statistical and grouping options) in ClueGO. Therefore, achieving a compromise between the adjustment of these parameters with the associated empirical knowledge of the user regarding his samples/input dataset will promote an enhanced exploration of this software tool. In addition, the functionality of this application can be expanded by joint analysis with the CluePedia application that offers additional information on pathways of interest to the user [21].

20.2.2.2 Disease–Gene Associations

The conclusion of the Human Genome Project (HGP) led to an unprecedented increase of studies related to uncovering the role of genetics in human disorders [22]. This event translated in a disparate growth in the number of publications and on the other side a limited and slow-paced biocuration of these newly discovered evidences. Currently, the DisGeNET database [23, 24] makes the effort to unify biomedical literature evidence from gene–disease associations, by matching in first instance diseases, conditions, and phenotypes using a dictionary mapping from the Medical Subjects Headings (MeSH) hierarchy for disease classification and by the use of the Unified Medical Language System (UMLS) metathesaurus [23]. The DisGeNET database collects supportive evidence from several public resources such as the Online Mendelian Inheritance in

Man (OMIM, https://www.ncbi.nlm.nih.gov/omim), the Comparative Toxicogenomics Database (CTD, http://ctdbase.org), Pharmacogenomics Knowledge Base (PharmGKB, https://www.pharmgkb.org), the Literature-derived Human Gene–Disease Network (LHGDN, http://www.dbs.ifi.lmu.de/~bundschu/LHGDN.html), UniProt/SwissProt, and by gathering text-mining information from the literature in order to rank gene–disease associations based on initial queries.

The Cytoscape application with the same database denomination—DisGeNET [24] can be used to query and analyze a network representation of human gene–disease associations. The application uses the underlying data from the DisGeNET database for the integration of gene–disease associations from several biomedical database resources and the generated network can be either displayed as gene-centric or disease-centric. The exploration of the different data sources and types of associations available in this application as an initial exploratory analysis can be helpful in the elucidation of complex human diseases with genetic context. Thus, this application can be further used to assist in the prediction of unknown gene–disease associations, and in this way infer good candidate genes in advance of experimental analysis [22].

20.2.2.3 Resolving Molecular Interactions (Protein–Protein Interaction, Metabolite–Reaction–Protein–Gene)

Physical interaction networks can show how a set of molecules bind to each other, thereby potentially revealing a functional association such as cascades in metabolic and signaling or of cellular structure and ultimately underpin the complex interplay between healthy and disease phenotypes [25]. Thus, in this section we propose some databases and respective application tools for the analysis of PPIs as well as those associated with metabolic reactions.

The STRING database [26] available at http://string-db.org collects molecular information to cover both known and predicted PPIs, which are inferred by physical and functional interactions. The database source data for molecular interactions relies mainly on primary interaction databases (e.g., IntAct [27], BioGRID, [28]), automatic extraction of information by text mining, from co-expression and high-throughput experiments, and as well derived from computational prediction. The current database version 10.0 contains 26 217 572 interactions with a confidence score greater than 0.9 for 9 643 763 proteins from 2 031 organisms. Moreover, the Cytoscape version of STRING is more amenable for creating or expanding existing networks within the Cytoscape framework.

GeneMANIA [29] available as a Cytoscape application and as an online version (http://www.genemania.org) [30] can be used to assist in gap-identification and gap-fill approaches due to its capability to predict gene function as a quick routine. This application tool identifies the most related genes to a query gene input using a guilt-by-association approach. Thus, it predicts protein function by assessing PPIs between the query protein and proteins with well-established functions [31]. The app uses a large database of functional interaction networks, indexing 2152 association networks containing 537 599 442 interactions mapped to 166 084 genes from nine organisms (online version on November 2016). The GeneMania app tool performs multiple searches across several publicly available databases and dissects many large-scale datasets in order to find relationship between genes regarding, for example, PPI networks, gene co-expression and co-localization data, shared protein domains, gene–gene interactions, pathway data, and prediction of functional associations [29]. In case of analysis of large gene set lists, due to a policy of memory preallocation for each user in the online version, there is a limitation on the number of input genes; therefore, we recommend the Cytoscape-based application instead.

The MetScape 3 application [32] for Cytoscape offers the possibility to perform a combined network analysis of metabolomic and gene expression profiling datasets. Since, it runs under the Cytoscape environment, selection of sub-networks and their respective visualization and application of custom layouts is possible and as well as the entire network functionalities, such as merging networks from diverse sources. The list of metabolites should have the KEGG compound IDs and genes the NCBI Entrez Gene IDs in order to generate the metabolic networks. Optionally, for narrowing down the analysis in case of dealing with large-scale datasets, a concept file can be uploaded into the app enclosing information regarding gene set enrichment analysis (GSEA) of gene expression data. The MetScape app integrates data sourced originally from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [17] and from the Edinburgh human metabolic network (EHMN) [33].

20.2.2.4 Transcription Factor(TF)-Driven Modules and microRNA–Target Regulation

Transcription factor (TF) is a molecule that controls the activity of a gene by determining whether the gene's DNA is transcribed into RNA [34]. A compendium on nonredundant TF and TF binding sites can be found at JASPAR [35] (http://jaspar.genereg.net/). The number of human TF ranges from 1500 to 2600, depending on source and stringency [34, 36]. Direct analysis of regulated events by TFs is valuable and might shed light on hidden elements that conventional pathway analysis cannot reveal. However, many TF binding sites and thus the corresponding regulated genes are hypothetical since

these predictions are not validated by experimental data. Therefore, network analysis involving TF elements has to be assessed carefully. microRNAs (miRs) exert post-transcriptional regulation of gene expression by binding to the 3′-untranslated regions (3′-UTR) of specific messenger RNAs (mRNAs) [37] and lead to their degradation and translation repression. Thus, several miRNAs regulate gene expression and their role should be scrutinized in the context of disease onset and progression.

The CyTargetLinker application [38] for Cytoscape extends biological networks by adding interactions associated with regulatory elements such as TF–target, miRNA–target, or drug–target. The regulatory interaction information is sourced from several public databases such as ENCODE (including both distal and proximal TF data) and TFe for TF–target data; MicroCosm and TargetScan for predicted miRNA–target data; miRTarBase and miRecords for validated miRNA–target data; DrugBank for drug–target interaction data. Then, the multisource regulatory data is converted to a supported format for Cytoscape in order to create Regulatory Interaction Networks (RegINs). The software tool in order to perform the linkage to regulatory information requires that the network aimed to be extended, contains in the network attributes preferably molecular identifiers pointing to Ensembl [39], NCBI gene [40], UniProt [41], miRBase [42], and DrugBank [43] accession numbers.

Through the use of the CluePedia [21] application for Cytoscape, the user can perform miR analysis, matching target genes by selecting different database sources, and then setting the threshold accuracy for each of them. The user can as well add its own list of candidate genes and interrogate the application for gene/miRNA enrichments. It generates an miRNA–target network that can be reused afterward for inline integration with ClueGO in order to uncover associated GO and pathway terms [21].

On the other hand, an online software tool for miR analysis is the DIANA-miRPath v3.0 [44] that allows both single and as well as integrative miR analysis of experimental data. Association with GO and pathway terms and its visualization, for example, in KEGG is also possible based on its newly incorporated reverse search module feature [44].

20.2.2.5 Pathway Visualization and Mapping

Humans are more amenable to detect differences across data comparisons if these are represented in a concrete form such as shapes and colors. Thereby, the translation of abstract data into diagrams that allows visual representation of the detailed steps of BP and concepts is a powerful tool to better understand and disseminate knowledge [45, 46]. We only discuss in more detail the tools for pathway biological data visualization that are widely used. A review and comparison of the main tools used for visualization and analysis of high-throughput omics molecular data acquisition by exploring networks and pathways was performed recently [47].

20.2.2.5.1 PathVisio Mapping

PathVisio [45, 46] is a standalone Java-based software application that allows creation, edition, and visualization of pathways, and ultimately leads to the inference of relevant pathways based on the results of a permutation test using an archive of pre-existent pathways. The application uses the pathway maps from WikiPathways [18] and Reactome [19] collections and is able to handle several gene, protein, and metabolite database identifiers as inputs, which are then cross-mapped through the BridgeDb [48] framework.

20.2.2.5.2 KEGG Mapper

The KEGG [17] is an integrated database resource of biological systems that integrates genomic, chemical, and systemic functional information. The database has several web-based tools for the analysis of the genome and metagenome sequences such as BlastKOALA and GhostKOALA, which perform automatic assignments of functions at the molecular level to genes using KEGG orthology (KO) [49]. In addition, analysis of data resulting from other high-throughput omics technologies can be performed, if the user holds a list of molecules of interest. Firstly, the list with external molecular identifiers needs to be converted to internal KEGG identifiers using the convert ID tool available in KEEG that is able to handle UniProt and NCBI-geneIDs identifiers. The data is then mapped into a collection of curated pathways, covering metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human disease, and drug development using one of the subset tools from KEGG mapper. The result is a list with an associated number of pathway hits based on the user input, subsequent to the selection of the pathway(s) of interest the user can color code (KEGG Search&Color Pathway utility) the molecules within for the purpose of representing deregulated molecules in a pathway across a differential expression dataset.

20.2.2.6 Data Harmonization: Merging and Mapping

The majority of analysis tools require database-specific accession numbers. This means that unique IDs have to be converted to other accession numbers compatible with specific databases. The UniProt database contains a feature to map IDs to other databases (Retrieve/ID mapping, http://www.uniprot.org/uploadlists) and BioMart (http://www.biomart.org) can also be used for ID conversion.

Multiomics datasets might not only contain protein and gene data, but also expression profiles of chemical compounds. While it is easy and straightforward to

combine protein/DNA/RNA expression data using common IDs, this is not the case for chemical compounds such as metabolites. However, this kind of data can be integrated by using "guilt by association," or in other words, exploiting that metabolites are usually generated by enzymes, and a change in metabolites can reflect a change in the protein/gene pattern. Nevertheless, this is an uncommon technique since it involves currently a manual search to identify the potential proteins, and has some inherent pitfalls such as uncertainty for which enzyme/isoform is responsible for the metabolic change. In addition, the same compound could also be used/generated by several proteins. Therefore, metabolic datasets are often treated as separate entities in multiomics studies, analyzed independently, and integrated only at the level of final outcomes/outputs. In order to integrate metabolomic data by mapping to proteins/genes, it is possible to apply the assumption that a downregulation is associated with a downregulation of upstream events, whereas an upregulation can be due to a downregulation of a downstream event.

20.2.3 Computational Drug Discovery

Computational drug discovery methods are based on bioinformatics and computational biology, which deal with the application of computer technology in the management of biological information and the modeling of BP. These approaches are essential for the elucidation of the molecular basis of human diseases and the identification of novel targets for drug discovery. The integration of data generated by various methods is performed by bioinformatics approaches and has created the discipline of systems biology. Systems biology might revolutionize the practice of medicine with respect to preventative and personalized health care by analyzing diseases using a "graphical network model." The models developed using systems biology aim to mathematically or quantitatively describe the differences between disease-perturbed protein and gene regulatory networks from healthy networks by using multiple temporal and spatial parameters of pathway markers. By studying the relationships and interactions between various parts of a biological system (e.g., metabolic pathways, organelles, cells, physiological systems, organisms) by using gene-knockout animal models and environmental variables in experiments, an overall model of the whole system can be developed and analyzed. Moreover, a goal of a systems biology modeling approach is to define the dynamic behavior of a biological system for the purpose of predicting new components, network links, and their behaviors within the system following a perturbation event, and comparing the changes in mRNA and protein expressions with the standard "healthy" model for disease molecular marker identification. The disease molecular markers (such as proteins produced as a result of cancer-specific genes) that are identified using this comprehensive systems approach can then become templates for disease diagnosis and disease status determination and vaccine or drug design [50].

20.2.3.1 High-Throughput Virtual Screening (HTVS)

At the heart of drug discovery is High-Throughput Virtual Screening (HTVS), a rapid computational method of evaluating millions of compounds against a specific molecular target, for the purpose of ranking the top compounds that may be good potential inhibitors or protein–protein binding disruptors. HTVS is a rapid protocol for molecular docking; molecular docking is a computational technique used for predicting whether one molecule will bind to another. The structure of the target molecule is needed to perform a docking, usually resolved by the use of X-ray crystallography, or not so regularly, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy. If the three-dimensional structure of the receptor is unavailable, computational methods, such as homology modeling, help predict the structure, using existing structural templates. A similarity search for the primary sequence information of the receptor finds the best templates with experimental structural information. The structural template is used to provide the backbone conformation for the receptor, and the residues mapped to this scaffold. The docking software tool requires the protein structure (e.g., PDB: Protein Data Bank archive, http://www.wwpdb.org) and a library collection with putative ligands (e.g., ZINC database, http://zinc.docking.org). The interdependencies between the search algorithm and the scoring function impacts directly in the rate of success of the software.

Most docking software solutions such as the AutoDock4 [51] uses a scoring function or a semi-empirical force field to search through a conformational landscape of a ligand in order to calculate the free energy of binding to a target site of a macromolecule. The scoring function approach based on the estimation of the molecular mechanics force fields has stable configurations as input and yields a coefficient score representative of the probability that the binding interaction is favorable, which is supported by a low configuration energy. The other approach is based on protein–ligand interaction data from large databases, such as the Protein Data Bank (PDB) archive [52], which evaluates the fit of the configurations according to statistical potential.

20.2.3.2 Advantages and Limitations of HTVS

Experimental and computational methods in drug discovery are highly complementary; experimental HTS of millions of compounds is expensive and time-consuming

and both the in vitro and the in vivo screening assays can generate false positives and false negatives because in HTS, the fast speed at which the compounds are tested causes misclassification of the compound due to faulty assays (false negatives) as inhibition is falsely observed or missed. False positives are generated due to the compounds' interference with the assay's properties, or inhibition of poor selectivity. These compounds tend to be over represented in the HTS hit list. Various experimental assays which rapidly but indirectly probe compound binding interactions are excellent triage assays in the search for ideal drug molecules; however, this process continues with ligand optimization for high potency and selectivity, and efficient synthesis. Therefore, there is a great advantage in elucidating the mechanism of binding and the mode of binding in the receptor binding site. A possible experimental method to verify the inhibition of the target is to obtain an X-ray structure of the complex, with the ligand bound to the target. An NMR structure is another possibility; however, it can be applied usually to proteins with MW < 30 kDa. The Fluorescence Resonance Energy Transfer (FRET) method might also be used to monitor protein–ligand interactions by fluorescence change. These experimental methods are time-consuming and therefore limited to the best candidates. Therefore, in order to minimize the size of the chemical libraries involved in experimental screening and validation, computational methods are essential to rank top chemicals for screening, serving as a rapid and cost-effective triage protocol.

On the other hand, in order to rapidly and accurately evaluate millions of molecules, HTVS protocols need to balance performance and speed. This is a function of the algorithm which runs the scoring function, as well as the power of the supercomputer that is running the calculation. The scoring function of molecular docking programs takes a stable configuration as input and returns a number indicating the likelihood that the specific configuration represents a favorable binding interaction. In order to increase speed, frequently scoring functions sacrifice flexibility of the receptor such that enough binding modes are evaluated for the best complex conformations. Most scoring functions are physics-based molecular mechanics force fields that estimate the energy of the configuration; however, they are semi-empirical which means that they still have parameters derived from training sets and these parameters of the scoring function should be carefully adjusted, and thresholds evaluated to optimize the correlation values. Also, these parameters have a substantial influence on accurate rankings of highly diverse compound libraries; the variability within the drug-candidate compounds, or their drug-likeness profiles might differ so much that there is no visible correlation. In summary, there is no standard and ideal HTVS protocol to handle all drug discovery approaches; therefore, HTVS protocols need to be adapted and the software customized to the specific drug discovery project.

20.3 Chronic Kidney Disease (CKD) Case Study

The etiology of CKD is not always known in the first instance of the patient diagnosis. However, any condition, disorder, disease, or chronically administered drug that promotes injuries to blood vessels or other primary kidney structures can potentially act as a risk factor for the development of kidney disease. Thereby the most prominent causes of CKD are diabetes and hypertension [53]. Nevertheless, other renal co-morbidities and associated risk factors such as acute kidney failure (AKI) and glomerulonephritis (GN) can lead to kidney damage and consequently to CKD [54].

In order to collect evidence on the molecular traits related with CKD, we retrieved gene associations of four selected renal disorders using the DisGeNET [23] database that integrates gene–disease information resulting from text-mining the literature and several database resources. Our search (on November 2016) within the DisGeNET database resulted in 186 molecule associations with Chronic Kidney Insufficiency (CRI), 606 molecule associations with Diabetic Nephropathy (DN), 176 molecule associations with Acute Kidney Failure (AKI), and 186 molecule associations with Glomerulonephritis (GN) (Figure 20.2 and

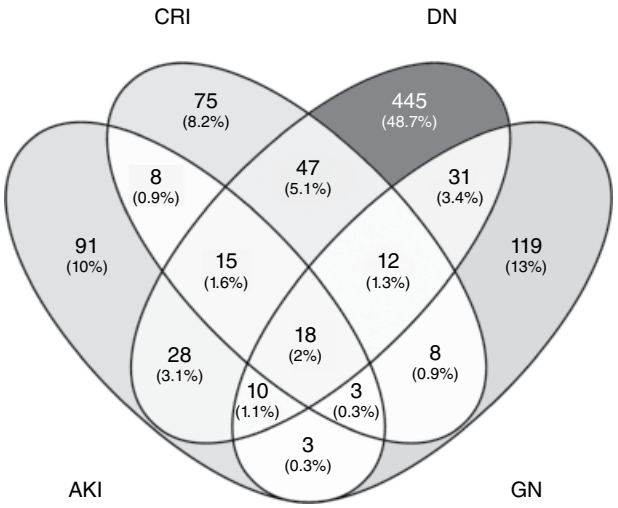


Figure 20.2 Molecular overlap of genes/proteins and microRNAs in CRI: Chronic Renal Insufficiency, DN: Diabetic Nephropathy, AKI: Acute Kidney Failure, and GN: Glomerulonephritis. Venn diagram generated in http://bioinfogp.cnb.csic.es/tools/venny/.

Table 20.4 Overlap of the associated genes in four related kidney diseases (CRI, Chronic Kidney Insufficiency; DN, Diabetic Nephropathy; AKI, Acute Kidney Failure; GN, Glomerulonephritis) using data sourced from DisGeNET.

Diseases	Total	Molecular elements
"AKI," "CRI," "DN," and "GN"	18	ACE, IGF1, SOS1, APOE, AGT, PPARG, CXCL8, TLR4, TNF, IL1B, REN, AGTR1, ALB, VEGFA, CCL2, HGF, IL6, PARP1
AKI, "CRI," and "DN"	15	HMOX1, NLRP3, FABP1, KL, HP, VCAN, TP53, HNF1B, CAT, BMP7, KLK1, BMP2, LCN2, EPO, FGF23
AKI, "CRI," and "GN"	3	CRP, ABCB1, IL17A
AKI, "DN," and "GN"	10	MMP2, EDN1, PPARA, IL10, BCL2, CFH, HLA-DRB1, MPO, IL18, DCN
CRI, "DN," and "GN"	12	PON1, TNFRSF11B, MYH9, NOS3, SERPINE1, VCAM1, CCL5, IL1A, ESR1, TGFB1, AGTR2, MMP9
AKI and "CRI"	8	SLC33A1, ATF3, CXCL12, CYP3A5, CSRP1, ATM, ADRB2, PPIAP10
AKI and "DN"	28	DECR1, CXCR4, CYBA, PTEN, MBL2, SOD1, HYOU1, MUC1, PIK3CB, SLC22A6, EDNRA, IGF1R, HSPA4, HIF1A, SELE, ADAMTS13, SLC22A8, MIR210, SLPI, ADAM10, TIMP2, NOX4, GPX4, TLR2, CDKN2A, MIR21, PNPLA2, HMGB1
AKI and "GN"	3	ELANE, IFNG, CFB
CRI and "DN"	47	NR3C2, XDH, KDR, FLT1, SOD2, LEP, GJA1, GSTM1, MTHFR, GFPT2, NPHS2, CDKN1B, MOK, CLU, AKR1B1, TRPC6, IL6R, CST3, NANOS3, VASH1, PTH, GLP1R, NFE2L2, MTOR, ATP6AP2, HFE, AOC3, RAPGEF5, INSR, DIANPH, UMOD, TCF7L2, AHSG, GH1, SIRT1, IRS2, LMNA, APOL1, KEAP1, HAVCR1, CYP24A1, GSTT1, GABPA, CTGF, UCP2, S100A9, DRD1
CRI and "GN"	8	MEFV, CFHR5, NGF, CD59, COL4A3, NOV, P2RX7, ANXA5
DN and "GN"	31	INS, MMP3, AXL, MAPK1, FOXP3, SELL, FAS, SMAD1, ICAM1, LTBP1, NPHS1, FN1, IL1RN, PLG, WT1, SELP, STAT3, ADM, CD2AP, SPP1, MIF, NFKB1, ITGB3, BECN1, ANKRD1, GAS6, JUN, TIMP1, IGAN1, STAT1, EPHX2

Source: Adapted from Bauer-Mehren et al. [23] and Pinero et al. [24].

Table 20.4). Based on data queries of the four clinical conditions, we found 18 overlapping genes: IGF1, HGF, CCL2, VEGFA, SOS1, IL1B, APOE, PPARG, AGT, IL6, TLR4, PARP1, REN, AGTR1, ACE, ALB, CXCL8, and TNF (Figure 20.2 and Table 20.4). Using these genes as input we can uncover relevant BP and pathways by displaying networks associated with physical molecular interactions, gene co-expression, and consolidated molecular pathways with the web-tool version of GeneMania (Figure 20.3). This resulted in a diverse array of pathways that appear to be linked with the four disorders, briefly the NOD-like receptor signaling (CCL2, IL6, IL8, IL1B, and TNF), the toll-like receptor signaling (TLR4, IL6, IL8, IL1B, and TNF), the renin–angiotensin system (REN, AGTR1, ACE, and AGT) and the cytokine–cytokine receptor interaction (VEGFA, HGF, CCL2, IL6, IL8, IL1B, and TNF). The network weighting was based on the GO BP that seems to point to the association with the inflammatory response (CCL2, IL6, IL8, IL1B, TNF, TLR4, APOE, AGT, and AGTR1), one of the three common pillars of disease pathogenesis alongside with oxidation and coagulation processes. Although such characterizations are simplistic, it is remarkable how many conditions appear to involve disturbances of these three disease pillars involved at some point in the continuum that characterizes their pathogenesis (Figure 20.3). Taken together, this confirms and highlights the recognized involvement of the innate immune system and as well the role of the renin–angiotensin system associated to progressive renal damage. Moreover, cumulative experimental evidence seems to support the involvement of both canonical and noncanonical NOD-like receptor and Toll-like receptor signaling pathways in the innate response pattern recognition by receptor activation and thereby it is associated with kidney inflammatory response in AKI and as well with CKD progression [55].

In order to investigate the expression of these molecules in a disease state we used multiomics datasets sourced from the iMode-CKD (URL: http://www.imodeckd.org/) consortium regarding the FP7 project "Clinical and system -omics for the identification of the Molecular Determinants of established Chronic Kidney Disease" and if it is available we provide the link to the original publication. The datasets and the respective

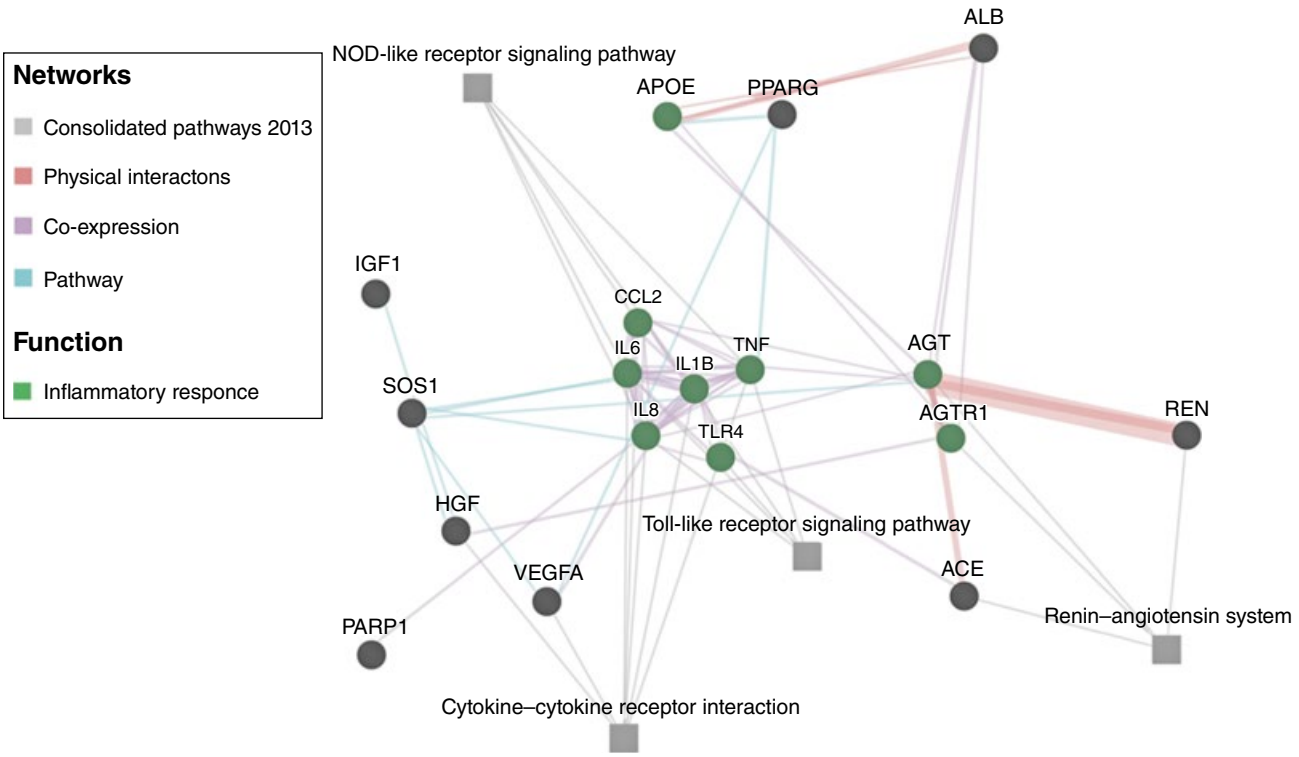


Figure 20.3 Molecular network associations of the 18 overlapping molecules from the four selected kidney disorders. The search parameters in the online based tool GeneMania were: organism Homo sapiens; Genes input: IGF1, HGF, CCL2, VEGFA, SOS1, IL1B, APOE, PPARG, AGT, IL6, TLR4, PARP1, REN, AGTR1, ACE, ALB, IL8, and TNF; Network weighting was based on the Biological process. GeneMania online application version 3.5.0.

acquisition omics platforms, cohort sizes, and group modifiers are summarized in Table 20.5. The molecular data used for a further integrative systems biology analysis from each omic platform is described in Table 20.6.

20.3.1 Dataspace Description: Demographics and Omics Platforms Information

Table 20.5 Patients with CKD stages II–IV were analyzed.

Clinical information				
Cohort	**No. stable**	**No. progressors**	**Fluid source**	**Platform**
GHENT	32	32	Urine	PRO
	204	23		MET
	N/A	N/A		MIR

Patients were frequency matched for age, sex, baseline eGFR and CKD stage. Stable: eGFR %slope/year > −1.5% and <1.5%; progressors: eGFR %slope/year >5% and <30%.

20.3.2 Dataspace Description: No. of Associated Molecules Per Omics Platform

Since the iMode-CKD project in the Ghent study evaluates CKD progression by comparing stable (eGFR %slope/year > −1.5% and <1.5%) patients with progressing

Table 20.6 Instance counting of the associated number of features per high-throughput acquisition omic technologies (PRO, proteomics; MET, metabolomics; MIR, microRNA transcriptomics) after applying each of the statistical criteria thresholds.

					Reg. trend	
	Initial	**Final**	**Adj.Pvalue**	**FC**	**Up**	**Down**
PRO	344	142	<0.05	1.5	34	108
MET	47	47	N/A	N/A	N/A	N/A
MIR	62	20	<0.05	2	17	3

N/A, not applicable; Reg. trend, regulation trend representing molecular expression directionality; Adj.Pvalue, adjusted *p*-value after correction for multiple testing.

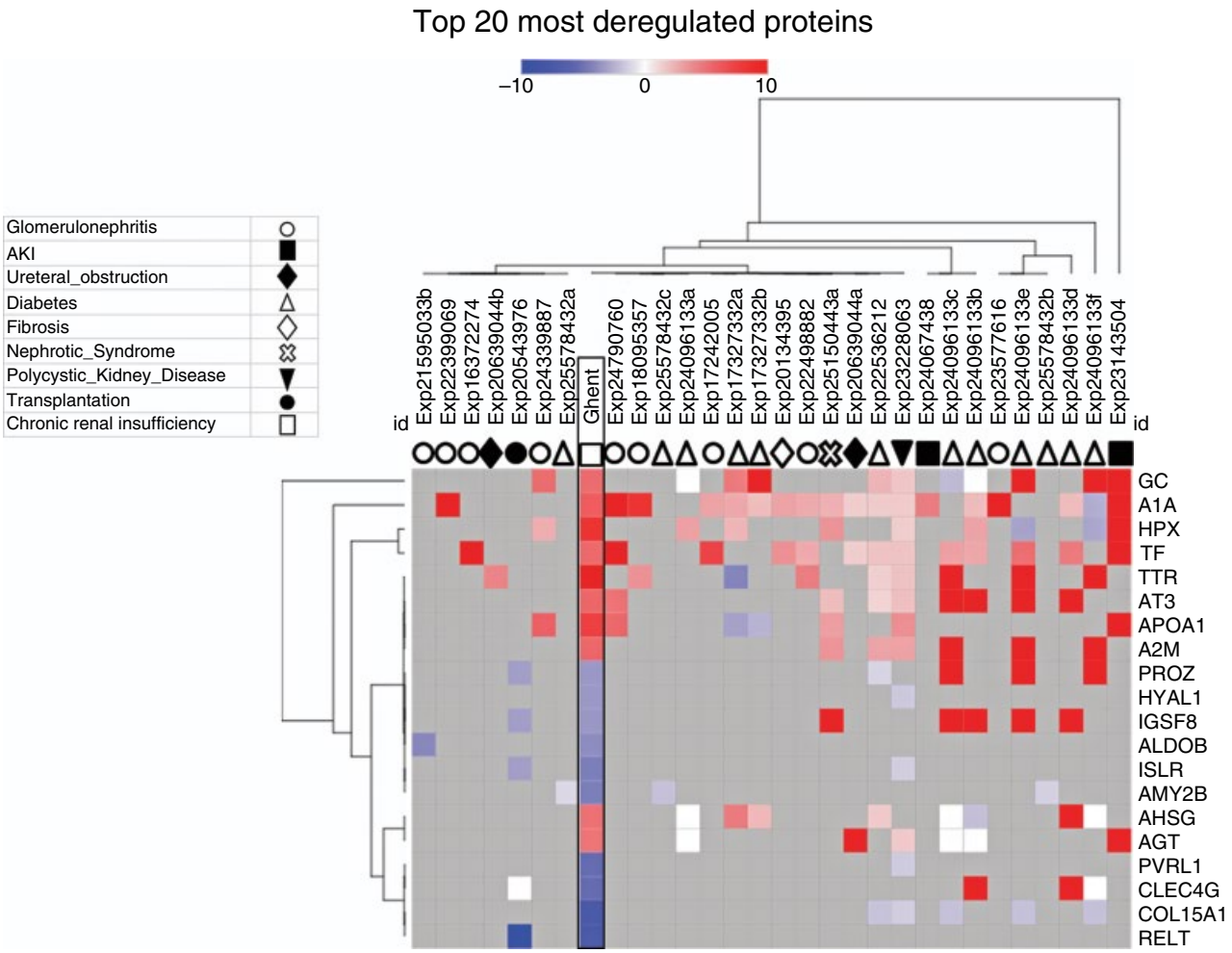


Figure 20.4 The 20 most deregulated proteins from the Ghent study are compared across several published studies related with chronic kidney disease. The Heatmap displays the hierarchical clustering by average linkage of the Euclidean distances. Morpheus web-tool available at https://software.broadinstitute.org/morpheus. The datasets comparisons were pulled from CKDdb (http://www.padb.org/ckddb/). The prefix "Exp" is related with the PubMed ID of the experimental study.

(eGFR %slope/year >5% and <30%) patients, we compared the expression in disease of the 20 most deregulated molecules from the proteomics platform with datasets from several related kidney diseases sourced from the literature available in the CKDdb (Figure 20.4). The expression profile of the urinary proteins derived of the Ghent study seems to be associated with a vast range of kidney disorders. A group of proteins seems to be consistently upregulated in progressors: vitamin D-binding protein (GC), alpha-1-antitrypsin (A1A, SERPINA1), hemopexin (HPX), serotransferrin (TF), transthyretin (TTR), antithrombin-III (AT3, SERPINC1), apolipoprotein A-I (APOA1), and alpha-2-macroglobulin (A2M) and they are associated with the complement and coagulation cascades (Figure 20.4).

20.3.3 Data Reduction by Principal Component Analysis (PCA)

We carried out dimensionality reduction through Principal Component Analysis (PCA) (Figures 20.5 and 20.6) using Multibase, an Excel add-in program to enable filtering (e.g., outlier's removal), rearrangement or removal of samples based on patterns of molecular expression across all the omics datasets. After plotting samples (Figure 20.5) and the variables (loadings) (Figure 20.6), we removed inconsistent datasets comparisons. We can observe that the dataset derived from the Ghent study from the iMode-CKD project and projected on PC1 presents a high contrast when compared with diabetes (including diabetic nephropathy datasets), literature-based datasets (pulled from the CKDdb database, http://padb.org/ckddb) that is projected on PC2 (Figure 20.5). This difference is mainly due to contrasting regulation in the signal-regulatory protein beta-1 (SIRPB1), Ig kappa chain C region (IGK, IGKC), prostatic acid phosphatase (ACPP), carboxypeptidase N subunit 2 (CPN2), antithrombin-III (AT3, SERPINC1), cadherin-2 (CDH2), immunoglobulin superfamily member 8 (IGSF8), CD9 antigen (CD9), nectin-1 (PVRL1, NECTIN1), haptoglobin (HP), and leucine-rich alpha-2-glycoprotein (LRG1) (Figure 20.6).

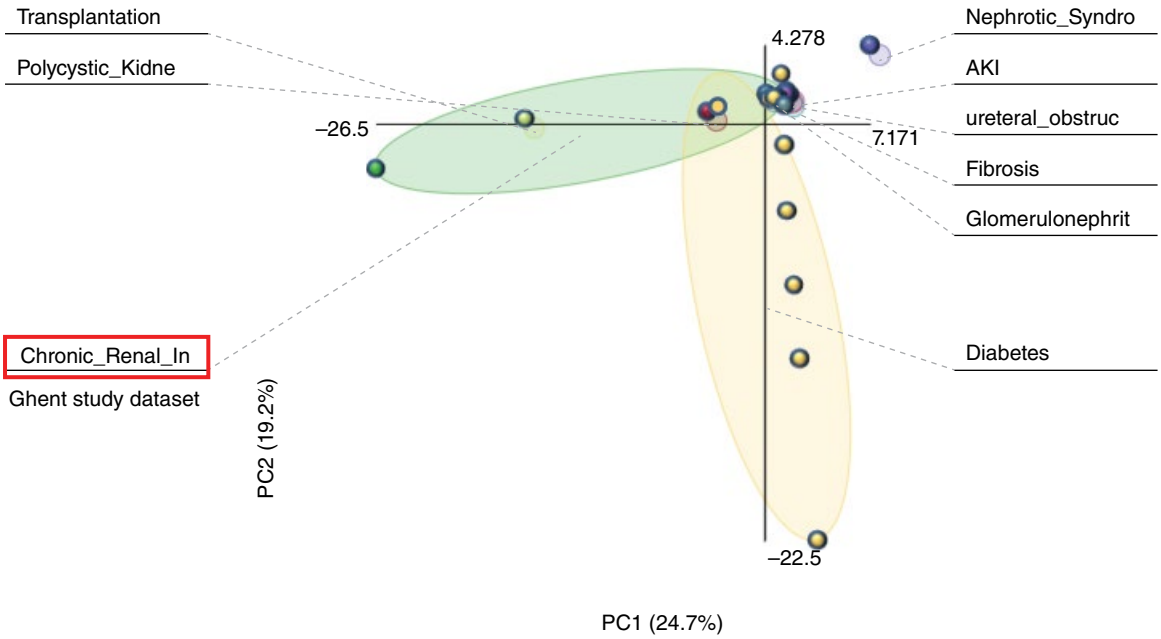
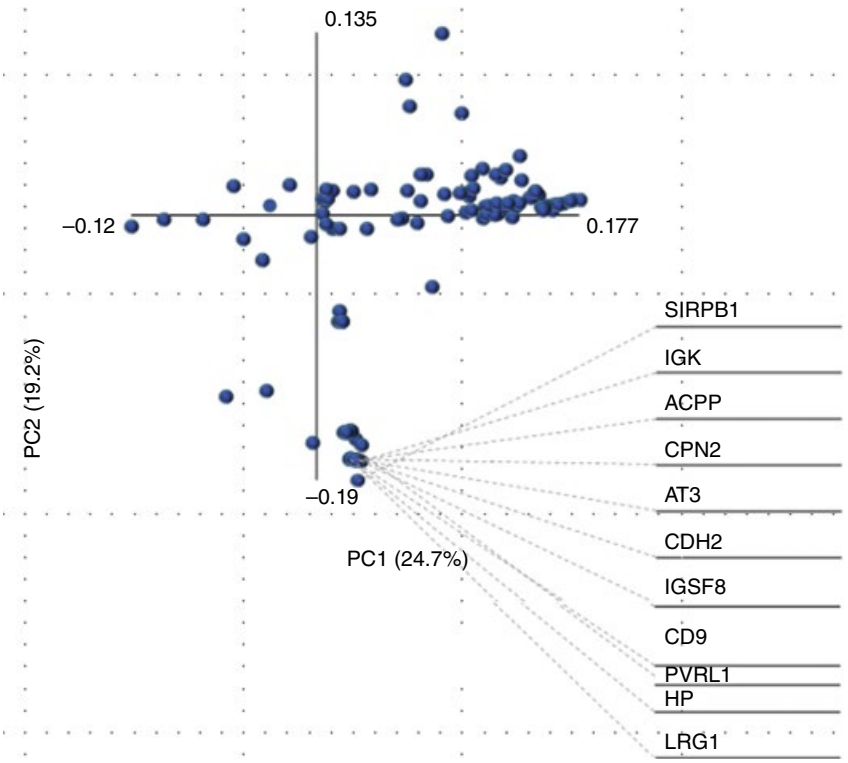


Figure 20.5 Variables (disorders) view of the Principal Component analysis (PCA) analysis. Both PC1 and PC2 explain 33.9% of the cumulative data variation. Two main clusters are formed the Chronic Renal Insufficiency (CRI) highlighted in the box on the PC1 and the Diabetes cluster on PC2.

Figure 20.6 Loadings (molecules) view of the Principal Component analysis (PCA) analysis with the top 11 molecules more accountable for discriminating within samples.



20.3.4 Gene Ontology (GO) and Pathway-Term Clustering

After applying data thresholding (p-value < 0.05 and Fold-change ≥ 1.5) across the entire dataspace, the Cytoscape application ClueGO v.2.2.5 was used to identify the associated BP and MF (Table 20.7), taking into account overexpressed and underexpressed molecules from the CKD comparisons between progressors and stable patients from the Ghent iMode-CKD study (Figure 20.7).

Table 20.7 Gene ontology (GO, BP, biological process; MF, molecular function, performed in the Cytoscape ClueGO [16] app.

GOTerm	Term PValue corrected with Bonferroni step-down	Group PValue corrected with Bonferroni step-down	% associated genes	Cluster	Genes cluster 1	Genes cluster 2	% genes cluster 1	% genes cluster 2	No. of genes
BP									
Platelet degranulation	17.0E−15	35.0E−18	13.91	Specific for cluster 1	[A1BG, A2M, AHSG, ALB, APOA1, ORM1, ORM2, SERPINA1, SERPINA3, TF]	[CD9, CLEC3B, EGF, FGA, ISLR, SERPINF2]	62.50	37.50	16
Acute inflammatory response	610.0E−12	35.0E−18	8.86	Specific for cluster 1	[A2M, AHSG, HP, ORM1, ORM2, SERPINA1, SERPINA3, SERPINC1, VTN]	[DEFB1, EPHB6, IL6ST, PRCP, SERPINF2]	64.29	35.71	14
Negative regulation of coagulation	6.5E−6	7.0E−6	12.28	Specific for cluster 2	[VTN]	[APOE, CEL, FGA, PROCR, SERPINF2, THBD]	14.29	85.71	7
Acute-phase response	7.8E−6	35.0E−18	11.86	Specific for cluster 1	[AHSG, HP, ORM1, ORM2, SERPINA1, SERPINA3]	[SERPINF2]	85.71	14.29	7
Glycosaminoglycan catabolic process	13.0E−6	110.0E−6	10.94	Specific for cluster 2		[ACAN, CSPG4, GLB1, HEXA, HYAL1, PGLYRP1, SDC4]	0.00	100.00	7
Collagen metabolic process	49.0E−6	33.0E−6	7.08	Specific for cluster 2		[CEL, COL15A1, COL18A1, COL6A1, PDGFRB, RETN, SERPINF2, TINAGL1]	0.00	100.00	8
Negative regulation of cell activation	84.0E−6	25.0E−6	5.49	Specific for cluster 2		[APOE, AXL, CD300A, CD84, CLEC4G, HAVCR2, PGLYRP1, THBD, TNFRSF14]	0.00	100.00	9
MF									
GOTerm	**Term PValue corrected with Bonferroni step down**	**Group PValue corrected with Bonferroni step down**	**% associated genes**	**Cluster**	**Genes cluster 1**	**Genes cluster 2**	**%Genes cluster 1**	**%Genes cluster 2**	**Nr. Genes**
Serine-type endopeptidase inhibitor activity	1.2E−9	1.0E−9	11.11	None specific cluster	[A2M, AGT, SERPINA1, SERPINA3, SERPINA7, SERPINC1]	[PEBP1, SERPINA5, SERPINF2, SPINT1, SPINT2]	54.55	45.45	11
Hydrolase activity, acting on glycosyl bonds	53.0E−6	53.0E−6	5.56	Specific for cluster 2		[AMY2A, AMY2B, GLB1, HEXA, HYAL1, MAN1A1, NEU1]	0.00	100.00	7
Hydrolase activity, hydrolyzing O-glycosyl compounds	31.0E−6	53.0E−6	7.14	Specific for cluster 2		[AMY2A, AMY2B, GLB1, HEXA, HYAL1, MAN1A1, NEU1]	0.00	100.00	7
Growth factor binding	38.0E−6	28.0E−6	5.93	Specific for cluster 2	[A2M, NGFR]	[ACVR1B, COL6A1, IGFBP7, IL6ST, PDGFRB, VASN]	25.00	75.00	8

Cluster 1 contains upregulated (FC ≥ 1.5 and p value < 0.05) proteins and cluster 2 downregulated (FC ≥ 1.5 and p value < 0.05) proteins.

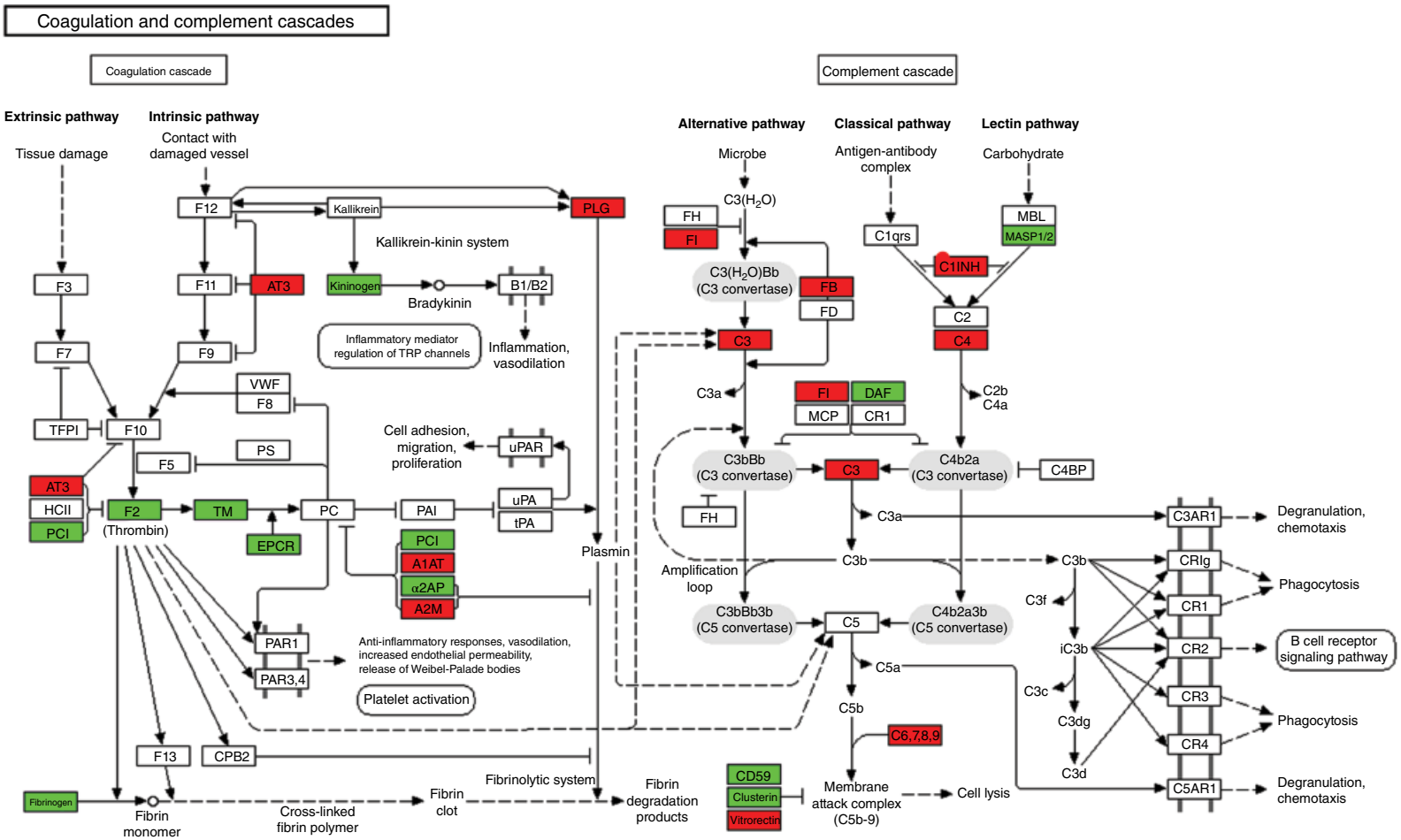


Figure 20.7 Using the KEGG [17] mapper tool for mapping the molecules into the human pathway for Complement and coagulation cascades—available at http://www.genome.jp/kegg-bin/show_pathway?hsa04610. The binary color rearrangement denotes molecular regulation from the proteomic study, which compares progressors with stable patients. Red objects: upregulation (fold-change, FC ≥ 1.5); Green objects: downregulation (fold-change, FC ≤ −1.5).

In general, the GO and pathway-term analysis resulted in more significant (p-value corrected Bonferroni step-down) functional groups with processes and pathways inactivated due to a downregulation of the associated molecules (Table 20.7).

The most significant GO term associated with the BP was the acute inflammatory response, and it was activated with all the genes belonging to cluster 1 (Table 20.7); A2M, AHSG, HP, ORM1, ORM2, SERPINA1, SERPINA3, SERPINC1, and VTN increased in expression. The process of granule secretion by the platelet (platelet degranulation) involved A1BG, A2M, AHSG, ALB, APOA1, ORM1, ORM2, SERPINA1, and SERPINA3; TF was also found triggered by the overexpression of the associated genes. Concomitantly, the acute-phase response process was found activated with the genes belonging to cluster 1 (AHSG, HP, ORM1, ORM2, SERPINA1, and SERPINA3) being increased in expression. In contrast, the glycosaminoglycan catabolic process was inactivated with the genes belonging to cluster 2 (ACAN, CSPG4, GLB1, HEXA, HYAL1, PGLYRP1, and SDC4) being decreased in expression. Moreover, by the analysis of the GO MF, it was determined that most of the downregulated proteins involved in glycosaminoglycan catabolism are enzymes (glycosidases) (Table 20.7).

20.3.5 Interactome Analysis: PPIs and Regulatory Interactions

20.3.5.1 Protein–Protein Interactions (PPIs)

Based on the assumption that proteins do not exert their functions in isolation, and that they are able to form functional clusters by, for example PPIs, it is possible to predict relevant BP/signaling pathways and deregulated functional modules in disease states.

Therefore, we performed PPI's network analysis using the Cytoscape GeneMania app (Figure 20.8). This resulted in 51 network nodes distributed in 8 functional network clusters (Figure 20.8), in which the cluster 1 has albumin (ALB) driving the interconnectivity of the nodes within the network and can be referred as a network hub.

Isolated nodes and small clusters (with less than two molecules bonding) were removed in order to highlight only relevant network features. Globally, the cluster 1 contains 32 nodes with a similar number of up- and downregulated molecules. The clusters 2 and 3 contain globally 5 and 4 nodes, respectively, that present in overall a downregulation trend. Clusters 4–8 form only clusters of two proteins and in its majority show decreased level of expression (Figure 20.8). The most upregulated

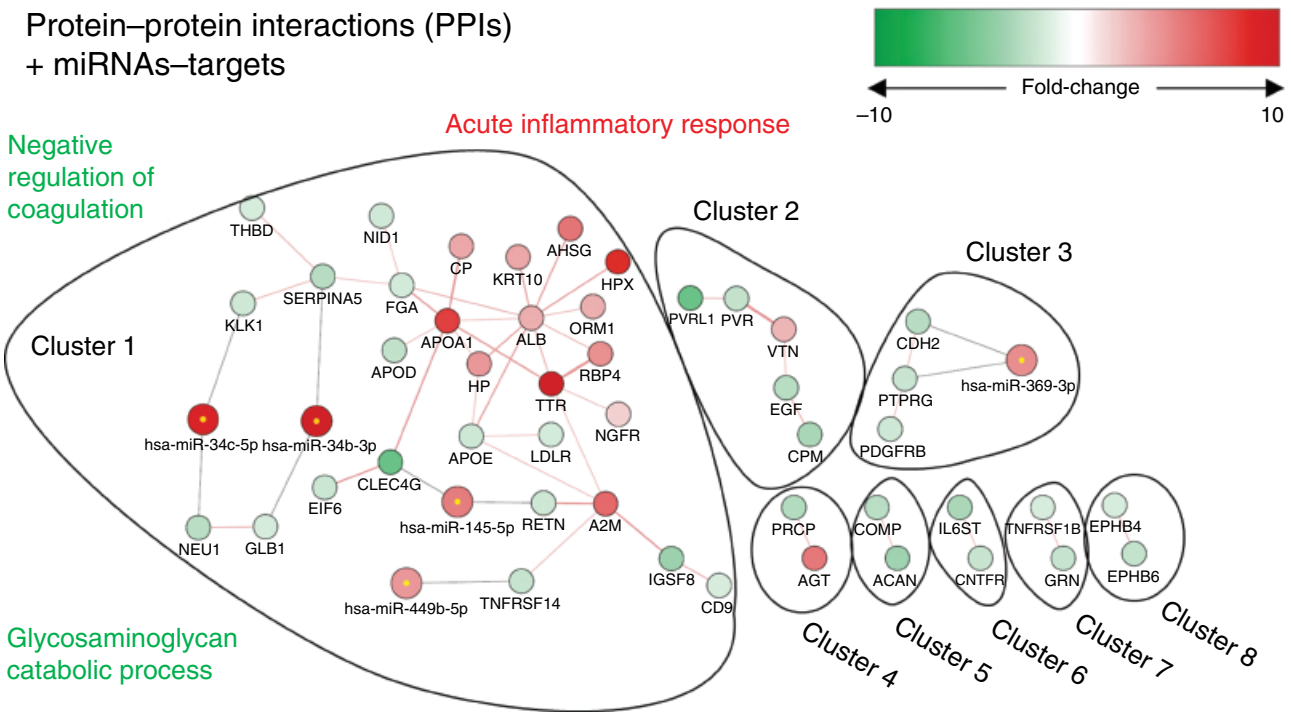


Figure 20.8 Protein–protein interactions (PPIs), 51 nodes, 8 clusters, no enrichment, plus miRNAs as regulatory elements. GeneMania app [29] for Cytoscape. The pink color stroke of the network edges represents PPIs and the black stroke color is linking regulatory elements, such as miRs. The molecular fold-change (FC) interval established was [–10, 10].

molecules within the displayed network (Figure 20.8) are proteins hemopexin (HPX), transthyretin (TTR), apolipoprotein A-I (APOA1), and the upstream regulatory elements miR-34c-5p and the miR-34b-3p.

20.3.5.2 Regulatory Interactions

Network analysis with regulatory interactions was performed using the CyTargetLinker [38] app for Cytoscape in order to uncover miRNA–target, TF–target, or drug–target interactions and then the generated network was merged with the former handling PPIs (Figure 20.8). For establishing the upstream interactions, regulatory data derived from the Regulatory Interaction Networks (RegINs) was used. Based on the essential role of miRNAs as posttranscriptional regulators of gene expression, we can presume that molecules affected by the same upstream factors are expected to display a similar expression regulation pattern. Thus, similarly modulated genes might be affected by the same upstream events.

Here, it was possible to pinpoint potential upstream regulators and their targets for the comparison of progressors with stable patients (Figure 20.8). Focusing on the cluster 1, the expression of kallikrein-1 (KLK1), plasma serine protease inhibitor (SERPINA5), sialidase-1 (NEU1), and beta-galactosidase (GLB1) are regulated by miR-34c-5p and miR-34b-3p, respectively. The C-type lectin domain family 4 member G (CLEC4G) and resistin (RETN) expression are upstream regulated by miR-145-5p; the tumor necrosis factor receptor superfamily member 14 (TNFRSF14) expression is being regulated by miR-449b-5p. Regarding the cluster 3, expression of cadherin-2 (CDH2) and the receptor-type tyrosine-protein phosphatase gamma (PTPRG) are both regulated by the miR-369-3p (Figure 20.8).

20.3.6 Interactome Analysis: Metabolic Reactions

The following data were used as input in the MetScape [32] Cytoscape application: the NCBI gene IDs, the metabolites KEGG IDs, and a file containing information regarding the gene set enrichment analysis (GSEA) from gene expression data. After selection of the suitable statistical and fold-change (FC) thresholds, here metabolite FC ≥ 1.1 and gene/protein FC ≥ 1.5 and for both an overall p-value < 0.05 we can visualize metabolic network associating, for example, a complete view over gene–enzyme–reaction–metabolite (Figure 20.9). In this example, with data sourced from the iMode-CKD project (Ghent study) we can observe the representation of decreased urinary levels of the putative gamma-glutamyltranspeptidase 3 (GGT3, GGT3P) (Figure 20.9a), platelet-derived growth factor receptor beta (PDGFRB), ephrin type-B receptor 6 (EPHB6), receptor-type tyrosine-protein phosphatase gamma (PTPRG) (Figure 20.9b), prostatic acid phosphatase (ACPP), lysosomal acid phosphatase (ACP2) (Figure 20.9c), activin receptor type-1B (ACVR1B) (Figure 20.9d) and increased levels of ceruloplasmin (CP) (Figure 20.9e). In the subnetwork metabolic cluster (Figure 20.9a), we can observe decreased levels of the enzyme gamma-glutamyltransferase (EC 2.3.2.2) encoded by CGT3 and the earlier is known by its transferase activity, being able to transfer gamma-glutamyl functional groups from donor molecules to acceptors. Here, gamma-glutamyltransferase is involved in the reactions R01687 and RE1473 (KEGG reactions database, http://www.genome.jp/kegg/reaction) having taurine and L-alanine (showing decreased levels in patients of the CKD progressors group, when compared to the stable group) as acceptors of gamma-glutamyl functional groups, forming 5-L-glutamyl-taurine and gamma-L-glutamyl-L-alanine, respectively (Figure 20.9a). Published studies report an association of decreased concentration for urinary L-alanine and taurine compounds in patients with primary focal segmental glomerulosclerosis (FSGS) [56, 57] and a decrease of urinary taurine in patients with advanced-stage of CKD [58].

20.4 Final Remarks

Integration of multidimensional data derived from omics profiles at a systems level with the currently available bioinformatics tools and database resources is not straightforward. It is challenging to correlate variations of the metabolite concentration with deregulated protein levels inferred from transcriptomics or proteomics data. It is also necessary to consider the epigenetic regulation of gene expression, elements associated with transcription regulation, posttranslational modifications (PTMs), and the interplay of protein degradation mediated by proteases and the ubiquitin–proteasome pathway. Thus, it would be possible to improve the current knowledge on the crosstalk between the molecular/cellular environments and the disease pathophysiology by a systems-level approach.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme FP7/2007–2013 under grant agreement FP7-PEOPLE-2013-ITN-608332.

Conflict of Interest Statement

K.C. is employed by Mosaiques Diagnostics and the authors have no conflict of interest.

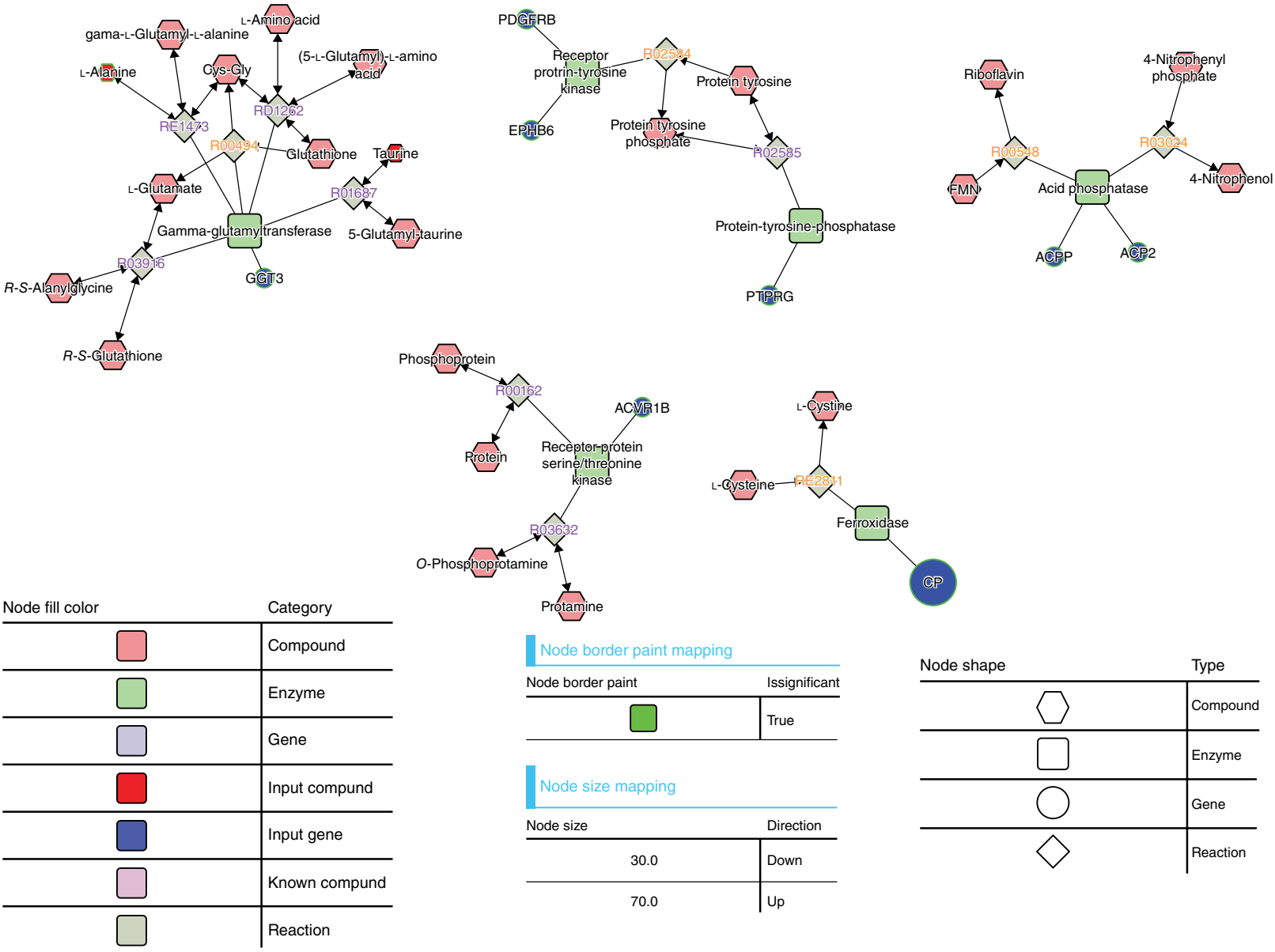


Figure 20.9 Reconstruction of metabolic pathways using network-based analysis in MetScape [32] app for Cystoscape. The following thresholding was applied metabolite FC ≥ 1.10 and *p* value < 0.05, gene–protein FC ≥ 1.5 and *p* value < 0.05. Association between gene–enzyme–reaction–metabolite is shown. The direction of the regulation (up or down) is represented by node size, in which smaller to greater nodes denote down- and upregulation, respectively. Statistically significant (*p* value < 0.05) molecules are surrounded by a border green paint.

References

1 Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet.* 2012; 380: 2095–2128.

2 Jha V, Garcia-Garcia G, Iseki K, et al. Chronic kidney disease: global dimension and perspectives. *Lancet.* 2013; 382: 260–272.

3 Kampen AH and Moerland PD. Taking bioinformatics to systems medicine. *Syst Med*. 2015; 1386: 17–41.

4 Papadopoulos T, Krochmal M, Cisek K, et al. Omics databases on kidney disease: where they can be found and how to benefit from them. *Clin Kidney J.* 2016; 9: 343–352.

5 Husi H, Sanchez-Nino MD, Delles C, et al. A combinatorial approach of Proteomics and Systems Biology in unravelling the mechanisms of acute kidney injury (AKI): involvement of NMDA receptor GRIN1 in murine AKI. *BMC Syst Biol.* 2013; 7: 110.

6 Fechete R, Heinzel A, Perco P, et al. Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. *Proteomics Clin Appl.* 2011; 5: 354–366.

7 Bhat A, Heinzel A, Mayer B, et al. Protein interactome of muscle invasive bladder cancer. *PLoS One.* 2015; 10: e0116404.

8 Neves J, Martins MR, Vilhena J, et al. A soft computing approach to kidney diseases evaluation. *J Med Syst.* 2015; 39: 131.

9 Husi H, Van Agtmael T, Mullen W, et al. Proteome-based systems biology analysis of the diabetic mouse aorta reveals major changes in fatty acid biosynthesis as potential hallmark in diabetes mellitus-associated vascular disease. *Circ Cardiovasc Genet.* 2014; 7: 161–170.

10 Holzinger A, Dehmer M and Jurisica I. Knowledge discovery and interactive data mining in bioinformatics—state-of-the-art, future challenges and research directions. *BMC Bioinformatics*. 2014; 15 Suppl 6: I1.

11 Meng L, Michaud GA, Merkel JS, et al. Protein kinase substrate identification on functional protein arrays. *BMC Biotechnol.* 2008; 8: 22.

12 Mohr S, Bakal C and Perrimon N. Genomic screening with RNAi: results and challenges. *Annu Rev Biochem.* 2010; 79: 37–64.

13 Saito R, Smoot ME, Ono K, et al. A travel guide to Cytoscape plugins. *Nat Methods.* 2012; 9: 1069–1076.

14 Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003; 13: 2498–2504.

15 Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25: 25–29.

16 Bindea G, Mlecnik B, Hackl H, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics.* 2009; 25: 1091–1093.

17 Kanehisa M and Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000; 28: 27–30.

18 Kutmon M, Riutta A, Nunes N, et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res.* 2016; 44: D488–D494.

19 Fabregat A, Sidiropoulos K, Garapati P, et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* 2016; 44: D481–D487.

20 Huang DW, Sherman BT, Tan Q, et al. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* 2007; 8: R183.

21 Bindea G, Galon J and Mlecnik B. CluePedia Cytoscape plugin: pathway insights using integrated experimental and in silico data. *Bioinformatics.* 2013; 29: 661–663.

22 Ozgur A, Vu T, Erkan G and Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics.* 2008; 24: i277–i285.

23 Bauer-Mehren A, Bundschus M, Rautschka M, Mayer MA, Sanz F and Furlong LI. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One.* 2011; 6: e20284.

24 Pinero J, Queralt-Rosinach N, Bravo A, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database (Oxford)*. 2015; 2015: bav028.

25 Keane H, Ryan BJ, Jackson B, Whitmore A and Wade-Martins R. Protein-protein interaction networks identify targets which rescue the MPP+ cellular model of Parkinson's disease. *Sci Rep.* 2015; 5: 17004.

26 Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* 2015; 43: D447–D452.

27 Orchard S, Ammari M, Aranda B, et al. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* 2014; 42: D358–D363.

28 Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A and Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* 2006; 34: D535–D539.

29 Montojo J, Zuberi K, Rodriguez H, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. *Bioinformatics.* 2010; 26: 2927–2928.

30 Warde-Farley D, Donaldson SL, Comes O, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* 2010; 38: W214–W220.

31 Oliver S. Guilt-by-association goes global. *Nature.* 2000; 403: 601–603.

32 Karnovsky A, Weymouth T, Hull T, et al. Metscape 2 bioinformatics tool for the analysis and visualization of metabolomics and gene expression data. *Bioinformatics.* 2012; 28: 373–380.

33 Ma H, Sorokin A, Mazein A, et al. The Edinburgh human metabolic network reconstruction and its functional analysis. *Mol Syst Biol.* 2007; 3: 135.

34 Jolma A, Yin Y, Nitta KR, et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature.* 2015; 527: 384–388.

35 Mathelier A, Fornes O, Arenillas DJ, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 2016; 44: D110–D115.

36 Vaquerizas JM, Kummerfeld SK, Teichmann SA and Luscombe NM. A census of human transcription factors: function, expression and evolution. *Nat Rev Genet.* 2009; 10: 252–263.

37 Ha TY. The role of MicroRNAs in regulatory T cells and in the immune response. *Immune Netw.* 2011; 11: 11–41.

38 Kutmon M, Kelder T, Mandaviya P, Evelo CT and Coort SL. CyTargetLinker: a Cytoscape app to integrate regulatory interactions in network analysis. *PLoS One.* 2013; 8: e82160.

39 Yates A, Akanni W, Amode MR, et al. Ensembl 2016. *Nucleic Acids Res.* 2016; 44: D710–D716.

40 Wheeler DL, Barrett T, Benson DA, et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 2007; 35: D5–D12.

41 UniProt C. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* 2012; 40: D71–D75.

42 Kozomara A and Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* 2014; 42: D68–D73.

43 Knox C, Law V, Jewison T, et al. DrugBank 3.0: a comprehensive resource for "omics" research on drugs. *Nucleic Acids Res.* 2011; 39: D1035–D1041.

44 Vlachos IS, Zagganas K, Paraskevopoulou MD, et al. DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.* 2015; 43: W460–W466.

45 Kutmon M, van Iersel MP, Bohler A, et al. PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput Biol.* 2015; 11: e1004085.

46 van Iersel MP, Kelder T, Pico AR, et al. Presenting and exploring biological pathways with PathVisio. *BMC Bioinformatics.* 2008; 9: 399.

47 Villaveces JM, Koti P and Habermann BH. Tools for visualization and analysis of molecular networks, pathways, and -omics data. *Adv Appl Bioinform Chem.* 2015; 8: 11–22.

48 van Iersel MP, Pico AR, Kelder T, et al. The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services. *BMC Bioinformatics.* 2010; 11: 5.

49 Kanehisa M, Sato Y and Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol.* 2016; 428: 726–731.

50 Tian Q, Price ND and Hood L. Systems cancer medicine: towards realization of predictive, preventive, personalized and participatory (P4) medicine. *J Intern Med.* 2012; 271: 111–121.

51 Huey R, Morris GM, Olson AJ and Goodsell DS. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem.* 2007; 28: 1145–1152.

52 Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res.* 2000; 28: 235–242.

53 Cryer MJ, Horani T and DiPette DJ. Diabetes and hypertension: a comparative review of current guidelines. *J Clin Hypertens (Greenwich).* 2016; 18: 95–100.

54 Evans PD and Taal MW. Epidemiology and causes of chronic kidney disease. *Medicine.* 2011; 39: 402–406.

55 Anders HJ and Lech M. NOD-like and toll-like receptors or inflammasomes contribute to kidney disease in a canonical and a non-canonical manner. *Kidney Int.* 2013; 84: 225–228.

56 Hao X, Liu X, Wang W, et al. Distinct metabolic profile of primary focal segmental glomerulosclerosis revealed by NMR-based metabolomics. *PloS One.* 2013; 8: e78531.

57 Sui W, Li L, Che W, et al. A proton nuclear magnetic resonance-based metabonomics study of metabolic profiling in immunoglobulin a nephropathy. *Clinics (Sao Paulo).* 2012; 67: 363–373.

58 Posada-Ayala M, Zubiri I, Martin-Lorenzo M, et al. Identification of a urine metabolomic signature in patients with advanced-stage chronic kidney disease. *Kidney Int.* 2014; 85: 103–111.

21

Application of Omics and Systems Medicine in Bladder Cancer

Maria Frantzi[1,], Agnieszka Latosinska[1,2,*], Murat Akand[3,4], and Axel S. Merseburger[5]*

[1] *Mosaiques Diagnostics GmbH, Hannover, Germany*
[2] *Biotechnology Division, Biomedical Research Foundation, Academy of Athens, Athens, Greece*
[3] *Department of Urology, School of Medicine, Selcuk University, Konya, Turkey*
[4] *Department of Urology, School of Medicine, Katholieke Universiteit Leuven, Leuven, Belgium*
[5] *Department of Urology, University of Lübeck, Lübeck, Germany*

21.1 Introduction

Bladder cancer (BC) is ranked second for incidence and mortality among cancers of the genitourinary system. Approximately 75% of patients experience non–muscle-invasive BC (NMIBC, cancer cells are confined to mucosa or submucosa) at the time of diagnosis, whereas the remaining 25% of cases exhibit muscle-invasive disease (MIBC, cancer cells have spread into muscle layer) [1]. Patients suffering from NMIBC have a high probability of disease recurrence and progression, with the probability of recurrence within 5 years' time to range from 31 to 78% and the probability of progression within 5 years' time to range from 0.8 to 45%, respectively [2]. Muscle invasion and metastasis result in poor prognosis, with reported 5-year survival rates to be between 46 and 63% for MIBC and decreasing to 15% for metastasized cancer cases (Table 21.1). The increased mortality as cancer progresses is attributed to the limited treatment options that are currently available, as also illustrated in Table 21.1. The treatment of choice for MIBC is total bladder resection, also known as radical cystectomy (RC), whereas for NMIBC transurethral resection of bladder tumor (TURBT) is applied, combined with either intravesical Bacillus Calmette–Guerin (BCG) or chemotherapy (mitomycin C, epirubicin, doxorubicin) instillation according to its risk profile. However, approximately 40% of BC patients are frequently not responsive to BCG therapy, as defined by the detection of muscle invasive disease, high-grade papillary tumors, and/or CIS lesions (carcinoma in situ), or by the presence of serious side effects that would impede treatment [3]. However, patients that progress to MIBC after BCG frequently show poorer prognosis than those with primary diagnosed stage-matched MIBC (with the 3 years' disease-specific survival rate to be reported at 37% in comparison to 67%) [4, 5]. Upon BCG failure, RC or palliative TURBT are the preferred options [3]. In addition to surgical intervention, several potentially useful drugs are being applied, targeting different molecular pathways. In clinical trials for drugs, findings for predicting treatment response have been found to be inconsistent, indicating for the most part significant but moderate, patient benefit [6–9]. There is accumulating evidence indicating that the high heterogeneity of bladder tumors may significantly affect the efficacy of the tested drugs. In an effort to improve on patient outcome, extensive research has been undertaken to address early detection of incident and recurrent disease along with earlier and more personalized therapeutic intervention guided by the tumor's molecular profile.

In this chapter, a comprehensive description of the BC pathology along with the associated clinical challenges is provided. Driven by the clinical needs and considering the high disease heterogeneity, the research progress made over the last years, focusing on the application of state-of-the art *-omics* approaches and systems medicine is summarized.

---

*Equal contribution.

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

Table 21.1 Current treatment modalities across the bladder cancer disease development and progression.

	Nonmuscle-invasive		Muscle-invasive		Metastatic
Stage	**Ta/CIS**	**T1**	**T2**	**T3/T4**	**M+**
Survival (5 years)	**98%**	**88%**	**63%**	**46%**	**15%**
Treatment	**Transurethral resection (TURBT) ± intravesical**		**Radical cystectomy**		**Palliative therapy**
	BCG or chemotherapy (Mitomycin C, epirubicin, doxorubicin)		**±Chemotheraphy** (methotrexate, vinblastine, adriamycin and epirubicin, or cisplatin + gemcitabine)		
	Cystectomy (partial/ radical)		**TURBT ± chemotherapy + radiotherapy**		**Chemotherapy**
Clinical trials	Anti-PD-L1		Anti-PD-L1		Anti-PD-L1, anti-CTLA-4, anti-FGFR, cell-cycle checkpoint inhibitors

CIS, carcinoma in situ; CTLA-4, cytotoxic T-lymphocyte-associated protein 4; FGFR, fibroblast growth factor receptor; PD-L1, programmed death-ligand 1; TURBT, transurethral resection of the bladder tumor.

21.2 Bladder Cancer Pathology and Clinical Needs

21.2.1 Epidemiological Facts and Histological Classification

BC is the most common malignancy of the urinary tract, the ninth most commonly diagnosed cancer worldwide, and the fifth most common in Western countries [10]. According to the data of cancer incidence and mortality estimates for 2012, there is an estimation of 429 000 newly diagnosed BC cases and 165 000 BC-related deaths in 2012 worldwide [10]. The age-standardized disease rate is highest in northern America, European countries, and northern Africa, whereas it is lowest in middle Africa, South America, and South and East Asia [11]. In Europe, BC is the fifth most common cancer with age-standardized rates of 14.4 per 100 000 cases, in both sexes [12]. Importantly, as life expectancy is increasing in Europe, the annual incidence of BC is projected to reach 164 500 by 2030 [13]. Although BC is more commonly seen in males with a median male-to-female ratio of 3.05 (incidence rates of 26.9 for men and 5.3 for women per 100.000 EU cases [12]), females, especially younger ones, are more frequently diagnosed with advanced stage BC when compared to males in the same age group [14].

Smoking is the main risk factor, which accounts for approximately half of all BC cases [15]. Other risk factors include the exposure to polycyclic aromatic hydrocarbons, aromatic amines, and chlorinated hydrocarbons, which are often used in dye, metal, and petroleum industries [15].

BC is also categorized into two morphological subtypes: (i) flat and (ii) papillary carcinomas. Flat carcinomas (carcinoma in situ (CIS)) appear as a flat lesion that is limited to the inner layer of the bladder lining (transitional epithelium), whereas the structure of the papillary tumors resemble a "cauliflower" (i.e., projection of the transitional epithelium into the lumen of the bladder). In the absence of timely treatment, both these lesions can result in the development of muscle-invasive disease.

Currently, the treatment choice is based on the assessment of the tumor stage using The Tumor-Node-Metastasis (TNM) classification system. The following factors are considered: (i) size of the tumor and level of infiltration into bladder wall (T, tumor, with the range from T0 to T4; for patients with NMIBC—stages Ta, CIS, T1 and for patients with muscle-invasive diseases—stages T2, T3, T4), (ii) number of affected lymph nodes (N, nodes), and (iii) presence of metastasis to other parts of the body (M, metastasis).

21.2.2 Current Diagnostic Means

In order to diagnose BC, patients are subjected to both cystoscopy and voided urinary cytology. Cystoscopy is applied for the visual examination of the interior of the bladder. One variant of this approach is fluorescence-guided cystoscopy (called blue light cystoscopy). Collection of the tissue specimens is performed during a subsequent TURBT. Based on the histopathological examination of the collected tissue specimens, the extent of tumor invasion and tumor grade are assessed. Overall, patients with BC, especially those with low-grade tumors (characterized as well-differentiated) have a more favorable

disease outcome; while patients with high-grade tumors (undifferentiated growths) have poor prognosis and their tumors are likely to spread beyond the bladder wall. However, due to the invasive nature of the examination, patient compliance is usually limited. Of note, detection of flat tumors is considered challenging, even for highly competent urologists. Moreover, regarding the biopsy results, the correct examination of the tissue is highly dependent on the experience of the pathologist. Usually, cystoscopic examination is assisted by voided urinary cytology analysis (i.e., assessment of the cells in urine samples). Recently, some noninvasive urinary tests have been developed as an alternative. Tumor size and the extent of the spread of the tumor cells into the body (lymph nodes and distant organs) is determined using imaging techniques. Intravenous or retrograde pyelography is initially applied. Specifically, retrograde pyelography is performed in cases of suspected allergy to venous contrast agents. However, CIS is very frequently not detectable when using these tools, requiring further confirmation with endoscopic biopsy. To date, computed tomography (CT) urogram and abdominal CT (when combined with needle biopsy) appears to be the optimal approach to assess the cancer dissemination to nearby and distant organs. Along these lines, abdominal and pelvic magnetic resonance imaging (MRI) is an alternative solution. In the case of advanced stage disease, lung metastasis can be assessed by using chest radiography, while bone metastasis may be detected with bone scintigraphy.

21.2.3 Treatment Options

The selection of the optimal treatment scheme depends on the type of the tumor, its clinico-pathological characteristics (stage, grade, foci number, and the level of tumor infiltration into the bladder wall), as well as presence of metastasis and the general health condition of a patient. In the case of NMIBC, TURBT is the first-line treatment option. TURBT relies on the removal of tumor during cystoscopy by a resectoscope. When muscle-invasive disease is diagnosed, patients are treated with RC together with the dissection of lymph nodes, based on the level of tumor infiltration to nearby structures and organs. The available adjuvant intravesical therapies include immunotherapy and chemotherapy. Specifically, patients suffering from early-stage BC that undergo TURBT are frequently treated with intravesical immunotherapy with BCG. Moreover, adjuvant intravesical chemotherapy using mitomycin C, epirubicin, or doxorubicin can be also considered. In the case of patients with advanced BC, systemic (intravenous) administration of chemotherapeutic agents is required. In addition, neoadjuvant treatment is preferable prior to RC. Clinical trials in patients with node-negative MIBC (N0) demonstrated a benefit associated to the use of MVAC chemotherapy (MVAC abbreviation stands for the drugs applied in treatment of BC: Methotrexate, Vinblastine, Doxorubicin, and Cisplatin) or gemcitabine and cisplatin prior to RC, in comparison to those who underwent RC alone [16]. For treatment prediction to neoadjuvant therapy prior to RC, gene expression profiles have been already assessed for their predictive ability [17, 18]. When managing patients with recurrent cancer, surgery (with or without chemotherapy and targeted therapy with biologic agents) or combination chemotherapy is advised.

21.2.4 Recurrence and Progression

High recurrence and progression rates place a heavy burden on patients with BC. Currently, the risk of recurrence and progression for patients with NMIBC is based on the system established by the European Organization for Research and Treatment of Cancer Genito-Urinary Cancer Group (EORTC–GUCG) [2]. This includes a scoring system and the risk tables, established based on the analysis of 2596 patients diagnosed with NMIBC (stages Ta and T1). The scoring system accounts for several factors including: (i) number of tumors, (ii) tumor size, (iii) prior recurrence rate, (iv) tumor stage, (v) grade, and (vi) presence of concurrent CIS. The final score is then used to assign the patient to a specific category and allows to calculate the probability of recurrence and progression at 1 year and 5 years [2]. According to the EORTC system, NMIBC is categorized into three risk groups including low-, intermediate- and high-risk. Low-risk tumors have characteristics of primary and solitary tumors with Ta stage, low-grade (or grade (G)1) and size <3 cm without concomitant CIS. Tumors with any of the characteristics of T1 stage, high-grade (or G3), concomitant CIS or multiple, recurrent and large (>3 cm) Ta G1–G2 tumors are categorized as high-risk. Intermediate-risk tumors consist of all tumors that are not defined in these two categories. Currently, specific treatment recommendations have been defined for these three groups [3]. Patients classified to the low-risk group are characterized by a reduced risk of disease progression and a low to moderate risk of recurrence, whereas patients that fall into the intermediate- and high-risk group have an increased risk for disease recurrence and a moderate to high risk for progression to MIBC. Based on the current treatment options, patients belonging to the low-risk group have good prognosis, whereas patients with intermediate- and high-risk NMIBC have less-favorable prognosis [3].

In addition, understaging of 35–62% Ta/T1 tumors based upon a large cystectomy series is reported [19–21]. Other studies indicate that second TURBT identifies that 24–49% of T2 tumors had been diagnosed initially

as non–muscle-invasive tumors [22, 23]. Progression to MIBC significantly decreases cancer-specific survival (CSS). In a review of 19 trials and 3088 patients, CSS after progression to MIBC from NMIBC was 35% after a median follow-up of 48–123 months, which was significantly worse compared to that of patients with MIBC without a history of NMIBC [5].

21.2.5 Molecular Classification

Insufficiency of the current scoring systems that are solely based on clinical and pathological variables may be in part attributed to the intrinsic heterogeneity of BC at the molecular level. Molecular research has revealed two genetically distinct pathways, based on which BC develops into the papillary and flat/nonpapillary forms [24, 25]. The papillary BC form histologically refers to the NMIBC cases that develop via urothelial hyperplasia and are associated with disruption on the PI3K-AKT-mTOR pathway and additional mutations in the FGFR3 and HRAS genes [26]. On the contrary, aggressive nonpapillary MIBC disease can be developed via flat dysplasia and CIS. MIBC tumors, derived from CIS, are characterized by genetic alterations in tumor suppressor genes that regulate cell cycle and apoptosis (TP53, CDKN2A, CCND1, CDKN1B and RB1) [26]. This genetic characterization in part explains the BC heterogeneity and the different evolution of the two BC types tumor [25]. However, recent omics studies support additional classification of BC tumors within the MIBC and NMIBC types [27–29].

21.2.6 Biomarkers for Bladder Cancer

Up to date, white light cystoscopy and voided urinary cytology consist of the typical diagnostic and monitoring means in BC. Apart from being an invasive procedure, cystoscopy frequently misses high-risk CIS tumors. In a meta-analysis performed by Mowatt et al. [30] in a total of 27 studies including 2949 patients, sensitivity of white light cystoscopy was reported to be 71% (49–93%, 95% CI) while specificity was 72% (47–96%, 95% CI). Sensitivity levels are pronounced for detecting less aggressive low-risk tumors (95%), while it decreases for more aggressive high-risk tumors (67%). Another option, so called photodynamic diagnostic (PDD), was also evaluated as an alternative to increase the tumor detection, and indeed exhibits increased sensitivity of 93% (80–100%, 95% CI) but decreased specificity of 57% (36–79%, 95% CI). In the case of the cytological examination, an average sensitivity of 44% and specificity of 99% were reported in the meta-analysis by Mowatt et al. [30].

In an effort to reduce the number of cystoscopies, noninvasive biomarkers have been proposed as an alternative. An overview of biomarker classification system along with the examples of Food and Drug Administration (FDA)-approved biomarker-based assays available for BC are outlined in Box 21.1.

Box 21.1 Biomarker Classification

Definition:

- **Biomarker**—Single feature associated with specific condition, detectable in biological fluids or tissues.
- **Biomarker profile**—Combination of individual markers by an established algorithm providing with the output associated with the specific condition.

Type of biomarkers and examples approved for clinical application:

- **Early detection**—Utilized for evaluation of the patients susceptible to risk factors or exhibiting some disease symptoms.
  Approved for clinical application: BladderChek (Matritech, Newton, MA)
- **Diagnostic biomarker**—Applicable for detection and identification of particular type of cancer.
  Approved for clinical application: BTA stat, BTA TRAK, NMP22, Immunocyt, UroVision
- **Prognostic biomarker**—Used for prediction of the course of disease including recurrence or progression.
  Approved for clinical application: N/A
- **Predictive biomarker**—Evaluation of response to treatment before starting therapy. Enables patient categorization into responders and nonresponders. To the best of our knowledge, this type of biomarker is not currently in use.
  Approved for clinical application: N/A
- **Surrogate endpoint**—Replace a clinical endpoint and measurement of clinical benefits and harms.
  Approved for clinical application: N/A

Type of biomarkers measurements:

- **Binary**—Evaluation of specific condition based on their presence or absence; for example, positive/negative.
- **Categorical**—Assessment of specific condition based on defined categories; for example, low/medium/high immunoreactivity (IHC-based studies).
- **Quantitative**—Categorize the specific condition based on the precise expression level of the protein (applicable in measurements in body fluids).
- **Multidimensional**—Examination of specific condition based on the molecular signature; for example, proteomic signature or peptide profile.

For almost 20 years now, several urinary-based biomarker tests have been approved by the FDA. These include (i) immunoassays to detect urinary proteins such as BC-associated antigen (BTA TRAK, BTA stat) and nuclear matrix protein NMP22 (ELISA and point of care test), (ii) immunocytofluorescence-based test, namely ImmunoCyt, and (iii) fluorescence in situ hybridization based assay (UroVysion assay). The intended use of the biomarkers is, according to the FDA, for either diagnosis or monitoring recurrence, always in conjunction with standard diagnostic procedures. The performance of each test, as summarized in the meta-analysis of 71 studies (3321 patients), indicate the highest sensitivity for ImmunoCyt assay [30]. However, urine cytology achieves the highest specificity levels. In particular, the ImmunoCyt assay was characterized by a sensitivity of 84% (77–91%; 95% CI) and a specificity of 75% (68–83%; 95% CI). FISH test presented with a sensitivity and specificity of 76% (65–84%; 95% CI) and 85% (78–92%; 95% CI), respectively, whereas for NMP22 the sensitivity was estimated at 68% (62–74%; 95% CI) and the specificity at 79% (74–84%; 95% CI) [30]. Although the performance of these biomarkers is not adequate to replace cystoscopy, a combination of the biomarkers resulted in an improved sensitivity value. Combination of cytology and NMP22 resulted in 63% sensitivity and 84% specificity, compared to 48% sensitivity and 86% specificity of cytology alone [31]. For high-grade tumors, the sensitivity was reported to be 94%, while for low-grade tumors it was 31%. Combination of cytology and BTA stat on the other hand had 73% sensitivity (with 91% sensitivity reported for high-grade and 42% for low-grade) [31]. Finally, cytology when applied with ImmunoCyt resulted in 65% sensitivity and 78% specificity (90% sensitivity was reported for high-grade and 50% for low-grade) [31]. Even though these tests were approved by the FDA, their incorporation into the clinical management of BC is still pending, indicating that there is room for improvement. There is a growing number of studies supporting the use of molecular markers to screen patients with low-risk disease and predict tumor progression in high-risk patients. In addition, biomarkers can also serve as a companion/complementary tool to guide the therapeutic decision-making through discrimination of the patients to those that are likely to respond to a therapy and nonresponders. In conclusion, noninvasive biomarkers could be used as a companion test to currently available means for disease diagnosis or follow-up (detection of disease recurrence and/or progression).

21.2.7 Considerations on Patient Management

In a recent paper evaluating the economic burden of BC in Europe, it was estimated that the associated healthcare costs for BC reached €2.87 billion in 2012, accounting for 5% of the total cancer health expenditure [32]. This number increased to €4.9 billion when productivity losses (due to morbidity and mortality) and informal care costs were also considered. Particularly, NMIBC is becoming a very expensive disease to manage, as NMIBC patients undergo life-long surveillance. Cystoscopies and TURBT procedures together constitute 53% of all BC-related costs in Northern European countries [33], while it has been estimated that monitoring of patients with BC covers for approximately 75% of postdiagnostic costs (i.e., surgery-related complications, annual examinations, diagnostic and laboratory testing). Moreover, the overall treatment costs increase upon disease progression. Specifically, the costs associated with the management of MIBC patients are three times higher in comparison to patients with NMIBC. Therefore, timely diagnosis of primary and recurrent cancer followed by earlier intervention are both crucial factors for decreasing the risk of progression and subsequently reducing the BC-associated costs.

21.2.8 Defining the Disease-Associated Clinical Needs

The intrinsic clinical features of BC, high recurrence, and progression rates mandate continuous patient monitoring. Considering the drawbacks of the current diagnostic and monitoring procedures, new biomarker tests are needed for timely diagnosis and/or surveillance. Also, as BC progresses, it becomes increasingly difficult to treat, as the therapeutic options are limited or the patients present with increased variability in their response. The high disease heterogeneity is therefore responsible for the limited response to current treatments. Means to stratify patients for drug development as well as monitoring tools to predict the response to therapy is the new research and clinical focus in BC. In each case, the clinically valid biomarkers apart from targeting a specific clinical context of use, should clearly demonstrate an added value over the stage-of-the-art. The clinical needs for BC as well as several targeted clinical contexts are summarized in Figure 21.1.

21.3 Systems Medicine in Bladder Cancer

Systems medicine refers to cross-disciplinary and innovative approaches that emphasize the translational value of research (i.e., combination of clinical applicability and biological relevance of findings) (Figure 21.2). Through the analysis of *-omics* datasets by using systems biology approaches, researchers aim to analyze BC at the molecular level, and particularly to decipher molecular key

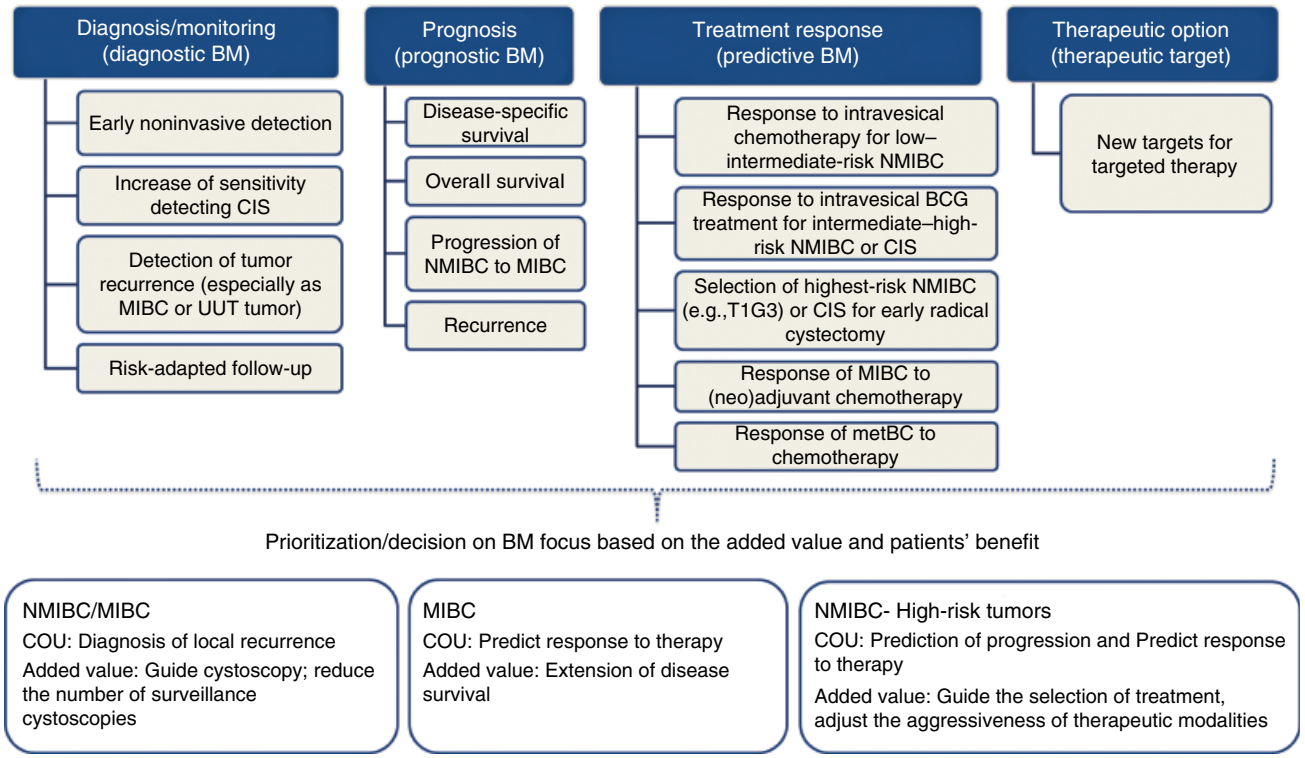


Figure 21.1 Clinical needs for bladder cancer. BCG, Bacillus Calmette–Guerin; BM, biomarker; CIS, carcinoma in situ; COU, context of use; metBC, metastatic BC; N(MIBC), (non)–muscle-invasive BC; UUT, upper urinary tract.

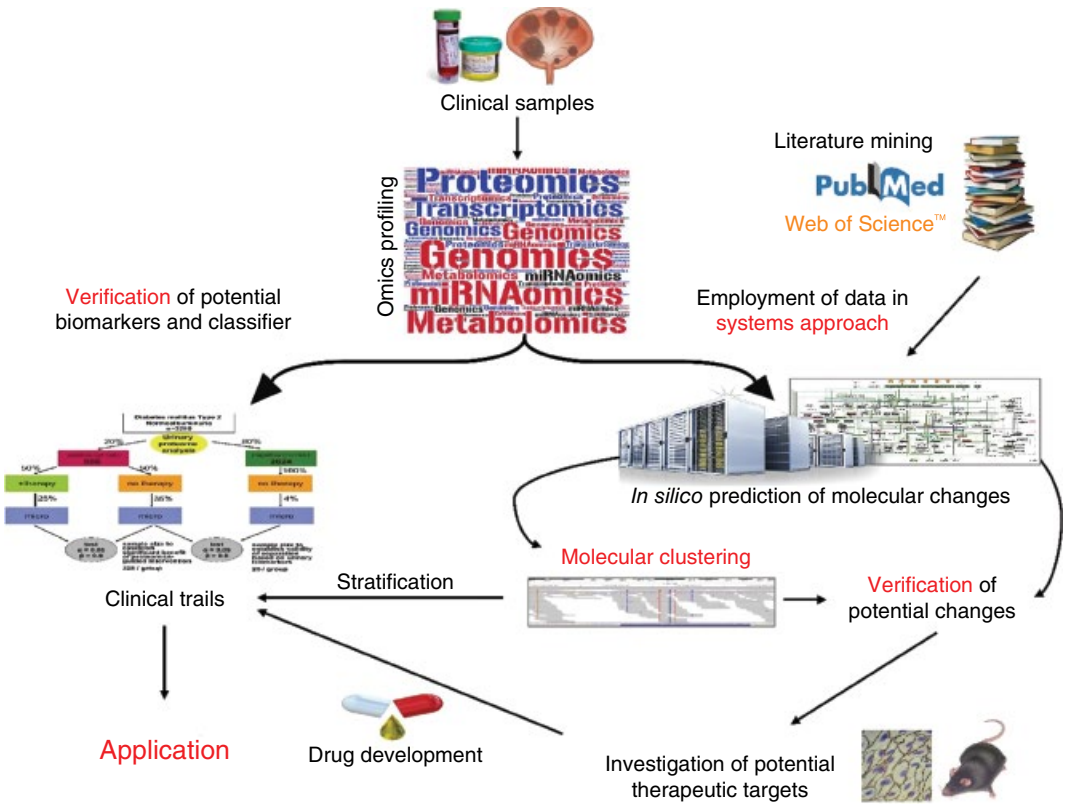


Figure 21.2 Graphical summary of the systems medicine-driven approach.

elements implicated in BC invasion. The concept generally refers to the combination/cross-correlation of high dimensional *-omics* (genomics/transcriptomics/proteomics, etc.) datasets and their interpretation by appropriately powered bioinformatics software and tools. Based on this knowledge, the desired aim is to tailor therapeutic approaches for BC by: (i) generating specific disease models and indicating novel potential drug targets and (ii) rendering risk profiles for treatment response. Moreover, based on the multilevel integration of the *-omics* results, more accurate diagnostic and/or prognostic biomarkers are expected to be established in order to support disease diagnosis and monitoring of recurrence, as well as patient stratification to assist clinical trials for appropriate drug testing and development.

The *-omics* datasets in combination with computational systems biology methodologies can provide information about the molecular changes occurring in BC and identify molecular targets at an individual level. Over the last 5 years, after the first genetic sequencing of BC tumor specimens [34], there has been rapid increase in data derived from whole genome sequencing [35], exome BC sequencing [36], and transcriptomics data [28]. Importantly, regarding the field of proteomics, developments in mass spectrometry instrumentation and bioinformatics software now enable high-throughput proteomics analysis [37]. To better exemplify the possibilities offered by the *-omics* technologies as well as their implication on the management of BC patients, most representative studies published over the last 5 years are presented. The literature review criteria are presented in Table 21.2. Studies were shortlisted based on the number of citations (at least 10 citations per year), with an exception for studies published in 2016.

Table 21.2 Overview of the literature review including the search criteria and the results obtained.

Search criteria	
Search engine	*Web of Science*
Resources	*Web of Science Core Collection*
Search terms	**Topic:** ("omic*" or "proteom*" or "transcriptom*" or "genom*" or "metabolom*" or "signature*" or "systems biology" or "systems medicine") *AND* **Topic:** ("bladder ca*" or "urothelial ca*" or "transitional cell*") *AND* **Topic:** ("drug" or "therap*" or "biomarker*" or "predict*" or "dignos*" or "prognos*")
Time span	2012–2016

Results			
Year	**No. manuscripts**	**No. original articles**	**No. articles (citation threshold)**
2012	119	86	14 (>40×)
2013	149	112	16 (>30×)
2014	176	140	27 (>20×)
2015	202	154	12 (>10×)
2016	146	120	N/A

*Denotes that the end of the word can vary so that the search is broad.

21.3.1 Omics Datasets for Biomarker Research

The main research applications are summarized below, categorized based on their intended clinical context of use in: (i) diagnostic biomarkers for disease detection/monitoring, (ii) prognostic signatures, (iii) predictive molecular profiles, and (iv) features for molecular subclassification.

21.3.1.1 Diagnostic Biomarkers for Disease Detection/Monitoring

To date, numerous studies have been conducted to develop noninvasive biomarker-based test to improve on the diagnosis of primary and recurrent/relapsed BC. Application of both proteomics and genomics profiling have been advocated. Regarding the proteomics advances in the BC field, recently, high-resolution capillary electrophoresis coupled to mass spectrometry (CE-MS) was applied to investigate the urinary proteome profile of BC patients and urological controls (total $n = 1357$) [38]. Two multimarker panels were established using support vector machine algorithms, aiming at detection of primary ($n = 721$) and relapsed BC ($n = 636$). A biomarker panel to detect the primary disease was comprised of 116 differentially excreted peptides between patients with primary BC and urological controls, while a biomarker panel targeting to identify recurrent disease, included 106 differentially excreted peptides between patients with relapsed BC and patients negative for recurrence (disease free interval of at least 1 year). Upon validation in the independent cohort (total $n = 481$), an accuracy of 87 and 75% was reported for detecting primary and relapsed BC. For the former panel, sensitivity of 91% and specificity of 68% were achieved, while for the latter panel sensitivity of 87% and specificity of 51% were obtained. Importantly, combination of urinary biomarkers with cytology increased the accuracy for detecting disease relapse (87%) [38]. In another study, Chen et al. performed a targeted proteomics analysis using LC-MRM/MS, aiming at the validation of previously discovered biomarkers for BC [39]. Protein concentration was measured in 156 urine samples collected from patients with BC, hernia, and urinary tract infection/hematuria. Emphasis was placed on the investigation of 63 proteins

that are usually detected in plasma samples. Urinary levels of 12 proteins were found to be increased in patients with BC in comparison to both control groups, indicating the potential diagnostic utility. As a result of this study, a six-biomarker based panel was established and allowed for discrimination between patients with BC and noncancerous control (diagnostic accuracy of 81%) [39].

In parallel, numerous biomarker-based panels have already been established using gene expression profiling datasets of urine and tissue samples. The performance of four-gene expression signatures was validated in the course of a multicenter, prospective blinded study [40]. A total of 789 urine samples were collected, out of which 525 were successfully analyzed using TaqMan gene expression arrays. All four panels exhibited a good diagnostic performance in the range of 90–92%. The highest diagnostic accuracy was obtained for the two-gene–based signature (namely GS_D2; AUC of 0.918, sensitivity of 81.48% and specificity of 91.26%). Of note, the diagnostic accuracy was associated with the size of the tumor ($p = 0.008$), while the number of tumors had no impact on the diagnostic accuracy [40]. In another study, as published by van der Heijden et al., 21 tumor samples from patients with progressive and nonprogressive T1G3 BC were analyzed using Illumina microarrays, aiming at the development of a gene signature to identify BC patients with a high risk of progression [41]. A total of 1294 genes were found to be differentially expressed between patients with nonprogressive and progressive BC. Ninety-four of these features were subsequently validated in the independent set of samples using qPCR ($n = 75$), confirming the differential expression of 15 genes. A five-gene–based signature enabled to discriminate between patients with progressive disease from nonprogressive BC with an AUC of 0.83 (sensitivity of 79%, specificity of 86%) [41].

Interestingly, a growing number of evidence is now collected to also support the diagnostic value of microRNA in the context of BC. Recently, in a meta-analysis by Zheng et al. [42], the potential of microRNA to diagnose MIBC was evaluated. In this meta-analysis, a total of 10 studies were assessed (total $n = 989$) [42]. Urinary (miR-124), blood (e.g., miR-200b, miR-541, miR-566, miR-543, miR-544, and others), and tissue (e.g., miR-100, miR-125b, miR-199b, miR-222, and others) microRNAs were investigated. Random-effect model was applied to assess the diagnostic performance of single and multiple miRNAs among the studies included in the meta-analysis. Forest plot analysis was performed to assess the mean values of the sensitivity and specificity per sub-group and study. Furthermore, a summary receiver operator characteristic (SROC) curve was conducted and resulted in a pooled AUC value of 0.80, with estimated overall sensitivity of 78% and overall specificity of 77% [42]. The results from the meta-analysis supported the further validation of the use of microRNAs for MIBC diagnosis.

Another analysis of urinary microRNA in a cohort of 131 patients (81/50 patients in discovery/validation set), aiming to improve the detection of BC recurrence was performed [43]. Based on the literature review, 12 microRNAs were selected for further investigation. Highest differentiation potential between patients with the history of BC (no recurrence at cystoscopy) and patients with BC was reported for the set of six biomarkers (i.e., miR16, miR200c, miR205, miR21, miR221, and miR34a), with an AUC of 0.85 [43]. Subsequent validation in an independent set of samples ($n = 50$) yielded a sensitivity of 88% and specificity of 48% [43], with the highest performance reported for pT1 stage (AUC of 0.92) and the lowest when aiming at detection of low-volume tumor (AUC of 0.69). Preliminary analysis revealed that using the established panel, the number of performed cystoscopies could be reduced by 30%. Thus, the urinary miRNA profiling gives an opportunity to reduce the morbidity and costs associated with management of BC patients during follow-up. However, these findings require prospective validation in a bigger cohort [43].

21.3.1.2 Prognostic Signatures

To improve the current clinical and pathological prognostic biomarkers for BC progression, different *-omic* methodologies have been already employed, based on either tissue datasets or those derived from systemic biological fluids (e.g., urine, and plasma). Genomic screening for prognostic factors related to genetic mutation and copy number alterations was performed in a study by Kim et al. in 2015 [44]. Next-generation sequencing was performed in tissue and matched blood samples obtained from 109 BC patients who underwent RC. Mutations in genes associated with chromatin remodeling were revealed in 83% of the cases, while genes regulating cell cycle were also identified with mutations in a percentage of 46%. In addition, TP53 mutation and mutations in the PI3K/AKT pathway were present in 57 and 35% of the patients, respectively. PI3K/AKT pathway included, among others, mutations in PIK3CA, PTEN, AKT1, and TSC1. Importantly, complete clinical and pathological follow-up data were available for 89 patients, enabling the prognostic assessment of the mutation status for high-grade BC tumors. PIK3CA mutation was significantly correlated with better recurrence-free survival (Hazard ratio (HR) = 0.39; $p = 0.032$) after adjusting for stage and lymph-node metastasis status. Moreover, mutations in cyclin-dependent kinase inhibitor 2A (CDKN2A) reported decreased recurrence-free and cancer-specific survival (HR = 5.76; $p < 0.001$ and HR = 2.94; $p = 0.029$, respectively).

In an effort to address the genotyping profile in relation to the BC prognosis, in another study including a total of 822 NMIBC patients [45], genotyping of biological material derived from plasma (or saliva whenever plasma was not available) was performed. In this study, screening of 1679 single-nucleotide polymorphisms (SNPs) was conducted for 251 corresponding genes. These genes were associated with inflammatory processes. In this study, the time until the first recurrence, as well as the time for progression, were investigated using the group of NMIBC patients with a median follow-up of 80.4 months. Based on the above assessment the combination of three SNPs (TNIP1, CD5, JAK3) in a multimarker model, as established based on Bayes A and Bayesian Lasso statistical algorithms, showed a better prognostic probability (posterior probability >90%) than any single SNP (assessed by Cox regression analysis). In parallel, CD3G SNP was significantly associated with disease progression, as indicated by an HR value of 2.69 (adjusted $p = 0.023$). Although the superiority of the multimarker modeling is proven in this study, additional validation is requested to prove the validity of the prognostic features, particularly because of the rather decreased number of the progression events in the investigated cohort [45].

The concept of cross-correlation of the available datasets to establish prognostic signatures was explored in the study by Riester et al. in 2012 [46], by employing a combination of microarray datasets. Originally, microarray analysis was performed in 93 frozen bladder tumors derived from patients guided for RC—with the clear majority having advanced BC ($n = 78$ MIBC, $n = 15$ NMIBC). The data were further correlated with 49 gene signatures, previously published in 6 studies, including a total of 578 patients. The combination of the prognostic gene signature led to the establishment of a 20-gene signature by using support vector machine algorithms. The prognostic signature was cross-validated in the previously published datasets and the prognostic performance was further compared for proving added value for prediction of the overall survival, over a previously validated nomogram consisting of clinical and pathologic variables, as published by the International Bladder Cancer Network (IBCN) consortium [46]. Among the rest, the 20-gene signature consisted of apolipoprotein B, transcription factors like ATF3, FADD, JUNB, kinases MAP3K1 and MAP2K1, profilin-1, platelet derived growth factor C, and chemokine CCL5. Upon validation in the publicly available datasets, it significantly improved the prediction of overall survival (mean $\Delta C = 0.14$, $p < 0.001$), but also when compared with the IBCN survival nomogram, it improved on its performance ($\Delta C = 0.08$, $p < 0.005$). Nevertheless, the results, although promising, need to be prospectively validated in an appropriate clinical study design [46].

In another study by Pignot et al. [47], the microRNA status was investigated in frozen tissue specimens from 166 BC patients (86 NMIBC and 80 MIBC), followed for a median time of 15 months [47]. The patient population was separated in a training ($n = 14$) and a validation set, consisting of BC patients ($n = 152$) and individuals contributed with noncancerous bladder tissue ($n = 11$) [47]. Out of 804 microRNAs that were screened, a three-microRNA signature (miR-9, miR-182, and miR-200b) was identified and further validated. Higher tissue expression of miR-9, miR-182, and miR-200b was significantly correlated with decreased recurrence-free survival ($p = 0.025$; $p = 0.021$; $p = 0.023$) and overall survival ($p = 0.0025$; $p = 0.024$; $p = 0.035$). Importantly, based on the three-microRNA signature expression levels, unsupervised hierarchical clustering enabled the classification of the BC patients into two groups [47]. Further assessment of the clustered BC patients by using the three-microRNA signature also revealed significant differences between the two clusters in terms of recurrence-free survival and overall survival ($p = 0.035$; $p = 0.015$) [47].

21.3.1.3 Predictive Molecular Profiles

Omics datasets have been extensively used in the investigation of mechanisms and/or signatures that are related to response to therapeutic schemes. A cross-correlation study was performed by Harryman et al. [48], using publicly available datasets from 91 epithelial cancer studies, representing a total of 12 different types of cancers, including a total of 5647 samples that provided DNA copy number alteration and/or mutation data [48]. Available data were analyzed using cBioPortal (assessment of the alteration frequency of a gene signature, survival analysis, etc.) and Oncomine software (analysis of drug resistance profiles). Based on the cross-correlation of the data, special emphasis was placed on genes related to the laminin–integrin axis and particularly a five-gene signature (ITGA3, ITGB4, LAMB3, PLEC, and SYNE3) was studied further. Kaplan–Meier survival analysis indicated a significant difference in survival between cases with and without copy number alterations in the gene signature for bladder ($p = 0.0143$) and cervical squamous cell carcinoma and endocervical adenocarcinoma ($p = 0.0432$). In parallel, genes included in the signature were queried in the Oncomine database to generate heat maps of gene expression for drug-resistant and drug-sensitive cells. Data were available for the four out of five genes (i.e., ITGA3, ITGB4, LAMB3, PLEC). Based on the heatmap plot analysis, a positive correlation was shown for the two histone deacetylase (HDAC) inhibitors, namely vorinostat and panobinostat, and topoisomerase II inhibitor Irinotecan [48].

21.3.1.4 Molecular Sub-Classification

Phenotypic diversity in bladder tumors, reflected by the presence of multiple subtypes, is generally delaying the translation of clinical trials to new standard treatments for patients. To address this point, substantial efforts are dedicated in improving the classification of BC patients based on molecular disease profiles rather than using the classical pathological assessment (e.g., Refs. [27–29, 35, 36, 49–51]).

Sjödahl et al. performed a gene expression profiling of tissue samples from 308 patients with BC (both NMIBC and MIBC). Extensive hierarchical clustering of the molecular tumor profiles indicated a presence of five distinct tumor subtypes (i) urobasal A, (ii) genomically unstable, (iii) urobasal B, (iv) squamous cell carcinoma (SCC)-like, and (v) highly infiltrated by nontumor cells [27]. These subtypes exhibited differences in both expression profiles and disease outcome. The most favorable prognosis was observed for urobasal A subtype whereas adverse prognosis was observed for urobasal B and the SCC-like subtypes. The main molecular differences were associated with the expression of cell-cycle genes, receptor tyrosine kinases (i.e., FGFR3, ERBB2, EGFR), cytokeratins, as well as cell adhesion genes [27]. In addition, differences in the FGFR3, PIK3CA, and TP53 mutation frequencies were observed. Specifically, significantly higher frequency of FGFR3 and PIK3CA mutations was observed for urobasal A subtype in comparison to the genomically unstable subtype, whereas significantly higher frequency of TP53 mutations was shown in the genomically unstable subtype, compared to the urobasal A subtype. In addition, there was no significant difference in the frequency of FGFR3 and PIK3CA between urobasal A and B tumors. Similarly, in the case of TP53, there was no significant difference in the mutation frequency between the genomically unstable and the urobasal B tumors. Importantly, it has been shown that the molecular subtypes are an intrinsic feature of the tumors, as the defined gene signatures show coordinated expression independently of tumor stage/grade [27].

Integrative analysis of 131 subjects with MIBC was conducted, as a part of The Cancer Genome Atlas initiative, aiming at the thorough characterization of disease-associated changes at the level of genome, transcriptome, and proteome [28]. Multiple mutations in genes related to regulation of cell-cycle, chromatin remodeling, and kinase signaling were found. Further analysis of mRNA, miRNA, and protein resulted in identification of distinct BC subtypes. Specifically, using RNA sequencing, a total of four expression subtypes were defined, including two subtypes (papillary-like and basal/squamous-like), which were also supported by the miRNA and protein data. Furthermore, this study allowed for the identification of numerous genomic alterations in the context of PI3-kinase/AKT/mTOR, CDKN2A/CDK4/CCND1, and RTK/RAS pathways, as well as ERBB2 (Her-2), ERBB3, and FGFR3, which are amenable in principle to therapeutic targeting [28].

In another study, Hedegaar et al. investigated the transcriptome of patients with NMIBC ($n = 460$; 345 Ta, 112 T1, 3 CIS), together with some patients from MIBC ($n = 16$) [29]. Total RNA sequencing analysis followed by unsupervised consensus clustering revealed the presence of three disease classes (classes 1, 2, and 3). Accordingly, the identified distinct molecular profiles differed regarding the clinical and histopathological characteristics as well as the progression-free survival. For the latter, Kaplan–Meier analysis was performed for assessing progression-free survival and indicated a better prognosis for tumors assigned into the classes 1 and 3 in comparison to class 2. Specifically, high-stage/-grade as well as high-risk tumors (EORTC system) fall into class 2 or 3. Interestingly, most of the MIBC cases were classified as class 2, indicating high similarity with high-risk NMIBC. Further analysis of differentially expressed genes between the defined classes indicated that tumors corresponding to class 1 were characterized by increased expression of early cell-cycle gene, while class 2 tumors exhibited elevated expression of late cell-cycle genes. Both classes exhibited high expression of uroplakins, while cytokeratins (KRT5, KRT15) were found mostly in class 3 tumors. In addition, class 3 tumors had a high expression of CD44 (stem cell and basal cell marker), while class 2 were enriched by ALDH1A1, ALDH1A2, PROM1, NES, and THY1 (cancer stem cell markers) [29]. In addition, enhanced expression of transcription factors involved in EMT activation (for class 2) and differentiation markers (for classes 1 and 2) were reported. Considering the molecular properties, it appears that class 1 and 2 tumors show luminal-like characteristics, although the level of aggressiveness varies, and class 3 tumors show basal-like characteristics. Following the above results on the molecular sub-classification, a 117-gene panel was established and further validated in four independent datasets for the molecular classification of NMIBC cancer [29]. The results in this study could demonstrate a potential upon the use of molecular features to stratify the patients with NMIBC into molecular sub-classes. This stratification of the patients could indicate those at higher risk for progression, thus better guide the current treatment schemes and adjust the monitoring frequency.

Similarly, in the study by Ross et al., the main focus was placed on the characterization of genomic profiles of advanced BC, aiming at identification of clinically

relevant genomic alterations (CRGAs–genetic alterations that are linked to available/under development drugs) [49]. DNA was extracted from formalin-fixed paraffin-embedded tissue sections from patients with recurrent/metastatic cancer ($n = 295$). For the vast majority of patients (>90%) at least one CRGA was found, and the most common target genes were CDKN2A, FGFR3, PIK3CA, and ERBB2. Different genetic alterations were observed, including fusions (for FGFR3), amplifications, and substitutions (for ERBB2) and others [49].

Hoadley et al. investigated similarities in molecular profiles across 12 cancer types (PanCancer-12 analysis), aiming to achieve an objective classification of cancer based on molecular features [50]. A total of 3527 specimens were evaluated using genomics, transcriptomics, and proteomics approaches. Classification was based on the following data types: DNA copy number, DNA methylation, mRNA expression, microRNA expression, protein expression, and somatic point mutation. Through the integrated classification, distinct subtypes were identified including 11 subtypes that have more than 10 samples assigned [50]. From all cancers analyzed, BC ($n = 120$) was among the most diverse tumor types; bladder tumors were clustered into 7 out of 11 major subtypes. However, most of the samples were mapped to three subtypes, that is, C1–LUAD (lung adenocarcinoma)-enriched ($n = 10$), C2-squamous-like ($n = 31$), and C8–BLCA (BC) ($n = 74$) [50]. However, further studies are required to validate these findings.

Following the same principle, Mak et al. aimed at molecular characterization of features associated with epithelial-to-mesenchymal transition (EMT) across 11 tumor types and identification of some therapeutic vulnerabilities [51]. To that end, global genomic and proteomic profiling of 1934 tumors led to the establishment of the EMT signature. The signature comprised of 77 genes, for which mRNA levels correlated with the levels of canonical EMT markers such as E-cadherin, vimentin, fibronectin, and N-cadherin. Functional analysis of genes included in EMT signature components indicated an association with cellular movement, growth, and proliferation and cell-to-cell signaling. Importantly, it has been noted that the expression of 20 potentially targetable immune checkpoint genes were correlated with the EMT scores for each cancer type. Considering the running clinical trials and the interest in cancer immunotherapy (e.g., PD-L1 inhibitors), information about the increased expression of immune checkpoint targets (PD1, PD-L1, CTLA4, OX40L, and PD-L2) in mesenchymal tumors appears to be of high clinical relevance. These findings support the use of EMT status as an additive to indicate which cancer patients might benefit from application of immune checkpoint inhibitors [51].

21.4 Outlook

Currently, extensive research is conducted in order to discover and validate BC biomarkers, for several clinical purposes (diagnosis and disease monitoring, prognosis, and patient stratification for prediction of treatment response to current therapies).

Starting from diagnosis, a great number of studies have been already reported, aiming at the development of noninvasive biomarkers with high sensitivity in BC. Until recently, in most of the studies, the results although significant, could not be efficiently introduced into the clinical settings. Successful implementation of single biomarkers is hampered by the increased intrinsic heterogeneity of BC as a result of distinct clonalities, and intratumor and interpatient variations. Thus, the focus has shifted to the evaluation and establishment of biomarker panels, consisting of multiple biomarkers that reflect more accurately the disease-associated changes and heterogeneity, in comparison to the single biomarkers. Toward this direction, novel diagnostic tests for timely and accurate disease detection and monitoring of recurrence are evaluated. The clinical concept relies on the principle that the application of accurate, noninvasive, and cost-effective biomarkers is expected to reduce the number of invasive and costly cystoscopies.

Increasing interest is also observed for the development of prognostic and predictive biomarkers to better stratify the patients and guide the treatment selection. Multidimensional molecular signatures based on *-omics* data (genomics, transcriptomics, proteomics, and metabolomics) may enable to obtain more accurate clinical information about the disease, despite the heterogeneity of BC.

Recent studies on BC allowed the classification of tumors based on molecular features. Thus, the subclonal evolution of BC and the increased heterogeneity of the disease were partially elucidated. It is expected that future studies will contribute to additional molecular characteristics of BC tumor types and the identified features will be of great value to the identification of novel drug targets. These developments open the way for targeted therapy of BC tumors.

Acknowledgments

This work was supported in part by a clinical scholarship from the European Urologic Scholarship Program (EUSP) and by Marie Curie EID program BCMolMed (PITN-GA-2012-317450).

References

1 Nielsen ME, Smith AB, Meyer AM, Kuo TM, Tyree S, et al. (2014) Trends in stage-specific incidence rates for urothelial carcinoma of the bladder in the United States: 1988 to 2006. *Cancer* 120: 86–95.

2 Sylvester RJ, van der Meijden AP, Oosterlinck W, Witjes JA, Bouffioux C, et al. (2006) Predicting recurrence and progression in individual patients with stage Ta T1 bladder cancer using EORTC risk tables: a combined analysis of 2596 patients from seven EORTC trials. *Eur Urol* 49: 466–465; discussion 475–467.

3 Babjuk M, Burger M, Zigeuner R, Shariat SF, van Rhijn BW, et al. (2013) EAU guidelines on non-muscle-invasive urothelial carcinoma of the bladder: update 2013. *Eur Urol* 64: 639–653.

4 Schrier BP, Hollander MP, van Rhijn BW, Kiemeney LA, Witjes JA (2004) Prognosis of muscle-invasive bladder cancer: difference between primary and progressive tumours and implications for therapy. *Eur Urol* 45: 292–296.

5 van den Bosch S, Alfred Witjes J (2011) Long-term cancer-specific survival in patients with high-risk, non-muscle-invasive bladder cancer and tumour progression: a systematic review. *Eur Urol* 60: 493–500.

6 Kim S, Ding W, Zhang L, Tian W, Chen S (2014) Clinical response to sunitinib as a multitargeted tyrosine-kinase inhibitor (TKI) in solid cancers: a review of clinical trials. *Onco Targets Ther* 7: 719–728.

7 Grivas PD, Daignault S, Tagawa ST, Nanus DM, Stadler WM, et al. (2014) Double-blind, randomized, phase 2 trial of maintenance sunitinib versus placebo after response to chemotherapy in patients with advanced urothelial carcinoma. *Cancer* 120: 692–701.

8 Pinto-Leite R, Arantes-Rodrigues R, Palmeira C, Colaco B, Lopes C, et al. (2013) Everolimus combined with cisplatin has a potential role in treatment of urothelial bladder cancer. *Biomed Pharmacother* 67: 116–121.

9 Galsky MD, Hahn NM, Powles T, Hellerstedt BA, Lerner SP, et al. (2013) Gemcitabine, Cisplatin, and sunitinib for metastatic urothelial carcinoma and as preoperative therapy for muscle-invasive bladder cancer. *Clin Genitourin Cancer* 11: 175–181.

10 Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, et al. (2014) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer* 136: 359–386.

11 Ferlay J, Shin HR, Bray F, Forman D, Mathers C, et al. (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer* 127: 2893–2917.

12 Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, et al. (2013) Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *Eur J Cancer* 49: 1374–1403.

13 GLOBOCAN 2012: estimated cancer incidence, mortality, and prevalence worldwide in 2012. International Agency for Research on Cancer. http://globocan.iarc.fr (accessed August 24, 2017).

14 Donsky H, Coyle S, Scosyrev E, Messing EM (2014) Sex differences in incidence and mortality of bladder and kidney cancers: national estimates from 49 countries. *Urol Oncol* 32: 40.e23–40.e31.

15 Burger M, Catto JW, Dalbagni G, Grossman HB, Herr H, et al. (2013) Epidemiology and risk factors of urothelial bladder cancer. *Eur Urol* 63: 234–241.

16 Booth CM, Siemens DR, Li G, Peng Y, Tannock IF, et al. (2014) Perioperative chemotherapy for muscle-invasive bladder cancer: a population-based outcomes study. *Cancer* 120: 1630–1638.

17 Williams PD, Cheon S, Havaleshko DM, Jeong H, Cheng F, et al. (2009) Concordant gene expression signatures predict clinical outcomes of cancer patients undergoing systemic therapy. *Cancer Res* 69: 8302–8309.

18 Smith SC, Baras AS, Lee JK, Theodorescu D (2010) The COXEN principle: translating signatures of in vitro chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer. *Cancer Res* 70: 1753–1758.

19 Hautmann RE, Gschwend JE, de Petriconi RC, Kron M, Volkmer BG (2006) Cystectomy for transitional cell carcinoma of the bladder: results of a surgery only series in the neobladder era. *J Urol* 176: 486–492; discussion 491–482.

20 Madersbacher S, Hochreiter W, Burkhard F, Thalmann GN, Danuser H, et al. (2003) Radical cystectomy for bladder cancer today—a homogeneous series without neoadjuvant therapy. *J Clin Oncol* 21: 690–696.

21 Stein JP, Lieskovsky G, Cote R, Groshen S, Feng AC, et al. (2001) Radical cystectomy in the treatment of invasive bladder cancer: long-term results in 1,054 patients. *J Clin Oncol* 19: 666–675.

22 Brauers A, Buettner R, Jakse G (2001) Second resection and prognosis of primary high risk superficial bladder cancer: is cystectomy often too early? *J Urol* 165: 808–810.

23 Herr HW (1999) The value of a second transurethral resection in evaluating patients with bladder tumors. *J Urol* 162: 74–76.

24 Czerniak B, Dinney C, McConkey D (2016) Origins of bladder cancer. *Annu Rev Pathol* 11: 149–174.

25 Knowles MA, Hurst CD (2015) Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nat Rev Cancer* 15: 25–41.

26 Netto GJ (2011) Molecular biomarkers in urothelial carcinoma of the bladder: are we there yet? *Nat Rev Urol* 9: 41–51.

27 Sjodahl G, Lauss M, Lovgren K, Chebil G, Gudjonsson S, et al. (2012) A molecular taxonomy for urothelial carcinoma. *Clin Cancer Res* 18: 3377–3386.

28 The Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* 507: 315–322.

29 Hedegaard J, Lamy P, Nordentoft I, Algaba F, Hoyer S, et al. (2016) Comprehensive transcriptional analysis of early-stage urothelial carcinoma. *Cancer Cell* 30: 27–42.

30 Mowatt G, Zhu S, Kilonzo M, Boachie C, Fraser C, et al. (2010) Systematic review of the clinical effectiveness and cost-effectiveness of photodynamic diagnosis and urine biomarkers (FISH, ImmunoCyt, NMP22) and cytology for the detection and follow-up of bladder cancer. *Health Technol Assess* 14: 1–331, iii–iv.

31 Yafi FA, Brimo F, Steinberg J, Aprikian AG, Tanguay S, et al. (2015) Prospective analysis of sensitivity and specificity of urinary cytology and other urinary biomarkers for bladder cancer. *Urol Oncol* 33: 66.e25–66.e31.

32 Leal J, Luengo-Fernandez R, Sullivan R, Witjes JA (2016) Economic burden of bladder cancer across the European Union. *Eur Urol* 69: 438–447.

33 Hedelin H, Holmang S, Wiman L (2002) The cost of bladder tumour treatment and follow-up. *Scand J Urol Nephrol* 36: 344–347.

34 Gui Y, Guo G, Huang Y, Hu X, Tang A, et al. (2011) Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat Genet* 43: 875–878.

35 Cazier JB, Rao SR, McLean CM, Walker AL, Wright BJ, et al. (2014) Whole-genome sequencing of bladder cancers reveals somatic CDKN1A mutations and clinicopathological associations with mutation burden. *Nat Commun* 5: 3756.

36 Guo G, Sun X, Chen C, Wu S, Huang P, et al. (2013) Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation. *Nat Genet* 45: 1459–1463.

37 Altelaar AF, Munoz J, Heck AJ (2013) Next-generation proteomics: towards an integrative view of proteome dynamics. *Nat Rev Genet* 14: 35–48.

38 Frantzi M, van Kessel KE, Zwarthoff EC, Marquez M, Rava M, et al. (2016) Development and validation of urine-based peptide biomarker panels for detecting bladder cancer in a multi-center study. *Clin Cancer Res* 22: 4077–4086.

39 Chen YT, Chen HW, Domanski D, Smith DS, Liang KH, et al. (2012) Multiplexed quantification of 63 proteins in human urine by multiple reaction monitoring-based mass spectrometry for discovery of potential bladder cancer biomarkers. *J Proteomics* 75: 3529–3545.

40 Ribal MJ, Mengual L, Lozano JJ, Ingelmo-Torres M, Palou J, et al. (2016) Gene expression test for the non-invasive diagnosis of bladder cancer: a prospective, blinded, international and multicenter validation study. *Eur J Cancer* 54: 131–138.

41 van der Heijden AG, Mengual L, Lozano JJ, Ingelmo-Torres M, Ribal MJ, et al. (2016) A five-gene expression signature to predict progression in T1G3 bladder cancer. *Eur J Cancer* 64: 127–136.

42 Zheng LF, Sun WY (2016) Meta-analysis of microRNAs as biomarkers for muscle-invasive bladder cancer. *Biomed Rep* 5: 159–164.

43 Sapre N, Macintyre G, Clarkson M, Naeem H, Cmero M, et al. (2016) A urinary microRNA signature can predict the presence of bladder urothelial carcinoma in patients undergoing surveillance. *Br J Cancer* 114: 454–462.

44 Kim PH, Cha EK, Sfakianos JP, Iyer G, Zabor EC, et al. (2015) Genomic predictors of survival in patients with high-grade urothelial carcinoma of the bladder. *Eur Urol* 67: 198–201.

45 Masson-Lecomte A, Lopez de Maturana E, Goddard ME, Picornell A, Rava M, et al. (2016) Inflammatory-related genetic variants in non-muscle-invasive bladder cancer prognosis: a multimarker Bayesian assessment. *Cancer Epidemiol Biomarkers Prev* 25: 1144–1150.

46 Riester M, Taylor JM, Feifer A, Koppie T, Rosenberg JE, et al. (2012) Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clin Cancer Res* 18: 1323–1333.

47 Pignot G, Cizeron-Clairac G, Vacher S, Susini A, Tozlu S, et al. (2013) microRNA expression profile in a large series of bladder tumors: identification of a 3-miRNA signature associated with aggressiveness of muscle-invasive bladder cancer. *Int J Cancer* 132: 2479–2491.

48 Harryman WL, Pond E, Singh P, Little AS, Eschbacher JM, et al. (2016) Laminin-binding integrin gene copy number alterations in distinct epithelial-type cancers. *Am J Transl Res* 8: 940–954.

49 Ross JS, Wang K, Khaira D, Ali SM, Fisher HA, et al. (2016) Comprehensive genomic profiling of 295 cases of clinically advanced urothelial carcinoma of the urinary bladder reveals a high frequency of clinically relevant genomic alterations. *Cancer* 122: 702–711.

50 Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, et al. (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158: 929–944.

51 Mak MP, Tong P, Diao L, Cardnell RJ, Gibbons DL, et al. (2016) A patient-derived, pan-cancer EMT signature identifies global molecular alterations and immune target enrichment following epithelial-to-mesenchymal transition. *Clin Cancer Res* 22: 609–620.

Index

Integration of Omics Approaches and Systems Biology for Clinical Applications, First Edition.
Edited by Antonia Vlahou, Harald Mischak, Jerome Zoidakis, and Fulvio Magni.

WILEY SERIES ON MASS SPECTROMETRY

John R. de Laeter • *Applications of Inorganic Mass Spectrometry*

Michael Kinter and Nicholas E. Sherman • *Protein Sequencing and Identification Using Tandem Mass Spectrometry*

Chhabil Dass • *Principles and Practice of Biological Mass Spectrometry*

Mike S. Lee • *LC/MS Applications in Drug Development*

Jerzy Silberring and Rolf Eckman • *Mass Spectrometry and Hyphenated Techniques in Neuropeptide Research*

J. Wayne Rabalais • *Principles and Applications of Ion Scattering Spectrometry: Surface Chemical and Structural Analysis*

Mahmoud Hamdan and Pier Giorgio Righetti • *Proteomics Today: Protein Assessment and Biomarkers Using Mass Spectrometry, 2D Electrophoresis, and Microarray Technology*

Igor A. Kaltashov and Stephen J. Eyles • *Mass Spectrometry in Structural Biology and Biophysics: Architecture, Dynamics, and Interaction of Biomolecules, Second Edition*

Isabella Dalle-Donne, Andrea Scaloni, and D. Allan Butterfield • *Redox Proteomics: From Protein Modifications to Cellular Dysfunction and Diseases*

Silas G. Villas-Boas, Ute Roessner, Michael A.E. Hansen, Jorn Smedsgaard, and Jens Nielsen • *Metabolome Analysis: An Introduction*

Mahmoud H. Hamdan • *Cancer Biomarkers: Analytical Techniques for Discovery*

Chabbil Dass • *Fundamentals of Contemporary Mass Spectrometry*

Kevin M. Downard (Editor) • *Mass Spectrometry of Protein Interactions*

Nobuhiro Takahashi and Toshiaki Isobe • *Proteomic Biology Using LC-MS: Large Scale Analysis of Cellular Dynamics and Function*

Agnieszka Kraj and Jerzy Silberring (Editors) • *Proteomics: Introduction to Methods and Applications*

Ganesh Kumar Agrawal and Randeep Rakwal (Editors) • *Plant Proteomics: Technologies, Strategies, and Applications*

Rolf Ekman, Jerzy Silberring, Ann M. Westman-Brinkmalm, and Agnieszka Kraj (Editors) • *Mass Spectrometry: Instrumentation, Interpretation, and Applications*

Christoph A. Schalley and Andreas Springer • *Mass Spectrometry and Gas-Phase Chemistry of Non-Covalent Complexes*

Riccardo Flamini and Pietro Traldi • *Mass Spectrometry in Grape and Wine Chemistry*

Mario Thevis • *Mass Spectrometry in Sports Drug Testing: Characterization of Prohibited Substances and Doping Control Analytical Assays*

Sara Castiglioni, Ettore Zuccato, and Roberto Fanelli • *Illicit Drugs in the Environment: Occurrence, Analysis, and Fate Using Mass Spectrometry*

Ángel Garciá and Yotis A. Senis (Editors) • *Platelet Proteomics: Principles, Analysis, and Applications*

Luigi Mondello • *Comprehensive Chromatography in Combination with Mass Spectrometry*

Jian Wang, James MacNeil, and Jack F. Kay • *Chemical Analysis of Antibiotic Residues in Food*

Walter A. Korfmacher (Editor) • *Mass Spectrometry for Drug Discovery and Drug Development*

Alejandro Cifuentes (Editor) • *Foodomics: Advanced Mass Spectrometry in Modern Food Science and Nutrition*

Christine M. Mahoney (Editor) • *Cluster Secondary Ion Mass Spectrometry: Principles and Applications*

Despina Tsipi, Helen Botitsi, and Anastasios Economou • *Mass Spectrometry for the Analysis of Pesticide Residues and their Metabolites*

Xianlin Han • *Lipidomics: Comprehensive Mass Spectrometry of Lipids*

Jack F. Kay, James D. MacNeil, and Jian Wang (Editors) • *Chemical Analysis of Non-Antibiotic Veterinary Residues in Food*

John R. Griffiths and Richard D. Unwin • *Analysis of Protein Post-Translational Modifications by Mass Spectrometry*

Henk Schierbeek (Editor) • *Mass Spectrometry and Stable Isotopes in Nutritional and Pediatric Research*

Joaquim Ros (Editor) • *Protein Carbonylation: Principles, Analysis, and Biological Implications*