Methods in Molecular Biology 1381

Springer Protocols

Robert Grützmann
Christian Pilarsky *Editors*

Cancer Gene Profiling

Methods and Protocols

Second Edition

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

For further volumes:
http://www.springer.com/series/7651

Cancer Gene Profiling

Methods and Protocols

Second Edition

Edited by

Robert Grützmann

Department of Surgery, Universitätsklinikum Erlangen, Erlangen, Germany

Christian Pilarsky

Department of Surgery, TU-Dresden, Dresden, Germany

Humana Press

Editors
Robert Grützmann
Department of Surgery
Universitätsklinikum Erlangen
Erlangen, Germany

Christian Pilarsky
Department of Surgery
TU-Dresden
Dresden, Germany

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-3203-0 ISBN 978-1-4939-3204-7 (eBook)
DOI 10.1007/978-1-4939-3204-7

Library of Congress Control Number: 2015958099

Springer New York Heidelberg Dordrecht London

Printed on acid-free paper

Humana Press is a brand of Springer
Springer Science+Business Media LLC New York is part of Springer Science+Business Media (www.springer.com)

Preface

Science is the facilitator between imagination and reality
—(Anonymous)

Since the last edition of this book, the pace of technology has accelerated. The major cancer genome projects have been finished and shown that next-generation sequencing is the winning technology for high-throughput expression profiling.

In this book we have brought together the experiences of leading scientists in the discipline of cancer gene profiling. We have included different techniques, since cancer genes can be profiled in different ways. Such different approaches are needed to understand the key stages of cancer development, as using only one technique would be insufficient. Therefore this book attempts to give an overview of the state-of-the-art methods, which will enable the reader to perform these experiments successfully. This book has been written for any student or practitioner with an interest in cancer gene profiling and can be used in any well-equipped research laboratory. It may also serve as a demonstration of the kind of analysis that is possible today and will be complementary to other textbooks in the area of biomedical research.

We would like to offer our gratitude to all the contributing authors and the staff of Humana Press—without their help this book would not have been possible. We would also like to thank our families for their love and patience.

Science is not just a profession—it should also be fun. This fun comes from the inception of an idea, which goes on to be proven through experimentation, or, as we found in a Chinese fortune cookie: "The impossible is only the untried."

We hope that you will not only be successful but also have fun using our book in your research.

Erlangen, Germany ***Robert Grützmann***
Dresden, Germany ***Christian Pilarsky***

Contents

Contributors

Ole Ammerpohl • *Institute of Human Genetics, University of Kiel and University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany*

Vladan Arsenijevic • *Department of Bioinformatics, Seven Bridges Genomics, Cambridge, MA, USA*

Daniel S. Atherton • *Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA; Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA*

Walter C. Bell • *Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA*

Katie Bullock • *NIHR Pancreatic Biomedical Research Unit, Molecular and Clinical Cancer Medicine, Royal Liverpool University Hospital, Liverpool, UK*

Hsueh-Ting Chu • *Department of Computer Science and Information Engineering, Asia University, Taichung, Taiwan; Department of Medical Research, China Medical University Hospital, China Medical University, Taichung, Taiwan*

Pedro R. Cutillas • *Center for Haemato-Oncology, Bart Cancer Institute, Queen Mary University of London, London, UK*

Ana-Maria Dancau • *Institute of Pathology, University Medical Hamburg-Eppendorf, Hamburg, Germany*

Brandi N. Davis-Dusenbery • *Department of Bioinformatics, Seven Bridges Genomics, Cambridge, MA, USA*

Lu Gao • *Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA*

B. Michael Ghadimi • *Department of General, Visceral, and Pediatric Surgery, University Medical Center Göttingen, Georg-August-University, Göttingen, Germany*

William Greenhalf • *NIHR Pancreatic Biomedical Research Unit, Molecular and Clinical Cancer Medicine, Royal Liverpool University Hospital, Liverpool, UK*

William E. Grizzle • *Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA; Department of Pathology, University of Alabama at Birmingham, Birmingham, Al, USA*

Andrea Haake • *Institute of Human Genetics, University of Kiel and University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany*

Feng Jiang • *Department of Pathology, University of Maryland School of Medicine, Baltimore, MD, USA*

Peter Jo • *Department of General, Visceral, and Pediatric Surgery, University Medical Center Göttingen, Georg-August-University, Göttingen, Germany*

Tatjana Crnogorac-Jurcevic • *Molecular Oncology Center, Bart Cancer Institute, Queen Mary University of London, London, UK*

Holger Kalthoff • *Institute for Experimental Cancer Research, Section for Molecular Oncology, University Hospital Schleswig-Holstein, Campus Kiel, Kile, Germany*

Julia Kolarova • *Institute of Human Genetics, University of Kiel and University Hospital Schleswig-Holstein, Campus Kiel, Kiel, Germany*

Glen Kristiansen • *Universitätsklinikum Bonn (AöR), Institut für Pathologie, Bonn, Germany*
Martina Mirlacher • *Institute of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
Lahiri Kanth Nanduri • *Department of Surgery, TU-Dresden, Dresden, Germany*
Dennis Otali • *Department of Pathology, University of Alabama at Birmingham, Birmingham, AL, USA*
Christian Pilarsky • *Department of Surgery, TU-Dresden, Dresden, Germany*
Anja Rabien • *Research Division, Department of Urology, Charité—Universitätsmedizin Berlin, Berlin, Germany*
Janine Roy • *Biotechnology Center, Technische Universität Dresden, Dresden, Germany*
Guido Sauter • *Institute of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
Michael Schroeder • *Biotechnology Center, Technische Universität Dresden, Dresden, Germany*
Katherine C. Sexton • *Department of Pathology, University of Alabama at Birmingham, Birmingham, Al, USA; Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA*
Victoria Shaw • *NIHR Pancreatic Biomedical Research Unit, Molecular and Clinical Cancer Medicine, Royal Liverpool University Hospital, Liverpool, UK*
Reiner Siebert • *Institute of Human Genetics, University of Kiel and University Hospital Schleswig-Holstein, Campus Keil, Kiel, Germany*
Ronald Simon • *Institute of Pathology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany*
Sanjay Tiwari • *Department of Radiology and Neuroradiology, Section Molecular Imaging, Christian-Albrechts-Universität zu Kiel, Kiel, Germany*
Christof Winter • *Faculty of Medicine, Department of Clinical Sciences, Oncology MV, University of Lund, Lund, Sweden*
Margot Zöller • *Tumor Cell Biology, University Hospital of Surgery, Heidelberg, Germany*

Chapter 1

Factors Affecting the Use of Human Tissues in Biomedical Research: Implications in the Design and Operation of a Biorepository

Daniel S. Atherton, Katherine C. Sexton, Dennis Otali, Walter C. Bell, and William E. Grizzle

Abstract

The availability of high-quality human tissues is necessary to advance medical research. Although there are inherent and induced limitations on the use of human tissues in research, biorepositories play critical roles in minimizing the effects of such limitations. Specifically, the optimal utilization of tissues in research requires tissues to be diagnosed accurately, and the actual specimens provided to investigators must be carefully described (i.e., there must be quality control of each aliquot of the tissue provided for research, including a description of any damage to tissues). Tissues also should be collected, processed, stored, and distributed (i.e., handled) uniformly under a rigorous quality management system (QMS). Frequently, tissues are distributed to investigators by tissue banks which have collected, processed, and stored them by standard operating procedures (SOPs). Alternatively, tissues for research may be handled via SOPs that are modified to the specific requirements of investigators (i.e., using a prospective biorepository model). The primary goal of any type of biorepository should be to ensure its specimens are of high quality and are utilized appropriately in research; however, approaches may vary based on the tissues available and requested. For example, extraction of specific molecules (e.g., microRNA) to study molecular characteristics of a tissue may require less clinical annotation than tissues that are utilized to identify how the molecular expression might be used to clarify a clinical outcome of a disease or the response to a specific therapy. This review focuses on the limitations of the use of tissues in research and how the design and operations of a tissue biorepository can minimize some of these limitations.

Key words Research, Human tissue, Biorepository, Prospective tissue collection, Tissue banking, IRB, HIPAA, Limitations

1 Introduction

Medical, as well as veterinary, research requires the appropriate use of high-quality tissues, including bodily fluids [1–6]. Biorepositories which provide tissue may focus on approaches chosen to meet specific research requirements [1–3]. Many factors affect the use of human tissues in research, and some of these factors may limit the

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_1, © Springer Science+Business Media New York 2016

use of specific tissues in some types of research. Specific variables affecting human tissues may be the cause, in part, of problems with poor reproducibility of some research based on human tissue. Biorepositories must be aware of biorepository sciences and use this knowledge to aid investigators in the selection of optimal tissues for their research. Of special importance is to minimize uses of specific human tissues that may add bias to research. This review briefly describes how different models of biorepositories address some of the major issues affecting the successful use of human tissues in biomedical research.

2 Bias Affecting the Use of Human Tissues in Research

Bias affects the evaluation of experimental results when there are confounding variables that cause changes in the experimental data that are attributed incorrectly to the experimental questions being evaluated [1–3, 7–10]. There are multiple potential causes of bias in research involving human tissues [7–10], and without careful records, incorrect conclusions based upon biased experimental data could be reported. Consider the comparison of molecular markers in the serum from patients with a specific disease (cases) compared with serum from controls without the disease. If the cases were evaluated based on samples of serum that were collected in the operating room (fasting sample), but the "matched" serum from controls were collected locally using a mobile van (non-fasting individuals), sensitive methods such as mass spectrometry might identify molecular differences in serum based on different collection and storage approaches; thus, investigators could conclude incorrectly that there were molecular changes in serum because of the disease. If there were no consideration of the potential for bias in the sample sets, incorrect conclusions that are secondary to bias might be published [7–10]. A review of experimental records and approaches might have identified that the differences in the specimens might be secondary to bias, and the conclusions of the study could have been verified on a subset of cases and controls handled more uniformly. The identification of biased published conclusions might not occur until efforts were made to confirm the original conclusions of the study. Reports of incorrect conclusions based on biased results emphasize the need for using standard operating procedures (SOPs) so that tissues are handled as consistently as practicable. Nevertheless, even using the same SOPs, there are likely to be inter-site differences in a study. There are multiple factors that affect tissue samples and potentially can cause bias; some of these are included in Table 1.

Table 1
Examples of potential sources of bias in tissue sets

1.	Population (e.g., racial mixture)
2.	Fed or fasting state of patients and controls
3.	Diurnal variations (i.e., time of collection)
4.	Stress
5.	Collection container (red top vs. separator)
6.	Time to processing
7.	Time to freezing (cold ischemia time)
8.	Temperature and length of storage
9.	Freeze–thaw cycles
10.	Different sites of sample collection

3 Models of Tissue Collection

Some approaches that are utilized to provide human tissues for biomedical research may be disorganized so that a designation of "catch as catch can" identifies how surgeons, pathologists, veterinarians, or other personnel collect, process, store, and provide tissues for research. Usually such tissues have not been provided under a quality management system (QMS) based on using SOPs, audits of operations, and quality control (QC) of the specimens; thus, these tissues may be inadequate for research because their diagnostic description may be wrong and they may have been handled inconsistently. In addition, the specimens could have been collected without approval of an Institutional Review Board (IRB) and their associated annotation may not meet the requirements of the Health Insurance Portability and Accountability Act (HIPAA).

3.1 Banking Model

One approach to supporting biomedical research with human specimens is to rely on an organized "*banking model*" in which SOPs are utilized for collecting, processing, storing, and distributing (i.e., handling) of specimens of human tissues. The types of tissues provided may vary depending upon the goals of the biobank. For example, only aliquots of paraffin blocks (e.g., paraffin slides) of one or more types of cancer may be available. Of note, some preparations of tissues including unfrozen fresh samples are frequently unavailable from a biobank. The advantages of a "banking model" are that numerous specimens may be available, can be provided when requested, and demographic and clinical data as well as information on clinical outcomes can be provided rapidly upon request. These advantages are based upon the storage of

specimens awaiting distribution and the aging of the specimens so that more outcomes have had an opportunity to occur.

An important disadvantage of a banking model is that the available tissue may not meet the exact needs of an investigator in that specimens are uniform and may be too small or may be processed or stored by approaches which exclude their use in a specific type of research. Also, some molecular and other features may degrade upon storage. Perhaps the greatest failure of the biobanking model is that a large proportion of specimens will never be used in research and, hence, the model is not cost-effective and may be ethically problematic (i.e., tissues from consented patients are not used as anticipated by the patients) [2, 3, 11].

3.2 Clinical Trial Model

The *clinical trial model* is a subtype of banking model. In this model, remaining tissues that were obtained to support assays associated with a clinical trial are stored for undefined future research. An advantage of this model is that tissue specimens frequently have extensive information, beyond that usually available on most patients, due to the participation of the patients in a clinical trial and the collection of information associated with current and prior therapies. The clinical trial model is a cost-effective approach for *beginning* a biobank, but it has the same advantages and disadvantages of the biobank model. An additional potential problem is that the informed consent may not indicate that tissues can be utilized for other types of research, so an IRB may not permit their use in all types of research.

3.3 Population-Based Models

Population-based or *epidemiology models* rely on collection, processing, and banking of tissues, typically collected for a specific goal and from a cohort of patients with defined characteristics and/or subpopulations. For example, a cohort may be representative of a specific population such as the UK Biobank [12] for which primarily biofluids have been collected from 500,000 individuals in the United Kingdom. One of the goals of the UK Biobank is to have data and some samples available from normal individuals before they develop disease processes such as Alzheimer's disease. Alternatively, a cohort may consist of patients with specific disease risks such as obesity, and specific epidemiological questions may be posed based on this risk. Most population and/or epidemiology banks are based on collecting biological fluids either at a single time point or longitudinally on the same patients such as by the Prostate, Lung, Colorectal and Ovarian (PLCO) study [13].

Some advantages of this model are the questions that can be proposed based on the samples' characteristics. For example, samples of the National Health and Nutrition Examination Survey have been used to evaluate environmental exposures such as decreases in lead levels in blood in the U.S. population following removal of lead from gasoline [14, 15].

3.4 Prospective Model

The *prospective model* is based upon focusing the collection of tissues on requests of specific investigators. In a model such as the Cooperative Human Tissue Network (CHTN), investigators indicate specifically the tissues they want, the parameters of the specimens (e.g., size and source) and the methods of processing and storage [16–18]. For example, an investigator may request tissues from ten triple-negative breast cancers from African American females. The investigator may specify that each specimen must be at least 0.15 g and frozen in optimal cutting temperature compound (e.g., OCT).

Advantages of the prospective collection model include that specimens are specifically what is requested and needed by investigators including fresh, non-frozen specimens; that there should not be artifacts or biases caused by long-term storage; and that most specimens collected are utilized in research so the model is very cost-effective. The obvious disadvantage of the prospective model is that specimens must be collected, so they are not immediately available, and some requests may take weeks to months to be fulfilled completely. Also, because the specimens are collected prospectively, data on clinical outcomes are not available when the specimens are provided, and the clinical outcomes may require years after collection to develop. In addition, rare diseases usually are difficult to provide.

3.5 Combination Biorepository Model

The *combined banking and prospective collection* model incorporates aspects of both the prospective collection and banking models and includes the major advantages of these two models. The operation of a combination biorepository is complicated, and there are numerous challenges including the requirement for a complex informatics system and an awareness of the limitations of the long-term storage of tissues. There is typically a strong focus on distribution of specimens to investigators, even from the banking component.

4 Factors Affecting Human Tissues and Limiting Their Usefulness in Research

There are multiple variables that may affect human tissues and limit their usefulness in specific research. In addition, differences in tissues may act as confounding variables and result in incorrect conclusions caused by biased data. Some of these factors are a characteristic of the individuals from whom tissues are obtained, while preoperative variability in tissues may be induced by past and current diagnostic and therapeutic approaches (e.g., screening approaches, neoadjuvant therapy). Other differences develop in tissues during and after their removal from patients, especially during handling of the tissues. Such features are frequently unknown to researchers; however, it is very important that both investigators

and biorepository personnel understand the differences in the morphological and molecular features of tissues and how these variables can be a source of bias in research [2, 3, 7].

As methods of screening for cancer have improved, tissues involved by cancer are typically smaller and are of lower stage. For example, many early breast cancers are now identified by mammography. The frequent use of this imaging technique has reduced the size and hence the stage of most breast cancers. More and more surgically treated neoplastic lesions of the breast are in situ disease, ductal carcinoma in situ (DCIS) or lobular carcinoma in situ (LCIS), or invasive lesions of less than 2 cm in diameter. This severely limits the amount of breast cancer available to support research in that all carcinoma in situ, as well as some small tumors, are completely processed for pathologic diagnosis including clinical studies to obtain therapeutic, predictive, and prognostic information; thus, other than fixed, paraffin-embedded diagnostic blocks, no tissues from such specimens may be available to support research [2, 3]. Similarly, screening with prostate specific antigen (PSA) has reduced the sizes of prostate cancers. Also, newer imaging approaches together with the use of fine-needle aspirates have almost eliminated the availability of tissue samples from some types of metastatic diseases, especially bone metastases of breast, prostate, and lung. Also, the availability of pristine tumor tissues is limited by neoadjuvant and intraoperative changes [2, 3].

4.1 Preoperative Population and Individual Differences in Tissues

An individual is separated from others by basic differences of age, race, ethnicity, and sex, all of which may affect their tissues; thus, there are preoperative differences in tissues that vary among all patients. For example, there may be large differences in risk factors, e.g., obesity and smoking, as well as different co-morbid conditions such as adult onset (Type 2) diabetes mellitus. In addition, family histories, stress levels, environmental exposures, diets, as well as many other variables, may all contribute to different effects on normal and diseased tissues. In general, the extent and characteristics of most of these effects on human tissues are unknown and/or uncharacterized. Also, at this time, many tumors have some type of therapy prior to surgical removal, i.e., neoadjuvant therapy. Specifically, patients with cancers of the rectum and esophagus typically are treated using radiation before surgical removal of tissues, and high risk cancers of the breast, pancreas, and other tumors may have neoadjuvant chemotherapy and/or radiation. Because selective types of malignant cells may be killed by preoperative therapy, the resulting tumor may not be characteristic of the original tumor; also, the specific changes induced by neoadjuvant therapy are unknown and uncharacterized. Similarly, systemic neoadjuvant therapy may destroy metastatic lesions and reduce the size of primary lesions so that accurate staging of the treated cancer is impossible, except for stage IV disease. Identification of patients

treated prior to surgery is important so that investigators will know that the tissues received from such patients have limitations on their usefulness for specific types of research [2, 3].

4.2 Intraoperative Changes

When the vasculature of tissues is compromised during surgery, enzymes which are more active at body temperature begin to degrade specific molecules, and more labile and unstable molecules may be undetectable in tissues immediately after surgery. In contrast, mRNAs of most genes that respond to stress and/or ischemia increase. Therefore, there are multiple molecular changes that occur during surgery, a period that is designated as intraoperative ischemia or warm ischemia. Of note, needle and other small biopsies and the collection of bodily fluids usually are subject to less molecular changes in that vascular damage at body temperature does not develop over the very brief times before these tissues are removed from the body.

Other intraoperative changes may affect tissues. For example, some surgery is performed with a cautery or laser to cut tissue by heat, and the heat may damage adjacent areas of tissue, and sometimes the damage is severe enough to be identified morphologically and molecularly. Examples of damage to tissues probably secondary to cautery damage are demonstrated in Fig. 1. In this figure, the phenotypic expression of fatty acid synthase (FASN) can be demonstrated in undamaged, but not in damaged prostate cancer.

Use of robotic surgery in some procedures, e.g., radical prostatectomies, has increased operative times and, hence, warm ischemia. It is important for investigators to be aware of such changes in surgical approaches; specifically, data obtained from current cases of prostate cancer may not be consistent with prior data because of the extended warm ischemia. In contrast, other tissues may be removed more rapidly. If available, operative times should be recorded to aid investigators in evaluating tissues; of note, bias may be introduced in studies if older plus more recent samples of prostate cancer are grouped together, so the influence of specimen age on experimental data should be evaluated.

4.3 Effects of Cold Ischemia

After tissues are removed from the body, the period of "in vitro ischemia" or "cold ischemia" begins, which continues until the tissues are stabilized (e.g., frozen, fixed in 10 % neutral buffered formalin [10 % NBF]). Cold ischemia may be accompanied by cooling the tissues by biorepositories to 2–5 °C, but in the clinical setting the tissues usually remain at room temperature—both of which are less than body temperature and hence, are consistent with the terminology of cold ischemia. In contrast, some in vitro studies of specimens maintained at room temperature before analysis refer to this in vitro time as being warm ischemia, so care should be taken in reviewing the literatures based on these different terminologies.

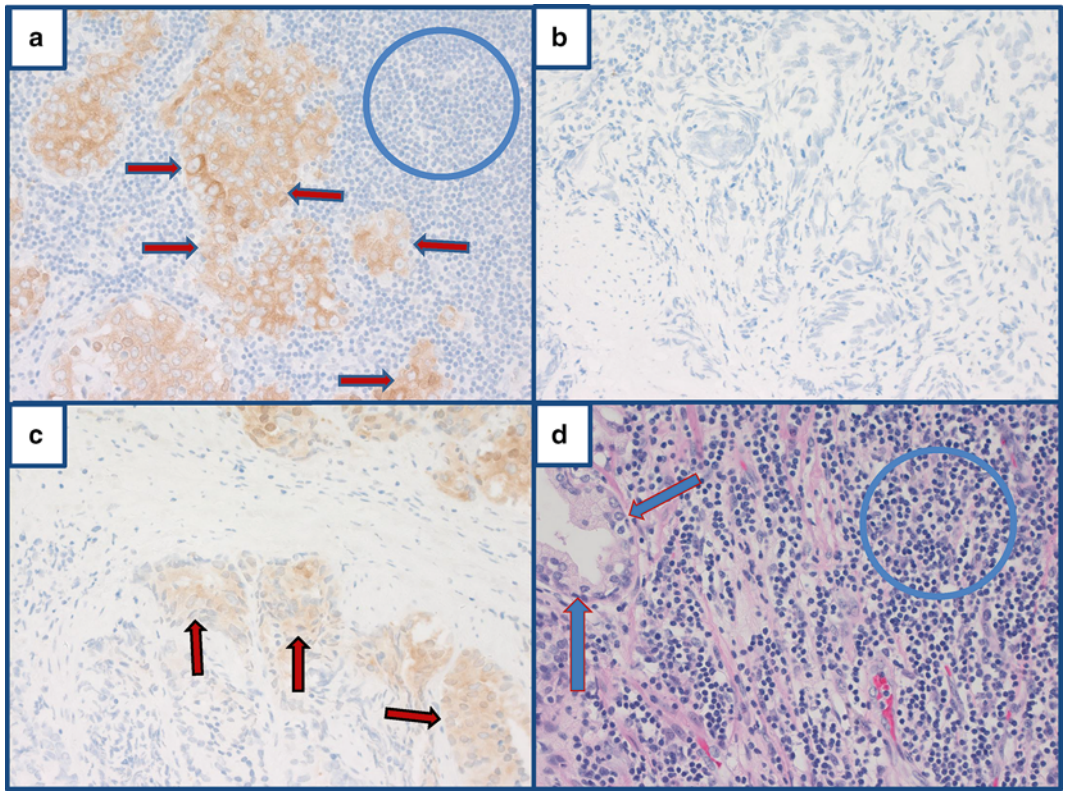


Fig. 1 Panels (**a**–**c**) (all original magnification ×200) demonstrate staining with fatty acid synthase (FASN) of an area of a lymph node with metastatic prostate cancer (*red arrows*, *blue outline*). Normal lymphocytes are within the *blue circle*. Panel (**b**) demonstrates severe tissue damage to a different area of the same lymph node in panel (**a**); this damage is most likely caused by a cautery burn and there is no staining of the apparently damaged tumor for FASN. Panel (**c**) Another area of the same lymph node with somewhat less damage demonstrates the partial loss of staining for FASN as damage to metastatic prostate cancer (*red arrows*, *black outline*) becomes more severe. Panel (**d**) (original magnification ×200, hematoxylin and eosin stain) demonstrates an uninvolved gland of the prostate (*blue arrow* and *red outline*) from a different case with prostate cancer. This area of the tissue contains an extensive area of stroma which contains numerous normal lymphocytes (some within *blue circle*)

The period of cold ischemia usually is less than 24 h. Because many less stable molecules are degraded and hence lost during warm ischemia, other molecules have begun to degrade, and stress/ischemic responsive molecules have already increased, the limited studies of molecular changes during cold ischemia indicate that this period is less damaging to molecular characteristics of tissues than warm ischemia [2, 3]. Also, like warm ischemia, the directions, extent, and time courses of changes in specific molecules usually are unknown and cannot be reliably predicted.

4.3.1 Molecular Changes During Cold Ischemia

The extent of molecular changes in tissues during cold ischemia likely varies with the category of molecule (e.g., mRNA), but also

changes are likely to vary with the specific molecules within each category or subtype. Changes in the mRNAs for molecules whose expressions are increased by ischemia (e.g., HIF1α) are likely to be more extensive and in different directions than changes in molecules that are not responsive to stress or ischemia, e.g., smooth muscle actin.

There have been multiple studies that have focused on the molecular and morphological changes that occur prior to stabilization of tissue specimens including changes in RNA and proteins.

Effects of Cold Ischemia on RNA

The quality of total RNA frequently is measured by electrophoresis based on the ratio and intensities of the 28S and 18S bands of ribosomal RNA (rRNA). The quality and intensity of these bands can be visually described, the 28S/18S ratio can be measured, or more quantitatively can be measured as the RNA integrity number (RIN) using the Agilent 2100 system [19]. RIN values range from 0 to 10 with 10 being the measure of intact rRNA with no degradation and 0 being completely degraded rRNA. Of note, because RIN is a measure of the stability of rRNA, it is only a pseudo-measure of the quality of mRNA and other types of RNA. Nevertheless, RIN is usually considered to be the primary measure that is thought to be correlated with the overall quality of RNA. Studies of the effects of cold ischemia on RNA have used 28S/18S observations or RIN as well as measures of the expression of specific genes using either real-time, reverse transcriptase quantitative polymerase chain reaction (RT-Q-PCR) assays, or other measures of expression such as gene chip assays of expression of transcripts.

RIN or Observation of 28S and 18S Bands as Measures of Cold Ischemia

There have been multiple studies that have evaluated changes in RIN in specimens left at "room temperature" (RT), usually considered to be 25 °C, or maintained at the temperature of wet ice (about 4 °C) for up to 24 h. These studies have almost uniformly concluded that there is no change or very small changes (<10 %) in RIN when specimens are maintained at RT or on ice [20–23]. Most studies have reported that RIN decreases slightly over 5–6 h after tissues are removed from the body, even if specimens are maintained at RT. For example, in 17 breast cancers with average RINs of 7.39 that were maintained at room temperature after operative removal, Hatzis et al. [20] reported that, on the average, RIN decreased 0.12 units/h or only about 2 %/h; frequently, in other studies of changes of RIN after tissue removal from a patient, the decreases in RIN usually are not statistically significant [20–25].

The examples usually cited as to demonstrating important effects on RIN of time following removal of tissue from a patient are Bray et al. [23], Optiz et al. [24], and Strand et al. [26]. Optiz et al. [24] saw changes in RIN (about 15 % at 100 min) only when specimens were thermally degraded by maintaining them at 60 °C;

however, they reported that they saw no or minimal changes in RIN over time when tissues were maintained either at RT or at 45 °C. The results of changes in RIN for up to 200 min compared to specimens frozen rapidly on removal from the patient reported by Bray et al. [23] were stated not to be statistically significant. Also, the study of Strand et al. [26] did not attempt to evaluate changes in RIN over time, but rather to evaluate methods to measure the quality of RNA with RIN being one of the best methods. In addition, most studies that have evaluated the effects of cold ischemia using 28S and 18S peaks have demonstrated only minor changes [27, 28].

Even though RIN or 28S and 18S peaks have not been demonstrated to change very much after specimens are removed from the body and maintained at RT or colder, these observations do not indicate that, during cold ischemia, levels of specific mRNAs do not change in expression and/or via degradation, because RIN, as indicated, is only a pseudo-measure of overall RNA quality.

Changes in Expression of mRNA During Cold Ischemia

The studies of changes in expression of mRNAs during cold ischemia typically have used gene chip technology in which changes in approximately 17,000–25,000 transcripts are evaluated at times of cold ischemia from 15 min to 24 h. These studies are frequently confusing because they sometimes use statistical criteria (>factor 2 change $p<0.05$), but sometimes describe any changes in transcripts independent of the extent of the changes or their statistical evaluation. Some investigators also refer to changes in transcripts as changes in genes, even though the transcripts have not been associated with known genes. In addition, these studies sometimes combine both increases and decreases in transcripts at multiple times after removal of tissues from patients; nevertheless, in the first 6 h after tissues are removed from patients, there are changes in very small proportions of transcripts, usually less than 2 % [21–25, 29, 30]. Studies of longer times of up to 24 h have seen changes increase from 1 % at 6 h to 4 % at 24 h [21] and 2 % at 1 h to <10 % at 6 h [30]. However, studies of changes in the first 30 min of removal of tissues from a patient have reported larger numbers of transcripts that change (e.g., 20 %) with 15 % of these transcripts increased during the initial 15 min [31]. Of note, most of these studies have been performed with the samples maintained at RT rather than cooled to wet ice temperatures. In general, studies of multiple specimens report much higher variability of transcripts among tissues from different patients than secondary to the effects of times of cold ischemia on the expression of transcripts.

Based on changes in RIN and mRNA, there is no clearly demonstrated need to stabilize specimens rapidly before 2 h after removal of tissues from the body; nevertheless, tissues should be stabilized as rapidly as practicable. Of note, tissues that are stabilized within 6 h after removal should be adequate for most

studies in translational research [32]; however, in studies of single genes, care should be taken to determine the likelihood of biased data.

Changes in Proteins During Cold Ischemia

There are fewer studies that have evaluated changes in proteins after tissues are removed from the body, except for those using immunohistochemistry and in situ hybridization to study clinical specimens. Using surface-enhanced laser desorption/ionization time of flight mass spectrometry (SELDI-TOF-MS), Spruessel et al. [32] reported that 30 % of peaks significantly changed >2-fold within 30 min of removal of the tissue from the body with most changes occurring within 15 min.

One of the most extensive studies of the effects of cold ischemia on the proteome is that of Gündisch et al. [33] who evaluated 11 non-malignant specimens of the intestine by reverse phase protein array (RPPA) with 23 antibodies including 7 to phosphoproteins. By RPPA, only phosphor-p44/p42 MAPK increased (none decreased) statistically at 30 and 60 min, but it was essentially unchanged (0.98 vs. 1.00) at 2 h. Similarly, 30 proteins were evaluated by RPPA in 17 liver specimens and of these, only 4 proteins increased significantly, 3/30 at 30 min, 1/30 at 60 min, 4/30 at 180 min, and 4/30 at 360 min. Comparing the four statistically significant proteins that changed, all were increased compared to the rapidly frozen specimens. Comparing 30 min with 360 min, 3 of the 4 proteins that were significantly increased were the same, and the changes at 180 and 360 min were similar, but greater than the changes at 30 min. They also reported that there was much greater variability of the proteins among patients than with time of ischemia. In this same study, when 4 of the 17 liver specimens were evaluated by liquid chromatography—mass spectrometry (LC-MS), of the 1254 proteins analyzed based on more than one peptide, only 18 (1.4 %) changed significantly (9 increased and 9 decreased) at 6 h of ischemia, and only one (cytokeratin 18) increased at 1 h. They also found that the effects of warm ischemia were more extensive than 30 min of cold ischemia [33].

Much of the research to understand the effects of cold ischemia on proteins has been performed by studies to improve the clinical evaluation of diagnostic, prognostic, and predictive biomarkers. In the clinical arena, the goal is to identify 100 % of patients who are eligible for a specific therapy; therefore, in the case of breast cancer, estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) critical biomarkers that aid in determining the response of patients to two therapeutic approaches—therapy targeted at ER or at HER2. Most clinical studies have focused on a component of cold ischemia which has been described as “delay to formalin fixation” (DFF), which primarily involves the time between when a specimen is received in pathology and the time the specimen is fixed in

10 % NBF [34]. As would be expected, the reported effects on peptide/protein biomarkers of DFF vary with the methodology of the experiment, including the temperature of cold ischemia (RT versus 4 °C), the peptide/protein, the tissues, patient variables, and operative variables including time of warm ischemia.

The effects of DFF and, hence, cold ischemia on clinical biomarkers seem to vary with the subtypes of molecules and, to a lesser extent, on molecules within the subtype. Also, because of the importance of ER, PR, and HER2 on choices of therapy, breast cancer is the focus of most studies. In the category of nuclear receptors, most studies have focused on ER, PR, and the proliferation markers (Ki67 and PCNA). One study [35] based on one specimen found that the ER and PR status of the tumor did change after storage at 4 °C for 4 days. In contrast, Khoury et al. [34] noted a decline in immunostaining scores of ER (5 cases) of 3 % at 2 h at RT, 9 % at 4 h and 20 % at 8 h and of PR (4 cases) of 11 % at 1, 2 and 4 h of RT and 15 % at 8 h. Although the changes in ER and PR were not statistically significant, the power of the study was low. When the study was repeated using various clones of ER and PR, ER clones did not change very much until after 2 h [36]. Pekmezci et al. [37] evaluated loss of ER or PR positivity between core needle biopsies versus the corresponding lumpectomy or mastectomy samples. They reported a 3.4 % (5/149) loss of ER and 7.1 % (9/126) loss of PR. Other studies have reported similar results. Of importance, in addition to different times of cold ischemia, there are other changes that occur that are different between a needle biopsy and a tissue excision including the effect of the biopsy on the surrounding breast tissue and much shorter times of warm ischemia for needle biopsies compared to excisions [38]. In studies of the effects of DFF on markers of proliferation, including Ki67 and PCNA, no changes were noted for even overnight delays in fixation [39–41].

Molecular features that are localized to the cell membrane may be more sensitive to DFF than molecules located in other intracellular locations (e.g., cytoplasm). The main cell membrane markers that have been studied include HER2 and E-cadherin. Based on HER2 expression as determined by fluorescent in situ hybridization (FISH), Khoury et al. [34] identified statistically significant decreases in FISH signals after 1 h of cold ischemia in 6 of 10 cases of breast cancer, but these changes were not associated with the clinical interpretation of the HER2 expression. In contrast, Portier et al. [42] in a cohort of 84 patients concluded that cold ischemia time of up to 3 h had no practicable effects on the clinical interpretation of HER2 by immunohistochemistry (IHC) or FISH. Similarly, Pekmezci et al. [37] in a study of 167 cases of invasive carcinoma found that HER2 status by immunohistochemistry did not change between core needle biopsies versus lumpectomies or mastectomies. When the effects of DFF on E-cadherin were

evaluated, E-cadherin changed in 2 of 10 cases, between 1 and 2 h of cold ischemia, causing a decrease in the average immunostaining score from 230 to 207 (10 % decrease). Of note, once the score decreased to 208 it stayed almost constant at 204 from 2 h to overnight so in most cases, E-cadherin is relatively stable during cold ischemia.

There have been only a few studies of the effects of cold ischemia on cytoplasmic markers. Khoury [40] observed no consistent changes in cytokeratins AE1/AE3 (cases, $n=10$) CK7 ($n=9$), and CK14 ($n=4$) or in CAM5.2, EMA, GCDFP-15, and mammaglobin over various times, including 8 h of cold ischemia or even maintenance of the specimens at RT overnight.

In general, results of studies of DFF are variable and very sensitive to the experimental design. Specifically, most studies ignore warm ischemia and some studies begin when a specimen is received in pathology and ignore the component of cold ischemia in the operating room; in contrast, other studies focus on local rather than radical excisions so that the time of warm ischemia is minimized, and the studies ensure that specimens are transferred rapidly to pathology, e.g., in less than 5 min [34, 36, 40]. Also, some studies ignore statistical evaluation of the results and/or base conclusions on small numbers of cases for which the power is low. Sometimes changes are statistically significant, but the extent of change may be too small to be clinically significant; in addition, in most cases, specimens were maintained at RT, so these studies must be considered with great care.

Just as with protein profiling, studies of the effects of cold ischemia on metabolomics have been limited. The study of Cacciatore et al. [43] used nuclear magnetic resonance which could reliably resolve 18 metabolites of which 15 changed (increased and decreased) significantly during 6 h of cold ischemia and 2 (glycine and aspartate) changed significantly during 30 min of cold ischemia.

In summary, there seems to be no optimal time to establish as a goal for the maximum time for cold ischemia. Most studies are performed at RT and have reported variable results in molecular changes during cold ischemia; in general, at shorter times of less than 1 h, only a small proportion of genes or proteins change, and most genes that change are stress/ischemia genes which are increased. Frequently, stress/ischemia are increased at 15 min or less and the increases sometimes continue for over several hours. Most studies that have evaluated the times of cold ischemia between 1 and 3 h seem to indicate minimal changes at the DNA, RNA, protein, or metabolomic levels. Somewhat larger changes seem to occur at 6 h of cold ischemia at RT.

Many investigators request specimens that are collected within 30 min or less after removal from the patient. This time would include the period of cold ischemia in the operating room and the

time of transport to pathology which usually is not considered in studies of cold ischemia. Also, the processing of multiple aliquots from a single specimen frequently consumes much more time than collecting and processing a single aliquot. Thus, in general, it is impracticable for biorepositories to meet a 30 min goal for the maximum time for cold ischemia for all specimens. Due to greatly increased costs, it is very problematic when relatively short goals for maximum times of cold ischemia are recommended in publications without scientific justification. This causes investigators to assume that there is scientific justification for recommending that tissues be stabilized in 1 h or less. Trying to meet such requests adds a large and unnecessary cost burden on biorepositories.

4.4 Effect of Fixation and Tissue Processing on Tissues

Fixation is a process by which tissues are stabilized at a point in time. The original goal of fixation was to maintain a consistent appearance and stable physical structure of tissues when embedded in wax and stored over many years. In part, fixatives were selected based on greatly reducing the solubility in aqueous solutions of various categories of molecules and preventing the destruction of tissues by microscopic organisms (e.g., bacteria and fungi). The primary fixative used in diagnostic pathology is 10 % neutral buffered formalin (10 % NBF) which was selected based upon the consistency of the cellular and stromal appearance of tissues after embedding the fixed tissue in wax (paraffin). The ability to analyze specific molecules in tissues was not a consideration in the selection of 10 % NBF as the diagnostic fixative of choice.

Benefits and approaches to fixation have been discussed extensively [44–46]. Importantly, the use of 10 % NBF has not changed in over five decades, and this approach to fixation has not been modified as new methods of molecular analysis such as immunohistochemistry or DNA/RNA analysis have been developed. When initially developed, fixation in 10 % NBF was based on greater than 24 h of fixation, but over the last several decades, shorter times of fixation in 10 % NBF have been used to deliver medical care more rapidly.

Histologic processing of tissue involves gradual removal of water from the fixed tissues and following the dehydration of the tissues, infiltration by a hydrophobic agent such as xylene, which facilitates the infiltration of the tissue by hydrophobic paraffin. Infiltration of tissues by paraffin permits cutting very thin (3–5 μm) and consistent sections from the tissue for microscopic evaluation following staining.

After fixation of cells and tissues in 10 % NBF, there is a decrease in immunorecognition of some specific molecules based on antibody–antigen choices; this decrease is prominent after about 24 h of fixation [48, 49]. Fixation in 10 % NBF results in covalent molecular additions of formaldehyde to specific amino acids of proteins and, over longer times of fixation, to intra-protein

and inter-protein cross-linking. It is hypothesized that these changes cause the decrease in immunorecognition and are reversed by antigen retrieval. For cells and cellular models, this decrease in immunorecognition can be reduced by transfer of the specimen after about 18 h of fixation in 10 % NBF to 70 % ethanol, which reduces loss of immunorecognition for at least 180 h, but not as long as 4 weeks [48, 49]. Of note, the decreased immunorecognition can be recovered in part by antigen recovery methods, including boiling of tissue sections in various solutions such as EDTA pH 9.0 and/or enzymatic treatment of tissues.

The loss of immunorecognition is greatly affected by histologic processing and the induction of a hydrophobic environment, and this process also is likely to be reversed partially by antigen retrieval via rehydration of the tissues [47, 49]. Of note, the effects on immunorecognition of fixation in 10 % NBF and histologic processing of tissue affect the evaluation of studies of the delay to formalin fixation. Fixation also can be too short (e.g., <6 h for thin aliquots [≤1–2 mm] of solid tissues). Specifically, the intensity of immunostaining for $p27^{Kip1}$ is optimal after 24 h [50].

4.5 Molecular Changes During Storage

In general, selection of storage methods balances costs of storage versus limitations in the methods of storage. Each method of storage, e.g., room temperature, −20 °C, −80 °C, liquid nitrogen vapor phase (LNVP) or liquid nitrogen, has limitations as to the maintenance of molecular stability of tissues on intermediate (1–2 years) to long-term storage (>2 years). As with most aspects of biorepository sciences, each category of molecule (e.g., peptides, proteins, mRNA, microRNA, and DNA) varies as to its stability in various preparations (solid tissue, DNA, RNA, biofluids, paraffin blocks) and within each category there also are variations in specific molecules. As with most areas of biorepository sciences, there are little actual data on molecular stability of specific molecules during each of these methods of storage.

4.5.1 Stability of RNA and DNA During Storage

From the standpoint of stability upon long-term storage, DNA is the most extensively studied category of biological molecules from humans and, in general, is the most stable type of molecule during storage. Studies at the National Institute of Standards and Technology (NIST) indicate that DNA undergoes few changes while frozen at ultra-cold temperatures or even, if stabilized, upon storage for several years on paper at room temperature [51]; of note, even when degraded, DNA may be useful for some types of assays [51].

DNA and RNA are susceptible to freeze–thaw cycles if specimens are not stored to prevent the formation of ice crystals [52, 53]; however, if non-degraded RNA can be extracted from a tissue, non-degraded DNA also can be successfully extracted. DNA from whole blood frequently is stored on paper treated to reduce

contamination by organisms. In this format, DNA can be stored at room temperature (RT). After 10 years storage at RT, the DNA will show changes of degradation that is not seen if the samples on paper are stored at −20 °C or colder; however, the degraded DNA stored at RT still is adequate for identification of individuals as well as for other studies [51]. In contrast, DNA in untreated urine was not stable and not useful for PCR detection of schistosomiasis after 18 months of storage at −80 °C [54].

There have been few studies of the stability of DNA, RNA, or protein in intact solid human tissues on long-term storage (>2 years) at −70 °C or colder. One study found that in most cases minimally degraded and useable DNA could be extracted from about 80 % of specimens stored at −80 °C or colder including specimens stored for more than 1 year [27].

In regard to RNA, specimens stored for more than 10 years at −80 °C or LNVP had RNAs that had RINs of 5.8 or 4.3, respectively, suggesting that useable DNA also could be obtained from these specimens ([55], Table 2). Of note, the RINs on this long-term storage suggest some degradation of RNA over time, but this extent of degradation is adequate for assay by RT-Q-PCR for most molecules [56, 57].

4.5.2 Stability of Peptides/Proteins on Long-Term Storage

There are few studies evaluating the stability of specific peptides/proteins on long-term storage. For proteins and peptides, the study of Potter et al. [58] suggests that effects of storage of serum at −80 °C for 2 or more years can cause changes in the concentrations of cytokines. The extent of these changes was found to increase with time of storage, and the changes caused bias in experimental results [58]. This is a very concerning result, and there needs to be many more studies of stability during storage of intact tissues.

Table 2
Yield of RNA and RNA integrity of specimens after at least 10 years of storage at −80 °C and LNVP[a]

	−80 °C		LNVP		
	Mean	**Standard deviation**	**Mean**	**Standard deviation**	***p*-value**
RIN	5.8	2.8	4.3	3.3	0.0002
Yield (μg)	7.1	6.8	4.4	5.4	0.006
DF[b]	39	37	56	42	0.003
3′/5′ GAPDH	2.6	6.2	2.2	5.8	0.9
3′/5′ ACTB	52	49	54	55	0.5

[a]Modified from 55
[b]Degradation factor (*see* [55])

Stability of Proteins in Paraffin Blocks

Paraffin blocks that are stored for several years frequently are used for immunohistochemistry to detect the phenotypic expression of peptides and proteins. As discussed, many factors affect paraffin blocks including variables associated with the preoperative and intraoperative periods, and with cold ischemia, fixation, and processing of fixed tissue to paraffin blocks. However, once tissues are processed to paraffin blocks, the peptides/proteins in paraffin blocks are considered to be stabilized based on the view that proteins in paraffin blocks are relatively stable over time [59].

In contrast to our prior results, our unpublished data indicate that some proteins are not stable in paraffin blocks on long-term storage of more than 10 years. We have found that E-cadherin, ɣ-tubulin, and sperm-related protein 17 had large decreases in immunorecognition if blocks of normal fallopian tubes were older than 10 years. There was less change in vimentin and cytokeratin 18 and little change in smooth muscle actin [60].

4.6 Effects of Heterogeneity of Tissues on Research

Some investigators who utilize solid human tissues in their studies homogenize these tissues without an adequate knowledge of the cellular features of the tissues. Specifically, a malignant tumor may be very heterogeneous, including focal areas of necrosis and/or fibrosis. Such areas may be mistaken for areas of viable tumor. In addition, tumors frequently are infiltrated by varying numbers of inflammatory cells or may be intermixed with uninvolved or normal tissues. Similarly, tissues obtained as normal or as uninvolved by the disease process may be intermixed with diseased cells and/or inflammatory cells. If such tissues are homogenized and analyzed for specific molecules, how will an investigator know from which cells the molecules of interest (e.g., microRNAs) are extracted [61]? For example, in panel A of Fig. 1, prostate cancer is metastatic to a lymph node; however, many more normal lymphocytes than malignant cells are in the lymph node. Similarly, panel D of Fig. 1 demonstrates that when uninvolved tissues (i.e., prostate) are obtained to compare with tumor or to be used in other studies, large numbers of normal inflammatory cells may be present in the tissue. This is similarly the case for uninvolved colonocytes in panel A of Fig. 2. In addition, other cells such as smooth muscle of the colon wall may be present in specimens provided for the study of colon cells (panel A of Fig. 2). Also, tumor specimens may contain variable amounts of necrosis (panel B of Fig. 2) as well as tissue damage (panels B and C of Fig. 1); this limits the amount of malignant cells for study and frequently increases the numbers of inflammatory cells. Thus, there must be some form of quality control of the specific solid tissue aliquots used in research. In some cases, the malignant cells of a tumor may be enriched by macrodissection [62] and/or other methods to study specific cellular populations.

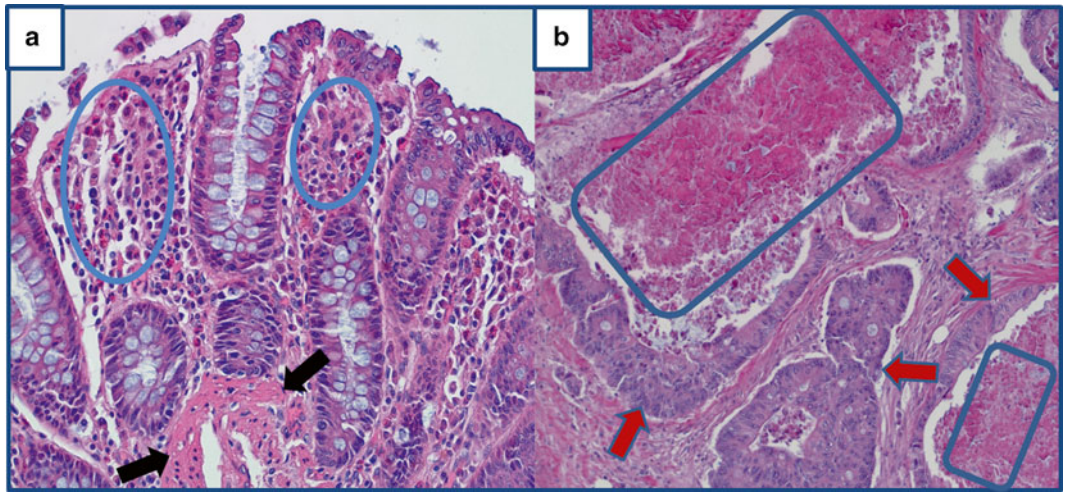


Fig. 2 Panel (**a**) (original magnification ×200) demonstrates typical inflammatory cells (within *blue ovals*) on each side of an uninvolved gland of the colon. Of note, the specimen also contains an area of colon wall (between *black arrows*) which is composed of smooth muscle cells. The number of inflammatory cells is about equivalent to the number of uninvolved cells of the colon. Panel (**b**) (original magnification ×200) is from a metastasis of colon cancer to the liver. There is extensive necrosis of the colon cancer (within *blue rectangles*). The stroma between viable appearing malignant colon cells (*red arrows*, *blue outline*) contains some inflammatory cells

5 Biorepository Operations

The primary goal of a human tissue biorepository supporting research should be to provide high-quality human tissues for research. There should be a strong emphasis on the distribution of tissues to investigators rather than long-term banking of tissues. This goal is facilitated by the efficient and cost-effective operations of the biorepository which should minimize factors that negatively affect the utilization of human tissues in research.

5.1 Identification of Tissue Aliquots

The unique and accurate identification of tissue specimens is of utmost importance if high-quality specimens are to be provided for support of biomedical research [1–3]. Any method of labeling specimens should minimize the chance of separation of labels from tissue aliquots; should avoid chances of mislabeling secondary to personnel error; and should minimize difficulties in reading labels (e.g., illegible handwriting). For most biorepositories, the most efficient approach is utilization of 1D or 2D bar codes. Bar codes link the tissue aliquots to a database containing defined information, including the source of the specimen as well as specimen characteristics, storage sites, and ultimate disposition. Times of the steps of collection and processing also aid to define each aliquot and its history [1–3]. In addition to the bar code, other

information sometimes is included on the label such as age, race, sex of the patient, specimen source, and investigator or bank assignment. The bar code software usually permits such modifications of labels; however, some governmental protocols may prohibit information other than the bar code from being on a label.

5.2 Collection and Processing Variables

Variables which may affect the collection and processing of tissues should be considered by biorepositories, including neoadjuvant therapy, which limits use of some specimens in specific types of research. It is critical for researchers who use human tissues to understand the potential limitations in the use of tissue aliquots imposed by such variables. Similarly, processing approaches which extend the time of cold ischemia should be minimized if practicable. Of note, the informatics system should maintain records of the times associated with the collection, processing, stabilization, and storage of specimens.

5.3 Difficult to Fulfill Requests

As more specific requirements are added by investigators to requests for tissues, the requests become much more challenging to fill [2, 3, 6]. Clearly, it is easier to collect and provide "any breast tumor" than fulfill a request for "poorly differentiated triple negative ductal breast carcinoma from an African American male less than 35 years old"; all breast cancers are uncommon in males, as are breast cancers from young patients. Also, molecular subtypes limit tissue availability. When any requirement is combined with specific racial and molecular characteristics, the requests can be met only rarely even by cooperation of multiple institutions. Requests for a large specimen (e.g., 5 g) of a tumor that is typically small (e.g., breast and prostate) are difficult to meet, though large amounts of the common tumors of the ovary and kidney frequently are available. Some cancers (e.g., breast, prostate) are requested by many investigators, so investigators requesting small sizes (e.g., 0.15 g) usually are served before those requesting large specimens (e.g., 2 g) [1–3]. Similarly, requests for many high demand tissues within 1 year (e.g., 200 cases) usually are difficult to complete, especially requests for large numbers of relatively rare tumors or tumors in great demand. Some tumors usually are not surgically removed (e.g., small cell undifferentiated [oat cell] carcinoma) so requests for such tumors are difficult to meet. The time necessary to fill a request for a specific tissue also may require extensive efforts because of difficult and complex protocols (e.g., collection of an intact spinal column) and such requests may not be met by a biorepository even with extensive resources. Of note, when additional requirements are added to a request for a specific tissue, there is almost an exponential increase in the difficulty of filling the request for each added requirement.

Biorepositories typically try to provide tissues via an equitable distribution plan; however, because of the extensive institutional

resources required to operate a biorepository, most biorepositories are closed to extramural investigators unless there is a collaboration with a local investigator or there is specific funding to provide for external distributions. Investigators should communicate with biorepositories to ensure their requests can be met and should not add restrictions and requirements that are not scientifically justified or may not be definitely needed by their specific research projects. Because multiple specimen requirements affect potential tissue availability, the lack of availability of specimens to support research should not be confused with their quality.

5.4 Time Interval Between Surgery and Tissue Stabilization

In general, collection of remnant diagnostic tissues should be reviewed by a pathologist or their designate (e.g., a physician assistant) to assure that the diagnostic integrity of the specimen is uncompromised. Investigators must be aware that requirements for very fast processing and/or freezing of tissues immediately following surgery are likely to decrease the availability of specimens for their research, frequently are not scientifically justified (*see* Subheading 4.3 Effects of Cold Ischemia) and may greatly increases costs to investigators. Specifically, freezing of multiple tissues in the operating room may be extremely expensive (>$1500 per specimen), is beyond the resources of most biorepositories, and may compromise the diagnosis of specimens. Nevertheless, the times associated with the surgical removal of tissues from the patient, the transport of the tissues from the operating room (OR) to the biorepository, of processing, and of stabilization (e.g., freezing) should be as rapid as practicable and should be documented in the informatics system (IS).

Biorepositories should try to maintain tissues at about 4 °C until diagnostic examination, and, subsequently, research specimens should be rapidly processed and stabilized. Of note, increased times of processing may develop when several tissues must be processed concomitantly. When this occurs, one aliquot of each tissue could be rapidly stabilized, e.g., frozen rapidly in LNVP. Subsequently, other aliquots could be collected, processed, and stabilized as rapidly as practicable based on the workload of the biorepository [1–3].

5.5 Education of Users of the Biorepository

Biorepository professional personnel should be available to provide educational services to researchers who need assistance. All requests for human tissues should be reviewed by pathologists or equivalently knowledgeable professional personnel who can discuss any needed modifications of requests with an investigator. For example, researchers sometimes need help in choosing specific tissues and methods of processing to support their studies. Investigators need to be aware that their requirements may reduce the availability of tissues, that tissues may vary in their usefulness for specific research, that biorepositories face constraints as to time

and services in handling of tissues, and that similar types of tissues may be molecularly and biologically different. Specifically, smooth muscle from a large artery will behave differently than smooth muscle from the uterine wall or from the wall of the colon.

5.6 Types of Tissues Collected and Services Provided

Biorepositories vary in their goals and characteristics and hence their ability to provide specific tissues to investigators. Investigators should identify which specific tissues and processes a biorepository can provide. Many services are limited by resources so that a biorepository may decide not to utilize its resources to deliver tissues to local investigators; similarly, culturing cells from tissues, development of xenografts and extracting DNA and/or RNA may not be supported by a biorepository because providing a wide range of services may inhibit the primary goal of the biorepository—to provide tissues to as many investigators as practicable. In addition, complex services add considerable costs for which a biorepository must be compensated.

5.7 Information on Specimen Aliquots

The typical information provided with a specimen to an investigator should be a patient's age, race and sex, the pathologic or patient diagnosis and, if applicable, a quality control description of the specific aliquot of tissue provided. Beyond this basic annotation, the extent of the information needed varies with the research and investigator. The annotation also may vary with the goals and resources of the biorepository. Specifically, if the goal of a biorepository is focused on population-based research of a disease (e.g., stroke), then extensive clinical, familial, and social information would be obtained from patients or their families before or at the time of tissue collection; however, if tissues are obtained to determine the biochemistry of a human enzyme found in the kidney, minimal clinical information initially may be required for such studies. UAB has found that only very basic information may be necessary for a large proportion of investigators. Thus, for typical biorepositories, obtaining extensive clinical information on every patient whose tissues are in the biorepository would be a waste of effort. It is more cost-effective to obtain such information from health records when the data are requested by investigators. Of note, the costs of the resources necessary to obtain such data should be passed on to the investigator through charge back mechanisms.

5.8 Storage of Frozen Specimens

If the goal of biospecimen storage is to maintain cell viability (e.g., that cells or tissue can be thawed and cultured in vitro or grown as xenografts), cells can be frozen in media plus 10 % DMSO and stored at the minimum temperature of LNVP. Even when stored at LNVP, if aliquots of solid tissue are frozen without additives, after thawing, cells will not be viable because during thawing, ice crystals lyse most cells. Freezing very small pieces (<1 mm^3) of solid

tissue in media plus 10 % DMSO may permit isolation of viable cells upon thawing, or the thawed tissue can be transplanted and grown in immunocompromised animals [63].

The optimal conditions for long-term storage (≥6 months) of tissue aliquots is an issue of controversy. Storage at −20 °C in a non-self-defrost freezer for more than a few months has been reported to cause changes in peptides; specimens should not be stored for even short period in self-defrost freezers [64]. The general belief for storage of tissues and bodily fluids is "the colder the better," based on storage of viable cells which requires storage at least in LNVP; however, data concerning if storage at LNVP is better for research than storage at −80 °C is very limited. Of note, one study indicates that bias occurred when serum from cases and controls was handled differently, including storage at −80 °C for different periods [7]. Separately, it has been reported there are no reproducible differences at the protein level between storage of solid tissues for 10 or more years at −80 °C versus at LNVP temperatures, and that storage at −80 °C is better than storage in LNVP for mRNA extraction and analysis ([55], Table 2).

5.9 Records of Collecting, Processing, and Storage of Tissues

A laboratory or biorepository should maintain records of the handling of tissues in order that the history of a tissue specimen can be reconstructed. It would be useful to note the length of the operation; however, most biorepositories may not have access to this information. If available, the time the specimen is removed from the patient, time and conditions (room temperature, on wet ice, etc.) of transport to pathology, the method of processing, sizes of aliquots, and the time to and type of stabilization (e.g., frozen in OCT) should be recorded. When paraffin blocks are constructed, the record should include the time until fixation and the length and type of fixation (e.g., 18 h in 10 % NBF). In addition, the model and maker of the tissue processor and the history of any changes to the time and chemistry of each step of the tissue processor (e.g., 70 % ethanol for 1 h) should be recorded.

5.10 Access to Different Populations and Types of Tumors

Tissues from local medical facilities may be limited as to the number of specimens of specific types that are available. For example, the biorepository may be located at a facility that treats only patients with cancer. Also, if investigators need tissues from African Americans or other racial and ethnic groups, investigators may have to obtain such tissues from different geographic regions whose populations include increased proportions of the racial/ethnic groups needed. Similarly, there may be needs for types of tumors or tumors with specific characteristics (e.g., triple negative breast cancer). To meet needs for large numbers of such specimens may be difficult for a single biorepository, especially if the types of tumors requested are rare (e.g., synovial cell sarcomas) or typically are not treated by surgery. Thus, relationships among

biorepositories could be developed that permit them to work together to aid in meeting difficult to meet requests for tissues. This is important because in general, biorepositories do not distribute adequate numbers of tissue to support biomedical research [65], and this approach expands the outreach of a biorepository.

6 Quality Management

A quality management system (QMS) or quality assurance (QA) program is a general approach to managing all activities of an organization, laboratory or biorepository at a high level of consistency and quality [1–3, 66–69]. The focus of the QMS should be to improve and standardize all activities so that each procedure or product is of a defined quality. Quality control (QC) is the component of the QMS that measures and monitors the attributes and performance of a process, product, or service compared to defined standards in order to verify that all defined standards of the product or service are met completely. QC should be one of the major components of a rigorous QMS [1–3, 66–69].

6.1 Standard Operating Procedures

A standard operating procedure (SOP) is a written detailed method that should permit a laboratory activity to be performed uniformly, day after day. If followed, a new employee who is trained should be able to use the SOP to perform the procedure equivalently to an employee experienced in the method. An SOP should be developed for all activities of a biorepository including equipment maintenance and monitoring. Each year, SOPs should be reviewed and revised if necessary. New or revised SOPs should be dated and the supervisor revising the procedure identified. Copies of the old SOP should be archived to permit review of prior versions so that differences in QC of products caused by changes in SOP can be identified. Employees must not deviate from the current SOPs.

In establishing a strong QMS, personnel assigned to monitor the QMS should ensure compliance of employees of the biorepository with all SOPs and regulatory requirements. They should aid supervisors in developing SOPs for the handling and shipping of specimens. When QMS problems or specimens of poor quality are identified, QMS personnel should inform higher levels of management and should initiate and participate in correction of any deficiencies.

6.2 Audits

Audits are planned evaluations of selected operations of the biorepository to ensure specific aspects of the QMS are being followed. Audits should be conducted periodically and should be documented as a critical component of the QMS [1–3, 66–69]. Audits may be as straightforward as a monthly review that freezer temperatures and liquid nitrogen levels are being monitored daily or

may be as complex as a quarterly review of specimen collections to determine why specific specimens were not obtained. QA personnel document audits and any deficiencies and should document problems to the management of the biorepository.

The QMS of the biorepository should describe how frequently audits are conducted, how audits are documented, and how deficiencies found on audits are corrected. Examples of potential audits could include the following:

- Adherence to SOPs for preparing specimen aliquots of breast carcinoma
- Evaluating maintenance, repair, and monitoring of freezers (e.g., performance of periodic maintenance of −80 °C freezers)
- Monitoring of compliance with shipping requirements
- Documentation of specimen disposition
- Adherence to SOPs for labeling of specimens
- Accuracy of specimen information
- Contact of internal investigators to inform them of available tissues
- Maintenance of records of training and adherence of staff to required training (e.g., training in biohazards)
- Monitoring of the accuracy of collected patient data and data management
- Monitoring the collection and prompt processing of pancreatic tissues
- Rate of obtaining needed tissues from autopsies.

6.3 Quality Control of Solid Tissues

It is important to ensure the quality, accurate diagnoses, and a detailed description of the actual aliquots of solid tissues distributed to investigators (i.e., to provide quality control of specimens [QC]). For a biorepository, QC is a critical component of the QMS. Biorepositories have utilized various types of QC to make sure that investigators understand the features of the tissues used in their research and their limitations. As discussed, many human tissues may be heterogeneous and this is very true of tumors; thus, tissues obtained from tumors have different proportions of malignant, inflammatory and uninvolved cells, amounts of desmoplasia, and/or areas of necrosis and/or mucin. Fibrosis within and next to tumors is frequently mixed with malignant cells or mistaken for actual tumor; also, some malignant cells may infiltrate uninvolved tissues diffusely so that some tumors such as prostate cancer may be hard to identify. Knowing the diagnosis of an overall tumor does not ensure that the aliquots of tissues provided to investigators are adequate for their research; thus, quality control which

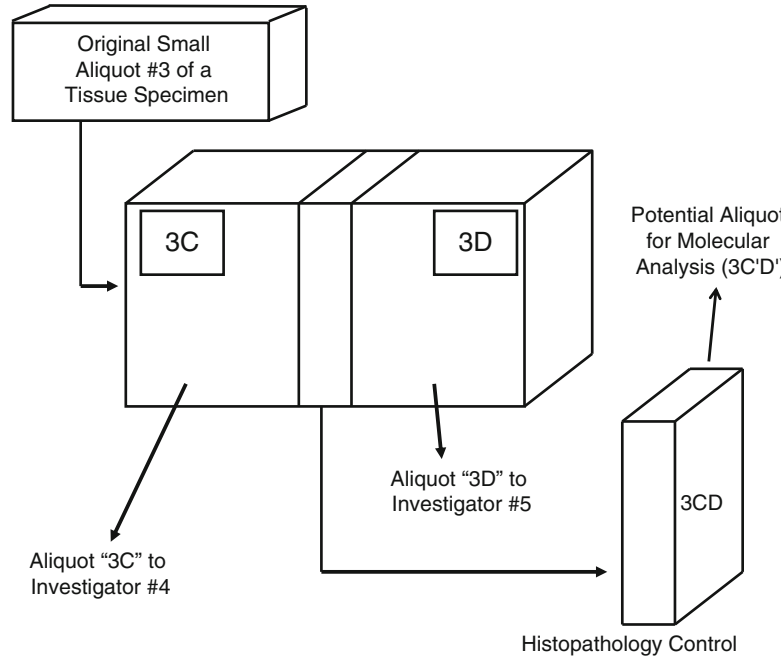


Fig. 3 This cartoon demonstrates the routine quality (QC) control (QC) of tissues provided for research by UAB. The QC is for the third specimen from a patient with a tumor. 3CD represents a small area between 3C (frozen for investigator 4) and 3D (frozen for investigator 5). This area is processed to a paraffin block from which a section is cut and stained by hematoxylin and eosin and used as the microscopic QC for both specimens 3C and 3D. If requested and/or needed, an additional frozen piece of tissue 3C′D′ can be cut before fixation for molecular analysis

includes detailed descriptions of each aliquot provided for research is necessary.

A minimum QC should be established by a biorepository for the specific specimens they distribute. It is recommended that the minimum QC be a microscopic diagnosis and description by a pathologist of the specific aliquot distributed. Optimally, except for very small aliquots, QC is based upon a mirror image aliquot of the tissue that is distributed to an investigator. Figure 3 demonstrates such a mirror image, 3CD that is processed to a paraffin block. The 3 indicates that this is the third specimen from a single surgery for colon cancer. The microscopic description of 3CD is the QC for both specimens 3C and 3D. Similarly, a piece from 3CD prior to fixation, i.e., 3C′D′, could be utilized for the molecular QC of aliquots 3C and 3D if requested. The microscopic examination is the basic QC performed by UAB on specimens distributed to investigators.

Via the use of similar QC, UAB has reported that 15 % or more of tissues collected for specific investigators cannot be

utilized for the investigators' specific research projects [1–3, 66]. Aliquots that appear grossly to be unaffected by a disease may be found to be affected upon microscopic examination. Also, tissues which appear diseased may be undergoing some other process or may not be appropriate for use in specific research. Specifically, focal areas of tumors (e.g., metastases of colorectal cancer to the liver) may be too necrotic for some research (Fig. 2). The quality control should specify the percent of the specimen that is tumor or is diseased. Of a tumor, the percent of malignant cells (tumor nuclei), the extent of necrosis/fibrosis as well as the percent of other factors such as mucin formation should be described [70]. Describing the proportion of cells of a tumor that is malignant is necessary because tumors may contain large numbers of inflammatory cells or may be mixed with uninvolved cells.

Quality control of solid tissue can be based on frozen sections of a tissue embedded in OCT which is necessary to obtain good frozen sections (Fig. 4). However, the OCT may affect some assays.

Investigators may request more extensive approaches to QC. For example, the quality control examination may include "molecular quality control" in which RNA, DNA, and protein are extracted from small aliquots followed by molecular

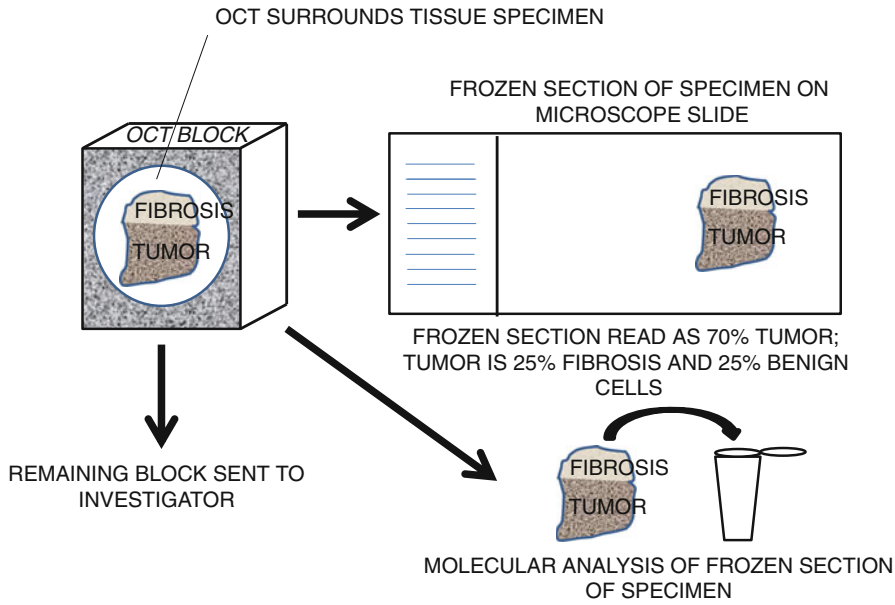


Fig. 4 This cartoon demonstrates an approach to quality control (QC) in which a specimen is embedded in optimal cutting media and a frozen section is cut as the QC control. If the QC section demonstrates acceptable tumor, the remaining specimen is provided to the investigator. If requested by the investigator, one or more frozen sections also can be cut for molecular analysis (e.g., determination of RIN). The added expense of molecular QC is added to the cost of the aliquot to the investigator

characterization of their quality (e.g., RIN as determined by the Agilent® 2100 system for quality of RNA). Molecular QC may be provided at an investigator's request or sometimes to determine the characteristics of tissues provided by a biorepository. As QC of tissue requested by researchers becomes more complex, the cost of the increased effort is charged to the investigator. As part of QC, investigators may also request macrodissection of specimens to increase the proportion of diseased cells. An investigator may request that even the reverse side of an aliquot be macrodissected (Fig. 5). Such approaches increase the price of specimen processing and typically reduce the amount of tissue available to the investigator for research.

6.4 QC in the Collection of Bodily Fluids

QMS of the biorepository should identify the parameters of collection, processing, storage and distribution of biofluids, and the biorepository should develop appropriate SOPs which minimize bias both in patient selection and tissue handling [71]. Most studies

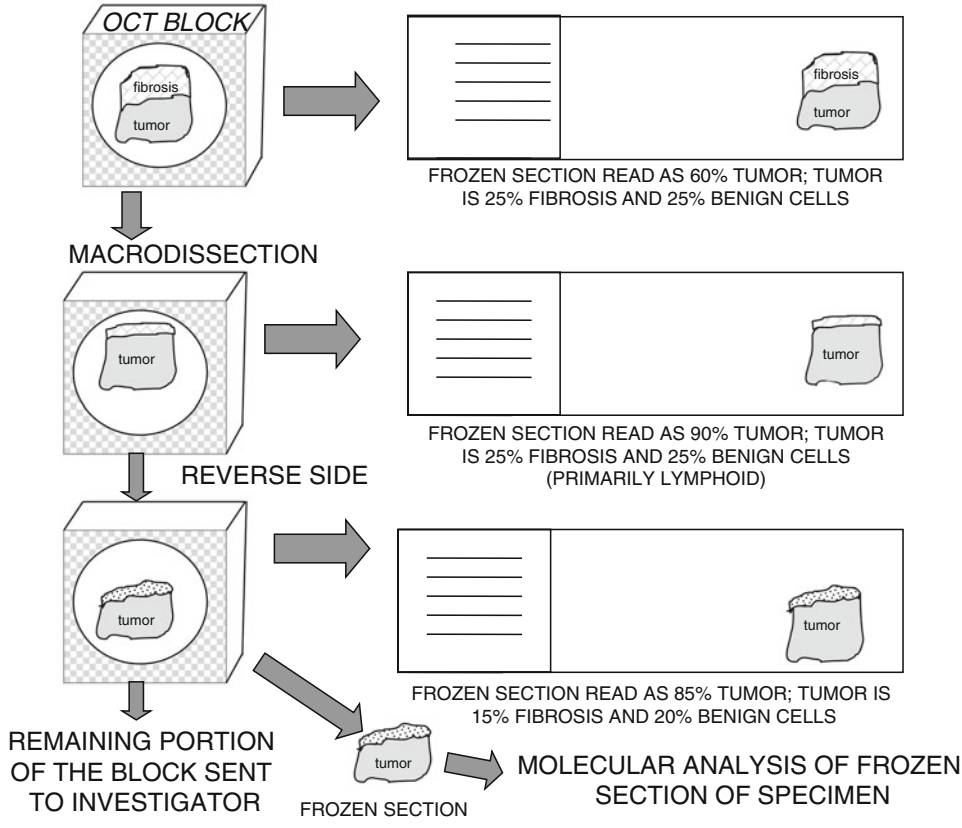


Fig. 5 This cartoon demonstrates an approach to quality control that is coupled to macrodissection of the specimen. Most of the area of fibrosis is removed, and the reverse side also can be evaluated to ensure that the amount of tumor on the reverse side is adequate. This is a relatively expensive approach to ensuring that the tissue used in research is optimal

require bodily fluids frozen within 4 h. Serum, plasma, buffy coats, and whole blood can be maintained at about 4 °C prior to freezing; however, freezing of whole blood before processing should be avoided if hemolysis is to be prevented, even though hemolysis may be acceptable for some assays [72].

6.5 Monitoring the Overall QMS and QC

Biorepositories should periodically survey the satisfaction of users/investigators, and the survey results should be considered carefully by QMS personnel. Investigators who indicate problems with specimens should be contacted and their problems discussed and corrected, if practicable.

If a biorepository provides specimens to extramural investigators, shipments should be monitored and a manifest should accompany each shipment. Of note, before shipping specimens via air, personnel must be trained in the requirements of the International Air Transport Association (IATA). In the U.S., training in the Department of Transportation (DOT) regulations is also necessary [73, 74].

6.6 Aids in Developing a QMS

Resources which can aid laboratories and biorepositories in establishing a QMS include the Best Practices of the International Society of Biological and Environmental Repositories (ISBER). These documents focus on optimizing biorepository operations and improving the quality of biorepository products and services [75–77]. The National Cancer Institute (NCI) has published more limited guidelines [78]. In addition, the Canadian Tumour Research Network (CTRNet) web site [79] provides their SOPs that have been tested based on their use in biorepositories.

Other references which may aid biorepositories are available from the International Organization for Standardization (ISO), especially ISO9001 [80]. The Federal Drug Administration's (FDA) Good Tissue Practices [81] and Good Manufacturing Practices (GMP) [82] also may be useful to biorepositories.

7 Regulatory and Ethical Issues in Tissue Repositories

7.1 Informed Consent and HIPAA Authorization

Biorepositories may obtain from patients remnant tissues that are no longer useful after diagnostic or therapeutic uses have been finished. The local IRB determines if patients from whom remnant tissues are obtained should be consented. Alternatively, the IRB may approve waiving the requirement for informed consent and HIPAA authorization; however, this approach is currently under review in the U.S. [83–86]. Biorepositories also may obtain aliquots from diagnostic paraffin blocks stored in hospital/pathology archives. In contrast, some biorepositories obtain tissues and bodily fluids from patients specifically to support research, and informed consent must be obtained from these patients.

At UAB, tissues may be collected from over 40,000 patients per year which would require a minimum of 20,000 h to obtain consent from all patients undergoing these operations. The cost for this would be over $500,000 dollars; thus, due to finances, all operative patients cannot be consented, and UAB usually obtains consent from most patients only after tissues are collected [86]. HIPAA authorization from patients for the utilization of protected health care information can be obtained at the time of informed consent. Unfortunately, this approach results in most of the pathology archival collections being from patients who have not been consented, so their use in research would have to be via waiver of informed consent. Specific additions to the informed consent might include permission to re-contact the patients, to obtain information from health records, and to provide the specimens for research by for-profit companies. There should be specific agreements (initials) of permission to obtain blood and other bodily fluids specifically for research.

7.2 Cost Recovery

It is ethical for biorepositories to recover any costs associated with handling tissues for research. This cost recovery can partially support and maintain the biorepository. The cost recovery necessary to determine the processing cost per specimen depends upon the biorepository, especially the number of specimens collected versus the number distributed for research. Some repositories have grant support and/or institutional support to provide resources for some expenses so handling fees may be reduced. If possible, the costs of handling human tissues to support research should be equivalent to the costs and maintenance of an experimental animal.

7.3 New Ethical/Regulatory Issues

Many ethical and regulatory issues are being considered which will have a large negative impact on biomedical research in general and biorepositories specifically. Frequently, ethical recommendations are not thoroughly evaluated as to the ethics of reducing some research because of the costs associated with new requirements or unanticipated problems caused by the new requirements. Specific ethical and regulatory issues which will reduce the amount and quality of research with human tissues include return to patients of research results [87–90]. Also, a requirement to require informed consent for the use of all tissues in biomedical research, even if they are anonymized, would devastate research with archival paraffin blocks which are touted as a great potential resource for biomedical research. The new U.S. requirements to include genomic sequence data in available databases, while at the same time maintaining patient confidentiality, are challenging. Similarly, the portion of the consent form for performing genomic sequencing should be designed with care. An important new ethical issue is the focus on the need to increase the distribution of tissues from biobanks [11].

8 Safety

Investigators and their personnel, as well as the personnel of a biorepository, may be subject to multiple injuries in performing their jobs [91–95]. Potential sources of injuries include biohazards and chemical, physical, fire, and electrical hazards. The chances of injuries must be minimized by a safety program which can be a part of the overall safety program of the associated institution; however, the specific risks of laboratories and biorepositories must be considered in this institutional safety program. The safety program is developed by a safety committee which reviews it annually. The safety program is administered by a safety officer who should provide aid to supervisors in safety issues and in development of SOPs, evaluates safety incidents and recommends to the safety committee any changes necessary in the safety program. There are multiple publications and web sites that can aid a biorepository in establishing a safety program [91–95].

The safety program should be developed so that the chance of injury to an employee is minimized via the use of engineering practices (e.g., safety equipment). Annual training in all areas of safety including biohazards and chemical hazards and SOPs that protect personnel should also be provided. In general, in developing the safety program for a laboratory or a biorepository, the safety of each employee should be considered based on their job/functions, as well as the areas in which he or she works. The safety program should also consider non-laboratory personnel who may enter the work area including maintenance and janitorial personnel. Development of an adequate safety plan should utilize the many resources and publications devoted to this complicated area [94, 95].

8.1 Biohazards

A biorepository must decide if tissues from patients infected with bloodborne pathogens will be collected (e.g., hepatitis B, HIV) or from patients at risk of such infections (e.g., i.v. drug abusers). Although many biorepositories have chosen not to collect such tissues, biorepositories might unknowingly distribute infected tissues to multiple laboratories. Thus, it is important for biorepositories to require that all their personnel and personnel receiving tissues from the biorepository be trained in biohazards and that the personnel handle all human tissues using universal precautions (i.e., as if they were known to be contaminated with human pathogens). Sites/investigators who receive tissues from a biorepository should sign an indemnification agreement that holds the biorepository not responsible for any injuries caused by the tissues or tissue products that are received. Also, the personnel of the biorepository must be trained in bloodborne pathogens and should be offered vaccinations for hepatitis B.

8.2 Chemical Hazards

Biorepositories may use and/or store hazardous chemicals including formaldehyde, xylene, and alcohols. All chemicals pose potential dangers to the personnel of laboratories, in general, or of a biorepository. Dangers are via direct contact, by exposure to toxic vapors, and/or as explosive or fire hazards. The safety program should require use of safety equipment (e.g., appropriate protective gloves for chemicals) to minimize the dangers of chemicals. The safety program must include yearly training in chemical safety. An inventory of all chemicals must be maintained along with their safety data sheets [95]. There are new approaches to chemical safety and training developed by the World Health Organization (WHO). These are discussed on the WHO web sites as well as on national web sites devoted to chemical safety [95].

Other laboratory hazards include physical hazards. Floors or steps that are in poor repair or are slippery due to contamination by fluids and/or wax/paraffin are a major safety concern. Also, burns due to hot or ultra-cold temperatures are a risk. Lighting should be adequate. Training in physical safety should focus on use of safety equipment (e.g., insulated gloves), prevention of inappropriate use of equipment and furniture, minimization of repetitive actions, prevention of improper approaches to lifting heavy objects, and on protection of personnel from abuse and or assaults by others. Standards for fire safety and for storage of liquid nitrogen are maintained by the National Fire Protection Association [96] and/or by local requirements.

9 Informatics

An informatics system (IS) of a biorepository should be selected based on the size and model of the biorepository, its operations, workflow, and business practices. In general, the IS should save time for the personnel of the biorepository so that data should be easily input and accessed with few required fields. An important component of the IS is the integration of a bar code for unique sample identification. There are multiple ISs for biorepositories that are available commercially. Of note, a "free" IS may require extensive resources to integrate the IS into the operations of the biorepository, and plans for such adoption should be evaluated carefully. Multiple time points of tissue collection, processing and storage, and information on specimen distribution and disposition should be recorded in the IS as well as requests for tissue by investigators and their contact information. The database should be developed to include data necessary for a "history" of any specimen.

9.1 Vocabulary of the IS and Biorepository

A flexible vocabulary is necessary in a database of a biorepository in order to follow both the storage and distribution of specimens. A diagnosis provided by a pathologist is very specific such as "well

differentiated, lobular carcinoma of the breast with extensive lobular carcinoma in situ." This is not problematic for entering this description into the IS when banking a specimen; however, the vocabulary must have the capability to match this specimen's diagnosis with a potential investigator request for "breast cancer." In addition, requests from investigators sometime incorporate age, race, and sex with the request as well as molecular subtypes such as triple-negative breast cancer. Thus, the design of the vocabulary of the IS must be more flexible than a vocabulary that just relies on the diagnostic vocabulary of the pathologist [97, 98]. If an IS contains identified patient and patient health care information (PHI), it must meet the security standards of HIPAA including the prevention of unauthorized access to the IS via its location on a secure server behind a firewall and by using strong individual access codes. Also, there should be extensive antiviral protection and nightly backup on an independent, physically separated, and similarly secure server. For a biorepository, the access codes should permit each specific user different forms of access. Some individual codes should permit read-only, others both data entry and editing of existing data, and administrative codes may permit access to patient identification. The codes of informatics personnel typically permit modification of the database fields. A major HIPAA requirement for an IS that contains PHI is audit trails. The audit trails monitor the use by all personnel who access the identified database, even for "read-only." The security of the IS of the biorepository should be approved by the privacy officer of the institution.

10 Future Directions of Tissue Repositories

Biorepositories now have access to multiple approaches to their design and operation based on the Best Practices of the International Society of Biological and Environmental Repositories (ISBER) [75–77], and of the NCI [78], and biorepositories should consider appropriate changes to operations based on these recommendations. Detailed records are required to develop a "history" of each human tissue stored in the biorepository. This will require a HIPAA compliant informatics program which can incorporate such data as well as meet HIPAA security standards. Biorepositories should also consider the advantages (improved operations) and disadvantages (e.g., costs) of obtaining accreditation (College of American Pathologists—CAP) and certification (e.g., CTRNet). Certification and/or accreditation in the future might be especially important in obtaining independent funding of a biorepository and accreditation by the CAP may be necessary for a biorepository to aid in individualized (personal) medicine.

The regulatory environment may change, which may greatly decrease research using paraffin blocks and add unacceptable costs

to biorepositories due to potential liability. There are multiple other challenges that are likely to affect biorepositories over the next 5 years. One change that would be beneficial is for a greater proportion of tissues collected by biobanks to be distributed for research. This will make biobanks more cost-effective and may be ethically required for patients consented for the use of their tissues in research.

Other challenges include the decreased availability of pristine samples due to the increasing use of neoadjuvant therapy which may affect tissues in unidentified ways. Also, due to medical care advances including improved screening, diagnosis and imaging, both the size of available tissues and the availability of tissues from metastatic lesions are decreasing. To address these issues, novel approaches to sampling such as nitrocellulose blotting [99] will have to be used, and patients will have to be consented for obtaining extra samples of tissues just for research. Also, assay methods using archival paraffin blocks will have to be better understood and improved. Similarly, research with human tissues, in general, will have to be adapted to the use of very small aliquots of tissues.

More research will shift from DNA and mRNA analysis to epigenetic control of transcription, post-translational regulation and proteomics as large databases are developed containing enough genetic data to cover, for example, the genomes of most typical cancers. Shifting to the study of proteomic and other non-DNA-based approaches will likely be more challenging than study of DNA and mRNA.

11 Summary

Access to high-quality human tissue specimens as well as associated clinical information is necessary to support biomedical research. As emphasized in this review, there are multiple factors that should be considered by investigators and addressed by biorepositories as to the use of human tissues in biomedical research. These issues must be considered in the design and operations of biorepositories. Foremost, the operations of various models of biorepositories should primarily emphasize that tissue specimens should be distributed to investigators and not sit unused over many years in a biobank. The quality of specimens must be maintained by a rigorous QMS including QC of the actual aliquots of specimens provided for research. Factors affecting the usefulness of human tissues in research must be better understood, requiring extensive research in biorepository sciences and becoming a component of an educational process provided to investigators by biorepositories. In addition, extensive effort is necessary to avoid bias in the conclusions of research studies utilizing human tissues. If research with human tissues is to be reproducible, the power of studies must be adequate and bias must be minimized.

Acknowledgement

This work was supported in part by the Cooperative Human Tissue Network (NCI 5U01CA044968-23), the Tissue Procurement Shared Facility of the UAB Comprehensive Cancer Center (P30CA13148), the Breast SPORE at UAB (NCI 5P50CA89019), the Pancreatic SPORE at UAB (NCI P50CA101955-05), the U54 MSM/TU/UAB Comprehensive Cancer Center Partnership (2U54CA118948), and the Pulmonary Hypertension Breakthrough Initiative (1R24HL123767-01 together with a grant from the Cardiovascular Medical Research and Education Fund). The authors report no conflict of interest.

References

1. Bell WC, Sexton KC, Grizzle WE (2010) Organizational issues in providing high-quality human tissues and clinical information for the support of biomedical research. In: Grützmann R, Pilarsky C (eds) Methods in molecular biology, vol 576, Cancer gene profiling. Springer, New York, NY, pp 1–30
2. Grizzle WE, Bell WC, Sexton KC (2012) Issues in collecting, processing and storing human tissues and associated information to support biomedical research. In: Srivastava S, Grizzle WE (eds) Translational pathology of early cancer. IOS Press BV, Amsterdam, pp 531–549
3. Grizzle WE, Sexton KC, Bell WC (in press) Issues in operating a human tissue resource supporting biomedical research in cell therapy: principles, methods and regulations, 2nd edn., Bethesda, MD: American Association of Blood Banks (AABB) (in press)
4. Grizzle WE (1989) Commentary. Use of human tissues in research. In: Swerdlow JL (ed) Matching needs, saving lives: building a comprehensive network for transplantation and biomedical research. Annenberg Washington Program, Washington, DC, pp 37–38
5. Grizzle WE, Sexton KC (1999) Development of a facility to supply human tissues to aid in medical research. In: Srivastava S, Henson DE, Gazdar A (eds) Molecular pathology of early cancer. Ios Press, Amsterdam, pp 371–383
6. Bell WC, Sexton KC, Grizzle WE (2009) How to efficiently obtain human tissues to support specific biomedical research projects. Cancer Epidemiol Biomarkers Prev 18(6):1676–1679
7. McLerran D, Grizzle WE, Feng Z et al (2008) Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: sources of sample bias. Clin Chem 54(1):44–52
8. Ransohoff DF (2009) Promises and limitations of biomarkers. Recent Results Cancer Res 181:55–59
9. Ransohoff DF (2007) How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines and study design. J Clin Epidemiol 60(12):1205–1219
10. Ransohoff DF (2005) Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 5(2):142–149
11. Cadigan RJ, Lassiter D, Haldeman K et al (2013) Neglected ethical issues in biobank management: results from a U.S. study. Life Sci Soc Policy 9:1. doi:10.1186/2195-7819-9-1
12. UK Biobank, http://www.ukbiobank.ac.uk/
13. Prostate, Lung, Colon and Ovary Study (PLCO), http://prevention.cancer.gov/plco
14. The National Health and Nutrition Examination Survey (NHANES), http://www.cdc.gov/nchs/nhanes.htm
15. Pirkle JL, Brody DJ, Gunter EW et al (1994) The decline in blood lead levels in the United States: the National Health and Nutrition Examination Surveys (NHANES). JAMA 272(4):284–291
16. Clausen KC, Grizzle WE, LiVolsi VA et al (1987) Availability of human tissues for research in cancer. Science 237(4810):10–11
17. LiVolsi VA, Clausen KP, Grizzle WE et al (1993) The cooperative human tissue network. An update. Cancer 71(4):1391–1394
18. Hewitt R, Grizzle WE, Watson P et al. (in press) Biobanking networks through 5 continents. In: Hainaut P, Vaught J Pasterk M, Zatloukal K (eds.), Biobanking of human biospecimens: principles and practices, New York, NY: Springer (in press)

19. Agilent Technologies, http://www.genomics.agilent.com/en/product.jsp?cid=AG-PT-106&_requestid=48027
20. Hatzis C, Sun H, Yao H et al (2011) Effects of tissue handling on RNA integrity and microarray measurements from resected breast cancers. J Natl Cancer Inst 103:1871–1883
21. De Cecco L, Musella V, Veneroni S et al (2009) Impact of biospecimens handling on biomarker research in breast cancer. BMC Cancer 9:409. doi:10.1186/1471-2407-9-409
22. Musella V, Verderio P, Reid JF et al (2013) Effects of warm ischemic time on gene expression profiling in colorectal cancer tissues and normal mucosa. PLoS One 8(1):e53406. doi:10.1371/journal.pone.0053406
23. Bray SE, Paulin FEM, Fong SC et al (2010) Gene expression in colorectal neoplasia: modifications induced by tissue ischemic time and tissue handling protocol. Histopathology 56:240–250
24. Opitz L, Salinas-Riester G, Grade M et al (2010) Impact of RNA degradation on gene expression profiling. BMC Med Genomics 3:36. doi:10.1186/1755-8794-3-36
25. Micke P, Ohshima M, Tahmasebpoor S et al (2006) Biobanking of fresh frozen tissue: RNA is stable in nonfixed surgical specimens. Lab Invest 86:202–211
26. Strand C, Enell J, Hedenfalk I et al (2007) RNA quality in frozen breast cancer samples and the influence on gene expression analysis – a comparison of three evaluation methods using microcapillary electrophoresis traces. BMC Mol Biol 8:38. doi:10.1186/1471-2199-8-38
27. Jewell SD, Srinivasan M, McCart LM et al (2002) Analysis of the molecular quality of human tissues: an experience from the Cooperative Human Tissue Network. Am J Clin Pathol 118(5):733–741
28. Ohashi Y, Creek KE, Pirisi L et al (2004) RNA degradation in human breast tissue after surgical removal: a time-course study. Exp Mol Pathol 77:98–103
29. Dumur CI, Sana S, Ladd AC et al (2008) Assessing the impact of tissue devitalization time on genome-wide gene expression analysis in ovarian tumor samples. Diagn Mol Pathol 17(4):200–206
30. Borgan E, Navon R, Vollan HKM et al (2011) Ischemia caused by time to freezing induces systematic microRNA and mRNA responses in cancer tissue. Mol Oncol 5:564–576
31. Spruessel A, Steimann G, Jung M et al (2004) Tissue ischemia time affects gene and protein expression patterns within minutes following surgical tumor excision. Biotechniques 36(6):1030–1037
32. Grizzle WE, Srivastava S, Manne U (2012) Translational pathology of neoplasia. In: Srivastava S, Grizzle WE (eds) Translational pathology of early cancer. IOS Press BV, Amsterdam, pp 7–20
33. Gündisch S, Hauck S, Sarioglu H et al (2012) Variability of protein and phosphoprotein levels in clinical tissue specimens during the preanalytical phase. J Proteome Res 11:5748–5762
34. Khoury T, Sait S, Hwang H et al (2009) Delay to formalin fixation effect on breast biomarkers. Mod Pathol 22:1457–1467
35. Apple S, Pucci R, Lowe AC (2011) The effect of delay in fixation, different fixatives, and duration of fixation in estrogen and progesterone receptor results in breast carcinoma. Am J Clin Pathol 135:592–598
36. Qju J, Kulkarni S, Chandrasekhar R et al (2010) Effect of delayed formalin fixation on estrogen and progesterone receptors in breast cancer: a study of three different clones. Am J Clin Pathol 134:813–819
37. Pekmezci M, Szpaderska A, Osipo C (2012) The effect of cold ischemia time and/or formalin fixation on estrogen receptor, progesterone receptor, and human epidermal growth factor receptor-2 results in breast carcinoma. Patholog Res Int 2012, 947041. doi:10.1155/2012/846941
38. Teicher I, Tinker MA, Auguste LJ et al (1985) Effect of operative devascularization on estrogen and progesterone receptor levels in breast cancer specimens. Surgery 98(4):784–791
39. Start RD, Cross SS, Clelland C et al (1992) Delay in fixation does not affect the immunoreactivity of proliferating cell nuclear antigen (PCNA). J Pathol 168:197–199
40. Khoury T (2012) Delay to formalin fixation alters morphology and immunohistochemistry for breast carcinoma. Appl Immunohistochem Mol Morphol 20(6):531–542
41. Di Tommaso L, Kapucuoglu N, Losi L et al (1999) Impact of delayed fixation on evaluation of cell proliferation in intracranial malignant tumors. Appl Immunohistochem Mol Morphol 7(3):209–213
42. Portier BP, Wang Z, Downs-Kelly E et al (2013) Delay to formalin fixation 'cold ischemia time': effect on ERBB2 detection by in-situ hybridization and immunohistochemistry. Mod Pathol 26(1):1–9. doi:10.1038/modpathol.2012.123
43. Cacciatore S, Hu X, Viertler C et al (2013) Effects of intra- and post-operative ischemia on

the metabolic profile of clinical liver tissue specimens monitored by NMR. J Proteome Res 12:5723–5729

44. Eltoum I, Fredenburgh J, Myers RB et al (2001) Introduction to the theory and practice of fixation of tissues. J Histotechnol 24(3): 173–190
45. Eltoum I, Fredenburgh J, Grizzle WE (2001) Advanced concepts in fixation: effects of fixation on immunohistochemistry, reversibility of fixation and recovery of proteins, nucleic acid, and other molecules from fixed and processed tissues, developmental methods of fixation. J Histotechnol 24(3):201–210
46. Grizzle WE, Fredenburgh J, Myers RB (2007) Fixation of tissues. In: Bancroft J, Gamble M (eds) Theory and practice of histological techniques, 3rd edn. Churchill Livingstone, Edinburgh, pp 53–74
47. Otali D, Stockard CR, Oelschlager DK et al (2009) Combined effects of formalin fixation and tissue processing on immunorecognition. Biotech Histochem 84(5):223–247
48. Otali D, He Q, Stockard CR et al (2013) Preservation of immunorecognition by transferring cells from 10% neutral buffered formalin to 70% ethanol. Biotech Histochem 88(3-4):170–180
49. Otali D, He Q, Grizzle WE (2013) The effect of antigen retrieval on cells fixed in 10% neutral buffered formalin followed by transfer to 70% ethanol. PLoS One 8(12):e82405. doi:10.1371/journal.pone.0082405
50. De Marzo AM, Fedor HH, Gage WR (2002) Inadequate formalin fixation decreases reliability of $p27^{Kip1}$ immunohistochemical staining: probing optimal fixation time using high-density tissue microarrays. Hum Pathol 33:756–760
51. National Institute of Standards and Technology, http://www.nist.gov/
52. Katepogu K, Chittor P, Kurumala D et al (2013) Freeze and thaw creates oxidative stress and DNA damage in frozen human spermatozoa. Indo Am J Pharm Res 3(9):7184–7191
53. Röder B, Fruhwirth K, Vogl C et al (2010) Impact of long-term storage on stability of standard DNA for nucleic acid-based methods. J Clin Microbiol 48(11):4260–4262
54. Fernández-Soto P, Tirado VV, Rodriguez CC (2013) Long-term frozen storage of urine samples: a trouble to get PCR results in Schistosoma spp. DNA detection? PLoS One 8(4):e61703. doi:10.1371/journal.pone.0061703
55. Auer H, Mobley J, Ayers L et al (2014) The effects of frozen tissue storage conditions on the integrity of RNA and protein. Biotech Histochem 89(7):518–528
56. Steg A, Wang W, Blanquicett C et al (2006) Multiple gene expression analyses in paraffin-embedded tissues by Taqman low density array: application to Hedgehog and Wnt pathway analysis in ovarian endometrioid adenocarcinoma. J Mol Diagn 8(1):76–83
57. Steg A, Vickers SM, Eloubeidi M et al (2007) Hedgehog pathway expression in heterogeneous pancreatic adenocarcinoma: implications for the molecular analysis of clinically available biopsies. Diagn Mol Pathol 16(4):229–237
58. Potter DM, Butterfield LH, Divito SJ et al (2012) Pitfalls in retrospective analyses of biomarkers: a case study with metastatic melanoma patients. J Immunol Methods 376(1-2): 108–112
59. Manne U, Myers RB, Srivastava S et al (1997) Re: Loss of tumor marker-immunostaining intensity on stored paraffin slides of breast cancer. J Natl Cancer Inst 89(8):585–586
60. Otali D, Maston V, Oelschlager DK et al. (2015) Loss of immunorecognition of selected molecules during long-term storage of paraffin blocks. To be presented at the 2015 AACR Annual Meeting, April 18–22, 2015, Philadelphia, PA
61. McNally LR, Manne U, Grizzle WE (2013) Post-transcriptional processing of genetic information and its relation to cancer. Biotech Histochem 88(7):365–372
62. Varley KE, Gertz J, Roberts BS et al (2014) Recurrent read-through fusion transcripts in breast cancer. Breast Cancer Res Treat 146(2):287–297
63. Garcia PL, Council LN, Christein JD et al (2013) Development and histopathological characterization of tumorgraft models of pancreatic ductal adenocarcinoma. PLoS One 8(10):e78183. doi:10.1371/journal.pone.0078183
64. Grizzle WE, Semmes OJ, Bigbee W et al (2005) The need for the review and understanding of SELDI/MALDI mass spectrometry data prior to analysis. Cancer Inform 1(1):86–97
65. Henderson GE, Cadigan RJ, Edwards TP et al (2013) Characterizing biobank organizations in the U.S.: results from a national survey. Genome Med 5(1):3. doi:10.1186/gm407
66. Grizzle WE, Sexton KC, Bell WC (2008) Quality assurance in tissue resources supporting biomedical research. Cell Preserv Technol 6(2):113–118
67. Grizzle WE, Gunter EW, Sexton KC et al (2015) Quality management of biorepositories. Biopreserv Biobank 13:183
68. Gunter EW, McQuillan GM (1990) Quality control in planning and operating the

laboratory component of the Third National Health and Nutrition Examination Survey (NHANES III). J Nutr 120:1451–1454

69. Gunter EW, Lewis BL, Konchikowski SM (1996) Laboratory methods used for the Third National Health and Nutrition Examination Survey (NHANES III), 1988-1994. As contained in NHANES III Reference Manuals and Reports, NCHS CD-ROM. Centers for Disease Control and Prevention, Hyattsville, MD
70. Cooperative Human Tissue Network (CHTN), http://www.chtn.nci.nih.gov/
71. Tuck MK, Chan DW, Chia D et al (2009) Standard operating procedures for serum and plasma collection: Early Detection Research Network consensus statement. Standard Operating Procedure Integration Working Group. J Proteome Res 8(1):113–117
72. Wenk RE (1998) Mechanism of interference by hemolysis immunoassays and requirements for sample quality. Clin Chem 44(12):2554
73. International Air Transport Association (IATA), http://www.ups.com/content/us/en/resources/ship/hazardous/biological_substances.html
74. United States Department of Transportation (DOT), http://www.ncsu.edu/ehs/dot/Bio_shipping.pdf
75. Aamodt RL, Anouna A, Baird P et al (2005) Best practices for repositories I: collection, storage and retrieval of human biological materials for research. Cell Preserv Technol 3:5–48
76. Pitt K, Campbell L, Skubitz A et al (2008) Best practices for repositories: collection, storage, retrieval and distribution of biological materials for research. Cell Preserv Technol 6(1):5–58
77. Campbell L, Betsou F, Garcia DL et al (2012) Best practices for repositories: collection, storage, retrieval, and distribution of biological materials for research. 3rd edn. Biopreserv Biobank 10(2):1–85
78. Office of Biorepositories and Biospecimen Research, National Cancer Institute: Best Practices for Biospecimen Resources, http://biospecimens.cancer.gov/bestpractices/
79. Canadian Tumour Repository Network (CTRNet), https://www.ctrnet.ca
80. International Organization for Standardization (ISO) – ISO 9001 and ISO 9001:2008, http://www.iso.org/iso/home/standards/management-standards/iso_9000.htm
81. Federal Drug Administration's (FDA) Good Tissue Practices (GTP, 21 CFR, Part 1271), http://www.fda.gov/BiologicsBloodVaccines/SafetyAvailability/ReportaProblem/GoodTissuePracticesGTPReporting/default.htm.
82. Federal Drug Administration's Good Manufacturing Practices, http://www.fda.gove/Drugs/DevelopmentApprovalProcess/Manufacturing/ucm169105.htm
83. Grizzle WE, Woodruff KH, Trainer TD (1996) The pathologist's role in the use of human tissues in research - legal, ethical, and other issues. Arch Pathol Lab Med 120(10):909–912
84. Aamodt R, Barnes M, Bledsoe MJ et al. (2007) Report of the Public Responsibility in Medicine and Research (PRIMR) Human Tissue/Specimen Banking Working Group. Part II: Tools for Investigators, IRBs and Repository Managers. The PRIM&R Human Tissue/Specimen Banking Working Group
85. Bledsoe MJ, Grizzle WE (2013) Use of human specimens in research: the evolving US regulatory, policy, and scientific landscape. Diagn Pathol 19(9):322–330
86. Hewitt R, Watson PH, Dhir R et al (2009) Timing of consent for the research use of surgically removed tissue: is postoperative consenting acceptable? Cancer 115(1):4–9
87. Grizzle WE, Knoppers BM, Zeps N et al (2011) What are the most oppressing legal and ethical issues facing biorepositories and what are some strategies to address them? Biopreserv Biobank 9(4):317–319
88. Bledsoe MJ, Grizzle WE, Clark BJ et al (2012) Practical implementation issues and challenges for biobanks in the return of individual research results. Genet Med 14(4):478–483
89. Bledsoe MJ, Clayton EW, McGuire AL et al (2013) Return of research results from genomic biobanks: cost matters. Genet Med 15(2):103–105
90. Bledsoe MJ, Clayton EW, McGuire AL et al (2013) Return of research results from genomic biobanks: a call for data. Genet Med 15(2):159–160
91. Grizzle WE, Polt SS (1988) Guidelines to avoid personnel contamination by infective agents in research laboratories that use human tissues. J Tissue Cult Methods 11(4):191–200
92. Grizzle WE, Fredenburgh J (2001) Avoiding biohazards in medical, veterinary and research laboratories. Biotech Histochem 76(4):183–206
93. Grizzle WE, Bell W, Fredenburgh J (2005) Safety in biomedical and other laboratories. In: Patrinos G, Ansorg W (eds) Molecular diagnostics. Academic Press (Elsevier), New York, NY, pp 421–428
94. Grizzle WE, Bell WC, Fredenburgh J (2010) General considerations concerning safety in biomedical research laboratories. In: Patrinos GP, Ansorg W (eds) Molecular diagnostics,

2nd edn. Academic Press (Elsevier), New York, NY, pp 563–572

95. Grizzle WE, Fredenburgh J, Sexton K et al. (in press) Safety issues in biorepositories. In: Hainaut P, Vaught J, Pasterk M, Zatloukal K. Human specimens: principle and practice. New York, NY: Springer (in press)
96. National Fire Protection Association (NFPA), NFPA Code-30, Flammable and Combustible LiquidsCode,http://www.nfpa.org/codes-and-standards/document-information-pages?mode=code&code=30
97. Edgerton ME, Morrison C, LiVolsi VA et al (2008) A standards based ontological approach to information handling for use by organizations providing human tissue for research. Cancer Inform 6:127–136
98. Edgerton ME, Grizzle WE, Washington MK (2010) The deployment of a tissue request tracking system for the CHTN: a case study in managing change in informatics for biobanking operations. BMC Med Inform Decis Mak 10:32
99. Gaston SM, Soares MA, Siddiqui MM et al (2005) Tissue-print and print-phoresis as platform technologies for the molecular analysis of human surgical specimens: mapping tumor invasion of the prostate capsule. Nat Med 11(1):95–101

Chapter 2

Tissue Microdissection

Anja Rabien and Glen Kristiansen

Abstract

The new opportunities of modern assays of molecular biology can only be exploited fully if the results can be accurately correlated to the tissue phenotype under investigation. This is a general problem of non-in situ techniques, whereas results from in situ techniques are often difficult to quantify. The use of bulk tissue, which is not precisely characterized in terms of histology, has long been the basis for molecular analysis. It has, however, become apparent, that this simple approach is not sufficient for a detailed analysis of molecular alterations, which might be restricted to a specific tissue phenotype (e.g., tumor or normal tissue, stromal or epithelial cells). Microdissection is a method to provide minute amounts of histologically characterized tissues for molecular analysis with non-in situ techniques and has become an indispensable research tool. If tissue diversity is moderate and negligible, manual microdissection can be an easy and cost-efficient method of choice. In contrast, the advantage of laser microdissection is a very exact selection down to the level of a single cell, but often with a considerable time exposure to get enough material for the following analyses. The latter issue and the method of tissue preparation needed for laser microdissection are the main problems to solve if RNA, highly sensitive to degradation, shall be analyzed. This chapter focuses on optimized procedures for manual microdissection and laser microdissection to analyze RNA of malignant and nonmalignant prostate tissue.

Key words Manual microdissection, Laser microdissection, RNA, Degradation, Cryosection, Prostate tissue, Cresyl violet

1 Introduction

In contrast to cell culture material, which is fairly homogenous, organic tissues have a far more complex architecture. Moreover, they are composed of fundamentally different cell types, e.g., epithelial cells, connective tissue, vessels, and various inflammatory cells. The quality of a tissue lysate for molecular analysis clearly depends on the tissue composition and the percentage of cells of interest in the total lysate. This is the more important since the final result (e.g., an expression profile or a protein concentration) cannot be clearly correlated to a tissue compartment. The only way

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_2,

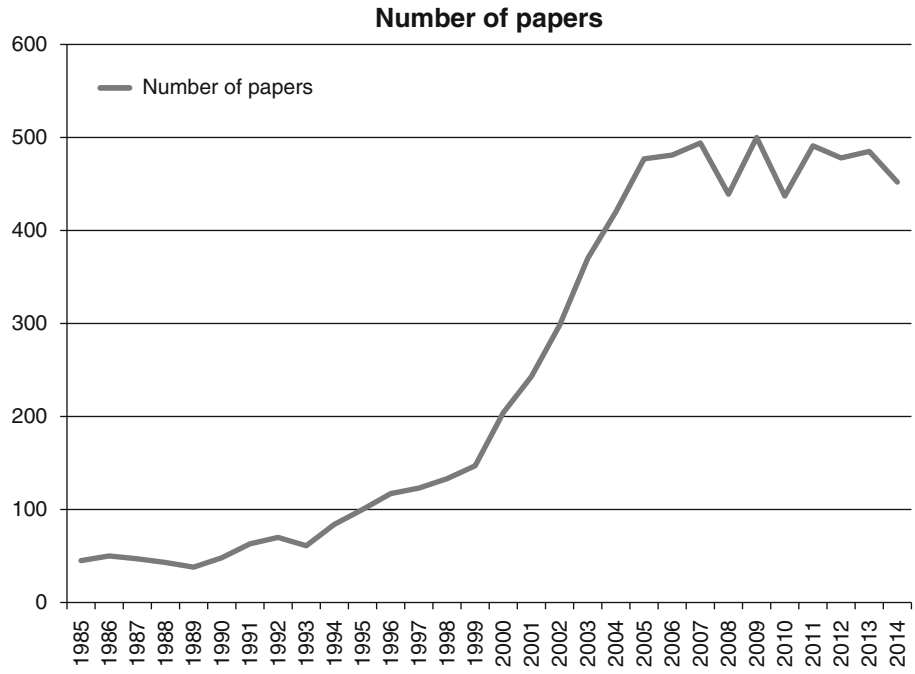


Fig. 1 Number of Pubmed-listed publications found by the search term, "microdissection" between 1985 and 2014. A steep increase can be seen in the late 1990s, a plateau with approximately 494 papers was reached in 2007 and remains relatively stable since

to solve this problem of expression-phenotype correlation is to sort the cells after characterization into various compartments or to even select single cells or cell types and to analyze them separately, using microdissection [1–5]. The published use of microdissection has enormously increased in the last decade, as Fig. 1 illustrates.

A contamination-free and very exact selection is the main advantage of laser-based microdissection, while manual microdissection is cheaper and often time-saving, simply gaining a more heterogeneous material. Often the analysis of tissue that can be assigned a single histological diagnosis (normal tissue, tumor tissue of a certain type or grade) is sufficient and requires manual microdissection only. This simple technique allows histologically characterizing and procuring tissue compartments as small as 1 mm^2. We have successfully used manual microdissection for the analysis of breast cancer and prostate cancer, which is a particularly heterogeneous neoplasm, characterized by the coexistence of morphologically diverse tumor growth patterns that are being mirrored in the Gleason grading system [6]. We found the following protocol for manual microdissection very helpful, as simple as it is.

Since the 1990s, laser microdissection systems are commercially available to precisely collect the material in focus up to single cells. The principle of laser microdissection is simple, cutting

microscopically selected tissue by a laser (ultraviolet—UV—or infrared), but the technology behind remains complex so that a laser microdissection system is still expensive to acquire. There are three main systems commercially available:

1. After cutting, cells can be catapulted into a collection tube with the PALM microlaser system (www.palm-microlaser.com) [7, 8].
2. Selected cells are transferred onto a film with the Arcturus laser capture microdissection system (www.moleculardevices.com) [9–11] or to a special cap by the patented mmi Isolation Cap technology from Molecular Machines & Industries (www.molecular-machines.com) [12].
3. Cut samples fall down into the lid of a collection tube, driven by gravitation, as provided by Leica Microsystems (www.leica-microsystems.com) [13, 14].

A short comparison of the laser microdissection techniques is given from Murray in 2007 [15] and different applications are given in Chaps. 2 and 3 of Methods of Molecular Biology 576 [16]. We use the Leica laser microdissection system (Fig. 2) to analyze expression of mRNA in prostate epithelium of adenocarcinoma,

Fig. 2 Laser Microdissection System Leica DMLA. The Leica CTR MIC electronics box is placed slightly distant from the microscope (*left side*). To the *right*, you see the Smartmove control in front and the laser cartridge behind. Pictures are transmitted from the camera (*on top*) to a computer providing the LMD software

intraepithelial neoplasia, "normal" epithelium, and atrophic glands without the surrounding stromal tissue. Distinction between prostatic epithelium and stroma is also important because genomic as well as proteomic analyses revealed significant expression differences [17, 18]. Since analyses of mRNA require a permanent protection from degradation during the whole procedure, we optimized our techniques to obtain high-quality mRNA from laser microdissected tissue.

2 Materials

2.1 Manual Microdissection

2.1.1 Tissue Sampling

1. Most crucial, when dealing with fresh tissues is the immediate cryoconservation, which should ideally take place in the operating theatre straight after removal of the tissue. Since this can be difficult to organise, alternatively the tissue can be transferred in a plastic bag on water ice from the operating theatre to the frozen section laboratory of the pathology department for further processing.
2. After grossing and sectioning of the organ/tissue by the pathologist a 5 mm slice of the region of interest, e.g., grossly tumor suspicious area is placed between two conventional glass slides and immediately immersed in liquid nitrogen for 1–2 min. Since glass slides often break off the frozen tissue, which is inconvenient, alternatively reusable metal slides can be used, which also have the advantage of an even faster freezing performance.
3. Wrapped in labeled aluminium foil these frozen slices can be stored at −80 °C for years.

2.1.2 Sectioning and Staining

1. Mount the deep frozen tissue slice to the specimen holder of the cryotome, either sterile saline or conventional OCT embedding medium can be used (Fig. 3, *centre*).
2. Cut serial sections, thickness can be modified according to the research question. For RNA extraction we have usually cut 30 sections at 12 μm, including thinner control sections at 4 μm after the 1st, the 10th, the 20th, and the 30th section for H and E staining. For DNA extraction, thinner sections [3–6], which yield a better morphology, can be sufficient.
3. Mount the sections on sterile glass slides (room temperature to easily take up the frozen tissue), and then leave to air-dry in the cryotome chamber (Fig. 3, *left-hand side*).
4. Process the thin sections for H&E staining as usual (hematoxylin staining, eosin staining, gradual dehydration, cover slipping).
5. Sort the sections for microdissection into a glass staining bench prior to staining.

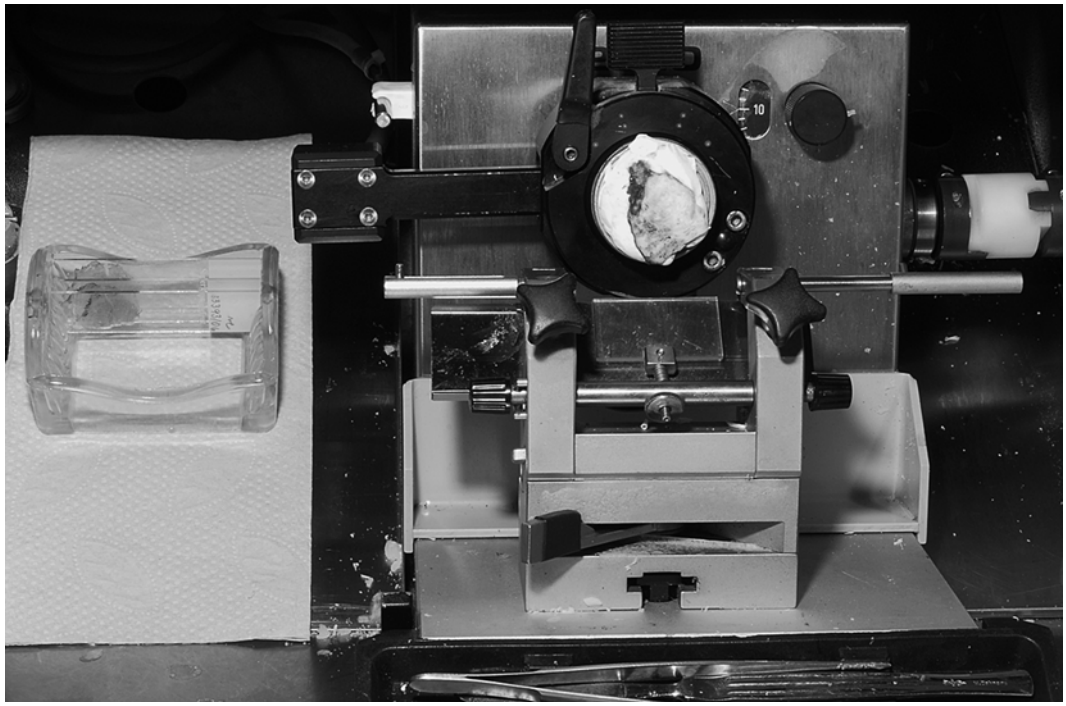


Fig. 3 View into the freezing chamber of a cryotome. In the *center*, the mounted tissue can be seen. Next to it, on the *left*, is a glass bench with the sections mounted on glass slides

6. Stain with hematoxylin for 1–2 min at 4 °C, and then immerse in cold (!) tap water (stored in the fridge) for another 2 min (*see* **Note 1**). This will yield a mild nuclear staining for microscopy.
7. Drop off superfluous water and store in a specimen box at −80 °C until microdissection.

2.2 Laser Microdissection

1. Liquid nitrogen, dry ice.
2. 2-Methylbutane.

2.2.1 Preparation of Cryosections

3. Membrane glass slides coated with polyethylene naphthalate (PEN) membrane 2.0 μm (Leica, Wetzlar, Germany).
4. Fully equipped cryostat.
5. Embedding medium: Jung, Leica OCT Cryocompound.
6. Superfrost Plus glass slides.

2.2.2 Staining Procedure and Storage

1. Cresyl violet acetate.
2. Mayer's acidic Hemalum solution.
3. Desiccator.
4. Eosin solution, 1 % alcoholic.
5. Absolute ethanol, ACS grade.
6. Xylol, ACS grade.
7. Mounting medium.

2.2.3 Laser Microdissection

1. Laser Microdissection System Leica DMLA (Leica, Fig. 2).
2. Software LMD 5.0 (Leica).
3. PCR tubes, 0.5 ml.
4. Lysis buffer: RNeasy Lysis Buffer (Qiagen, Hilden, Germany)/1 % β-mercaptoethanol.

3 Methods

3.1 Manual Microdissection

1. Mark on the H&E-stained control sections appropriate regions of interest, for example normal tissue, tumor tissue, and hyperplastic tissue with a water-resistant pen. This is illustrated in Fig. 4: On the left-hand side, an H&E-stained slide with markings is seen. On the right-hand side a serial section slide is shown after microdissection of the respective areas.
2. The most important prerequisite for manual microdissection is a person with a calm and skilled hand in order to retrieve the marked areas from the frozen tissue slides using conventional

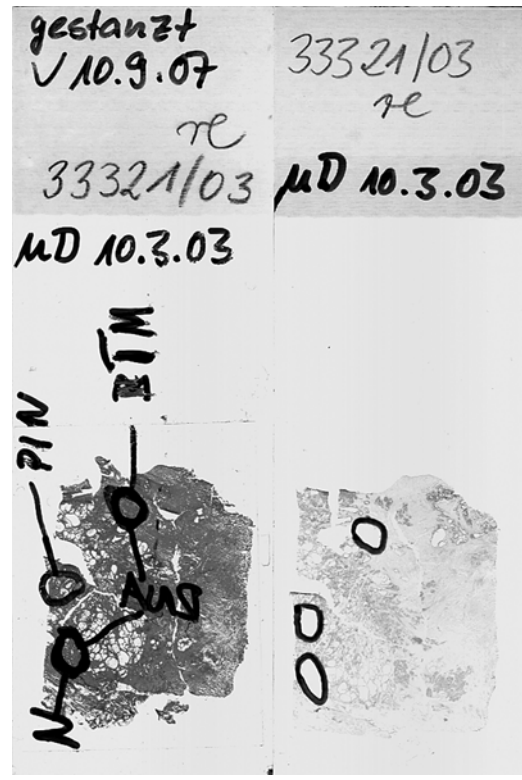


Fig. 4 Comparison of the H&E-stained control slide (*left*) with an adjacent tissue section (*right*) stained with hematoxylin alone after microdissection. The *marked areas of the left slide* have been removed on the *right slide*

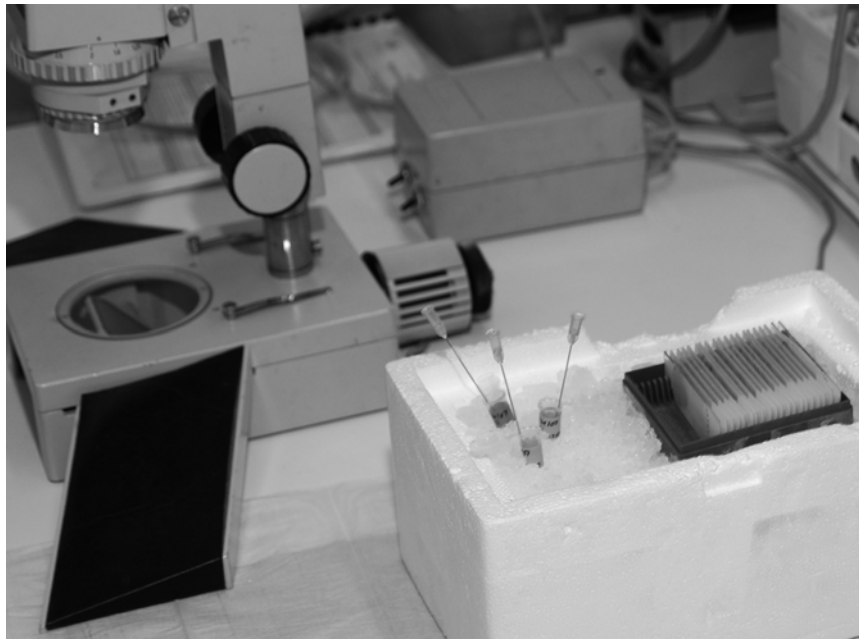


Fig. 5 Typical set up for microdissection: microscope, slides, and sample tubes on ice

sterile injection needles. This also best accomplished in a secluded laboratory to ensure a concentrated workflow. Technically, we found a cheap binocular microscope, commonly used in biological laboratories with a scanning magnification of 10×–60× comfortable to work with.

3. Put the box with the deep frozen hematoxylin-stained tissue sections on water ice next to the microscope. The typical work setting is shown in Fig. 5.
4. Compare every specimen under the microscope with the marked control slide and identify the respective areas of interest.
5. Retrieve these tissue areas using sterile injection needles, using one needle per area to avoid contamination (Fig. 6).
6. The retrieved tissue has to be transferred immediately into labelled (e.g., Eppendorf) tubes filled with lysis buffer. Once in lysis buffer, RNases should be inhibited.
7. Store at −80 °C or process lysates as appropriate for your assay.

3.2 Laser Microdissection

3.2.1 Preparation of Cryosections

1. The whole procedure is to be performed with gloves and material with a high degree of purity to avoid contamination with RNases.
2. As soon as possible after surgery, a slice of tissue is shock frozen in methylbutane in a bath of liquid nitrogen (*see* **Note 2**) and stored at −80 °C.

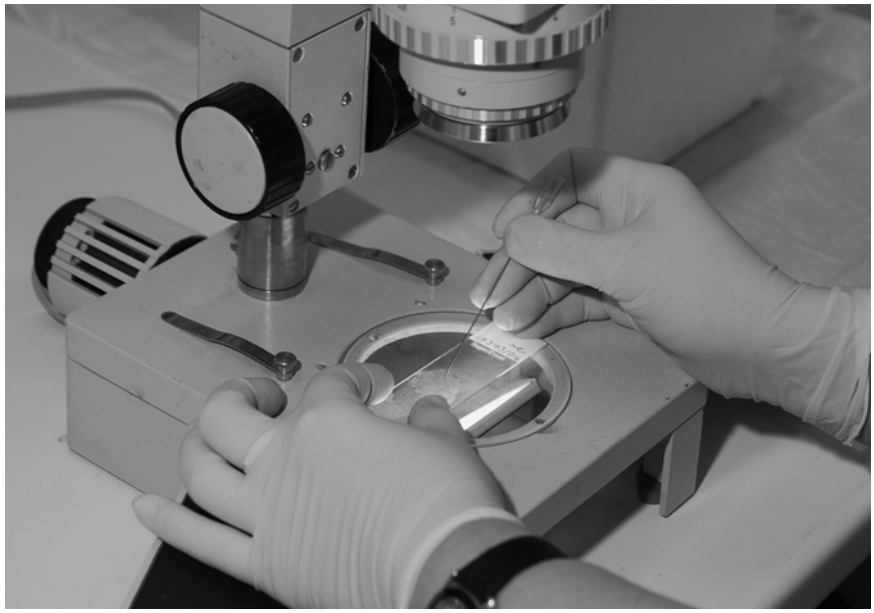


Fig. 6 The process of manual microdissection: one hand holds the glass slide, the other retrieves the regions of interest with an injection needle from the slide. Time is critical, for after thawing, the slide will dry within minutes

3. Before cryosectioning, PEN glass slides are UV irradiated under a sterile hood for cell culture for at least 30 min to inhibit RNase activity (*see* **Note 3**).
4. Specimen is allowed to acclimate to about –23 °C (prostate tissue, *see* **Note 2**) in the cryostat for 10–15 min and is embedded in Jung medium on a metallic plate. The tissue is fixed and cut with a microtome blade. Cut slices of 5 μm (*see* **Note 4**) are placed onto the PEN membrane of the slides which are immediately stored in a box in dry ice up to staining or storage at –80 °C. Control sections to be stained with hemalum/eosin are mounted on Superfrost Plus glass slides.

3.2.2 Staining Procedure and Storage

1. 1 % (w/v) cresyl violet acetate (*see* **Note 5**) is dissolved in absolute ethanol at room temperature overnight in a shaker. The solution is filtrated before use (pore size 0.2 μm).
2. Cryosections are air-dried for 1 min on ice (not necessary after storage at –80 °C) and fixed in –20 °C precooled 75 % ethanol for 2 min. Excessive ethanol is knocked off on an absorbent paper.
3. The slides are dipped into the 1 % cresyl violet acetate solution at room temperature for 20 s. Excessive staining solution is knocked off on an absorbent paper.
4. The slides are briefly dipped into 75 % ethanol before incubation for 30 s in 100 % ethanol. Excessive ethanol is knocked off

on an absorbent paper. The tissue is air-dried for 10 min at room temperature.

5. The slides are stored in a desiccator in the dark, at least for 90 min before usage (*see* **Note 5**).
6. Control slides are stained in hemalum solution for 5–10 min, incubated in warm tap water for 5 min, and rinsed with distilled water. After counterstaining with eosin for 2–5 min, the slides are rinsed in tap water and in distilled water. They are dipped into 70 % ethanol, 80 % ethanol, 96 % ethanol, 3× absolute ethanol (can be denatured), 2× xylol, and for at least 5 min in the third bath of xylol. The tissue is mounted with mounting medium (*see* **Note 6**) and covered by a 24 mm×40 mm glass. The slides are completely dried after 2 days.

3.2.3 Microdissection

The laser microdissection system should be explained and configured for the customer by the technical service of the supplier. In case of the Leica microdissection system, the electronics box is placed to the left of the microscope (Fig. 2). The system is connected to a computer providing the LMD software.

1. The Leica CTR MIC electronics box is turned on before starting the computer. The laser cartridge has a separate switch (Fig. 2).
2. A collection tube is fixed in the tube holder (Fig. 7). The cap of the collection tube is filled with 70 µl lysis buffer/1 % β-mercaptoethanol (*see* **Note 7**). When inserting (beginning obliquely from the right or the left side), the collection device (*see* **Note 8**) must snap into the mounting brackets.
3. The slide, thoroughly dried in the desiccator, is clamped in the specimen holder with the PEN membrane to the bottom (Fig. 8).
4. By use of the LMD Software, the holder with the collection tube can be chosen. Magnification, lamp brightness, movement, and focusing are adjusted with the Smartmove control (Fig. 2).
5. The laser is calibrated ("Laser," "Calibrate") utilizing an area of the PEN membrane without tissue.
6. The cells to be excised are selected at the monitor by drawing a line around and cutting ("Draw + Cut," Fig. 9, *see* **Note 9**). The line can be closed automatically ("Close Line"). Several pieces can be marked and are cut consecutively ("Multiple Shapes," Fig. 9). Bridges of tissue are eliminated by cutting while drawing ("Move + Cut"). The corresponding hemalum/eosin stained slide is used to better discriminate between different types of morphology.
7. Correct location of the pieces of tissue should be checked (Fig. 10); therefore switch from "Specimen" to "Collector."

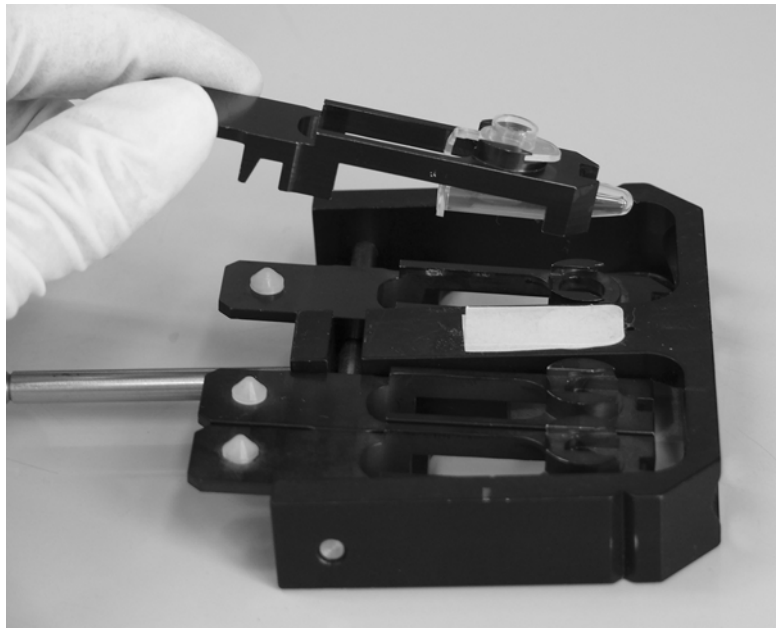


Fig. 7 Collection device with fourfold holder. From below, an opened collection tube is put through the big hole of a holder. After pushing the cap into its retainers, the tube is fixed underneath. The holder is placed in the collection device. Note covering of the "No Cap" position in the middle with white stickers (*see* **Note 8**)

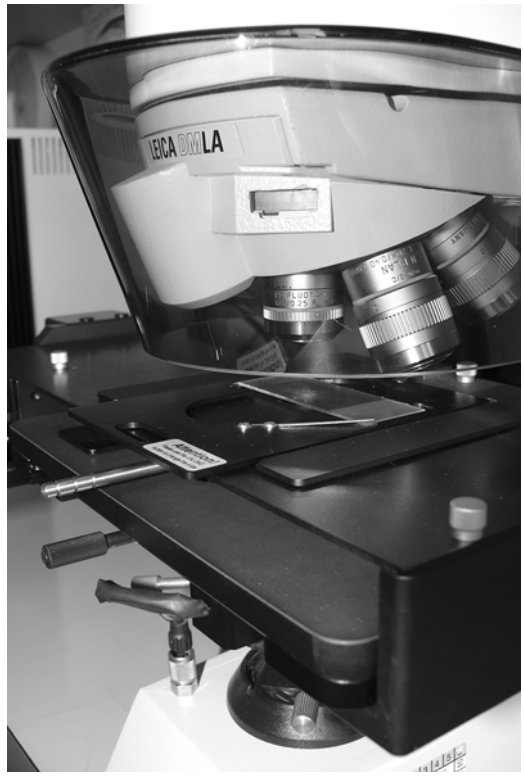


Fig. 8 Close up view of the microdissection stage. The slide is fixed with a cleat upside down in the specimen holder. Underneath, a cover plate occludes the motorized collection device. A UV stray light shield around the holder of the lenses protects from laser radiation

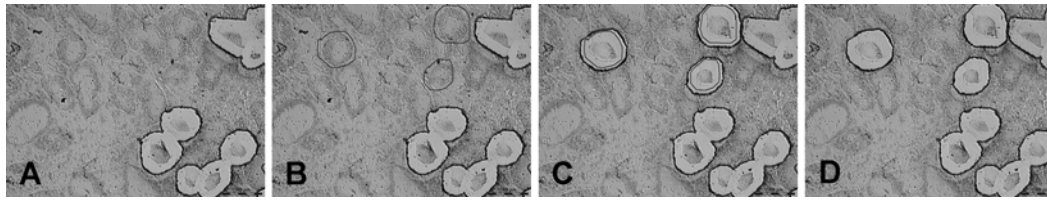


Fig. 9 Cut series of prostate tissue. (**a**) 200-fold magnification is used to select epithelium of prostatic adenocarcinoma. Cuts are shown to the *right*. (**b**) Further glands are selected by drawing ("Multiple Shape"). (**c**) The selected tissue is cut. Note that the laser beam burns a considerable band around. (**d**) The *marked lines* are removed by clearing the list of cut areas

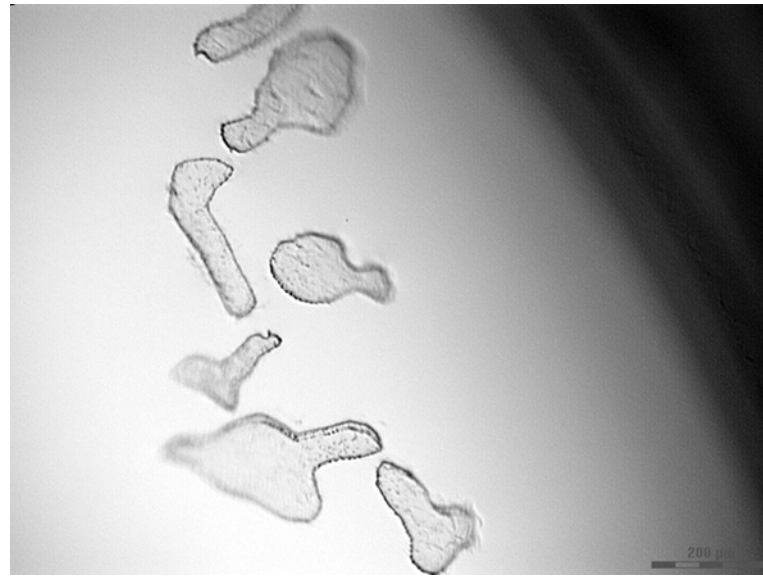


Fig. 10 Cap control. Cut tissue can be seen swimming in the buffer the cap is filled with (100-fold magnification)

8. Areas and object numbers (*see* **Note 9**) from one run of cuts are listed and should be exported, e.g., to Microsoft Office Excel, because the list must be cleared to avoid repeated cutting of the same areas.
9. To remove the slide and the collection device (*see* **Note 10**), click "Unload" so that the cap of the tube is protected from hitting.
10. The tube is carefully removed from the holder as follows: Detach the tube, then inch out the lid (Do not turn!) and put the tube onto the cap for closing. Until use, tubes should be stored at −80 °C.

4 Notes

1. We found it astonishing to learn that even the use of unsterile tap water (for blueing the sections) did not markedly diminish the RNA quality of the results. We have also attempted DEPC pre-treated sterilized solutions for all these staining steps, but since this did not significantly improve RNA integrity, we kept our protocol simple.
2. For mRNA analysis, we recommend cryoconservation instead of formalin fixation and paraffin embedding of the tissue, as degradation of RNA is much more likely using the latter method. The tissue can also be shock frozen in liquid nitrogen, but methylbutane in a bath of liquid nitrogen (−160 °C) is better conserving. For cryosectioning, the temperature of the cryostat should be optimized according to the instrument and to the type of tissue, sometimes also from sample to sample.
3. Alternatively to PEN membrane slides, polyethylene terephthalate (PET)-coated slides can be used, but they need a metallic frame (Leica). UV-irradiated membrane slides can be stored dust-free, dry in the dark for up to 1 week.
4. Per patient, we used five sections of only 5 μm, because prostate tissue proved to be difficult to cut due to inserts and consistency. Nevertheless, sections of up to 20 μm can be cut with the Leica microdissection system. Current systems have higher laser power and higher repetition rates than ours.
5. In comparison to methyl green staining (DAKO, Hamburg, Germany) and hemalum staining (Hollborn), cresyl violet staining resulted in a better RNA quality measured with a 2100 bioanalyzer (Agilent Technologies, Palo Alto, CA). Stained cryosections can be stored dust-free, dry (in the desiccator) in the dark for up to 1 week.
6. The hemalum/eosin controls can be mounted with an organic mounting medium, alternatively to Eukitt, e.g., Vitro-Clud, R. Langenbrinck, Emmendingen, Germany, or Entellan, Merck, Darmstadt, Germany.
7. RNeasy Lysis Buffer (Qiagen) with 1 % β-mercaptoethanol can be stored for up to 4 weeks.
8. To avoid light effects, the "No Cap" position of the collection device can be masked translucently (e.g., with white stickers, Fig. 7), and the eyepieces of the microscope can be covered.
9. Adjust the laser ("Laser Control") for the magnification you want to cut with to obtain a comfortable speed and beam, avoiding bridges but saving tissue. The laser beam burns a considerable band of the tissue (Fig. 9). For a first general survey, we used the 2.5× lens, but excision was optimal with the 20×

lens to discriminate within the heterogeneous prostatic tissue. Information about the number of cut cells (objects) is available if the diameter of an object is given under "Settings" and "Object Counting." For our prostate tissue we calculated 20 μm per cell. Program settings for a distinct type of tissue can be saved and restored in the next session ("Restore Application Configuration"). Images of the tissue can be saved in a database (IM500).

10. We recommend not to exceed 1–1.5 h of cutting for one slide, depending on air humidity, to keep the tissue dry. To collect epithelium of the heterogeneous prostate tissue, we need at least five tubes of each type of tissue. A single cell should average 10–15 pg of total RNA (among them 1–3 % mRNA).

Acknowledgments

G.K. is grateful to Britta Beyer and Eva Polzin[+] for sectioning and performing the manual microdissection in his lab. G.K. also thanks Christoph Weber for excellent photography and Alfred E. Neumann for fruitful discussions. A.R. is very grateful to Cornelia Stelzer and Sabine Becker for technical assistance and photography for the laser microdissection part.

References

1. Montironi R, Mazzucchelli R, Scarpelli M (2003) Molecular techniques and prostate cancer diagnostic. Eur Urol 4:390–400
2. Gillespie JW, Ahram M, Best CJ, Swalwell JI, Krizman DB, Petricoin EF, Liotta LA, Emmert-Buck MR (2001) The role of tissue microdissection in cancer research. Cancer J 7:32–39
3. Xu LL, Stackhouse BG, Florence K, Zhang W, Shanmugam N, Sesterhenn IA, Zou Z, Srikantan V, Augustus M, Roschke V, Carter K, McLeod DG, Moul JW, Soppett D, Srivastava S (2000) PSGR, a novel prostate-specific gene with homology to a G protein-coupled receptor, is overexpressed in prostate cancer. Cancer Res 60:6568–6572
4. Dahl E, Kristiansen G, Gottlob K, Klaman I, Ebner E, Hinzmann B et al (2006) Molecular profiling of laser-microdissected matched tumor and normal breast tissue identifies karyopherin alpha2 as a potential novel prognostic marker in breast cancer. Clin Cancer Res 12:3950–3960
5. Grützmann R, Foerder M, Alldinger I, Staub E, Brümmendorf T, Röpcke S (2003) Gene expression profiles of microdissected pancreatic ductal adenocarcinoma. Virchows Arch 443:508–517
6. Kristiansen G, Pilarsky C, Wissmann C, Kaiser S, Bruemmendorf T, Roepcke S et al (2005) Expression profiling of microdissected matched prostate cancer samples reveals CD166/MEMD and CD24 as new prognostic markers for patient survival. J Pathol 205:359–376
7. Micke P, Ostman A, Lundeberg J, Ponten F (2005) Laser-assisted cell microdissection using the PALM system. Methods Mol Biol 293:151–166
8. Schutze K, Niyaz Y, Stich M, Buchstaller A (2007) Noncontact laser microdissection and catapulting for pure sample capture. Methods Cell Biol 82:649–673
9. Espina V, Milia J, Wu G, Cowherd S, Liotta LA (2006) Laser capture microdissection. Methods Mol Biol 319:213–229
10. Espina V, Wulfkuhle JD, Calvert VS, VanMeter A, Zhou W, Coukos G et al (2006) Laser-capture microdissection. Nat Protoc 1:586–603
11. Espina V, Heiby M, Pierobon M, Liotta LA (2007) Laser capture microdissection technology. Expert Rev Mol Diagn 7:647–657

12. Anslinger K, Mack B, Bayer B, Rolf B, Eisenmenger W (2005) Digoxigenin labelling and laser capture microdissection of male cells. Int J Legal Med 119:374–377
13. Kolble K (2000) The LEICA microdissection system: design and applications. J Mol Med 78:B24–B25
14. Vega CJ (2008) Laser microdissection sample preparation for RNA analyses. Methods Mol Biol 414:241–252
15. Murray GI (2007) An overview of laser microdissection technologies. Acta Histochem 109: 171–176
16. Grützmann R, Pilarsky C (eds) (2003) Methods in molecular biology 576. Humana, Totowa, NJ, 2005–2008
17. Ornstein DK, Gillespie JW, Paweletz CP, Duray PH, Herring J, Vocke CD et al (2000) Proteomic analysis of laser capture microdissected human prostate cancer and in vitro prostate cell lines. Electrophoresis 21:2235–2242
18. Richardson AM, Woodson K, Wang Y, Rodriguez-Canales J, Erickson HS, Tangrea MA et al (2007) Global expression analysis of prostate cancer-associated stroma and epithelia. Diagn Mol Pathol 16:189–197

Chapter 3

Tissue Microarrays

Ana-Maria Dancau, Ronald Simon, Martina Mirlacher, and Guido Sauter

Abstract

Modern next-generation sequencing and microarray technologies allow for the simultaneous analysis of all human genes on the DNA, RNA, miRNA, and methylation RNA level. Studies using such techniques have lead to the identification of hundreds of genes with a potential role in cancer or other diseases. The validation of all of these candidate genes requires in situ analysis of high numbers of clinical tissues samples. The tissue microarray technology greatly facilitates such analysis. In this method minute tissue samples (typically 0.6 mm in diameter) from up to 1000 different tissues can be analyzed on one microscope glass slide. All in situ methods suitable for histological studies can be applied to TMAs without major changes of protocols, including immunohistochemistry, fluorescence in situ hybridization, or RNA in situ hybridization. Because all tissues are analyzed simultaneously with the same batch of reagents, TMA studies provide an unprecedented degree of standardization, speed, and cost efficiency.

Key words TMA, Tissue microarrays, High-throughput in situ analysis, IHC, Immunohistochemistry, FISH, Fluorescence in situ hybridization, Translational research

1 Introduction

The demand for analyses of newly discovered genes in diseased tissues, especially human tumors, has grown massively during the last years. To identify the most significant ones among all the emerging candidate cancer genes, it is desirable to analyze many genes in a significant number of well-characterized tumors. Hundreds of tumors must often be analyzed per gene to generate statistically meaningful results. This leads to a massive workload in involved laboratories. Moreover, traditional analysis of multiple genes results in a critical loss of precious tissue materials since the number of conventional tissue sections that can be taken from a tumor block does usually not exceed 200–300. The tissue microarray (TMA) technology does significantly facilitate and accelerate tissue analyses by in situ technologies [1, 2]. When this technology was developed in 1997, the term "microarray" was generally used for small structures organized in an array-like fashion. However, with the advent of DNA-array technologies, such as cDNA arrays

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_3,

or oligonucleotide arrays, the term microarray has more and more become specific for a kind of array made from homogeneous spots that have been placed on a glass surface by automated arraying machines, i.e., array spotters. It is to note that the TMA technology is substantially different from such "spotted" arrays, and represents miniaturized pathology that requires pathologist's skills for analysis in first place.

In this method, minute tissue cylinders (typical diameter: 0.6 mm) are removed from hundreds of different primary tumor blocks and subsequently brought into one empty "recipient" paraffin block. Sections from such array blocks can then be used for simultaneous in situ analysis of hundreds to thousands of primary tumors on the DNA [3], RNA [4], and protein level [5]. The cylindrical shape and the small diameter of the specimen taken out of the donor block maximizes the number of samples that can be taken out of one donor block and minimizes the tissue damage inferred to it. Studies have shown that tissue samples with a diameter of as little as 0.6 mm allow a reliable analysis and yield representative data for research and possibly also for diagnostic purposes [6, 7]. The possibility of using such small tissue cores is important for pathologists, since they can now give researchers access to their material, and, at the same time, retain their tissue blocks. Punched tissue blocks remain fully interpretable for all morphological and molecular analyses that may subsequently become necessary, provided that the number of punches is reasonably selected. Dozens of punches can be taken from one tumor without compromising interpretability. Only few tissues require the use of larger tissue spots per tumor. These tissues include several normal tissues such as blood vessels where it is necessary to have the entire wall arrayed, or normal tissues that have important structures dispersed within the tissue such as glomeruli in kidney or Langerhans islets in pancreas. Inexperienced pathologists sometimes also require core diameters of >0.6 mm for cancers. However, this results in an unnecessary waste of tissue, additional study costs, and increased workload during interpretation.

Virtually all tissues are suitable to be placed into a TMA, including also frozen tissues [8, 9] and core needle biopsies [10]. Therefore, the range of TMA applications is very broad. One of the most distinct advantages of TMAs is that one set of tissues (which has been reviewed by one pathologist) with available clinical data can now be used for almost an unlimited number of studies. The TMA technique is not limited at all to cancer research, although this still is the predominant application. The normal expression pattern of gene products can optimally be tested on TMAs containing all kind of normal tissues. Similarly as for patient tissues, TMAs can be made from suspension cells [11] and other experimental tissues such as engraft tumors or tissues from animal models [12, 13]. There are also applications for

TMAs in diagnostic molecular pathology. Here, TMAs can be used as positive control sections or for the inexpensive high-throughput testing of predictive markers such as HER2 overexpr ession/amplification. In addition to the "classical" TMA technique described here, alternative techniques have been described that may be useful in specific situations, including the hypodermic needle technology [14], the resin technology [15], or the cutting-edge technology [16].

2 Materials

2.1 Sample Collection

1. Standard routine histology microscope for review of tissue sections.
2. Colored pens to mark representative areas on the slides, e.g., red for tumor, blue for normal, and black for premalignant lesions.
3. Sufficient working space especially for large-scale projects that require extensive sorting of thousands of sections and blocks.

2.2 Preparing Recipient Blocks

1. PEEL-A-WAY Embedding Paraffin Pellets, melting point: 56–58 °C (Polysciences Inc., PA, USA).
2. Slotted processing/embedding cassettes for routine histology, e.g., EMS cat. # 70070 (Electron Microscopy Sciences Inc., PA, USA).
3. Stainless steel base molds for processing/embedding systems, e.g., EMS cat. # 62510–30 (Electron Microscopy Sciences Inc., PA, USA).
4. Filter/filter papers.
5. Oven for paraffin melting (70 °C).

2.3 TMA Making

1. Pre-manufactured empty paraffin recipient blocks.
2. Illuminated magnifying lenses and supplies (e.g., Luxo U wave II/70, cat. # 27950, Luxo Inc, Switzerland) (optional).

2.4 TMA Sectioning

1. Standard routine histology microtome and supplies (e.g., Leica SM2400, Leica Microsystems Inc., IL, USA).
2. Slide label printer (e.g., DAKO Seymour glass slide labeling system, product code S3416; DAKO A/S, Denmark) or special slide marker (e.g., Securline Marker II, Precision Dynamics Corporation, CA, USA).
3. Boxes for slide storage.
4. Refrigerator for slide storage.
5. Paraffin Sectioning Aid-System (Instrumedics Inc., NJ, USA; cat. # PSA) containing Ultraviolet Curing Lamp,

Adhesive-Coated PSA Slides, TPC Solvent, TPC Solvent can, Hand roller, Tape windows (optional).

2.5 TMAs from Frozen Tissues

1. OCT Tissue-Tek compound embedding medium (Sakura BV, The Netherlands).
2. Dry ice to keep punching needles and recipient block in optimally cooled condition.
3. Freezer for frozen tissue storage (–70 °C).

3 Methods

3.1 TMA Manufacturing

3.1.1 Sample Collection

Although a device is needed to manufacture TMAs, it must be understood that most of the work (approximately 95 %) is traditional pathology work that cannot be accelerated by improved (i.e., automated) tissue arrayers. This preparatory work is similar to what is needed for traditional studies involving "large" tissue sections. The major difference is the number of tissues involved, which can be an order of magnitude higher in TMA studies than in traditional projects. The different tasks related to sample collection are described below:

1. Exactly define the TMA that is to be made (*see* **Note 1**). Include normal tissues of the organ of interest and of a selection of other organs as well.
2. Generate a list of potentially suited tissues.
3. Collect all slides from these tumors from the archive.
4. One pathologist must review all sections from all candidate specimens to select the optimal slide. If possible, tumors should be reclassified at that stage according to current classification schemes and tissue areas suited for subsequent punching should be marked (*see* **Note 2**). Different colors are recommended for marking different areas on one section (for example, red for tumor, black for carcinoma in situ, blue for normal tissue). Collect the tissue blocks that correspond to the selected slides.
5. These blocks and their corresponding marked slides must be matched and sorted in order of their appearance on the TMA.
6. Define the structure (outline) of the TMA and compose a file that contains the identification numbers of the tissues together with their locations and real coordinates (as they need to be selected on the arraying device). As a distance between the individual samples, 0.2 mm is recommended. To facilitate navigation on the TMA we recommend arranging the tissues in multiple sections (e.g., quadrants). The distance between the quadrants may be 0.8 mm (*see* **Note 3**). In most laboratories

capitalized letters define quadrants, whereas small letters and numbers define the coordinates within these quadrants. Examples of a TMA structure (outline) and data file containing the necessary information for making a TMA are given in Fig. 1 and Table 1.

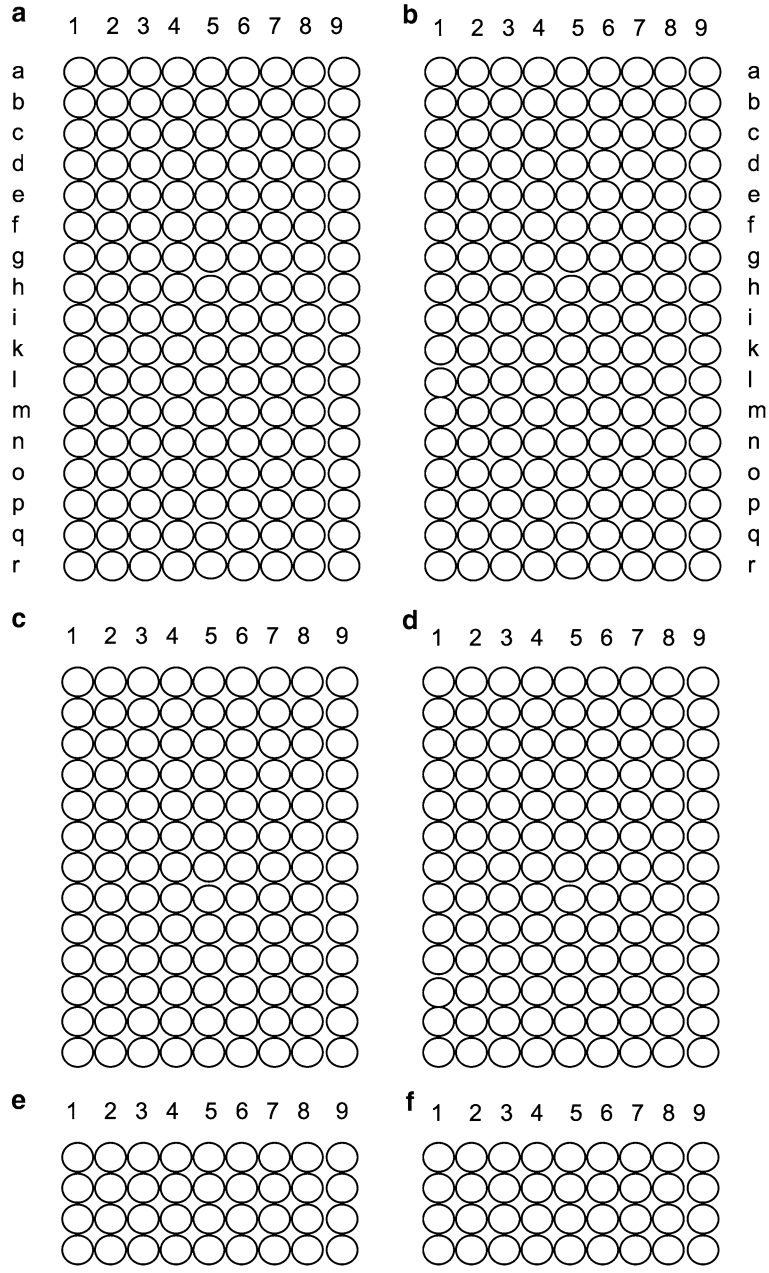


Fig. 1 TMA outline example. The division of the TMA into multiple subsections facilitates the navigation during microscopy

Table 1
Example file for TMA construction

loc	coord	loc	coord	loc	coord
A 1a	0/0	A 2a	0/800	A 3a	0/1600
A 1b	800/0	A 2b	800/800	A 3b	800/1600
A 1c	1600/0	A 2c	1600/800	A 3c	1600/1600
A 1d	2400/0	A 2d	2400/800	A 3d	2400/1600
A 1e	3200/0	A 2e	3200/800	A 3e	3200/1600
A 1f	4000/0	A 2f	4000/800	A 3f	4000/1600
A 1g	4800/0	A 2g	4800/800	A 3g	4800/1600
A 1h	5600/0	A 2h	5600/800	A 3h	5600/1600
A 1i	6400/0	A 2i	6400/800	A 3i	6400/1600
A 1k	7200/0	A 2k	7200/800	A 3k	7200/1600
A 1l	8000/0	A 2l	8000/800	A 3l	8000/1600
A 1m	8800/0	A 2m	8800/800	A 3m	8800/1600
A 1n	9600/0	A 2n	9600/800	A 3n	9600/1600
A 1o	10,400/0	A 2o	10,400/800	A 3o	10,400/1600
A 1p	11,200/0	A 2p	11,200/800	A 3p	11,200/1600
A 1q	12,000/0	A 2q	12,000/800	A 3q	12,000/1600
A 1r	12,800/0	A 2r	12,800/800	A 3r	12,800/1600

3.1.2 Preparing Recipient Blocks

In contrast to normal paraffin blocks, tissue microarray blocks are cut at room temperature (*see* **Note 4**).

1. The paraffin is melted at 60 °C, filtrated, and poured in a stainless steel mold.
2. A slotted plastic embedding cassette (as used in every histology lab) is then placed on the top of the warm paraffin.
3. Recipient paraffin blocks are then cooled down for 2 h at room temperature and for 2 additional hours at 4 °C. Blocks are then removed from the mold. It is important not to cool down the paraffin on a cooling plate because of the risk of block damage.
4. Quality check of the recipient blocks is important because they must not contain air bubbles.

Large recipient blocks (for example $30 \times 45 \times 10$ mm) are easier to handle than the small blocks (for example $25–35 \times 5$ mm) that are typically used in routine histology labs.

3.1.3 TMA Making

Only if all this preparatory work has been done, a tissue-arraying device can be employed. Several tissue-arraying systems are now commercially available (http://www.estigen.com; http://www.biegler.com, http://www.alphelys.com, http://www.veridiamtissuearrayers.com, http://www.pathologydevices.com, http://www.alphametrix.de, and http://www.chemicon.de) but many groups also use homemade tissue arrayer. Using manually operated devices, excellent TMAs can be expected only after a significant training period, mostly including several hundred, if not a few thousand punches. A patient and enduring personality as well as keen eyesight are important prerequisites for operators of the manual tissue arrayers. Automated tissue arrayers are available but these devices are expensive and they neither accelerate nor significantly improve TMA making.

The TMA manufacturing process consists of three steps that are repeated for each sample placed on the TMA:

1. Generating a hole in an empty (recipient) paraffin block.
2. Removing a cylindrical tissue sample from a donor paraffin block (*see* **Note 5**).
3. Placing the cylindrical tissue sample in the premade hole in the recipient block.

Exact positioning of the tip of the tissue cylinder at the level of the recipient block surface is crucial for the quality and the yield of the TMA block. Placing the tissue too deeply into the recipient block results in empty spots in the first sections taken from the TMA block. Positioning the tissue cylinder not deep enough causes empty spots in the last sections taken from this TMA (*see* **Note 6**).

As soon as all tissue elements are filled into the recipient block, the block is heated at 40° for 10 min

3.1.4 Array Sectioning

Regular microtome sections may be taken from TMA blocks using standard microtomes. However, the more samples a TMA block contains, the more difficult regular cutting becomes. As a consequence, the number of slides of inadequate quality increases with the size of the TMA, and in turn, fewer sections from the TMA block can effectively be analyzed.

Using a tape sectioning kit (Instrumedics) facilitates cutting and leads to highly regular non-distorted sections (ideal for automated analysis). In addition, the tape system may prevent arrayed samples from floating off the slide, if very harsh pretreatment methods are used. However, the sticky glued slides have the disadvantage of increased background signals between the tissue spots in IHC analyses. The tissue samples themselves do not show increased nonspecific background in IHC. The use of the tape sectioning system is described below:

1. An adhesive tape is placed on the TMA block in the microtome immediately before cutting.

2. A 3–5 μm section is cut. The tissue slice is now adhering to the tape.
3. The tissue slice is placed on a special "glued" slide (stretching of the tissue in a water bath or on a heating plate is not necessary).
4. The slide (tissue on the bottom) is then placed under UV light for 35 s. This leads to polymerization of the glue on the slide and on the tape.
5. Slides are placed into TPC solution (Instrumedics) at room temperature for 5–10 s. The tape can then be removed gently from the glass slide. The tissue remains on the slide.
6. Slides are dried at room temperature.

Using the tape system can cause inhomogeneous immunostainings when certain automated immunostainers are being used. In our experience, this especially applies for Ventana devices.

3.1.5 TMAs from Frozen Tissues

Fejzo and Slamon reported manufacturing of TMAs from frozen tissues using a commercially available tissue array device [17, 8].

1. Recipient blocks are made from OCT that is frozen down in a Tissue-Tek standard cryomold. The resulting OCT block is mounted on top of a plastic biopsy cassette. As long as the recipient OCT block is sized exactly like a paraffin recipient block (for which the arrayer had been constructed) no modifications of the arrayer are necessary to mount the block.
2. The recipient block must be surrounded with dry ice to prevent melting.
3. Tissue biopsies (diameter 0.6 mm; height 4–5 mm) are then punched from OCT-embedded tumor tissues and placed into the recipient OTC array block using a commercial tissue microarrayer. There are some main differences compared to the procedure described for paraffin blocks. It is important that the tissue in the needle is kept frozen during the procedure and that the needle will not be damaged (*see* **Note 7**). The frozen TMAs often become more irregular and distorted than TMAs from formalin-fixed material. Therefore a larger space between samples is recommended (e.g., 1 mm).
4. 4–10 μm sections of the whole block are cut from the array block. A cryostat microtome (Microm GmbH, Germany) can be used with or without the Basic CryoJane Tape Transfer System and slides (Instrumedics).

3.2 TMA Analysis

3.2.1 General Considerations

TMAs are suited for all types of in situ analysis methods, including immunohistochemistry (IHC), fluorescence in situ hybridization (FISH) [3], and RNA in situ hybridization (RNA-ISH) [4]. Examples of stained TMA sections are shown in Fig. 2. The most

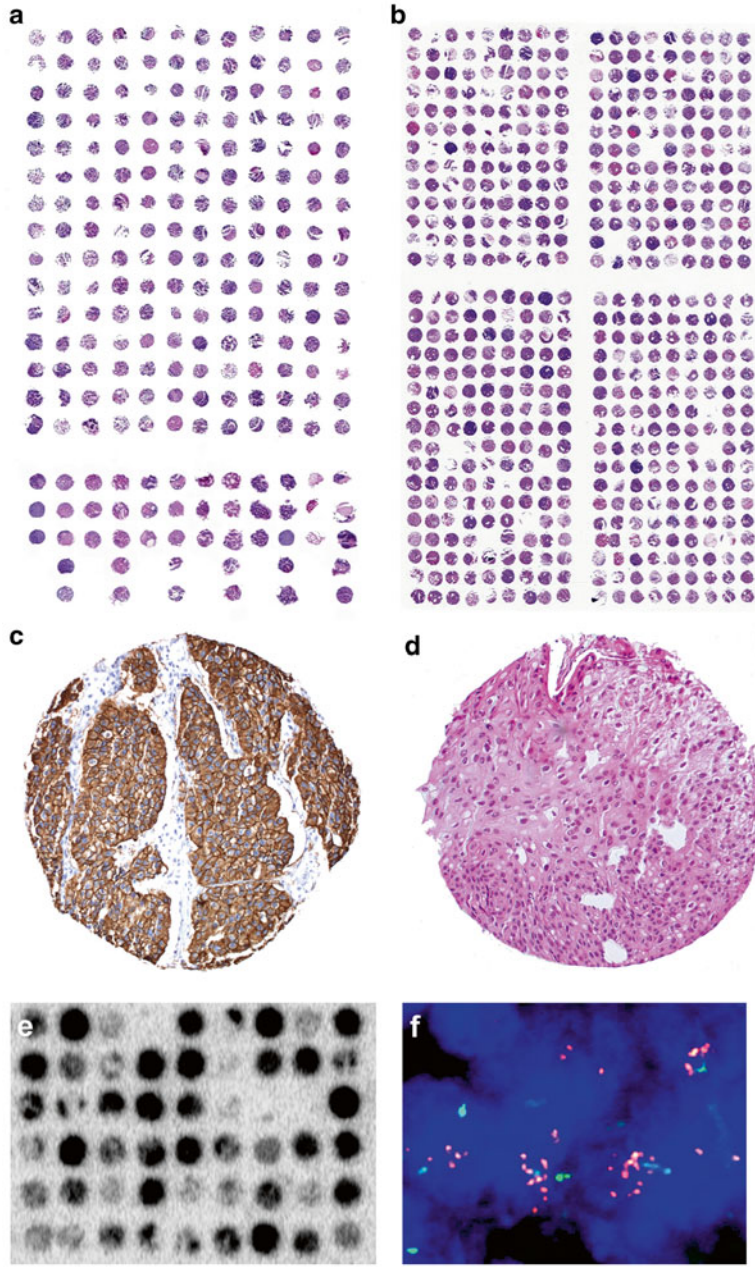


Fig. 2 Examples of stained tissue sections. Hematoxylin and eosin (H&E)-stained sections of (**a**) a TMA from frozen tissue containing 228 tissue spots. Each tissue spot measures 0.6 mm in diameter. Missing samples result from the sectioning/staining process or indicate samples that are already exhausted. Note that the spot to spot distance is larger on the frozen TMA as compared to the paraffin TMA. (**b**) A TMA from formalin-fixed, paraffin-embedded tissues containing 540 tissue spots. (**c**) A spot showing immunohistochemical analysis using an antibody directed against the Her2/neu protein in a breast cancer sample. (**d**) Magnification of an H&E-stained 0.6 mm tissue spot of a breast cancer. (**e**) RNA in situ hybridization on a frozen TMA made from normal and malignant kidney tissues. A radioactively labeled oligonucleotide was used as a probe against vimentin mRNA. The black staining intensity indicates the level of mRNA in each tissue spot. (**f**) FISH analysis of centromere 11 (*green signals*) and the CCND1 gene (*red spots*) in cell nuclei (*blue staining*) of a tissue spot (630×). The high number of CCND1 signals indicates a gene amplification

significant difference compared to traditional large section studies is the high level of standardization that can be achieved in TMA experiments. All slides of one TMA study are typically incubated in one set of reagents assuring identical concentrations, temperatures, and incubation times. Other minor variables that may have an impact on the outcome of in situ analyses such as the age of a slide (time between sectioning and use) or section thickness are also fully standardized, as long as all tissues of one study are located on the same TMA section. As a result of this unprecedented standardization within each experiment, surprising interassay variations can occur, if experiments are repeated under slightly different conditions.

3.2.2 Immunohistochemistry

In general, the same rules apply for IHC analysis on TMA as on large sections. The small size of the arrayed tissues on a TMA facilitates the staining interpretation since predefined criteria can be applied to a well-defined tissue area. This reduces interobserver variation of IHC interpretation. For many immunohistochemical tumor analyses, the following information can be recorded:

- Percentage of positive cells.
- Staining intensity (0, 1+, 2+, 3+).
- Subcellular localization of the staining (membraneous, cytoplasmatic, nuclear).
- Tissue localization of the staining (tumor cells, stroma, vessels).

For statistical analyses, tumors can be classified into three or four groups based on the percentage of positive cells and the staining intensity. For example:

Negative	No staining
Weak positivity	1+ in 1–70 % or 2+ in ≤30 % of cells
Moderate positivity	1+ in >70 %, or 2+ in 31–70 % or 3+ in ≤30 % of cells
Strong positivity	2+ in >70 % or 3+ in >30 of cells

This scoring system is based on the fact that the staining intensity is less reliable than the fraction of stained tumor cells. It is important to note that fixation-related variations of the staining intensity are an inherent feature of IHC experiments. For example, a 2+ staining intensity can indicate true medium-level protein expression under optimal fixation conditions, but might as well result from "understaining" of high-level expression in case of suboptimal fixation. In contrast, the fraction of stained tumors cells is virtually not affected by the fixation quality. Our scoring system was defined to accommodate these facts. For example, a 1+ cancer

with 70 % stained tumor cells is considered "moderate" because we consider the high fraction of stained cells more important than the low staining intensity, which might be due to poor immunoreactivity. Vice versa, a 2+ cancer with ≤30 % stained cells is considered "weak," as it has only a small fraction of positive tumors cells, although the tissue sample was probably optimally immunoreactive. However, the thresholds of 30–70 % were selected almost arbitrarily based on the experience that they provide solid results in terms of detecting significant associations between molecular markers and clinic-pathological tumor features in many previous studies.

Some of the arrayed tissues may show falsely negative or inappropriately weak IHC staining intensity due to variations in tissue processing (e.g., fixation medium and time). The large number of tissues included in a TMA will often compensate for this phenomenon, which is also encountered in large section IHC analyses. At least a fraction of tissue spots yielding false-negative IHC staining results can be identified in control experiments assessing the antigen integrity of the samples, e.g., IHC detection of tissue type-specific antigens like cytokeratins or vimentin. For tissues with a reasonable proliferative activity, Ki67 (MIB1) is an optimal quality control antibody (*see* **Note 8**).

It is highly recommended to use freshly cut sections for IHC analysis. The time span between sectioning and immunostaining should be less than 2 weeks. Studies have shown that staining intensity decreases significantly with time for many antibodies [18, 19].

3.2.3 FISH

Because biopsies are all treated individually at the time when they are removed, fixed, and subsequently paraffin-embedded, one must expect a certain degree of heterogeneity with respect to protein and nuclear acid preservation.

The proof of this assumption is best illustrated in the outcome of FISH analyses. Like in large section studies, TMA FISH analyses yield interpretable results in only about 60–90 % of the analyzed tumors (depending on the quality and size of the FISH probe) at the first attempt. Again like in large section studies, it is possible to achieve interpretability in a fraction of initially non-informative cases by changing experimental conditions. For example, an increased proteinase concentration for slide pretreatment will result in interpretable signals in some initially non-informative cases at the cost of overdigestion of some previously interpretable samples. In general, we do not attempt to improve the fraction of FISH-informative cases by changing experimental conditions. Because of the high number of tumors on our TMAs (usually >500), we rather tolerate a fraction of non-interpretable tumors than using too many precious TMA sections for additional experiments.

3.2.4 Summary

The TMA methodology is now an established and frequently used tool for tissue analysis. The equipment is affordable and easy to use in places where pathology expertise is available. Basically all kinds of in situ analyses, like IHC and DNA/RNAc in situ hybridization may be adapted to TMAs with only slight (if any) modifications of the respective large section protocols.

4 Notes

1. Often TMA users realize that one critical control tissue has been forgotten only after completion of the TMA block.
2. It is advisable to have a freshly HE stained section if the actual block surface is not well reflected on the available stained section.
3. For unequivocal identification of individual samples on TMA slides, it is important to avoid a fully symmetrical TMA structure.
4. Therefore, a special type of paraffin is needed with a melting temperature between 55 and 58 °C ("Peel-A-Way" paraffin, *see* Subheading 2).
5. The hole may also be drilled into the paraffin block. Drilling is superior to punching because less force is applied to the paraffin block, making it more stable especially when large arrays are made. We have constructed our own arrayer where we have replaced the punch needle with a drill.
6. However, a location of the tissue cylinder which is too superficial is less problematic than a position which is too deep, since protruding tissue elements can—to some extent—be leveled out after finishing the punching process. The use of a magnifying lens facilitates precise deposition of samples, especially for beginners. With the use of a glass slide protruding tissue cylinders are then gently pressed deeper into the warmed TMA block.
7. This can be done by precooling the needle with a piece of dry ice before punching and while dispensing the tissue core into the recipient block. Needles may easily bend or break. To prevent needle breakage, coring must be performed slowly with minimal pressure.
8. MIB1, which must lead to strong staining in all mitoses, is often falsely negative in suboptimally processed tissues.

References

1. Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G, Kallioniemi OP (1998) Tissue microarrays for high-throughput molecular profiling of tumor specimens. Nat Med 4(7):844–847

2. Bubendorf L, Kononen J, Koivisto P, Schraml P, Moch H, Gasser TC, Willi N, Mihatsch MJ, Sauter G, Kallioniemi OP (1999) Survey of gene amplifications during prostate cancer progression by high-throughout fluorescence in situ hybridization on tissue microarrays. Cancer Res 59(4):803–806
3. Kallioniemi A (2010) DNA copy number analysis on tissue microarrays. Methods Mol Biol 664:127–134
4. Veeck J, Dahl E (2010) RNA expression analysis on formalin-fixed paraffin-embedded tissues in TMA format by RNA in situ hybridization. Methods Mol Biol 664:135–150
5. Simon R, Mirlacher M, Sauter G (2010) Immunohistochemical analysis of tissue microarrays. Methods Mol Biol 664:113–126
6. Hoos A, Cordon-Cardo C (2001) Tissue microarray profiling of cancer specimens and cell lines: opportunities and limitations. Lab Invest 81(10):1331–1338
7. Tennstedt P, Koster P, Bruchmann A, Mirlacher M, Haese A, Steuber T, Sauter G, Huland H, Graefen M, Schlomm T, Minner S, Simon R (2012) The impact of the number of cores on tissue microarray studies investigating prostate cancer biomarkers. Int J Oncol 40(1):261–268
8. Fejzo MS, Slamon DJ (2010) Tissue microarrays from frozen tissues-OCT technique. Methods Mol Biol 664:73–80
9. Hu Z, Chang E, Hodeib M (2010) An alternative technology to prepare tissue microarray using frozen tissue samples. Methods Mol Biol 664:81–91
10. Datta MW, Kajdacsy-Balla AA (2010) Tissue microarrays from biopsy specimens. Methods Mol Biol 664:103–111
11. Zhao S, Natkunam Y (2010) Building "tissue" microarrays from suspension cells. Methods Mol Biol 664:93–101
12. Simon R, Struckmann K, Schraml P, Wagner U, Forster T, Moch H, Fijan A, Bruderer J, Wilber K, Mihatsch MJ, Gasser T, Sauter G (2002) Amplification pattern of 12q13-q15 genes (MDM2, CDK4, GLI) in urinary bladder cancer. Oncogene 21(16):2476–2483
13. Abbott RT, Tripp S, Perkins SL, Elenitoba-Johnson KS, Lim MS (2003) Analysis of the PI-3-Kinase-PTEN-AKT pathway in human lymphoma and leukemia using a cell line microarray. Mod Pathol 16(6):607–612
14. Pires AR, de Souza SR (2010) Hypodermic needle without recipient paraffin block technique. Methods Mol Biol 664:53–61
15. Howat WJ, Wilson SJ (2010) Resin technologies: construction and staining of resin TMAs. Methods Mol Biol 664:63–72
16. Tran TH, Lin J, Sjolund AB, Utama FE, Rui H (2010) Protocol for constructing tissue arrays by cutting edge matrix assembly. Methods Mol Biol 664:45–52
17. Schoenberg Fejzo M, Slamon DJ (2001) Frozen tumor tissue microarray technology for analysis of tumor RNA, DNA, and proteins. Am J Pathol 159(5):1645–1650
18. Bertheau P, Cazals-Hatem D, Meignin V, de Roquancourt A, Verola O, Lesourd A, Sene C, Brocheriou C, Janin A (1998) Variability of immunohistochemical reactivity on stored paraffin slides. J Clin Pathol 51(5):370–374
19. Jacobs TW, Prioleau JE, Stillman IE, Schnitt SJ (1996) Loss of tumor marker-immunostaining intensity on stored paraffin slides of breast cancer. J Natl Cancer Inst 88(15):1054–1059

Chapter 4

Gene Expression Analysis in the Age of Mass Sequencing: An Introduction

Christian Pilarsky, Lahiri Kanth Nanduri, and Janine Roy

Abstract

During the last years the technology used for gene expression analysis has changed dramatically. The old mainstay, DNA microarray, has served its due course and will soon be replaced by next-generation sequencing (NGS), the Swiss army knife of modern high-throughput nucleic acid-based analysis. Therefore preparation technologies have to adapt to suit the emerging NGS technology platform. Moreover, interpretation of the results is still time consuming and employs the use of high-end computers usually not found in molecular biology laboratories. Alternatively, cloud computing might solve this problem. Nevertheless, these new challenges have to be embraced for gene expression analysis in general.

Key words Next-generation sequencing, RNA-seq

1 Introduction

Gene expression analysis (transcriptomics) is widely practiced in modern molecular biology. DNA microarrays were used to analyze the gene expression of various samples. Although large studies were successfully performed to identify differential gene expression patterns between tumor and normal tissues using microarrays [1–4], microarray-based transcriptomics has several disadvantages which include cross-hybridization between genes of similar sequence [5, 6]; unreliability in detecting all genes, especially those with low level of expression [7]; and the lack of information about the exact length and the sequence of RNAs being analyzed. However, these problems can be overcome using NGS. Next-generation sequencing is a converging technology like the smartphone and it enables researchers to analyze nucleic acid sequence and its abundance on a common platform. Whereas NGS was primarily used in the first few years to understand the mutanome of a given disease or the inter/intra species variation, it is now expanding its technological reach into other fields of nucleic acid-based research. A major driver of this

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_4, © Springer Science+Business Media New York 2016

phenomenon is the reduction in costs of the analysis of a human genome indicated by the term "1000$ genome" [8]. Moreover with the human cancer genome project, existing NGS systems are underutilized and its potential can be used for other ventures too [9]. NGS uses the known approach for sequencing cDNA libraries to identify the transcribed part of the genome, the so-called expressed sequence tags, but generates libraries in vitro excluding the tedious step of subcloning [10]. This is a radical change in the field of gene expression analysis and therefore RNA sequencing (RNA-Seq) has regained interest in the scientific community for transcriptomic studies. The increasing utility of NGS is demonstrated by its increased occurrence in scientific manuscripts, while gene expression analysis based on Affymetrix GeneChips decreases (in 2010 Pubmed contained 219 citations for the term "GeneChip" and 131 for "RNA-Seq," in 2014 the numbers were 143 and 1623, respectively).

RNA-Seq gives the opportunity to investigate the changes in the nucleic acid sequences in several aspects. It can detect somatic mutations and is effective in the discovery of recurrent mutations in cancer [11]. RNA-Seq can additionally be used for gene expression profiling with high sensitivity, since the detection rate of transcripts is only limited by the number of reads produced. However, as every available RNA molecule is sequenced the gene expression profiling is biased to highly expressed genes. In contrast, older techniques like microarrays tend to limit the signal of highly expressed genes as the generated expression profiles overrepresent the genes with intermediate expression [12, 13].

Finally, RNA-Seq is a one-stop shop for the determination of mutations, allele frequency, RNA splicing, discovery of new classes of RNA molecules, and gene expression (mRNA, but also all types of noncoding RNAs). Furthermore, another very interesting feature of RNA-Seq is the possibility to detect chimeric RNA molecules. For the very first time researchers can identify translocations using this type of information. Translocations are notoriously hard to identify, while the partners of the translocations are not known, conventional cytogenetic analysis lacks the resolution to identify these translocations. Therefore, chimeric RNA analysis might produce a landscape of translocations in cancers with unprecedented resolution [14].

One major driver of gene expression in cancer is the hypermethylation of DNA in the coding areas of the human genome [15, 16]. It can be assumed that hypermethylation of genes is an early event in tumorigenesis [17]. Nowadays hypermethylation in a genome can easily be analyzed using bead based arrays like the Illumina's Infinium HumanMethylation450 BeadChip which measures the methylation status of over 450,000 sites [18–20]. Additionally, NGS can also be used to identify new markers and gain insights into the basic changes that take place during tumor development by interrogating the methylome [21–23]. Although the sequencing of the complete human methylome is still

expensive and time intensive, the anticipated reduction in costs will make human genome-wide methylation analysis feasible for a wider range of studies.

Despite changes in the technological landscape, the key challenge of gene expression analysis however lies in the use of the right type of tissue. Characterization of the tissue in question by classic histopathology in combination with microdissection is a good starting point, but the protocols for sequencing minute amounts of tissue samples have to be improved. Microdissection can be performed manually or with a laser-based system. While manual microdissection gives the researcher an opportunity to produce samples for analysis much faster but with lesser accuracy, the selection of instruments can be important for projects using laser assisted microdissection [24, 25]. Furthermore, the combination of microdissection and modern molecular analysis identifies a large number of gene expression changes. This leads to the question about the clonality and evolution of tumors. This makes it worthwhile to analyze parts of the cancer tissue indicating that the whole tumor analysis may be a part of the molecular biology heritage [26]. Another challenge is the interpretation of the large amount of data created by NGS. Right now, trained bioinformaticians using centralized hardware with a set of independent programs for different steps of analysis are needed (Fig. 1). As an

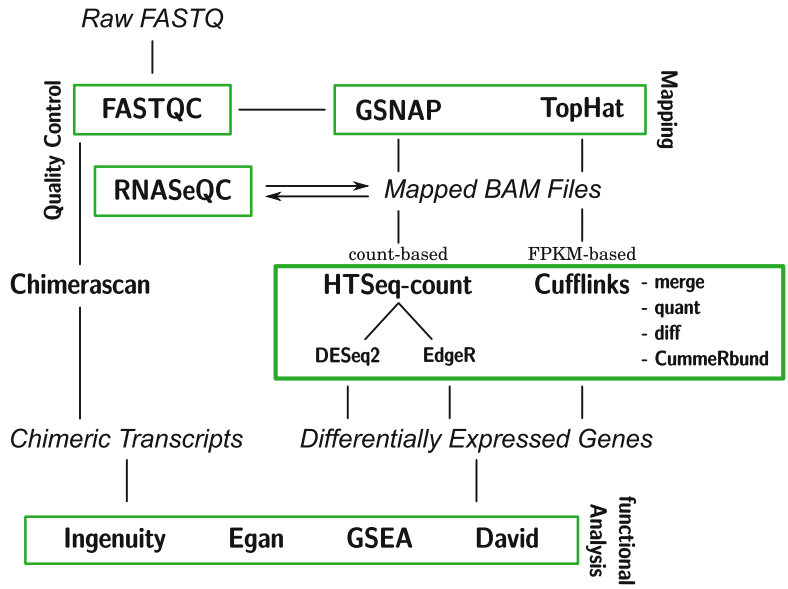


Fig. 1 A general work flow for the analysis of differential gene expression of RNA-Seq data after sequencing. Raw data are processed in the FASTQ file format (a text-based file format including the sequence and assigned quality scores). After quality control the sequences are mapped to the genome (different mappers will produce different results) and differential gene expression can be assigned. A separate routine is used for the assignment of chimeric RNAs

Table 1
A small compendium of the most used software for RNA-Seq analysis

Software	Use	Website
FastQC	Quality control of FASTQ Files	http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
RNASeQC	Quality control mapping	http://www.broadinstitute.org/cancer/cga/rna-seqc
GSNAP	Mapping	http://research-pub.gene.com/gmap/
TopHat	Mapping with bowtie	http://ccb.jhu.edu/software/tophat/index.shtml
HTSeq-count	Summarization	http://www-huber.embl.de/users/anders/HTSeq/doc/count.html
Cufflinks	Summarization, differential expression analysis	http://cole-trapnell-lab.github.io/cufflinks/
DESeq2	Differential expression analysis	via bioconductor.org
EdgeR	Differential expression analysis	via bioconductor.org
Cummerbund	Gene set and graphical analysis	http://compbio.mit.edu/cummeRbund/ via bioconductor.org
Ingenuity	Pathway analysis	http://www.ingenuity.com/
Egan	Pathway analysis	http://akt.ucsf.edu/EGAN/
DAVID	Functional annotation	http://david.abcc.ncifcrf.gov/
GSEA	Gene Set Enrichment Analysis	http://www.broadinstitute.org/gsea/

alternative, professional companies have sprung up to cope with the number crunching using expert software (Table 1).

Despite the massive technological developments, the analysis of RNA-Seq data still requires a high amount of computing power. The computational complexity of the RNA-Seq pipeline (Fig. 1) does not only depend on the hard- or software used, but also on the sequencing depth. Table 2 gives an overview of the running time of crucial steps in the analysis of RNA-Seq data. *Chimerascan*, a software for the identification of chimeric transcripts, includes an

Table 2
Execution time for different RNA-Seq mappers and differential analysis pipelines on Dell PowerEdge E710 server with 2× Intel Xeon Processor X5650 2,67GHz, 6 Cores each and 144GB RAM

Mapping	*Tophat*				*GSNAP*			
# Reads	30 Mio, 75 bp single end		100 Mio, 100 bp paired end		30 Mio, 75bp single end		100 Mio, 100 bp paired end	
# Threads	18	5	18	5	18	5	18	5
Time (hh:mm)	02:21	03:02	17:58	30:16	00:20	00:46	24:06	43:26
Differential analysis	*FPKM based*				*Count based*			
# Reads	30 Mio, 75 bp, single end, 5 vs. 5							
#Threads	5							
Time	~7 h				~5 h			

internal mapping step like *tophat*; therefore their running times correlate. Overall, it takes around 15–37 h for the analysis of 10 single-end RNA-Seq datasets with 30 million reads (75 bp length) using five threads from initial FASTQ files to a list of differentially expressed genes.

In conclusion, to generate meaningful and reliable data based on the questions that were raised beforehand and to use the available resources, as economically as possible bioinformaticians should work closely with their scientific partner.

References

1. Grützmann R, Boriss H, Ammerpohl O, Lüttges J, Kalthoff H, Schackert HK, Klöppel G, Saeger HD, Pilarsky C (2005) Meta-analysis of microarray data on pancreatic cancer defines a set of commonly dysregulated genes. Oncogene 24(32):5079–5088. doi:10.1038/sj.onc.1208696
2. Alldinger I, Dittert D, Peiper M, Fusco A, Chiappetta G, Staub E, Lohr M, Jesnowski R, Baretton G, Ockert D, Saeger HD, Grutzmann R, Pilarsky C (2005) Gene expression analysis of pancreatic cell lines reveals genes overexpressed in pancreatic cancer. Pancreatology 5(4–5):370–379. doi:10.1159/000086537
3. Grutzmann R, Pilarsky C, Ammerpohl O, Luttges J, Bohme A, Sipos B, Foerder M, Alldinger I, Jahnke B, Schackert HK, Kalthoff H, Kremer B, Kloppel G, Saeger HD (2004) Gene expression profiling of microdissected pancreatic ductal carcinomas using high-density DNA microarrays. Neoplasia 6(5):611–622. doi:10.1593/neo.04295
4. Volinia S, Calin GA, Liu C-G, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM (2006) A microRNA expression signature of human solid tumors defines cancer gene targets. Proc Natl Acad Sci U S A 103(7):2257–2261. doi:10.1073/pnas.0510565103
5. Okoniewski MJ, Miller CJ (2006) Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. BMC Bioinformatics 7:276. doi:10.1186/1471-2105-7-276
6. Royce TE, Rozowsky JS, Gerstein MB (2007) Toward a universal microarray: prediction of gene expression through nearest-neighbor probe sequence identification. Nucleic Acids Res 35(15):e99. doi:10.1093/nar/gkm549
7. Klebanov L, Yakovlev A (2007) How high is the level of technical noise in microarray data? Biol Direct 2:9. doi:10.1186/1745-6150-2-9

8. Kedes L, Campany G (2011) The new date, new format, new goals and new sponsor of the Archon Genomics X PRIZE competition. Nat Genet 43(11):1055–1058. doi:10.1038/ng.988
9. Ledford H (2015) End of cancer-genome project prompts rethink. Nature 517(7533):128–129. doi:10.1038/517128a
10. Adams MD, Dubnick M, Kerlavage AR, Moreno R, Kelley JM, Utterback TR, Nagle JW, Fields C, Venter JC (1992) Sequence identification of 2,375 human brain genes. Nature 355(6361):632–634. doi:10.1038/355632a0
11. Shah SP, Köbel M, Senz J, Morin RD, Clarke BA, Wiegand KC, Leung G, Zayed A, Mehl E, Kalloger SE, Sun M, Giuliany R, Yorida E, Jones S, Varhol R, Swenerton KD, Miller D, Clement PB, Crane C, Madore J, Provencher D, Leung P, DeFazio A, Khattra J, Turashvili G, Zhao Y, Zeng T, Glover JNM, Vanderhyden B, Zhao C, Parkinson CA, Jimenez-Linan M, Bowtell DDL, Mes-Masson A-M, Brenton JD, Aparicio SA, Boyd N, Hirst M, Gilks CB, Marra M, Huntsman DG (2009) Mutation of FOXL2 in granulosa-cell tumors of the ovary. N Engl J Med 360(26):2719–2729. doi:10.1056/NEJMoa0902542
12. Mastrokolias A, Dunnen JT, Ommen GB, Hoen PAC, Roon-Mom WMC (2012) Increased sensitivity of next generation sequencing-based expression profiling after globin reduction in human blood RNA. BMC Genomics 13(1):28. doi:10.1186/1471-2164-13-28
13. Mooney M, Bond J, Monks N, Eugster E, Cherba D, Berlinski P, Kamerling S, Marotti K, Simpson H, Rusk T, Tembe W, Legendre C, Benson H, Liang W, Webb CP (2013) Comparative RNA-Seq and microarray analysis of gene expression changes in B-cell lymphomas of Canis familiaris. PLoS One 8(4):e61088. doi:10.1371/journal.pone.0061088
14. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale A-L (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 12(1)
15. Kulis M, Esteller M (2010) DNA methylation and cancer. Adv Genet 70:27–56. doi:10.1016/B978-0-12-380866-0.60002-2, B978-0-12-380866-0.60002-2 [pii]
16. Nakhasi HL, Lynch KR, Dolan KP, Unterman RD, Feigelson P (1981) Covalent modification and repressed transcription of a gene in hepatoma cells. Proc Natl Acad Sci U S A 78(2):834–837
17. Wissmann C, Wild PJ, Kaiser S, Roepcke S, Stoehr R, Woenckhaus M, Kristiansen G, Hsieh J-C, Hofstaedter F, Hartmann A, Knuechel R, Rosenthal A, Pilarsky C (2003) WIF1, a component of the Wnt pathway, is down-regulated in prostate, breast, lung, and bladder cancer. J Pathol 201(2):204–212. doi:10.1002/path.1449
18. Slieker RC, Bos SD, Goeman JJ, Bovée JV, Talens RP, van der Breggen R, Suchiman HED, Lameijer E-W, Putter H, van den Akker EB, Zhang Y, Jukema JW, Slagboom PE, Meulenbelt I, Heijmans BT (2013) Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. Epigenetics Chromatin 6(1):26. doi:10.1186/1756-8935-6-26
19. Sánchez-Vega F, Gotea V, Petrykowska HM, Margolin G, Krivak TC, Deloia JA, Bell DW, Elnitski L (2013) Recurrent patterns of DNA methylation in the ZNF154, CASP8, and VHL promoters across a wide spectrum of human solid epithelial tumors and cancer cell lines. Epigenetics 8(12)
20. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S (2013) ChAMP: 450k Chip Analysis Methylation Pipeline. Bioinformatics. doi:10.1093/bioinformatics/btt684
21. Murphy PJ, Cipriany BR, Wallin CB, Ju CY, Szeto K, Hagarman JA, Benitez JJ, Craighead HG, Soloway PD (2013) Single-molecule analysis of combinatorial epigenomic states in normal and tumor cells. Proc Natl Acad Sci U S A 110(19):7772–7777. doi:10.1073/pnas.1218495110
22. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462(7271):315–322. doi:10.1038/nature08514
23. Stirzaker C, Taberlay PC, Statham AL, Clark SJ (2013) Mining cancer methylomes: prospects and challenges. Trends Genet 30:75–84. doi:10.1016/j.tig.2013.11.004
24. Vandewoestyne M, Goossens K, Burvenich C, Van Soom A, Peelman L, Deforce D (2013) Laser capture microdissection: should an

ultraviolet or infrared laser be used? Anal Biochem 439(2):88–98. doi:10.1016/j.ab.2013.04.023

25. Kristiansen G (2010) Manual microdissection. Methods Mol Biol 576:31–38. doi:10.1007/978-1-59745-545-9_2

26. Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med 366(10):883–892

Chapter 5

Quantitative DNA Methylation Profiling in Cancer

Ole Ammerpohl, Andrea Haake, Julia Kolarova, and Reiner Siebert

Abstract

Epigenetic mechanisms including DNA methylation are fundamental for the regulation of gene expression. Epigenetic alterations can lead to the development and the evolution of malignant tumors as well as the emergence of phenotypically different cancer cells or metastasis from one single tumor cell. Here we describe bisulfite pyrosequencing, a technology to perform quantitative DNA methylation analyses, to detect aberrant DNA methylation in malignant tumors.

Key words DNA methylation profiling, Epigenetics, Bisulfite pyrosequencing

1 Introduction

Deciphering epigenetic alterations contributes to our understanding of the molecular basis of different clinical phenotypes as well as characteristics of malignant tumors. Like genetic information epigenetic information is heritable, but epigenetic information is not encoded in the DNA sequence itself. Epigenetic information has a major impact on gene expression and chromatin structure. Since epigenetic modifications are reversible, they allow the adaptation of gene activity to the environmental conditions. Epigenetic mechanisms include modifications of bases in the DNA, of histones, and of other proteins of the chromatin, the expression of non-coding RNAs (ncRNA) as well as positioning of genes in the nucleus. The probably nowadays best studied epigenetic modification is the methylation of DNA. In this post-replicative process particularly cytosine residues located in CG dinucleotides become methylated by DNA-methyltransferases [1–3]. Especially in cancer cells, a large number of epigenetic alterations in the DNA methylome can be observed. Consequently, multiple drugs have been developed targeting epigenetic modifications in cancer, including drugs suppose to demethylate DNA [4].

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_5,

It has been estimated that 70–80 % of the approximately 2.7×10^7 cytosines in the human genome which are located in CG dinucleotides are methylated [5]. This corresponds to 3–5 % of all cytosines in the human genome [6]. The function of DNA methylation depends on the localization of the methylated cytosine in the genome. High methylation of repeated sequences and "molecular parasites" prevents DNA recombination events and ensures genome stability. DNA methylation of regulatory elements (i.e. promoters and enhancers) controls gene expression and DNA methylation of single alleles depending on their parental origin controls parental imprinting [7–17]. Germline alterations in the DNA methylation pattern limited to single regions of a genome might be associated with developmental disorders [18], while extensive alterations in the DNA methylome in somatic cells including aberrant DNA methylation of regulatory elements and hypomethylation of repetitive sequences are a typical hallmark of cancer contributing to genomic instability [12, 19, 20].

Consequently, strategies to analyze alterations in the DNA methylation pattern got into the focus not only of scientific questions but also of clinical diagnostics and applications. For example, altered DNA methylation of imprinting centers on chromosome 11p might be causative for the Silver-Russell syndrome (SRS) or the Beckwith-Wiedemann syndrome (BWS). Furthermore, depending on which imprinting center (IC1 or IC2) is affected in BWS altered DNA methylation is associated with a significantly increased risk for either Wilms' tumors or hepatoblastomas, respectively [21, 22].

DNA methylation analysis might also be of interest for individualized antitumor therapies. A promising chemotherapeutic drug for the treatment of glioblastomas is temozolomide, which induces alkylation of the DNA and disturbs DNA replication in tumor cells. However, lack of DNA methylation of the O-6-methylguanine-DNA methyltransferase-activity gene (MGMT, a member of the family of DNA repair enzymes) in glioblastomas corresponds to MGMT activity which reverses the alkylation of the DNA and therefore confers resistance to temozolomide therapy [23–25]. Thus, knowing the DNA methylation state of MGMT allows identifying patients which will benefit from temozolomide therapy. Furthermore, several drugs acting by inhibition of DNA methyltransferases or other epigenetic key players are already used for cancer therapy [4, 19, 20, 26].

Understanding the impact of alterations in the DNA methylome for cancer research and clinical management, numerous techniques with different foci and fields of applications have been developed. Some of them focus on a small number of selected sequences or even single CG loci, others focus on the genome-wide analysis of the DNA methylome. However, most of the protocols for performing quantitative DNA methylation analysis

available today do rely on the chemical modification of cytosine residues in the DNA in a methylation-dependent manner [27]. Because of length restrictions we cannot discuss multiple techniques in detail in this chapter. Therefore, we here focus on bisulfite pyrosequencing (BSPS), which is a quite universal and robust technique for targeted DNA methylation analysis in a quantitative manner suiting many requirements in research and clinical diagnostics.

1.1 The Principles of Bisulfite Conversion of DNA and of Bisulfite Pyrosequencing

Bisulfite pyrosequencing is a sequencing-by-synthesis technology allowing the analysis of DNA methylation at multiple CG loci. In the first step, a bisulfite conversion of the DNA has to be performed. During this procedure, a sulfonation of unmethylated cytosines results in cytosinsulfonate which becomes converted into uracilsulfonate by hydrolytic deamination. A final desulfonation step results into uracil [28]. In a subsequent PCR, the uracil will be further "converted" to thymidine. Since the bisulfite conversion prefers unmethylated cytosine residues, while methylated cytosines are much less affected, cytosine residues located in the resulting PCR products correspond to methylated cytosine residues in the sequence of interest (Fig. 1). Bisulfite-converted DNA is the analyte used for bisulfite pyrosequencing.

For performing bisulfite pyrosequencing three primers are needed, two for amplifying the target sequence of interest and one for the subsequent sequencing reaction. When designing these primers, it is mandatory to keep in mind that after bisulfite

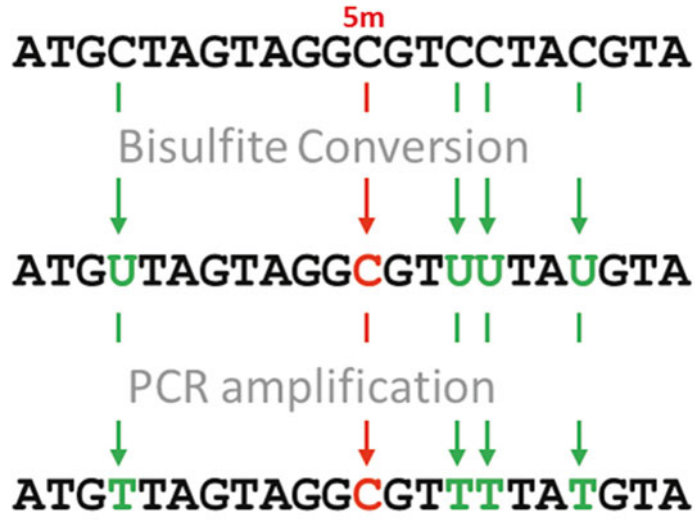


Fig. 1 Bisulfite conversion and subsequent PCR amplification. Unmethylated cytosine residues in the DNA sequence will be converted into uracil by bisulfite treatment. In a subsequent PCR reaction, uracil will be converted into thymine

conversion the original DNA strands are not complementary anymore. Thus, in most situations the primer set has to be designed specifically either for the sense or for the antisense strand. Furthermore, for performing pyrosequencing purified single-stranded DNA is needed. Therefore, one strand of the PCR product containing the target sequence has to be biotinylated at its 5′-end (modified PCR primers can be ordered from several vendors). The sequencing primer has to be designed being complementary to that strand of the PCR product carrying the biotin label. When designing the PCR primers manually, you should consider the following general aspects:

1. Length of the PCR product (usually between 200 and 600 bp, might depend on the quality of DNA subjected to analysis).
2. Melting temperature of primers (please consider that the sequencing reaction will be performed at 28 °C, thus the sequencing primers have to anneal below this temperature but not at high temperature usually used in PCR).
3. CG content of the primers.
4. CG dinucleotides in the primer sequence should be avoided due to their unknown methylation status (if unavoidable, order a primer mix containing primers binding to the methylated or unmethylated sequence or, alternatively, order a primer containing a mismatch at the position of the respective cytosine).
5. Limitation of sequencing length to ~80–100 bp.
6. Repeats in the sequence to analyze (homopolymers) might interfere with correct quantification and might cause sequencing errors.

However, it is strongly recommended to make use of suitable software packages like those delivered with the pyrosequencer by the manufacturer. Besides packages for, e.g. identifying organisms or analyzing genetic variations also packages for DNA methylation analysis or for designing suitable PCR and sequencing primers are available (i.e. PyroMark Assay Design Software, Qiagen).

The final primer set should be validated whether it contains known SNPs or other known sequence alterations. These might cause differences in primer binding or PCR efficiency between both alleles resulting in selective allele amplification, allelic drop outs, and consequently misleading results. SNP information is available from several databases (e.g. dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) or F-SNP (http://compbio.cs.queensu.ca/F-SNP/)).

In the PCR amplification reaction, a high number of PCR cycles should be performed ensuring most primer molecules becoming incorporated (usually >35–45 cycles). Otherwise, in particular free biotinylated PCR primers might interfere with the

subsequent sequencing reaction. After amplifying the DNA sequence of interest, the PCR products will be purified and denaturized. Here, the protocol takes advantages from the fact that one of the amplification primers is labeled with biotin. The PCR products will be bound to streptavidin-coated sepharose beads. Using a washing device (aspiration tool) belonging to the pyrosequencing system, the beads loaded with the PCR product will be immobilized to a membrane filter by applying a vacuum to the filter. This setup is used to purify the PCR products by sticking the filters into washing buffers and to generate single-stranded DNA. For this purpose, the PCR products are incubated in an aqueous solution of sodium hydroxide. While the biotinylated PCR strand still binds to the sepharose beads and to the filter, the opposite strand is removed in this step.

In general, the bisulfite pyrosequencing reaction is based on the sequential addition of nucleotides to the DNA in the presence of a mixture of enzymes and substrates, "converting" every integrated nucleotide into a light signal (Fig. 2).

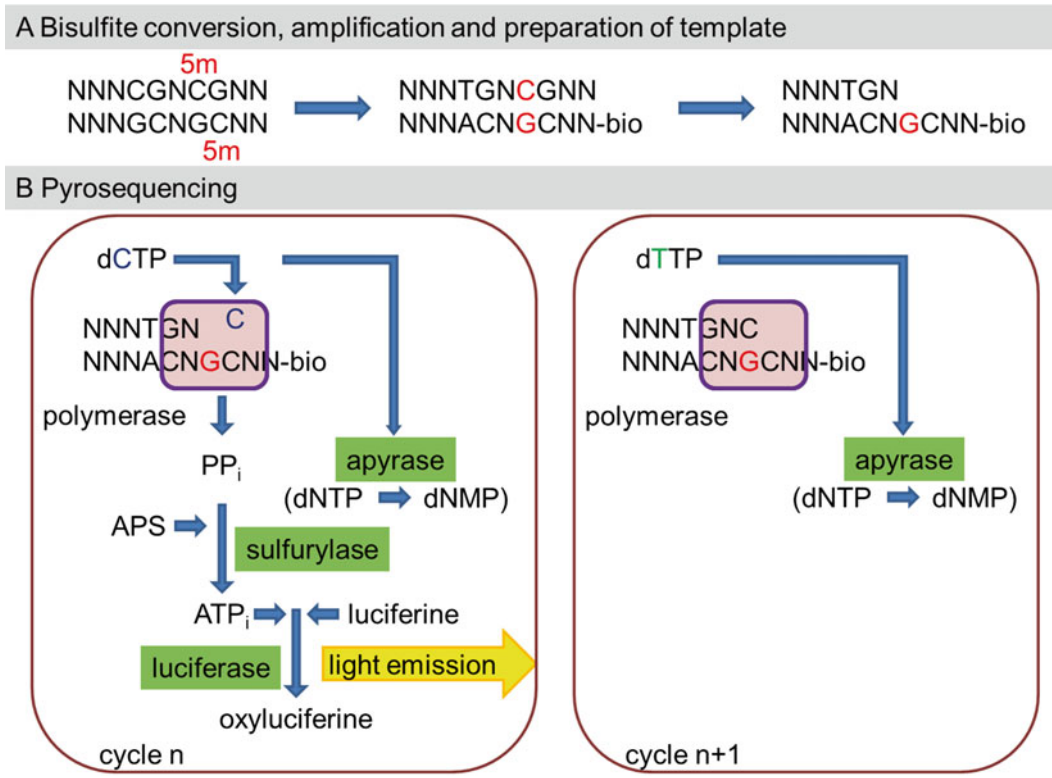


Fig. 2 Bisulfite pyrosequencing. After bisulfite conversion and purification of the DNA, the sequencing primer will be annealed to the single-stranded DNA template (**a**) before the pyrosequencing reaction can be initiated (**b**). *5mC* 5-methyl cytosine, *bio* biotin, *PPi* pyrophosphate, *APS* adenosine 5′ phosphosulfate, *dNTP* desoxy nucleoside triphosphate

After the initial annealing of the sequencing primer and the addition of a defined mix of enzymes and substrates, the nucleotides will be sequentially injected to the reaction according to the known sequence of the DNA to analyze. When using the manufacture's software package for designing the primer set this sequence to analyze, which might also contain additional injections for control purposes, is also provided by the software. If the injected nucleotide is complementary to the DNA strand to be sequenced, the DNA polymerase in the reaction mixture will prolong the sequencing primer. During this reaction, pyrophosphate will be released, giving the technique its name.

Together with adenosine 5′-phosphosulfate (APS) the pyrophosphate will be enzymatically processed by a sulfurylase to generate ATP, which afterward acts as substrate for luciferase converting luciferine to oxiluciferine. During the luciferase catalyzed reaction a light signal is emitted, which is recorded by a camera. The signal intensity correlates with the number of base repeats in the DNA sequence. In a parallel step, excessive nucleotides will be removed by an apyrase from the reaction. Finally, the next nucleotide is injected. Injection of a nucleotide which cannot be incorporated into the newly synthesized strand (i.e. the base is not complementary to the template DNA) does not result in light emission and no peak can be detected. When the sequencing reaction reaches a position in the DNA that contains a cytosine which is potentially methylated, dCTP and dTTP will be injected consecutively. The resulting signal ratio obtained from these both injections is used by the software to calculate the amount of methylated or unmethylated cytosines at that position. Due to limitations of the reaction mix the maximum read length is usually restricted to 100–150 bp, depending on the sequence to analyze. The typical read length is usually around 80 bp. To a certain amount, this decreasing enzymatic efficiency can be compensated by the software package used; however, this compensation is limited. As controls a fully methylated DNA (either generated by in vitro methylation of genomic DNA using, e.g. M.Sss I methyltransferase or using commercially available methylated DNA, e.g. Millipore #S7821) and an unmethylated DNA (e.g. generated by whole genome amplification of genomic DNA, e.g. by Illustra GenomiPhi (GE Healthcare)) should be processed in parallel to detect putative issues of this procedure. Furthermore, a genomic control (DNA which is not bisulfite-converted) is recommended, to ensure specificity of the primer set for bisulfite-converted DNA to exclude putative background from unconverted DNA remaining in the reaction. Including a control not containing any template DNA would reveal hidden contaminations of, e.g. kit reagents with DNA template.

2 Materials

2.1 Bisulfite Conversion

1. Sodium bisulfite or sodium metabisulfite.
2. Hydrochinone.
3. Purified water.
4. 0.2 ml PCR vials with tightly (!) closing lids.
5. PCR thermocycler.
6. 50 ng–2 μg DNA (DNA to analyze, methylated and unmethylated control DNAs).

Alternatively, kits for bisulfite conversion are commercially available making this highly critical step more convenient and reproducible. Bisulfite conversion kit systems are available from several manufacturers. For the pyrosequencing protocols outlined below, both the EpiTect-Kit (Qiagen) as well as the bisulfite conversion kit Gold (ZymoResearch) have been successfully tested in our hands. Bisulfite conversion using these kits can be performed according to the manufacturer's instruction, also with low amounts of DNA. To our experience applying different commercially available kit systems or protocols for bisulfite conversion of the DNA can affect the outcome of the analysis. Therefore, it is recommended using the same procedure (and kit system) for optimization as well as for the subsequent analyses.

2.2 Bisulfite Pyrosequencing

1. DNA to analyze (bisulfite-converted).
2. Appropriate primers for amplification and sequencing (concentration adjusted to 20 pmol/μl). These can be designed either using the PyroMark Assay Design software package (Qiagen), or an alternative software package available from other sources (e.g. Methprimer, available at http://www.urogene.org/cgi-bin/methprimer/methprimer.cgi). Alternatively, primers can also be designed manually (*see* **Note 1**).
3. Annealing Buffer (Qiagen).
4. Binding buffer (Qiagen).
5. PSQ 96 Sample Prep Thermoplate Low, U-shaped wells (PSQ96; Qiagen).
6. PSQ 96 Reagent Cartridge (Qiagen).
7. Denaturing solution (0.2 M NaOH).
8. 70 % Ethanol.
9. 1× Washing buffer (dissolve 6.5 g Tris-base in 5 l water).
10. PCR plate, 96 well, V-shaped wells (Eppendorf).
11. Adhesive tape (e.g. BioRad).
12. Streptavidin coated Sepharose Beads (GE Healthcare).

13. Heating device for 96-well plates.
14. VacuumPrep workstation coming with the sequencer (aspiration tool; Qiagen).
15. Plate shaker MS3 basic (IKA).
16. Pyrosequencer (e.g. PyroMark ID, Qiagen).
17. AmpliTaq Gold polymerase system (10 U/µl; Life Technologies).
18. dNTP Mix (5 mM of each nucleotide).
19. Purified water.
20. Wizard SV Gel and PCR Clean-Up System (Promega; only when performing bisulfite conversion manually).
21. DNA loading buffer (e.g. 3 ml glycerol, 25 mg bromophenol blue, add water to 10 ml).

Optional the PyroMark PCR kit (Qiagen) can be used. This kit substitutes for the **items 17, 18,** and **21** above.

3 Methods

3.1 Bisulfite Conversion

When using a commercially available kit system (highly recommended), please follow the instruction of the kit's manufacturer. If you prefer a conventional setup, you might follow the protocol outlined below. This protocol makes use of the "Wizard SV Gel and PCR Clean-Up System" available from Promega to isolate the DNA after bisulfite conversion. Besides the sample of interest also a negative (unmethylated DNA, e.g. generated by whole genome amplification), a positive control (fully methylated DNA) as well as a water control (containing no DNA sample) should be included and processed in parallel (*see* **Note 2**).

Freshly prepare the following solutions [29]:

1. 100 mM Hydrochinon solution (50 mg in 10 ml water, freshly prepared).
2. 3 M Sodium hydroxide solution (400 mg in 3.3 ml water, freshly prepared).
3. 0.5 M EDTA solution (adjust pH to 8.0 by adding sodium hydroxide).
4. Denaturation-buffer (3-fold, freshly prepared):
   (a) 3.0 µl sodium hydroxide solution.
   (b) 0.5 µl EDTA solution.
   (c) 6.5 µl water.
5. Bisulfite solution (saturated, freshly prepared):
   (a) 5.0 g sodium metabisulfite.
   (b) 7.0 ml water.

(c) 100 µl hydrochinon solution.

(d) 1.2 ml sodium hydroxide solution.

(e) Adjust pH to 5.0, heat to 50 °C to dissolve as much bisulfite as possible (saturated solution, the bisulfite might not dissolve completely).

6. Mix 10 µl denaturation-buffer (3-fold) with 20 µl DNA (~1 µg DNA).
7. Incubate for 15 min at 40 °C (e.g. in a PCR machine).
8. Add 180 µl bisulfite solution (prewarmed), close PCR lid tightly.
9. Incubate for 5 min at 98 °C.
10. Incubate for 6–10 h at 50 °C with heating up temperature to 98 °C for 3 min each hour.
11. Re-isolate the converted DNA using the "Wizard SV Gel and PCR Clean-Up System" according to the manufacturer's instruction. If necessary adapt the elution volume according to the DNA amount.
12. Adjust the final concentration of the DNA to ~50 ng/µl (depends on starting amount and elution volume in **step 12**).

3.2 Amplification of the Target Sequences

Besides including the controls stated above (positive, negative, water control), it is also recommended adding a sample with (unconverted) genomic DNA at this step (same amount of DNA as used for the bisulfite-converted DNA samples). Prepare the PCR-mix using standard 0.2 ml PCR-tubes. When processing several samples in parallel, prepare a mastermix and add 10 % of each component as "reserve." The components for a single sequencing reaction contain:

1. 1.0 µl bisulfite-converted DNA (~50 ng/µl).
2. 18.3 µl water.
3. 2.5 µl PCR buffer II.
4. 1.0 µl $MgCl_2$ solution (50 mM, comes with the AmpliTaq polymerase).
5. 1.0 µl dNTP mix (5 mM of each nucleotide).
6. 0.5 µl forward primer (20 pmol/µl).
7. 0.5 µl reverse primer (20 pmol/µl).
8. 0.2 µl AmpliTaq Gold polymerase (10 units/µl).

When using the PyroMark PCR kit instead of AmpliTaq polymerase the following scheme applies:

1. 1.0 µl bisulfite-converted DNA (~50 ng/µl).
2. 8.0 µl water.

3. 12.5 µl PyroMark MasterMix.
4. 2.5 µl Coral Load.
5. 0.5 µl forward primer (20 pmol/µl).
6. 0.5 µl reverse primer (20 pmol/µl).

The subsequent PCR amplification can be performed according to the following scheme:

Step	Temperature	Duration	PCR
1	95 °C	5 min	Denaturing step
2 3 4	95 °C Tm-5 °C 72 °C	45 s 45 s 45 s–1 min	45× PCR amplification cycles
5	72 °C	5 min	Final elongation step
6	4 °C	Inf.	Storage

The temperature in **step 3** (primer hybridization) should be adjusted according to the calculated melting temperature (Tm) of the primer minus 5 °C. The duration of the PCR elongation step (**step 4**) depends on the PCR product (e.g. its length). Usually, 45 s works fine. However, further optimization of the annealing temperature in **step 3** and the duration of the amplification step might be necessary. If using the PyroMark kit instead of AmpliTaq denaturing temperatures can be reduced to 94 °C and the duration time of all three steps of the amplification cycles can be shortened to 30 s. When using a different polymerase, possibly the elongation temperature and duration times have to be individually optimized also. Due to variations in speed, ramping rates as well as temperature stability it is recommended to perform assay optimization and subsequent analyses using the same thermocycler block.

After the PCR has been finished, take 5 µl of the PCR product, add 3 µl DNA loading buffer and load the sample to a standard agarose gel. Addition of loading buffer can be omitted when using the PyroMark kit. After separation and staining of the DNA (e.g. by ethidium bromide) only one strong specific PCR band should be visible on the gel. Any additional signal on the gel might interfere with the subsequent analysis. Therefore, it is mandatory to optimize PCR conditions. Of course, the control containing unconverted DNA as well as the water control should not result in any PCR amplification (*see* **Note 3**).

3.3 Purification of the PCR Product and Preparation of a Single-Stranded Sequencing Template

1. When using the PyroMark Assay Design software package for designing the primer sets, the software also provides the core sequence to analyze. However, we recommend spending some time analyzing both the provided sequence to analyze as well as the unconverted DNA sequence. To verify complete bisulfite conversion of the DNA, manually introducing additional

cytosines into the provided sequence to analyze at positions which contain cytosine residues not part of CG dinucleotides in the unconverted DNA sequence is highly recommended. For this purpose a specific function introducing bases into the sequence to analyze is provided by the Q-CpG software package, the particular positions are shown in bold letters by the software. When preferring performing the primer design manually, you also have to prepare the sequence to analyze by yourself (*see* **Note 4**).

2. Start the pyrosequencer, open the adequate software package (usually the Q-CpG software) and enter the assays to perform, the sequences to analyze and the plate designs according to the manual of the software package. Based on this information the software also calculates the volumes of enzyme, substrate, and nucleotide mixes needed to perform the assays and the plate layout.
3. Heat up a heating device to 85 °C (the device usually comes with the pyrosequencer).
4. Transfer 40 μl of each PCR amplification product into a well of a PCR 96-well plate.
5. Prepare a mastermix containing 5 μl sepharose beads (shake/vortex before use!) and 40 μl binding buffer for each sample or control to become sequenced.
6. Add 43 μl mastermix to each DNA sample to be analyzed on the 96-well plate.
7. Seal the plate tightly (e.g. by applying adhesive sealing tape) and mix for ~10 min at high speed on a plate shaker.
8. Prepare sequencing mastermix containing 40 μl annealing buffer and 0.5 μl sequencing primer (20 pmol/μl) for each sample (add 10 % of each component as a "reserve" for pipetting).
9. Prepare the sequencing plate (PSQ96) by pipetting 40 μl sequencing mastermix to each well needed. Take care that on the PSQ96 plate the same positions are used as on the PCR plate containing the sepharose beads, as the loaded beads will be transferred into the corresponding well on the sequencing plate.
10. Set up the vacuum/aspiration tool. Adjust the vacuum to ~ −70 mmHg. Prepare the dishes of the aspiration tool. One each is filled with (a) 70 % ethanol, (b) denaturing solution, and (c) washing solution.
11. Remove the PCR plate containing the sepharose beads from the shaker and instantaneously aspirate the solution using the aspiration tool. The DNA bound to the beads sticks to the filter membrane.

12. Transfer the filters for 5–10 s into the ethanol bath, followed by 5–10 s into the bath containing the denaturing solution and 5–10 s into the washing bath. When removing the filters from a bath turn around the handhold slowly (~180°), so that the filters show upwards for ~1 s. By this procedure, solution that remained in the handhold of the vacuum tool will be removed.
13. Transfer the filters on top of the wells of the plate containing the sequencing mix (PSQ96). Do not transfer the filters into the mix, since this would result in the aspiration of the mix. Subsequently, turn off the vacuum, transfer the filters (without vacuum applied!) into the sequencing mix and shake the filters gently to release the beads containing the DNA into the wells.
14. Incubate the plate at 80 °C for 2–5 min on the heating device. Finally, cool down the plate to room temperature.
15. Transfer the volumes of the mixes into a sequencing cartridge as determined by the software. Place the sequencing plate and the cartridge into the pyrosequencer, close the plate frame and the lid and start the sequencer.
16. The system will perform the sequencing automatically. Collected data will be continuously presented, thus the progress can be traced on the screen.
17. After finishing sequencing the results can be evaluated using the Q-CpG software (see below).

3.4 Data Analysis

After the run has been finished, the results are presented as a pyrogram by the Q-CpG software package (*see* **Note 5**).

In Fig. 3, the major components of a pyrogram are presented. It shows the analysis of a short sequence from the NFE2L3 gene. The upper pyrogram belongs to a methylated control sample (3a), while the lower one belongs to an unmethylated control generated by WGA (3b). The left (y-) axis of a pyrogram (3–1) indicates the intensity of a signal (indicated by the red line). If the sequence to analyze (3–2) contains a sequential series of the same base the nucleotide injected during one sequencing cycle will be incorporated consecutively several times, which finally results in increased signal intensity (3–3). Therefore, the signal intensity (peak height) is a direct measurement of how many consecutively repeats of a base are contained in the sequence. Since the sequence to analyze (3–2) is known before the sequencing reaction is started, the expected peak height at each position of the pyrogram can be calculated in advance by the software. The software can optionally indicate the expected peak height as grey bar in the background of each peak (3–4). A peak which is significantly lower or higher as expected would argue for a divergence between the DNA sequenced

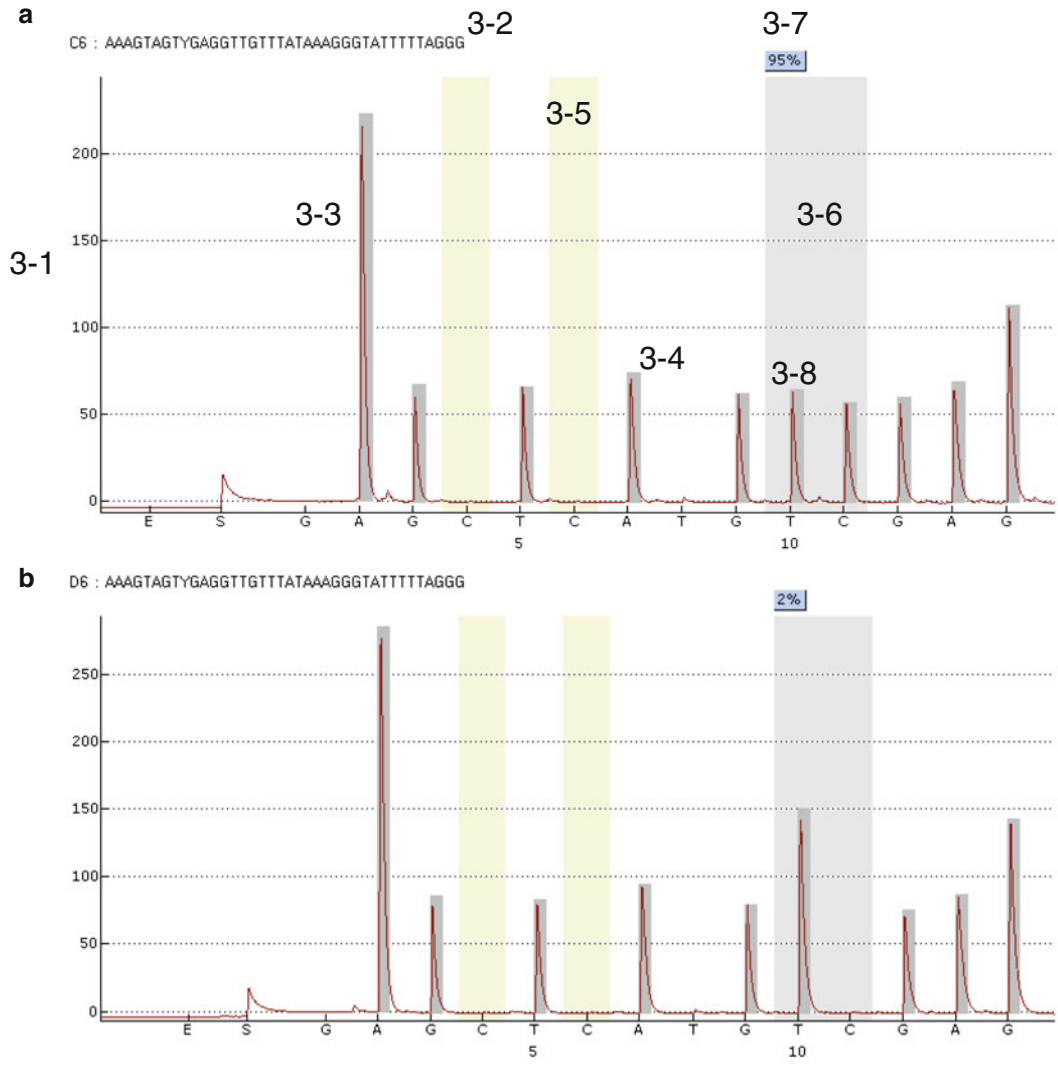


Fig. 3 The pyrogram: interpretation of the primary results. 3–1 scale (signal intensity). Peak heights (originating from a single base) below 15 indicate a bad quality of the pyrogram due to an unfavorable signal:noise ratio. 3–2 sequence to analyze, 3–3 the height of a peak correlates with nucleotides incorporated in a single sequencing cycle and the how often a base is present in the DNA sequence (here: four consecutive adenines), 3–4 *grey bars* beyond the peaks indicate the expected signal intensity, 3–5 *yellow bars* indicate positions introduced to detect incomplete bisulfite conversion, 3–6 putatively methylated cytosines are highlighted by large *grey bars*, 3–7 the methylation value is shown by small numbers on top of the pyrogram, 3–8 a thymine preceding a CG dinucleotide is integrated into the grey bar indicating a CG dinucleotide since the signals do overlap

and the sequence expected (the "sequence to analyze", (3–2)). Positions at which additional cytosines have been added into the sequence to analyze for verifying complete bisulfite conversion are highlighted in yellow (3–5). At the corresponding positions the original sequence (before bisulfite conversion) contains a cytosine

which is not part of a CG dinucleotide and therefore cannot become methylated (taking the rare non-CG methylation sites not into consideration). Consequently, all of these cytosine residues have to be converted into thymidine. If still cytosine is detectable, bisulfite conversion was not complete. In this case, quantitative DNA methylation analysis is erroneous. Positions in the pyrogram corresponding to CG loci in the sequence to analyze are indicated as grey bars (3–6). At these positions, dCTP and dTTP are injected consecutively. The ratio of the peak heights corresponds to the DNA methylation value (3–7). Please note that a thymidine preceding a CG dinucleotide also results in a T-corresponding peak even if the cytosine is fully methylated (3–8). In particular if a homopolymer (multiple copies of one nucleotide) precedes a CG dinucleotide correct quantification of the DNA methylation is error prone.

The DNA methylation data can be exported into a tab delimited format which can be easily processed by other software packages.

4 Notes

1. After bisulfite conversion of the target DNA, the resulting DNA strands are not complementary anymore. Thus, the primer sets have to be designed specifically either for one strand of interest or for both strands individually. Mixing up both strands would result either in no or unpredictable results.
2. To collect proper DNA methylation values, complete bisulfite conversion of the DNA is mandatory. Else unconverted cytosines become interpreted as methylated cytosines resulting in high DNA methylation values. An example is shown in Fig. 4a. The presence of peaks at control positions in the pyrogram (arrows, highlighted in yellow) indicate incomplete bisulfite conversion. For comparison, Fig. 4b shows a completely converted DNA. If bisulfite conversion turns out to be incomplete, fragmentation of the DNA (heat, ultrasound, enzymatic) prior to bisulfite conversion might improve the conversion rate. However, the sequence to be amplified for bisulfite sequencing has to be left intact (e.g. no restriction sites are allowed in this sequence to analyze).
3. Unspecific binding of the sequencing primer results into unexpected or missing peaks already at the beginning of the pyrogram. Furthermore, already at the beginning of the pyrogram the expected peak heights (indicated by grey bars beyond the peaks) are often not reached (Fig. 5).
4. An insufficient assay design might result in the PCR amplification and/or sequencing of unconverted DNA since the primers

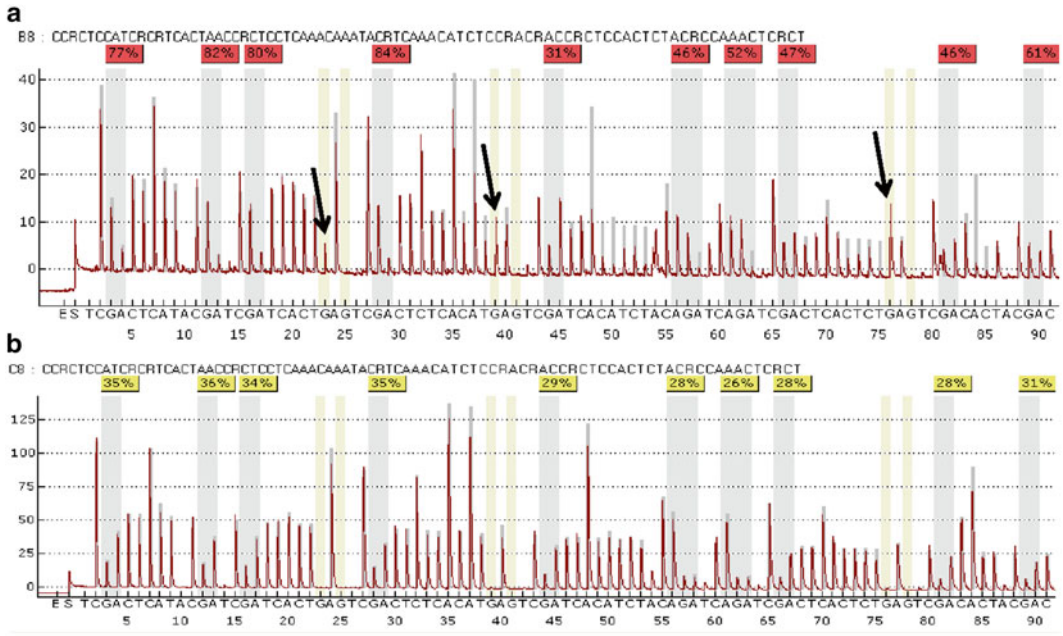


Fig. 4 Incomplete bisulfite conversion. The presence of peaks at control positions in the pyrogram (*arrows*, highlighted in *yellow*) indicate incomplete bisulfite conversion. (**a**) Poor bisulfite conversion, (**b**) complete bisulfite conversion of the DNA

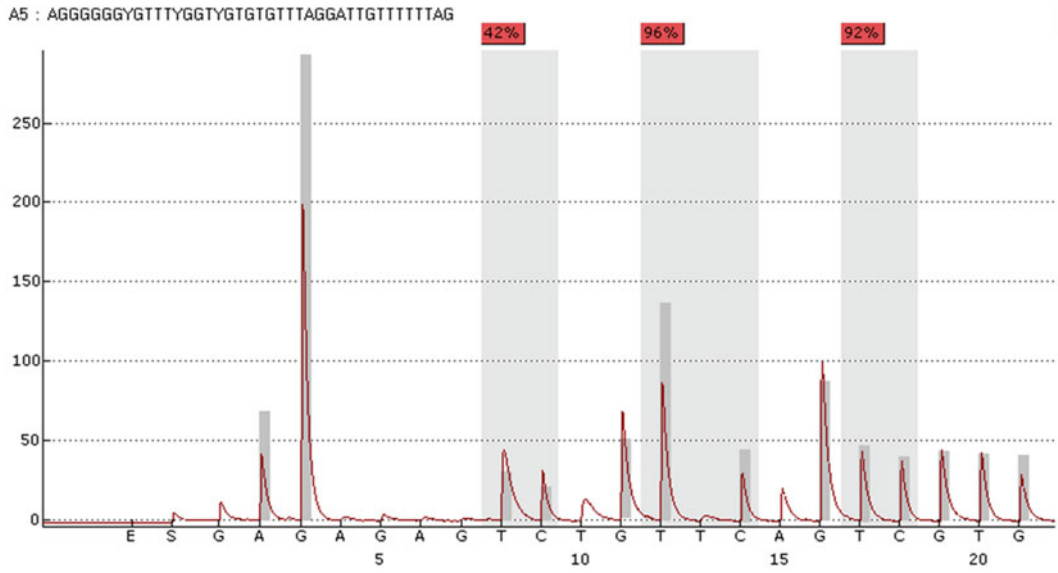


Fig. 5 Unspecific binding of the sequencing primer. Unspecific binding of the sequencing primer might result into unexpected peak heights (e.g. positions 10 and 15)

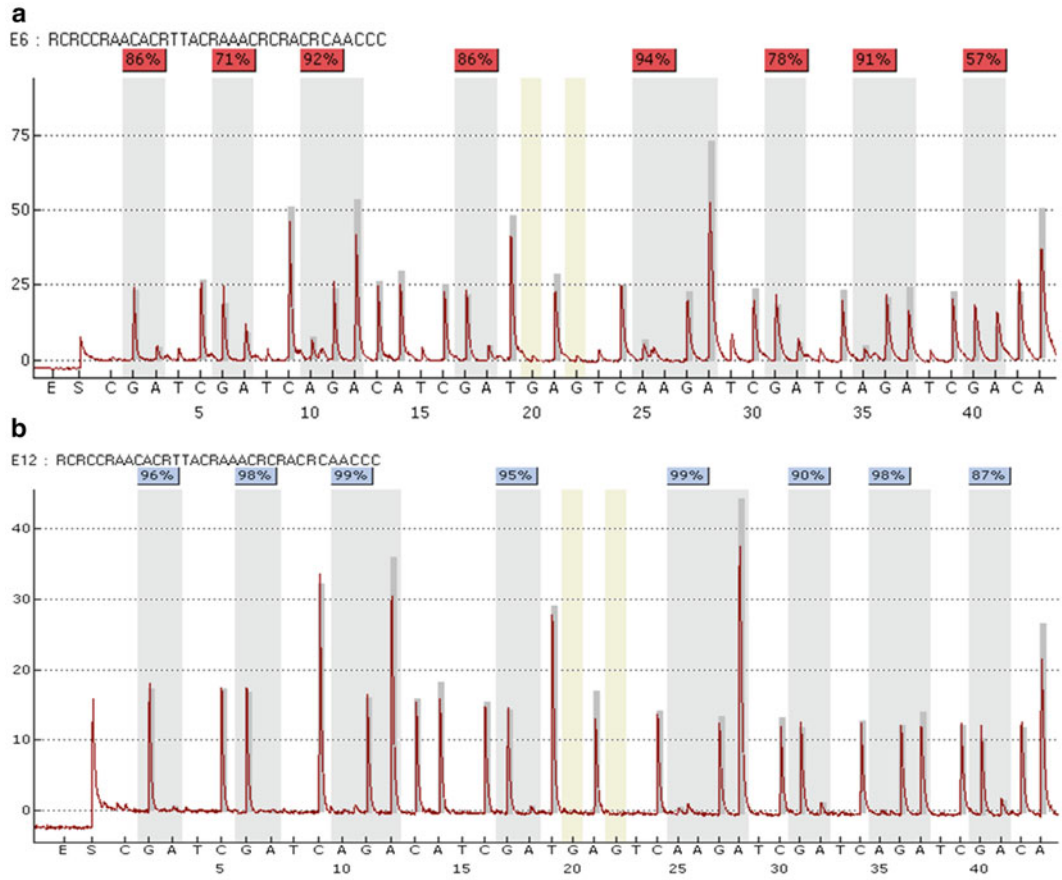


Fig. 6 An insufficient assay design can result in the amplification and sequencing of unconverted DNA (**a**), which can interfere with quantification of DNA methylation. Without an appropriate control containing unconverted DNA this putative source of errors might remain undetected. (**a**) Pyrogram obtained from unconverted DNA, (**b**) pyrogram obtained from bisulfite-converted DNA

are not specific for bisulfite-converted DNA. In this case pyrosequencing of a control containing only (unconverted) genomic control DNA will be helpful, since it results in a specific sequencing pattern similar to the expected pattern (Fig. 6a). Because without a specific control containing only genomic/unconverted DNA this problem is difficult to detect (Fig. 6b the corresponding bisulfite-converted sample), it is highly recommended to include such a control.

Methylated cytosines are a hotspot for C to T transitions [30]. Consequently, positions in the DNA containing methylated cytosines often contain SNPs. If a sample carries such SNP heterozygous, the theoretically maximal DNA methylation value could reach 50 %, if the sample is homozygous for such SNP no DNA methylation could be detected at all.

Therefore, when designing a pyrosequencing assay, the resulting sequence to analyze should be checked whether it contains already known SNPs using available databases.

5. Bisulfite pyrosequencing is mainly based on the assumption that the DNA methylation status of the DNA to analyze has no major impact on PCR amplification subsequent to bisulfite conversion. However, a high number of CG loci in the sequence to analyze as well as primer binding sites containing CG loci might result in the preferential amplification or sequencing of originally either methylated or unmethylated DNA leading to erroneous results. Thus, if highly accurate DNA methylation analysis is a concern, we would recommend performing a DNA methylation analysis of a serial dilution of fully methylated DNA (e.g. either commercially available methylated DNA or M.SssI-treated DNA) and unmethylated DNA (e.g. generated by whole genome amplification). A subsequent comparison of the bisulfite pyrosequencing analysis with the expected results might reveal a potential bias of the PCR amplification or the sequencing reaction. Furthermore, a serial dilution might be helpful for correcting the results obtained from the analysis.

References

1. Bestor TH (2000) The DNA methyltransferases of mammals. Hum Mol Genet 9(16): 2395–2402
2. Brenner C, Fuks F (2006) DNA methyltransferases: facts, clues, mysteries. Curr Top Microbiol Immunol 301:45–66
3. Cheng X, Blumenthal RM (2008) Mammalian DNA methyltransferases: a structural perspective. Structure 16(3):341–350. doi:10.1016/j.str.2008.01.004
4. Gowher H, Jeltsch A (2004) Mechanism of inhibition of DNA methyltransferases by cytidine analogs in cancer therapy. Cancer Biol Ther 3(11):1062–1068
5. Wigler MH (1981) The inheritance of methylation patterns in vertebrates. Cell 24(2): 285–286
6. Clark SJ, Harrison J, Frommer M (1995) CpNpG methylation in mammalian cells. Nat Genet 10(1):20–27. doi:10.1038/ng0595-20
7. Gama-Sosa MA, Wang RY, Kuo KC, Gehrke CW, Ehrlich M (1983) The 5-methylcytosine content of highly repeated sequences in human DNA. Nucleic Acids Res 11(10):3087–3095
8. Hsieh J, Gage FH (2004) Epigenetic control of neural stem cell fate. Curr Opin Genet Dev 14(5):461–469. doi:10.1016/j.gde.2004.07.006
9. Lopez-Serra L, Esteller M (2008) Proteins that bind methylated DNA and human cancer: reading the wrong words. Br J Cancer 98(12):1881–1885. doi:10.1038/sj.bjc.6604374
10. Morgan HD, Santos F, Green K, Dean W, Reik W (2005) Epigenetic reprogramming in mammals. Hum Mol Genet 14 Spec No 1:R47–R58. doi:10.1093/hmg/ddi114
11. Plass C, Smiraglia DJ (2006) Genome-wide analysis of DNA methylation changes in human malignancies. Curr Top Microbiol Immunol 310:179–198
12. Robertson KD (2001) DNA methylation, methyltransferases, and cancer. Oncogene 20(24):3139–3155. doi:10.1038/sj.onc.1204341
13. Robertson KD (2005) DNA methylation and human disease. Nat Rev Genet 6(8):597–610. doi:10.1038/nrg1655
14. Simonsson S, Gurdon JB (2005) Changing cell fate by nuclear reprogramming. Cell Cycle 4(4):513–515
15. Smith SS (1991) DNA methylation in eukaryotic chromosome stability. Mol Carcinog 4(2):91–92
16. Smith SS, Crocitto L (1999) DNA methylation in eukaryotic chromosome stability

revisited: DNA methyltransferase in the management of DNA conformation space. Mol Carcinog 26(1):1–9

17. Turek-Plewa J, Jagodzinski PP (2005) The role of mammalian DNA methyltransferases in the regulation of gene expression. Cell Mol Biol Lett 10(4):631–647
18. Soejima H, Wagstaff J (2005) Imprinting centers, chromatin structure, and disease. J Cell Biochem 95(2):226–233. doi:10.1002/jcb.20443
19. Brueckner B, Kuck D, Lyko F (2007) DNA methyltransferase inhibitors for cancer therapy. Cancer J 13(1):17–22. doi:10.1097/PPO.0b013e31803c7245
20. Ghoshal K, Bai S (2007) DNA methyltransferases as targets for cancer therapy. Drugs of Today 43(6):395–422. doi:10.1358/dot.2007.43.6.1062666
21. Cooper WN, Luharia A, Evans GA, Raza H, Haire AC, Grundy R, Bowdin SC, Riccio A, Sebastio G, Bliek J, Schofield PN, Reik W, Macdonald F, Maher ER (2005) Molecular subtypes and phenotypic expression of Beckwith-Wiedemann syndrome. Eur J Hum Genet 13(9):1025–1032. doi:10.1038/sj.ejhg.5201463
22. Shuman C, Beckwith JB, Smith AC, Weksberg R (1993) Beckwith-Wiedemann Syndrome. In: Pagon RA, Adam MP, Ardinger HH et al (eds) GeneReviews(R). Seattle (WA),
23. Hegi ME, Diserens AC, Godard S, Dietrich PY, Regli L, Ostermann S, Otten P, Van Melle G, de Tribolet N, Stupp R (2004) Clinical trial substantiates the predictive value of O-6-methylguanine-DNA methyltransferase promoter methylation in glioblastoma patients treated with temozolomide. Clin Cancer Res 10(6):1871–1874
24. Parkinson JF, Wheeler HR, Clarkson A, McKenzie CA, Biggs MT, Little NS, Cook RJ, Messina M, Robinson BG, McDonald KL (2008) Variation of O(6)-methylguanine-DNA methyltransferase (MGMT) promoter methylation in serial samples in glioblastoma. J Neurooncol 87(1):71–78. doi:10.1007/s11060-007-9486-0
25. Yachi K, Watanabe T, Ohta T, Fukushima T, Yoshino A, Ogino A, Katayama Y, Nagase H (2008) Relevance of MSP assay for the detection of MGMT promoter hypermethylation in glioblastomas. Int J Oncol 33(3):469–475
26. Mund C, Brueckner B, Lyko F (2006) Reactivation of epigenetically silenced genes by DNA methyltransferase inhibitors: basic concepts and clinical applications. Epigenetics 1(1):7–13
27. Ammerpohl O, Martin-Subero JI, Richter J, Vater I, Siebert R (2009) Hunting for the 5th base: techniques for analyzing DNA methylation. Biochim Biophys Acta 1790(9):847–862. doi:10.1016/j.bbagen.2009.02.001
28. Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci U S A 89(5):1827–1831
29. Kladde MP, Xu M, Simpson RT (1999) DNA methyltransferases as probes of chromatin structure in vivo. Methods Enzymol 304:431–447
30. Lutsenko E, Bhagwat AS (1999) Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. a model, its experimental support and implications. Mutat Res 437(1):11–20

Chapter 6

Single-Nucleotide Polymorphism to Associate Cancer Risk

Victoria Shaw, Katie Bullock, and William Greenhalf

Abstract

Genetic heterogeneity explains variation in predisposition for cancer. Whole-genome analysis allows risk to be quantified, giving better targeted screening and quantification of the personalized risk posed by environmental factors. Array-based approaches to whole-genome analysis are rapidly being overtaken by next-generation sequencing (NGS). In this review the different platforms currently available for NGS are compared and the opportunities and risks of this approach are discussed: including the informatics packages required and the ethical issues. Methods applicable to the personal genome machine (PGM) are given as an example of workflows.

Key words Single-nucleotide polymorphism (SNP), Next Generation Sequencing (NGS)

1 Introduction

In 2010 when the first edition of this book was produced arrays were the established method for identifying single nucleotide polymorphisms (SNPs). The big question being if there was a clinical utility to identifying such variants and if so which platform (Affymetrix or Illumina) should be used. It was clear even then that next-generation sequencing (NGS) was an alternative that would probably replace arrays but the world was split between the optimists who saw this as just around the corner and the realists who accepted that NGS would remain the tool of the super-rich for many years to come and arrays would be the workhorse the rest of us would rely upon. In just 5 years we realists have been shown up as being ridiculously pessimistic and the optimists now appear to have been if anything unduly conservative. NGS is no longer the preserve of a technocratic elite and is available very widely. Today's optimists talk casually of germline and even tumor whole genome sequencing: but questions remain on how this can be best used clinically.

Identification of cancer-causing mutations such as those in the Rb gene [1], the APC gene [2], mismatch repair genes [3–5], and the STK11 locus [6, 7] does have clear clinical significance, but

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_6,

carriers of such mutations are rare. As discussed in our last version of this chapter the majority of genetic predisposition results from complex interactions of multiple genes, with each other and with environmental exposure [8, 9], and these weaker associations will usually involve polymorphisms that are relatively common in the general population and only become significant risk factors when in combination. Weaker associations are not readily amenable to analysis by linkage studies, as non-affected carriers and those not carrying high-risk alleles cannot be distinguished.

Association of single base changes, occurring as a result of transition or transversion mutations, with a disease-related allele will be maintained over many generations; these changes may even be functionally related to the disease allele. There are approximately 5–10 million SNPs in the human genome occurring every 400–1000 base pairs [10]. It has been estimated that approximately 500,000 SNPs are required to genotype an individual of European ancestry [11]. Half a million SNPs is a large number, but arrays of this scale are readily available, so as described in our previous chapter, routine whole-genome analysis by SNP arrays is feasible. This will show associations with cancer and remains an option—but the same SNPs can also be assayed with NGS giving greater reliability; as the output is a sequence each base of which can be analyzed for quality control rather than hybridization which can give only a binary response.

2 The Pros and Cons of Large Information-Rich Data Sets for Genotyping

NGS is based on the parallel sequencing of individual DNA molecules. This can be done at random positions in a genome, mapping the small sequences produced against previous data for the same species so that eventually the whole genome will have been resequenced. Massively parallel resequencing of DNA makes sequencing of a whole human genome eminently feasible giving data on every SNP [12] and the price for such a sequence is dropping to a point ($1000 or less) where this would not be impractical for an individual patient [13]. However, cost is still an obstacle for routine use with large numbers of patients, in particular considering the considerable additional costs for data storage and analysis. A further consideration is the question of data protection. A patient may be perfectly happy for a specific clinically relevant set of information to be obtained and used for their care, but the genome is a veritable Pandora's box and so gaining individual and public acceptance for storage of such potentially dangerous data (however noble the intention) is an obstacle to widespread use of whole-genome sequencing [14]. A solution is to target sequencing so that less data needs to be stored and the potential outcomes of data analysis can be better defined. Targeting does involve an additional

cost, but the greater depth of sequencing (the number of parallel sequences obtained at a specific base) allows for greater levels of multiplexing, giving sequencing of the DNA of many patients on a single sequencing run: more than offsetting the cost of enriching specific sequences.

The biggest problem of NGS, but the most compelling argument for its use in SNP characterization rather than identification of mutations, is that all NGS is inherently error prone. PCR errors and misalignments will give the calling of false variants with even the most sophisticated of SNP calling software. The error relates to any base change at any position, the probability of detecting a variant is therefore the sum of all probabilities for all bases sequenced (in NGS this will be low probabilities multiplied by tens of millions of bases). The probability of a specific recognized allelic change at a SNP coinciding with an erroneous call will be one-third of the probability of an error at that specific base. Furthermore, the probability of an error at a specific base is quantifiable on the basis of variations in conserved bases in the same sequence.

In this chapter the different platforms that can be used to sequence targeted variable region will be discussed, these can be used for both identification of SNPs and small tandem repeats (STRs) [15]. The genotype data obtained could be used to identify associations with cancer risk and to assess risk of individuals who may be under consideration for inclusion in screening programs. A brief discussion will also be made of the use of the same technology using DNA from tumors in order to identify somatic changes [16].

3 Different Platforms (Table 1)

The discussion of NGS will be restricted to smaller scale sequencing (producing less than 100GB of sequence) as these are more suitable for targeted sequencing in a clinical setting. The different sequencing platforms available in this range include offerings from Thermo Fisher/Life Technologies, Illumina and Roche diagnostics.

Thermo Fisher base their personal genome approach on proton release after base addition to individual sequences distributed on a microchip. The two platforms are the personal genome machine (PGM) from Ion Torrent (which can reliably give up to 2GB of sequence) and the ion proton (giving 10GB of sequence). Perhaps the most significant problem of the proton release approach is in accurately quantifying the length of homopolymer regions leading to false calls of insertions or deletions, so PGM struggles with highly biased sequences such as *Plasmodium* [17]. For calling of SNPs this is less of a problem, but it is worth noting that for high A/T- or G/C-rich regions bias should be expected.

Illumina use incorporation of different fluorescent bases to give synthesis-based sequencing in flow cells. The MiSeq platform

Table 1
Comparison of different platforms for SNP analysis from different companies

Company	Thermo Fisher		Illumina	Roche
Machine	PGM	Ion proton	MiSeq	GS FLX titanium
Technology	Sequencing by synthesis using proton release on a chip with emulsion PCR	Sequencing by synthesis using proton release on a chip with emulsion PCR	Sequencing by synthesis using fluorescent nucleotides in flow cells	Pyrosequencing and emulsion PCR
Cost of machine[a]	$	$$$	$$	$$$$
Base reads	2GB	10GB	15GB	8GB
Total run time	2 h	4 h	27 h	10 h
DNA requirements[b]	100–1000 ng	100–1000 ng	50–1000 ng	5–30 μg
Overall accuracy[c]	+	++	+++	+

Note: A comparison of the different platforms currently available
[a]The cost of each machine will vary according to where you are and your relationship with the individual companies: However, the PGM is the cheapest at well below $100,000 while the GS FLX Titanium probably cannot be obtained for less than $300,000 (prices will reduce as new machines become available)
[b]In practice amplification steps can be applied meaning all machines can be used with fentogram levels of DNA in the starting material
[c]As discussed in the chapter accuracy is highly dependent on the nature of the sequence. None of the machines will be expected to give a Phred score (Q score) of less than 20. MiSeQ is the most robust with Phred scores unlikely to be below 30 even in difficult-to-sequence regions

which gives up to 15GB of sequence is the format that will be discussed here, but Illumina also supply NextSeq (giving 120GB) and HiSeq 2500 (giving 1000GB). This technology offers reduced bias in comparison with PGM [18], but at a cost in reagents and time of run.

Roche have adopted a pyrosequencing method on beads for their NFS machines. Their GS FLX Titanium system can give 8GB of sequence and is probably the fastest of the personal genome machines (albeit at a higher reagent cost).

4 Sample Types

Successful application of next generation sequencing requires good quality DNA, which means that where possible freshly obtained blood should be used. However, this is not always possible or convenient. Archived samples are often only available as formalin fixed paraffin embedded samples, these can be applied to all platforms, but DNA quality is very variable depending on the type of tissue,

age of the sample, and how the tissue was fixed. The most commonly used method for DNA quantification is absorbance at 260 nm (an absorbance value of 1 at 260 nm corresponds to a DNA concentration of 50 μg/ml) but this is sensitive to single standed (ss) DNA, RNA, protein, and reagent contamination from DNA preparation methods. DNA can be run through a low strength agarose gel (1 %) to ensure that the DNA is intact but in practice PCR-based quality control is essential before using the DNA.

5 Single-Cell Sequencing

Identification of variants associated with cancer predisposition generally involves germ line analysis and so the assumption can be made that all cells from an individual carry the same sequence information: lineage associations between common SNPs and the disease are the result of very small but incremental effects of each variant, or a close linkage of the SNP with a more penetrant mutation which is maintained during meiosis. This contrasts with somatic mutations where lineages branch off from each other upon each mutational event until a mixed population is produced where different disease mutations, each important in its own individual host cell, represent a small percentage of the total sequence in the population. Analysis of tumor samples is therefore complex and sequencing is likely to miss mutations that occur later during tumorigenesis.

Numerous software fixes are available to deal with mixed populations, such as qPure [19]. These software fixes were generally born from array-based analysis and rely on loss of heterozygosity (LOH) in heterozygous SNPs. Assuming that there is LOH in the tumor, the cancer cells will have 0 or 100 % of each allele at that SNP locus, whilst stromal tissue will have no LOH (50 % of each allele). If the tumor cellularity was 50 % then the B allele frequency (BAF) would be 0.33 or 0.66. qPure uses an algorithm based on the average gap between BAF (the d-score), in a perfect world with 50 % cellularity the d-score for heterozygous SNPs would be 0.66–0.33 = 0.33. Clearly we do not live in a perfect world and level of LOH will vary, but taking d-score as the mean of the greatest distance in clusters of BAF, a correlation has been shown with tumor cellularity and proportion of known cancer mutations in control experiments with spiked samples [19].

Knowing the proportion of tumor cells would allow identification of mutations present in all cancer cells, but sophisticated modeling and further improvements in sequencing technology are required to solve the problem of identifying mutations present in small subpopulations. Alternatively, single nuclei can be used to give parallel sequences from pure populations [20–33]. This would also allow associations with SNPs to be tracked: following the lineages with variants that predispose for or characterize cancer.

6 Whole-Genome Amplification

A critical step for single-cell sequencing is the ability to amplify picogram levels of DNA to microgram levels, the simplest being degenerate oligonucleotide primer (DOP) PCR. Coverage can be greatly improved by using linkers added to fragmented DNA which contain the primer sites for subsequent PCR (ligated adaptors), but this still gives uneven depth and bias across the genome. Much better is multiple displacement amplification (MDA), this uses Φ29 DNA polymerase (or equivalent polymerase) and random oligonucleotide primers, the polymerase displaces DNA from previous extensions avoiding the need to repeatedly denature the DNA and allowing high-fidelity polymerization at moderate temperatures (e.g., 30 °C). The time allowed for polymerization is critical in the MDA process, greater times give greater bias and errors, shorter time gives inadequate quantities of DNA: in practice the optimum time needs to be determined depending on input DNA and sequencing modality [23, 34].

7 Sequence Capture

The most commonly used approaches to sequence enrichment involve targeted PCR or array-based hybridization. The PCR approach can then be subdivided into parallel and mixed amplification. Mixed amplification is the conventional form of PCR with targeted primers competing with each other in a single pool. This leads to considerable bias in the amplification and inevitable underrepresentation of some sequences. This may not be a problem if the targeted region is relatively small with hundreds and not thousands of primer pairs, but where thousands or even tens of thousands of sequences are required this is not suitable. Parallel (which includes digital) PCR allows noncompetitive amplification, the best established of these approaches being Raindance.

7.1 Sequence Capture (Microdroplet PCR-Based Enrichment)

The Raindance system works on the basis of encapsulating a library of primers so that individual droplets include individual primer pairs that are targeted at a particular sequence. These are fused to droplets containing fragmented template DNA and a PCR reagent mix on microfluidic chip. The PCR takes place within the fused droplets producing millions of individual PCR reactions within a single PCR tube. The droplets are then dispersed to give amplicons. We would recommend processing of these before subsequent sequencing.

After purification (e.g., with Qiagen MinElute columns) the product should be quality controlled for example with a Bioanalyzer instrument (Agilent, Santa Clara, California). We use a 2100 analyzer with the Agilent High Sensitivity DNA Kit. The threshold for inclusion will vary according to the amplicons expected, but

overrepresentation of primer at below 100 bp or a peak in the size distribution below 130 bp would normally fail.

After repairing the ends of the products, for example with the NEBNext End Repair Module (New England Biolabs) and repeat purification, the product should be concatenated (with ligase mixes made in house or kits such as the NEB Quick Ligation Kit). This prevents end effects that would otherwise bias the sequencing results and gives a more equilibrated range of fragment sizes after subsequent fragmentation. The re-purified product can be fragmented by any of many techniques (for example Covaris sonication).

7.2 Sequence Capture (Haloplex)

Haloplex relies for its sequence specificity on oligonucleotide panels that hybridize within the target sequence. Input DNA is endonuclease fragmented and then circularized using the panel sequences which brings the ends together with primer sites and barcodes (allowing multiplexing on subsequent NGS runs). The circular product is then purified on streptavidin beads and can be amplified using primers specific for the incorporated primer sequence.

Sequence capture panels specific for any desired region can be created using the Haloplex design tool (Life Technologies, Paisley, UK). Sequence capture can be performed in batches. During the process individual barcodes can be added to each amplicon to allow multiple individuals to be sequenced in a multiplex.

The product should be quality controlled for example with a Bioanalyzer instrument (Agilent, Santa Clara, California). A threshold of 300 pg/L is taken for inclusion and any product with an inadequate trace (e.g., overrepresentation of primer at below 100 bp or a peak in the size distribution below 130 bp) would also be excluded.

7.3 Sequence Capture (SureSelect)

SureSelect target enrichment involves hybridization of biotinylated bait sequences (custom-synthesized RNA oligonucleotides) to fragmented DNA that has been built into a library by the addition of appropriate linkers. The hybrids are then captured on streptavidin coated beads which are separated from non-hybridized sequences using a magnet. After RNA digestion the enriched library is PCR amplified using primers specific for the linkers. By tiling the baits, i.e., designing a combination of overlapping bait series that cover the area of interest, high specificity for the target region can be achieved.

8 Methods for the PGM

In our laboratory the PGM from Ion Torrent (Thermo Fisher) is the workhorse for SNP analysis; below is an outline of the protocols followed for this analysis, from the point of obtaining DNA.

8.1 Emulsion PCR Using the One Touch 2 200 Kit

1. Install the recovery tubes and router into the centrifuge of the One Touch 2.
2. Insert a new amplification plate, thread the disposable tubing through the catch and pinch valve, and insert the needle into the port in the lid of the centrifuge.
3. Attach reagent sippers to both ports at the front of the One Touch 2.
4. Invert the One touch reagent oil three times to mix.
5. Half fill a Reagent Tube with One Touch oil and install on the left front port of the One Touch 2.
6. Invert the Recovery Solution three times to mix.
7. Quarter fill the other Reagent Tube with Recovery solution and install on the front right port of the One Touch 2.
8. Ensure the waste container is empty.

8.1.1 Preparation of Reagents

1. Defrost the OT2 Reagent Mix, once defrosted vortex for 30 s.
2. Vortex the OT2 PCR Reagent B for 1 min, and centrifuge for 2 s to remove droplets from the lid. If the reagent is clear proceed to the next step; if cloudy heat for 1 min at 75 °C. Repeat the 30-s vortex and check for cloudiness. Repeat until clear.
3. Centrifuge Ion PGM Template OT2 Enzyme Mix for 2 s to remove any droplets in the lid, and place on ice.
4. Prepare the library to a concentration of 100 pM. Dilute the library dependent on the type of sample and application; for AmpliSeq DNA for SNP analysis this would be 2 μL of library to 23 μL of molecular biology-grade water.
5. Vortex the diluted library for 5 s and then centrifuge for 2 s to remove any liquid from the lid.
6. In a 1.5 ml Eppendorf Lo-Bind tube add the following reagents in the order they are written:
   (a) 25 μL Nuclease-free water.
   (b) 500 μL Ion PGM Template OT2 Reagent Mix.
   (c) 300 μL Ion PGM OT2 PCR Reagent B.
   (d) 50 μL Ion PGM Template OT2 Enzyme Mix.
   (e) 25 μL diluted library.
7. Vortex this amplification solution for 5 s and then centrifuge for 2 s to remove any liquid from the lid.

8.1.2 Prepare the ISPs

1. Vortex the ISPs at maximum speed for 1 min.
2. Centrifuge for 2 s to remove any droplets from the lid.
3. Mix by pipetting up and down.
4. Add 100 μL of ISPs to the 900 μL amplification solution prepared in the 1.5 mL Lo-Bind tube.

8.1.3 Fill and Install the One Touch Plus Reaction Filter Assembly

1. Pipette 100 μL of the amplification solution into the sample port.
2. Add 1000 μL reaction oil through the same port.
3. Add a further 500 μL reaction oil through the same port.
4. Invert the filter assembly and install it into the One Touch 2.

8.1.4 Run the One Touch 2

1. Ensure the centrifuge lid is closed.
2. Touch "Run" on the home screen.
3. Select the One touch kit that has been used.
4. Follow the on-screen prompts.
5. Samples must be progressed within 16 h of the One Touch 2 finishing. The One Touch 2 takes 4–5 h to complete.
6. Follow the on-screen prompt to centrifuge the sample.
7. Immediately after centrifugation, remove the router and discard, remove the two sample tubes, and place in an appropriate rack.
8. Remove all but approximately 50 μL from each tube. Resuspend the pellets in the 50 μL solution and combine into one 1.5 mL Lo-Bind tube. Add 1 mL of One Touch Wash solution. At this point the sample can be stored for up to 3 days at 2–8 °C.

8.1.5 Clean the One Touch 2

1. Remove the needle from the centrifuge and the tubing from the pinch valve and catch.
2. Place the needle in an empty Falcon tube.
3. Remove the used filter adapter and replace with a cleaning adapter.
4. Ensure at least ¼ of a tube of One Touch Oil remains.
5. Press "clean" from the home screen.
6. After cleaning empty the waste container.

8.1.6 Enrich the Sample

1. Centrifuge the 1 mL solution for 2.5 min at 15,500 × *g*, and remove all but 100 μL being careful not to disturb the pellet.
2. Optional: Assess the quality of the DNA.
3. Load a pipette tip (provided) onto the tip arm of the ES.
4. Add 10 μL neutralization solution to a 200 μL PCR tube and place in the sample hole of the ES.

8.1.7 Prepare Melt Off Solution

In a Lo-Bind tube, add 280 μL Tween solution (provided) and 40 μL 1 M NaOH (not provided).

8.1.8 Prepare Dynabeads

1. Vortex the vial containing the Dynabeads to resuspend them in the solution for 30 s.
2. Pipette up and down to resuspend.

3. Add 130 μL MyOne Bead Wash to a new Lo-Bind tube.
4. Add 13 μL Dynabeads to the bead wash.
5. Place the tube on a magnetic rack for 2 min.
6. Leaving the tube on the magnetic rack; remove the supernatant without disturbing the pellet of Dynabeads.
7. Add a fresh 130 μL of MyOne Bead Wash to the Dynabeads and resuspend.
8. Vortex for 30 s and then centrifuge for 2 s to remove any liquid from the lid.

8.1.9 Prepare the 8-Well Strip

1. Place the 8-well strip (provided) onto the One Touch ES with the square tab on the left.
2. Reading L-R, in well 1 add the sample (100 μL).
3. In well 2 add the Dynabeads resuspended in bead wash (130 μL).
4. In wells 3–5 add 300 μL of One Touch Wash Solution.
5. In well 7 add 300 μL of the melt-off solution.
6. Leave wells 6 and 8 empty.

If the ES is not switched on turn it on. Press run. The run will last approximately 35 min.

Once the run has finished invert the PCR tube to mix the sample with the neutralization solution. There should be approximately 200 μL in the PCR tube. If less than 200 μL contact technical support.

The enriched ISPs can now be stored at 2–8 °C for up to 3 days.

Before use the machine must be cleaned following the cleaning protocol with either chlorite solution or dH_2O. Chlorite cleaning must be carried out if reagents are left on sippers overnight or every 1000 flows, whichever is sooner.

8.1.10 Initialize the PGM

1. Thaw dNTPs on ice.
2. Check that the pressure on the argon or nitrogen tank is >500 psi.
3. Rinse wash bottle 2 twice with dH_2O.
4. Prepare 1 ml of 1 M NaOH and from this also make 1 mL of 100 mM NaOH.
5. Fill wash bottle 2 to the mould line with 18 MΩ water. Add an entire bottle of wash 2 solution (provided in kit) to the bottle. Add 70 μl of previously prepared 100 μM NaOH. Invert the bottle five times to mix.
6. Add 350 μL 100 mM NaOH to wash bottle 1 only.
7. To wash bottle 3, add 50 mL of wash 3 solution only.

8. Change gloves and add new sipper tubes and wash bottles to the PGM.
9. Press initialize from the main menu and follow the on-screen prompts. The system will automatically pH the wash buffers; this takes approximately 30 min.
10. Use the labels provided to label 4 50 mL falcon tubes with a dNTP. Check that the dNTPs are fully defrosted, and vortex to mix. Keep dNTPs on ice.
11. Transfer 20 μL of each dNTP into the equivalently labeled Falcon tube.
12. Change gloves and insert new dNTP sippers to the front of the machine.
13. Attach the falcon tubes with the dNTPs in on the correct port at the front of the PGM. The symbol on the front of the PGM matches the symbol on the label: L-R

    G,C,A,T.
14. Follow the on-screen prompts to complete the initialization. The PGM will add wash 2 buffer to each of the dNTP tubes. If each reagent reaches its target pH after approximately 30 min the machine will display "PASSED" in green.
15. Press "next" to exit initialization and return to the home screen.

8.1.11 Sequencing Protocol for 316 and 318 Chips

1. Thaw sequencing primer on ice.
2. Vortex Control Ion Sphere Particles for 5 s and pulse spin in a centrifuge to remove any liquid from the lid.
3. Add 5 μl Control Ion Sphere Particles directly to the ~220 μL enriched sample from the One Touch ES.
4. Centrifuge the tube at 15,500 × *g* for 2 min.
5. Remove the supernatant leaving ~15 μL in the bottom of the PCR tube. Ensure that the pellet is not disturbed. N.B. the pellet is not visible.
6. Ensure that the sequencing primer is fully thawed, vortex it for 5 s, and then pulse spin for 5 s to remove any droplets from the lid.
7. Add 12 μL sequencing primer to the ISPs and mix well by pipetting up and down.
8. Program a thermal cycler for 1 cycle at 95 °C for 2 min and then 37 °C for 2 min using the heated lid option.
9. Flick the sequencing polymerase tube a couple of times then pulse spin for 5 s.
10. Once thermal cycling is complete, add 3 μL of sequencing polymerase to the ISPs and incubate at room temperature for 5 min.

8.1.12 Perform Chip Check

1. On the Home screen of the PGM press "Run." Follow the on-screen prompts.
2. Use a chip used for previous runs for the cleaning chip.
3. Remove gloves and touch the grounding plate. Unwrap a new chip and place in the chip socket, and close the clamp.
4. Perform the chip Check. Scan the barcode from the chip packaging when prompted. Inspect the PGM for leaks.

8.1.13 Load the Chip: Weighted Bucket Loading

1. Following chip check, remove the new chip from the chip socket.
2. Tilt the chip at a 45° angle, so that the loading port is at the bottom of the chip.
3. Insert a pipette tip into the loading port and aspirate as much liquid as possible. Discard aspirated liquid.
4. Load the chip into the chip centrifuge upside down with the tab facing inward.

   N.B. the weighted buckets are labeled with "IN" and "OUT"; ensure that they are correctly orientated in the chip centrifuge.
5. Centrifuge for 5 s to remove any residual liquid.
6. Carefully remove the chip from the centrifuge bucket and clean any liquid using an absorbent wipe.
7. Place the chip right-side up in the bucket with the chip tab facing inwards.
8. With the bucket on a flat surface, slowly and carefully load the sample into the loading port, there should be approximately 30 μl of sample to load. Insert the bucket in the correct orientation into the centrifuge.
9. Centrifuge for 1 min.
10. Remove and replace the chip so that the chip tab is now pointing outwards (do not flip the weighted buckets).
11. Centrifuge for 1 min.
12. Remove the chip from the centrifuge and firmly tap the chip tab on the bench.
13. Tilt the chip at a 45° angle, insert a pipette tip into the loading port, remove as much liquid as possible, and discard it.
14. Spin the chip again with the chip tab facing outward to pool any remaining liquid at the loading port.
15. Remove the chip from the centrifuge, with the chip at a 45 °C angle, insert a pipette tip, and remove any remaining liquid and discard.
16. Once chip loading is complete press next on the PGM and install the loaded chip in the chip socket and clamp.

17. If a run has been planned, select browse and choose the plan, or select no plan.
18. Check the settings for the run, e.g., number of flows, reference sequence, sample name, and barcode set, and ensure that it is correct for the sequencing run or alter it.
19. When prompted inspect the PGM and chips for leaks.
20. The PGM will calibrate the chip again and provided the chip passes, sequencing will automatically start within 1 min.
21. The run will be complete 3–5 h depending on the chip type used; once completed press next to return to the main menu. Leave a used chip in the PGM when not in use.

8.2 Data Output from the PGM

The Torrent Suite of software takes data from each run to give an output of the number and length of reads. A graphical representation of the quality of reads and a fingerprint of the chip loading can also be provided (*see* Fig. 1).

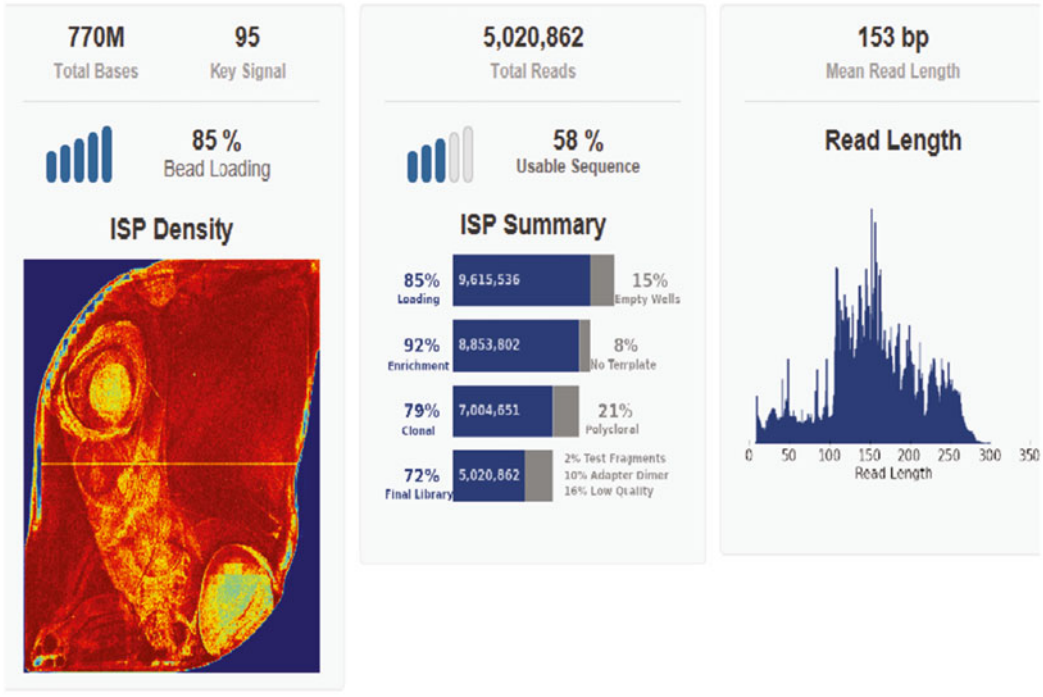


Fig. 1 The loading on the chip is shown with the droplets (beads) covering 85 % of the chip and only 8 % of the chip having no template. The *brighter* (*yellow*) *regions* are overloaded areas but 79 % of the reads are clonal (21 % having multiple templates in each droplet). 72 % of the reads are usable (over 5 million reads) and the distribution of the reads shows a peak at approximately 150 bases

9 General Requirements

9.1 Data Quality and Protection

When SNPs are called on arrays great care has to be taken to ensure that the data is robust. The Minimum Information About a Microarray Experiment (MIAME) greatly helped in this respect [35].

MIAME established standards that allow data from array experiments from different groups to be compared. This is obviously equally desirable for SNP data from NGS. Below are MIAME requirements as they could be applied to NGS.

1. The raw data for each sequence must be accessible:
   For arrays this is straightforward (e.g., CEL files for Affymetrix arrays); for NGS this is more complicated, mainly because the raw data will be in vast files in a variety of different formats. It is questionable as to whether general access to such files would be of general use, or will ever be generally accessible.
2. The final processed data for SNP calling should be accessible and transparent:
   This is more easily made available for NGS data (e.g., VCF files).
3. The essential sample annotation should be available.
   Information on the sample is included here. For example, was blood DNA used or were paraffin-embedded samples used. As discussed this aspect has potentially important data protection issues.
4. The experimental design including sample data relationships should be given.
   Describing which raw data file relates to which sample and any repeat sequences from libraries.
5. The essential laboratory and data processing protocols should be available.
   This will include all the quality control and quality systems (as described below).

So although somewhat more complex, technically NGS data for SNP analysis can be made to comply with the same standards of transparency as array data. The problem is whether such transparency will infringe the rights to anonymity of the patients or research subjects. In Europe publication of SNP data can be legally considered as publication of an individual's identity [36]. There can be little doubt that making NGS data available without explicit consent could be considered as a breach of data protection [37, 38]. Regardless of the legal position such data has the potential to be exploited to the patient's detriment, insurance issues being just one consideration. This emphasises the requirement to plan research or clinical assessment so that the data stored is as targeted as possible while giving an analysis which is fit for purpose.

9.2 Quantifying Output Quality

The probability of a false call is made up of a series of elements: the first being the chance of a miss call due to the machines detection method. This is usually expressed in terms of a Phred (or Q) score, which is the negative log (to the base 10) of the probable error call on a base multiplied by 10, so a Phred score of 20 means there is a 1 % chance of an error in the call at this specific base ($-\log(0.01) \times 10$) [39]. SNP calling algorithms can be set to ignore calls for individual reads which have a Phred score below a certain threshold (e.g., 30). This needs to be adjusted for SNP analysis rather than SNP discovery, for example if there are two possible alleles at a specific base and the minor allele is identified, then if the Phred is 30 there is a 1 in 1000 chance that a major allele would be called as something else, but a 1 in 3000 chance that a major allele would be called as the minor allele. It would be quite reasonable on this basis to accept a call at a Phred of 25 or over (for calling a minor or major allele).

After the data from the sequencer is converted into base calls it is aligned to a reference sequence, this is usually in the form of a Sequence Alignment Map (SAM) or a Binary Alignment Map (BAM). If the alignment is erroneous false variants will be called. This risk is reduced by quality control of the sequence, usually requiring trimming of individual sequences before alignment (as errors increase exponentially as the sequence extends), for example by the DynamicTrim module of SolexaQA. However, variations in repetitive sequences and pseudogenes are difficult to distinguish from polymorphisms. SNP alleles that coincide with sequences in pseudogenes (for example) need to be treated with caution.

There are many software options available for sequence alignment; these differ on the nature of indexing of the reads or the reference sequence. Most commonly by either conventional hash functions into hash tables (SOAP, MAQ, etc.) or by a Burrows Wheeler Transform (SOAP2, BWA, Bowtie, etc.). All of which function more quickly and more accurately when the reference sequence is restricted. Sequences containing too many errors will effectively be discarded as they will not be aligned. Definition of ambiguous reads vary but generally reads with a mapping quality score below 10 can be excluded from further analysis [40]. This will seriously bias against difficult to sequence regions and contribute to lower depths of read. Reads containing indels can be realigned using IndelRealigner from GATK version 2.0, and variants called using the GATK UnifiedGenotyper [41], increasing calls. Including reads with errors risks calling false positive SNPs, but the use of NGS for SNP calling rather than identification allows greater flexibility on sequence quality control (e.g., 10 errors in a 200 base alignment is a critical problem if they are all called as variants, but are of less concern if the purpose of the alignment is to identify a known variant which is one of the other 190 bases sequenced).

SAMtools can be used to format data for visualisation with the Integrative Genome Viewer (IGV), an example is shown in Fig. 1. Note that the SNP is heterozygous and each allele is represented by approximately 50 % of all reads (as expected) but other variants from the reference are also represented albeit at much lower frequencies. This is not an issue when the depth of read is adequate, but as the amount of sequence covered increases and/or the number of barcoded indexes is increased (increasing the number of samples that can be evaluated in a single sequencing run) the depth will inevitably fall. When depth of read reaches single figures distinguishing SNP alleles from errors becomes difficult.

There are numerous tools to quality control the SNP calls. These rely on databases of known SNPs such as dbSNP, comparing the calls at known SNP positions against variation at other bases where there is no recognized SNP (SOAPsnp is an example of such a tool); known linkage disequilibrium between SNPs can also be used. SAMTools and GATK both have routines that compare data from multiple individuals within a study, checking for inconsistencies with Hardy-Weinberg equilibrium for the SNPs in the test populations. These recalibration tools are extremely powerful in epidemiological studies; they are of less use when the purpose of the exercise is to assess individual risk based on a pattern of SNPs. Specific tools to identify high- or low-risk profiles are still in the development stage at the time of writing this chapter.

The outputs of the SNP calling algorithms are typically a VCF file. These contain the proportion of minor and major alleles in the reads and the depth of reads at the position that is called, these could be used to assess individual risk, but the VCF file needs to be tailored to the specific purpose of the ultimate analysis. For example, if a high-risk pattern of SNPs is being assessed absence of specific minor alleles cannot normally be considered to indicate a homozygous major allele, as the depth of sequence at that position will not be included within the VCF (which only include the positions which have a variation from the reference).

NGS normally relies on an amplification process to produce a library of DNA fragments. Phred values only indicate the probability of read errors; they do not take into account PCR errors which introduce false variations during the manufacture of libraries. Typically such errors are present in only a small proportion of reads at any given site (1 % or less) which is of little concern if germline variations are being investigated in homogenous samples. This becomes a concern when the sample has mixed sequences, a heterozygote present in 2 % of a population being indistinguishable from a PCR error of 1 %. This is a common problem when looking for somatic mutations but is not usually encountered when searching for germline SNPs.

References

1. Mateu E, Sanchez F, Najera C et al (1997) Genetics of retinoblastoma: a study. Cancer Genet Cytogenet 95(1):40–50
2. Kartheuser A, West S, Walon C et al (1995) The genetic background of familial adenomatous polyposis. Linkage analysis, the APC gene identification and mutation screening. Acta Gastro-Enterologica Belgica 58(5–6):433–451
3. Kohonencorish MRJ, Doe WF, Stjohn DJB, Macrae FA (1995) Chromosome 2p linkage analysis in hereditary nonpolyposis colon-cancer. J Gastroenterol Hepatol 10(1):76–80
4. Froggatt NJ, Koch J, Davies R et al (1995) Genetic-linkage analysis in hereditary nonpolyposis colon-cancer syndrome. J Med Genet 32(5):352–357
5. Green RC, Narod SA, Morasse J et al (1994) Hereditary nonpolyposis colon-cancer—analysis of linkage to 2p15-16 places the Coca1 locus telomeric to D2s123 and reveals genetic-heterogeneity in 7 Canadian families. Am J Hum Genet 54(6):1067–1077
6. Nakagawa H, Koyama K, Tanaka T et al (1998) Localization of the gene responsible for Peutz-Jeghers syndrome within a 6-cM region of chromosome 19p13.3. Hum Genet 102(2):203–206
7. Mehenni H, Blouin JL, Radhakrishna U et al (1997) Peutz-Jeghers syndrome: confirmation of linkage to chromosome 19p13.3 and identification of a potential second locus, on 19q13.4. Am J Hum Genet 61(6): 1327–1334
8. Schaid DJ, McDonnell SK, Blute ML, Thibodeau SN (1998) Evidence for autosomal dominant inheritance of prostate cancer. Am J Hum Genet 62(6):1425–1438
9. Bonadona V, Lasset C (2003) Inherited predisposition to breast cancer: after the BRCA1 and BRCA2 genes, what next? Bull Cancer 90(7):587–594
10. Botstein D, Risch N (2003) Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. Nat Genet 33(Suppl):228–237
11. Nicolas P, Sun F, Li LM (2006) A model-based approach to selection of tag SNPs. BMC Bioinformatics 7:303
12. Wheeler DA (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876
13. Sheridan C (2014) Illumina claims $1,000 genome win. Nat Biotechnol 32(2):115
14. Hoffman S, Podgurski A (2013) The use and misuse of biomedical data: is bigger really better? Am J Law Med 39(4):497–538
15. Fordyce SL, Mogensen HS, Borsting C et al (2015) Second-generation sequencing of forensic STRs using the Ion Torrent HID STR 10-plex and the Ion PGM. Forensic Sci Int Genet 14:132–140
16. Xu H, DiCarlo J, Satya RV, Peng Q, Wang Y (2014) Comparison of somatic mutation calling methods in amplicon and whole exome sequence data. BMC Genomics 15:244
17. Quail MA, Smith M, Coupland P et al (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics 13:341
18. Vancampenhout K, Caljon B, Spits C et al (2014) A bumpy ride on the diagnostic bench of massive parallel sequencing, the case of the mitochondrial genome. PLoS One 9(11), e112950
19. Song S, Nones K, Miller D et al (2012) qpure: a tool to estimate tumor cellularity from genome-wide single-nucleotide polymorphism profiles. PLoS One 7(9), e45835
20. Saliba AE, Westermann AJ, Gorski SA, Vogel J (2014) Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res 42(14):8845–8860
21. Trombetta JJ, Gennert D, Lu D, Satija R, Shalek AK, Regev A (2014) Preparation of single-cell RNA-Seq libraries for next generation sequencing. Curr Protoc Mol Biol 107:4.22.1–4.22.17
22. Streets AM, Zhang X, Cao C et al (2014) Microfluidic single-cell whole-transcriptome sequencing. Proc Natl Acad Sci U S A 111(19):7048–7053
23. Korfhage C, Fisch E, Fricke E, Baedker S, Loeffert D (2013) Whole-genome amplification of single-cell genomes for next-generation sequencing. Curr Protoc Mol Biol 104:Unit 7.14
24. Kim KI, Simon R (2014) Using single cell sequencing data to model the evolutionary history of a tumor. BMC Bioinformatics 15:27
25. Kohler GA, Lau PC (2013) Preface. Genomics, microbiomes, single cell analysis and next-generation sequencing. J Microbiol Methods 95(3):400
26. Potter NE, Ermini L, Papaemmanuil E et al (2013) Single-cell mutational profiling and clonal phylogeny in cancer. Genome Res 23(12):2115–2125

27. Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14(9):618–630
28. Shirai M, Taniguchi T, Kambara H (2014) Emerging applications of single-cell diagnostics. Top Curr Chem 336:99–116
29. Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28(11):1420–1428
30. Yilmaz S, Singh AK (2012) Single cell genome sequencing. Curr Opin Biotechnol 23(3):437–443
31. Navin N, Hicks J (2011) Future medical applications of single-cell sequencing in cancer. Genome Med 3(5):31
32. Navin N, Kendall J, Troge J et al (2011) Tumour evolution inferred by single-cell sequencing. Nature 472(7341):90–94
33. Leung ML, Wang Y, Waters J, Navin NE (2015) SNES: single nucleus exome sequencing. Genome Biol 16(1):55
34. Hittelman A, Sridharan S, Roy R et al (2007) Evaluation of whole genome amplification protocols for array and oligonucleotide CGH. Diagn Mol Pathol 16(4):198–206
35. Brazma A (2009) Minimum information about a microarray experiment (MIAME)—successes, failures, challenges. ScientificWorldJournal 9:420–423
36. Masca N, Burton PR, Sheehan NA (2011) Participant identification in genetic association studies: improved methods and practical implications. Int J Epidemiol 40(6):1629–1642
37. Suh KS, Sarojini S, Youssif M et al (2013) Tissue banking, bioinformatics, and electronic medical records: the front-end requirements for personalized medicine. J Oncol 2013:368751
38. Dove ES, Joly Y, Tasse AM et al (2014) Genomic cloud computing: legal and ethical points to consider. Eur J Hum Genet. doi:10.1038/ejhg.2014.196
39. Altmann A, Weber P, Bader D, Preuss M, Binder EB, Muller-Myhsok B (2012) A beginners guide to SNP calling from high-throughput DNA-sequencing data. Hum Genet 131(10):1541–1554
40. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9(4):357–359
41. DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498

Chapter 7

Exosomes in Cancer Disease

Margot Zöller

Abstract

Cancer diagnosis and therapy is steadily improving. Still, diagnosis is frequently late and diagnosis and follow-up procedures mostly are time-consuming and expensive. Searching for tumor-derived exosomes (TEX) in body fluids may provide an alternative, minimally invasive, yet highly reliable diagnostic tool. Beyond this, there is strong evidence that TEX could become a potent therapeutics.

Exosomes, small vesicles delivered by many cells of the organism, are found in all body fluids. Exosomes are characterized by lipid composition, common and donor cell specific proteins, mRNA, small non-coding RNA including miRNA and DNA. Particularly the protein and miRNA markers received much attention as they may allow for highly specific diagnosis and can provide hints toward tumor aggressiveness and progression, where exosome-based diagnosis and follow-up is greatly facilitated by the recovery of exosomes in body fluids, particularly the peripheral blood. Beyond this, exosomes are the most important intercellular communicators that modulate, instruct, and reprogram their surrounding as well as distant organs. In concern about TEX this includes message transfer from tumor cells toward the tumor stroma, the premetastatic niche, the hematopoietic system and, last but not least, the instruction of non-cancer stem cells by cancer-initiating cells (CIC). Taking this into account, it becomes obvious that "tailored" exosomes offer themselves as potent therapeutic delivery system.

In brief, during the last 4–5 years there is an ever-increasing, overwhelming interest in exosome research. This boom appears fully justified provided the content of the exosomes becomes most thoroughly analyzed and their mode of intercellular interaction can be unraveled in detail as this knowledge will open new doors toward cancer diagnosis and therapy including immunotherapy and CIC reprogramming.

Key words Exosome, Next-generation sequencing, Pancreatic cancer

1 Introduction

Cancer incidence is still increasing and for many cancers, the prognosis is still poor, which relies in part on late diagnosis, but also on high radiation and drug resistance and metastasis formation, which frequently set a cornerstone for curative therapy [1–3]. Hope has been created that diagnosis may become facilitated by tumor-derived exosomes (TEX). Exosomes, small vesicles, are recovered in all body fluids and are characterized by selective proteins, lipids, mRNA, and miRNA related to their donor cell, which

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_7,

could facilitate a minimally invasive cancer diagnosis [4]. Exosomes are also hotly discussed as a potential therapeutics [5–7].

Exosomes are small 40–100 nm vesicles, which derive from the fusion of the intraluminal vesicles (ILV) of multivesicular bodies (MVB) with the plasma membrane [4, 8]. Their homogeneous size is one of the major criteria to differentiate exosomes from apoptotic blebs, microparticles and microvesicles, which vary in size [9]. Exosomes are composed of a lipid bilayer; selected membrane and cytosolic proteins, mRNA, small non-coding RNA and DNA [10]. Exosomes are found in all body fluids [11, 12]. Due to their presence in body fluids and their expression of selected markers, exosomes are suggested to be optimal candidates for non-/minimally invasive diagnosis [13]. Exosomal proteins, mRNA, and miRNA are function-competent [14, 15] and are transferred into target cells [15, 16], which qualifies them as most important intercellular communicators [10] and thereby as potentially powerful therapeutics [17, 18].

2 The Origin of Exosomes and Exosome Recovery

Exosomes are small 40–100 nm vesicles delivered by many cells in the organism and abundantly by tumor cells [19]. Exosomes derive from the fusion of ILV in MVB, an intracellular compartment containing multiple vesicles [20] with the plasma membrane. MVBs are assembled from vesicles sorted from the trans-Golgi network or from internalized membranes. MVB can fuse with lysosomes for protein degradation [21] or they release their ILV by fusing with the plasma membrane. The released vesicles are termed exosomes [22]. The latter process was first described for the release of the transferrin receptor during reticulocyte maturation and was considered as a mode to eliminate obsolete proteins [23]. Exosome research became highly stimulated, when it was noted that antigen-presenting cells (APC) release exosomes derived from the major histocompatibility (MHC) class II compartment, a subset of MVB. These exosomes functioned similar to APCs and stimulated T cells in vitro and in vivo [24]. Exosome research received an additional boost with the discovery that they contain mRNA and microRNA. This so-called shuttle RNA is transferred into target cells, where the horizontal transfer of exosomal genetic material between cells can induce exogenous gene expression and mediate RNA silencing [25]. By the transfer of function-competent proteins, lipids, RNA and miRNA into target cells, exosomes are prone for intercellular communication. To make use of this powerful system, it is a prerequisite to know about its constituents.

2.1 Endosomes and Multivesicular Bodies

The relative abundance of proteins, lipids, mRNA, and miRNAs differs between exosomes and donor cells. This implies active sorting into MVB. Indeed, the sorting of proteins into MVB is a highly regulated process that can proceed via different pathways.

First described has been the ESCRT (endosomal sorting complex required for transport)-dependent mechanism, which sorts ubiquitinylated proteins into ILV of MVB, which can become degraded, when MVB fuse with lysosomes or alternatively are exocytosed [26]. The sorting of cargo into ILV, which are delivered as exosomes, is not fully elucidated. The ESCRT complex is composed of several subcomplexes, ESCRT I, II and III that work coordinately. Tsg (tumor susceptibility gene) 101 in the ESCRT complex I binds ubiquitinated proteins and recruits ESCRT II. ESCRT III becomes recruited via ESCRT II or Alix (ALG-2-interacting protein X). ESCRT III recruits a deubiquitinating enzyme that removes the ubiquitin tag from the cargo proteins prior to sorting into MVB [27]. To dissociate from the membrane the assembled ESCRT III complex requires energy, which is provided by the ATPase vacuolar protein sorting 4 (Vsp4). Silencing of 23 components of the ESCRT machinery revealed particularly silencing of serine/arginine-rich splicing factor 5, signal transducing adaptor molecule1 and Tsg101 to be accompanied by reduced exosome secretion, whereas Vsp4 silencing increased exosome secretion [28]. ESCRT-dependent exosome biogenesis involves syndecans [29]. Syndecan, transmembrane heparan sulfates, interact with syntenin, which interact with CD63 and Alix. Silencing of syntenin or syndecan decrease the number of released exosomes and reduces accumulation of the abundant exosome components Alix, heat shock protein (HSP) 70 and CD63 [29]. For a comprehensive overview I recommend a recent review by Schuh and Audhya [30].

Though the ESCRT system possibly is dominating, not all proteins require the ESCRT complex for incorporation into exosomes. Alternatively, proteins in detergent-resistant membrane complexes can become incorporated into MVB [31]. In this ESCRT-independent pathway, tetraspanins play a fundamental role [31, 32]. Tetraspanins and other proteins with high affinity for cholesterol and sphingolipids are partitioned into membrane domains (tetraspanin-enriched membrane microdomains, TEM), which according to their physical properties are prone for internalization [33, 34]. There is evidence that the scission/fission process for the internalization of TEM complexes differs from that of the ESCRT complex with dynamin and the intersectin 2 complex playing a major role [35]. The functional relevance of tetraspanins for exosome generation has been demonstrated for defective exosome secretion in bone marrow (BM) dendritic cells (DC) from CD9 knockout mice [36].

Lipid affinity also can account for MVB incorporation. Thus, proteolipid (PLP) exosome secretion is not affected by silencing of major components of the ESCRT complex [37]. The PLP-positive exosomes colocalize in the endosomal compartment with flotilin and glycosylphosphatidylinositol. They are rich in cholesterol

and ceramide and depend on ceramide production by neutral sphinomyelinase-2 [37]. Sphingosine-1-phosphatase and diaglycerol also plays a major role in MVB biogenesis [38]. Finally, phospholipase D, which hydrolyzes phosphatidylcholine regulates exosome secretion [39] and lysophosphatidic acid, enriched in exosomes, promotes inward budding of vesicles and interacts with Alix and HSP70 [40].

2.2 Vesicle Traffic and Release of Exosomes

The major actors in vesicle traffic are Rab proteins [41], a subfamily of small GTPases [42]. Rab proteins reversibly associate via geranylgeranyl modifications with membranes, where they regulate vesicle budding, tethering, and fusion [43]. Though the engagement of different rabs varies with the cell type, rab4 and rab5 mostly are recovered on early endosomes, rab11 is engaged in juxtanuclear recycling endosome traffic, rab7 and rab9 are late endosome and MVB markers [44]. Rab35 and rab11 possibly play a role in endocytic recycling [45]. Rab GTPase-activating proteins (GAP) are engaged in exosome release [46].

Rab proteins also regulate the subcellular localization of MVB via the interaction with actin and microtubules [47]. Thus, rab11 recruits myosin as well as dynein [48]. Invadopodia, which are rich in actin, are key docking sites for MVB and exosome release. This polarized exosome release is inhibited by N-WASP, critical for actin polymerization in invadopodia [49]. Maturation of late endosomes is governed by moving along microtubules, which is dynein-dependent. Instead, movement of MVB toward the plus end of microtubules for docking on the plasma membrane requires kinesin [50]. Finally, rab25 regulates MVB docking or tethering [46] and rab27b exosome release [51]. This is accomplished by the regulation of SNARE proteins (soluble-N-ethylmaleimide-sensitive fusion protein-attachment protein receptors) pairing on transport vesicles (v-SNARE) with SNARE-binding partners (t-SNARE) [52, 53]. For exosome secretion also the intracellular Ca^{2+} level [54] as well as the intracellular and extracellular pH [55] play key roles (Fig. 1).

3 Exosome Characterization

Exosomes are composed of a lipid bilayer, which contains transmembrane proteins. The small plasma also contains proteins, mRNA, small non-coding RNA including miRNA and DNA. Some of the components are shared by all exosomes, while others are selective for the donor cell type. Besides the characteristic size, the marker profile of exosomes and their lipid composition allows to differentiate exosomes from microvesicles that derive by direct budding from the plasma membrane. However, the exosome composition is not a mirror of the donor cells, as some components are

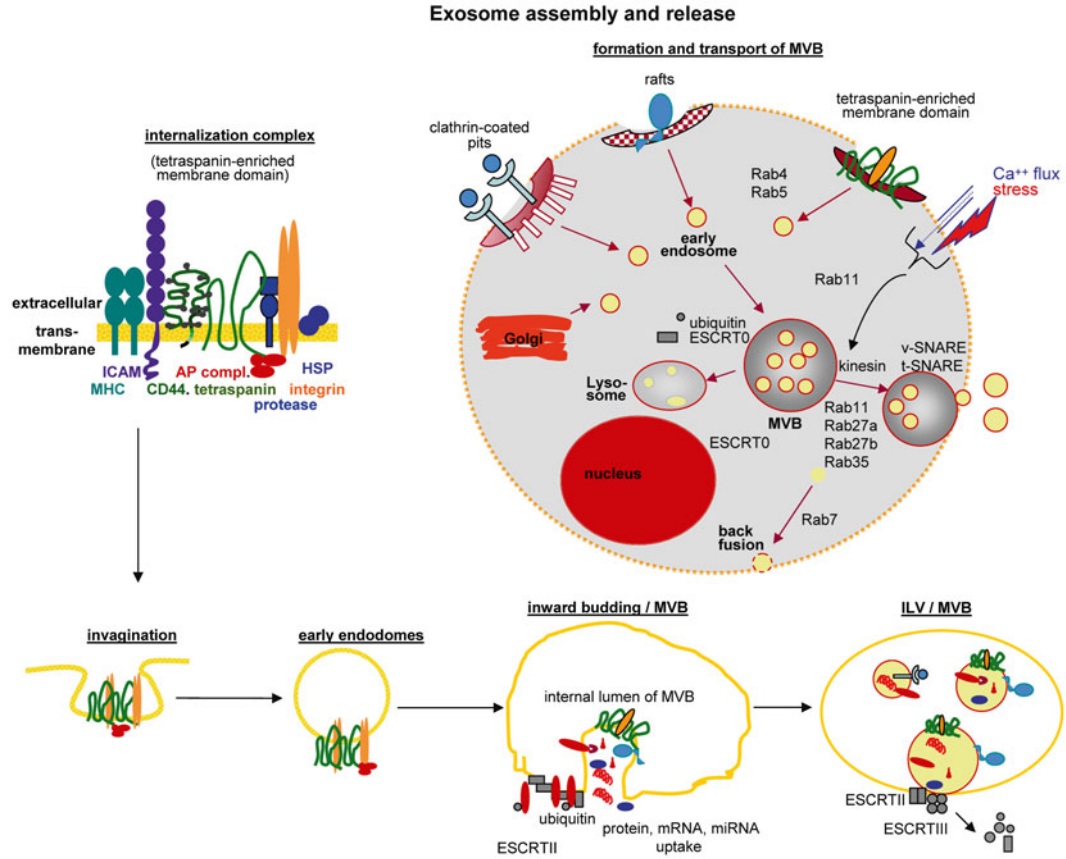


Fig. 1 Exosome assembly and release. Exosomes derive from inward budding of membrane microdomains like rafts, clathrin-coated pits, or tetraspanin-enriched microdomains. After fission and scission, the early endosome proteins become monoubiquitinated and are guided by the ESCRT machinery and Rab4 and Rab5 toward MVB, where they are loaded with cytoplasmic proteins, mRNA, and miRNA. By an energy-dependent process the invaginated early endosomes dissociate from the MVB membrane and then are termed ILV. MVB move along microtubules, which requires kinesin, rab25, and rab27 toward the cell membrane. SNARE proteins facilitate the release of the ILV, which are then called exosomes

enriched, while others are not recovered in exosomes, although they may be abundant in the donor cells (Fig. 2). This implies, outlined below, that exosome loading includes selection processes.

3.1 The Composition of Exosomes

3.1.1 Lipids

Lipids play an important role in membrane rigidity and stability as well as in fusion and budding processes. Exosomes contain sphingomyelin, phosphatidylcholine, phosphatidylethanolamine, phosphatidylserine, ganglioside GM3, and phosphatidylinositol, prostaglandins, and lysobisphosphatidic acid [56, 57], with enrichment of sphingomyelin, cholesterol, GM3, and phosphatidylserine [56]. Sphingomyelin and cholesterol allow the tight packing of lipid bilayers and increase overall rigidity and stability [58]. GM3 also increases the stability of exosomes [59].

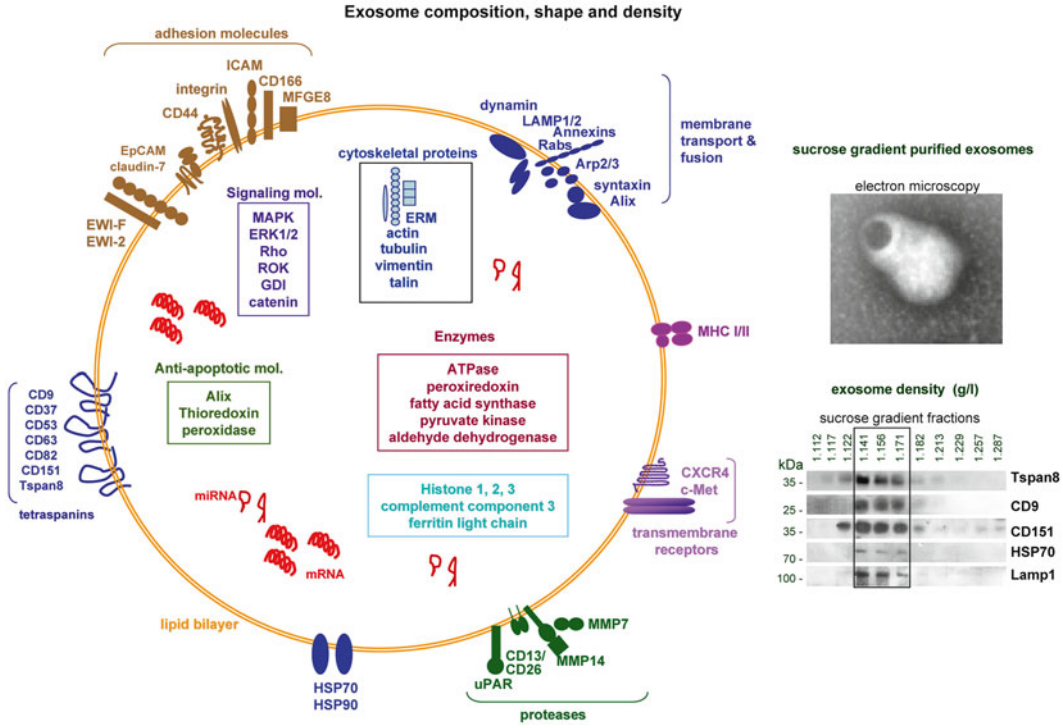


Fig. 2 Exosome constituents. Exosomes are composed of a lipid bilayer enriched in cholesterol, sphingomyelin, GM3, and phosphatidylserine. Constitutive membrane components are tetraspanins, adhesion molecules, proteases, and transmembrane receptors according to the donor cell membrane, a variety of fission, scission and vesicle transport molecules and selectively recruited cytoplasmic proteins as well as mRNA and small non-coding RNA including miRNA. Exosomes are 50–100 μm in size and, according to the lipid composition, have a density of 1.14–1.17 g/l

Conical-shaped phosphatidylserine helps to assemble the curved vesicular shape of exosomes and facilitates their fusion and fission [60]. Lysobisphosphatidic acid is involved in intracellular fusion and budding [56].

The rationale for studying exosomal lipids is not only because they carry bioactive lipids, but also that lipid second messengers such as phosphatidic acid, diglycerides, and ceramides are involved in exosome biogenesis. Thus, proteins of the ESCRT machinery interact with various lipids or lipid-related enzymes. Vps4 interacts with an oxysterol binding protein [27] making a link with cholesterol metabolism. Genome-scale screens for genetic interactions that affect Golgi/endosome/vacuole sorting unveil a key role for lipids in general, and more specifically for sterols and fatty acids [61]. Notably, too, packaging of miRNA into exosomes requires the neutral sphingomyelinase (nSMase2) for generating the lipid mediator ceramide, which triggers an exosome biogenesis pathway independent of the ESCRT machinery [37]. Furthermore, because of their high cholesterol content, exosomes can modify upon

transfer the lipid homeostasis of target cells [62]. For a detailed analysis, I recommend an excellent review that additionally outlines the lipid composition as a mode to differentiate between exosomes and microvesicles [63].

3.1.2 Proteins

Extensive investigations of exosome proteins have been carried out using mass spectrometry (MS)-based proteomic analyses, Western blotting (WB), and immunoelectron microscopy [64]. MS-based proteomic studies provided a high-throughput vesicular proteome dataset in various cell types and body fluids, where over 4000 proteins have been identified so far [65]. Proteomic studies on exosomes of various origins suggest controlled protein-sorting, as exosomes from different cell types contain some common vesicular proteins. Common vesicular proteins are tetraspanins (CD9, CD63, CD81, CD82, CD151, Tspan8), which are constitutive components of exosomes [34, 66]. Tetraspanins are enriched 7–124-fold in exosomes as compared to the parental cells [49] and are frequently used to differentiate exosomes from other extracellular vesicles [10]. Additional molecules most abundantly recovered in exosomes are adhesion molecules, proteases, MHC molecules, HSP60, HSP70 and HSP90, the ESCRT components TSG101 and Alix, annexins, the cytoskeleton proteins actins, cofilin-1, ERM proteins (ezrin/radixin/moesin), profilin-1, and tubulins, metabolic enzymes and cytosolic signal transduction molecules and ribosomal proteins [67–70]. In general, plasma membrane and cytoplasmic proteins are more commonly sorted into exosomes than nuclear and mitochondrial proteins [64, 71]. Constitutive exosomal proteins derived from the cell membrane are mostly located in internalization prone membrane domains, which besides other has been well explored for tetraspanins. These studies additionally revealed that with invagination of membrane microdomains protein complexes are maintained such that, e.g. tetraspanins coimmunoprecipitate with integrins in the plasma membrane and exosomes [35, 72]. Similarly, the transferrin receptor is recruited by its association with HSP. Corresponding observation also accounts for cytosolic proteins that frequently are associated with transmembrane proteins or are located close to the outer membrane of MVB [73, 74]. Thus, common proteins are mainly structural vesicle component or are involved in vesicle biogenesis and vesicle trafficking.

Exosomes also contain cell type-specific proteins, where for diagnosis surface markers are particularly important. Flow cytometry is an easy to handle and reliable detection method [75]. Most knowledge has so far been collected for tumor and tumor stem cell-associated markers. Thus, melanoma TEX contain the tumor-associated antigen MART1, epithelial cell-derived TEX contain epithelial cell adhesion molecule, EpCAM [76, 77], glioblastoma TEX contain EGFRVIII [78] and TEX of

docetaxel-resistant prostate cancer cells contain (multidrug resistance gene 1 (MDR-1)) [79]. Others described the transfer of the oncoprotein MET, mutant KRAS, and tissue factor [80–82].

3.1.3 miRNA

miRNA is a class of small noncoding RNA, which endogenously regulates gene expression at the posttranscriptional level [83]. miRNA range in size from 19 to 25 nucleotides. They regulate translation and degradation of mRNA through base pairing to complementary sites mostly in the 3′ untranslated region [84]. MiR constitute only 1–3 % of the human genome, but control about 30 % of the coding genes [85], most miR controlling multiple mRNA [86]. miRNA in the serum may derive from necrosis, apoptosis [87] or be actively released in microvesicles [88]. Free extracellular miRNA is associated with AGO, which accounts for the stability of the free miRNA [89].

Exosomes also contain miRNA, and additionally mRNA, rRNA, tRNA, mitochondrial DNA, and short DNA sequences of retrotransposons [90–92], which are protected from degradation due to the double lipid membrane [25, 89]. Using microarray- and "next-generation" sequencing (NGS)-based systemic approaches, significant quantities particularly of mRNA and miRNA have been recovered in exosomes. As already mentioned, the horizontal transfer of exosomal mRNAs and miRNAs can lead to epigenetic reprogramming of recipient cells [93]. Furthermore, disease-related differences in exosomal mRNAs and miRNAs may be a useful diagnostic tool [25, 94].

The biogenesis of miRNAs starts in the cell nucleus where DNA containing miRNAs is transcribed by RNA polymerase II to generate primary miRNAs (pri-miRNAs) [95]. The pri-miRNAs are processed by a microprocessor complex, consisting of the RNase type III endonuclease Drosha and an essential cofactor (DiGeorge syndrome critical region 8)/Pasha (protein containing two double-stranded RNA binding domains), to generate a precursor miRNA (pre-miRNA) [96]. The pre-miRNA is exported to the cytoplasm by the exchange factor of the guanine Ran nucleotide (GTP-binding nuclear protein Ran) and the exportin-5 receptor. Exported pre-miRNA is then processed by another RNase type III endonuclease known as Dicer, releasing a ~22-nucleotide miRNA duplex. One strand of the RNA duplex is selected to be subsequently loaded into the RNA-induced silencing complex (RISC) along with AGO2 and GW182 [97]. GW182 and AGO2, both main components of the RISC are recovered in MVB [90, 98]. Though it is still unanswered whether pre-miRNAs are readily incorporated into exosomes and later processed to mature miRNAs, it was repeatedly reported that pre-miRNAs as well as mature miRNAs are released in exosomes [99, 100].

Mature miRNA binds mostly to the 3′ UTR of target mRNA and represses translation. The mechanism of mRNA silencing depends on the degree of complementarities. With perfect base pairing between miRNA and mRNA, mRNA is cleaved by AGO. Upon imperfect binding, protein translation is repressed [101].

Expression of miRNA is measured using miRNA microarray, miRNA real-time qRT-PCR, and NGS. qRT-PCR can profile large sets of miRNA simultaneously and is suited for quantification. miRNA platforms allow for simultaneous analysis of all known miRNA. Only NGS provides a means to identify accurately a variety of small RNA species [102].

The increased knowledge on miRNA greatly fostered progress in oncology, where miRNA could be linked to prognosis, disease progression, local recurrence, and metastasis [103–105]. miRNA plays an important role in epithelial–mesenchymal transition (EMT) [105, 106], maintenance of CIC [107, 108], as well as tumor invasion, migration, and angiogenesis [109]. EMT is regulated by the mir-200 family, miR-141, miR-429, and miR-205. The expression level of miR-200 negatively correlates with znc finger E-box-binding homeobox (ZEB)1 and 2, which inhibit E-cadherin expression [110]. In PaCa, downregulation of miR-30 correlates with EMT, targets being vimentin and snail-1 [111]. Examples for the involvement of miR in CIC control are the tumor suppressor miR-34 that regulates Notch and Bcl2 [112] and miR-21 that correlates with chemoresistance [113]. Instead, miR-9, regulating E-cadherin expression, is suggested to be of major importance for metastasis-associated mobility and invasiveness [114]. miR-34a overexpression can inhibit metastasis by regulating CD44 [115] and miR-340 suppresses invasion and metastasis by regulating c-Met and via c-Met metalloproteinases (MMP) 2 and 9 [116].

3.1.4 Sorting into Endosomes

Though the exosome composition mostly allows identifying the donor cells, the relative abundance of proteins, lipids, mRNA, and miRNAs differs between exosomes and donor cells, which implies active sorting into MVB.

Protein sorting depends on mono-ubiquitination and ESCRT. Thus, Vps4 is involved in the secretion of shedding vesicles enriched in arrestin-domain-containing proteins (ARRDC), where secretion of these vesicles apparently is driven by interaction of the PSAP motif of ARRDC with the UEV motif of Tsg101. Ubiquitination of ARRDC also seems to be important for the secretion of ARRDC-bearing shedding vesicles [117]. Sorting of proteins also is facilitated by oligomerization [118]. Acylation or myristoylation additionally support recruitment into exosomes [118, 119]. In line with this, protein recruitment into tetraspanin networks and other internalization prone detergent-resistant membrane domains can also be decisive, where raft microdomains

enriched in sphingolipids, which form ceramide, play an important role [37, 119, 120]. Notably, recruitment via tetraspanin-enriched or raft microdomains is accompanied by exosomal recovery of protein complexes rather than singular molecules, that may have an impact on exosome targeting and the crosstalk with target structures [121].

Besides proteins and lipids, miRNA is selectively recruited into exosomes. Recruitment may be guided by a zip code in the 3′-UTR [122]. miRNA recruitment is facilitated by physical and functional coupling of RISC to components of the sorting complex. GW182 containing GW bodies, sorted into MVB, promote continuous assembly/disassembly of membrane-associated miRNA-loaded RISC [90, 98]. Furthermore, a specific EXOmotif (GGAG) was identified that controls miRNAs loading into exosomes by binding to the heterogeneous ribonucleoprotein A2B1 (hnRNPA2B1). HnRNPA2B1 is mostly sumoylated in exosomes, and this modification is essential for the transport function of hnRNPA2B1 through binding to an RNA transport signal (RTS or A2RE) present in the 3′UTR and containing the EXOmotifs identified in exosomal miRNAs [123, 124]. Annexin-2 is another protein that plays a role in RNA sorting into exosomes by binding specific RNAs [125].

Thus, neither proteins nor RNAs are randomly loaded into exosomes. Though there are first hints on the sorting mechanisms, many questions remain to be answered.

4 Exosome Isolation

Exosomes being released in the extracellular space are purified from cell culture supernatants and biological fluids like plasma or serum, urine, saliva, and milk. Commonly used isolation methods involve a series of differential centrifugations to remove cells and large debris followed by ultracentrifugation at $100{,}000 \times g$ for 1–2 h to pellet the exosomes [126]. According to their lipid bilayer, exosomes have a characteristic flotation density of 1.13–1.21 g/l, which allows for further enrichment by sucrose gradient centrifugation after ultracentrifugation. Size exclusion chromatography, preferably high-performance liquid chromatography provides another means of purification [127, 128]. An alternative method, asymmetrical field flow fractionation with in-lie ultraviolet absorbance, dynamic light scattering, and multi-angle light scattering, was recently described. It is suggested to purify exosomes beyond the methods described so far and should allow for large-scale isolation in clinical application [129]. Additionally, based on charge neutralization with 0.1 M acetate exosomes can be precipitated and are readily resolubilized in acetate-free buffer at neutral pH [130]. Capturing exosomes via antibodies recognizing molecules highly expressed on exosomes provides an alternative.

Immunoaffinity chromatography or magnetic bead-activated cell sorting are used [131, 132]. Finally, many commercial kits are available.

The preferred method depends on the special request and the amount of available material. Commercially available kits as well as MACS are rapid isolation procedures with a comparably high yield. However, up to now, the purity may be suboptimal. Sequential centrifugation with a final ultracentrifugation step is work intensive, but the yield is comparably high. When combined with a 0.22 μm filtration step, at least very large vesicles are excluded and the sterility of the sample is guaranteed. Sucrose gradient after ultracentrifugation still counts as the golden standard, although it may have minor contaminations by other vesicles, as these are mostly larger, but can be small and also vary in density. Nonetheless, this purification procedure is accepted for clinical use [133, 134]. Though the yield is low, sucrose gradient centrifugation is also recommended (as far as sufficient material is available) when aiming for extensive proteome, mRNA, or miRNA analyses.

Exosome purity can be analyzed by transmission electron microscopy for shape and size. Flow cytometry using exosome-specific marker antibodies can be used for exosome characterization including cell-type-specific exosomes [135]. WB serves the same goal. When combined with sucrose density fractionation, it additionally confirms the nature of exosomes compared to other extracellular vesicles and apoptotic bodies. A variety of commercially available ELISA-based kits allows multiple testing of individual samples [136]. qRT-PCR is the standard procedure to define the mRNA and miRNA content.

As stated, exosomes are secreted by most cells of the organism and are supposed to be of importance in embryogenesis, and nearly all physiological and pathological processes of multicellular organisms including plants [137]. I tried to lay the ground with this short and quite superficial introduction for understanding by no means fully explored multitude of functions that are attributed to these small vesicles. For the remaining of this review I will restrict to TEX. Building on the characterization of exosomes, the power of TEX as a diagnostic tool will be elaborated. Progressing with their functional activity, I will finish with a brief outlook for their therapeutic use in the future.

5 Cancer and Cancer-Initiating Cell Exosomes as a Diagnostic Tool

5.1 Why Exosomes Can Offer a Reliable Diagnosis?

For many types of cancer the mortality rate is still high, which accounts in particular for pancreatic cancer (PaCa), which is the most deadly cancer with an overall 5-year survival rate below 1 % and a mean survival time of 4–6 months [138]. Late initial diagnosis, which prohibits resection, as well as chemotherapy and radiation

resistance and early metastatic spread account for this unfavorable prognosis [139]. Nonetheless, even for cancer with a more favorable diagnosis, metastatic spread, chemotherapy, and radiation-resistance are mostly the cornerstone for curative therapy, where particularly early metastatic growth may escape imaging procedures. The detection of migrating tumor cells in the blood, though very sensitive, is burdened by a high rate of false positives as many migrating tumor cells die in advance of finding a site for metastatic settlement and growth. Serum protein markers as CEA, AFP, PCA and in the case of PaCa, CA19-9, are frequently reliable only in combination with additional screening parameters. Free serum miRNA has been used in several instances and could reveal promising results. In our hands, sensitivity and specificity of exosome-free serum was below the level to be recommended for a diagnostic or prognostic procedure [140]. Whether this was due to the exclusion of exosomes remains to be answered. Irrespective of this point, the risk of diluting out relevant miRNA by abundant "unselected" mRNA and miRNA from apoptotic bodies, microvesicles or "free" miRNA appears high. Instead, exosomes are found in all body fluids and, importantly, are characterized by two marker profiles, proteins and miRNA that both can easily be evaluated by flow cytometry or ELISA-based assays and qRT-PCR, respectively. In concern of the exosome source, high quantities are found in serum and plasma, with no significant differences between the two sources. The lowest amount of exosomes is found in the urine [141, 142]. Thus, urine may be a suitable source only for tumors of the urogenital tract. Similarly, sputum exosomes should preferentially be taken into account in suspicion of lung, esophageal, or pharyngeal cancer. Notably, too, exosomes are very stable [143], which allows for follow-up studies including stored serum samples. Thus, exosomes have considerable advantages in tumor diagnosis [144, 145].

5.2 The Marker Panel Selection

I am much concerned about panel selection as not every tumor cell and TEX thereof express all potentially interesting protein markers and carry all potentially relevant miRNA. This accounts for the individual tumor/TEX thereof and even more for tumors/TEX from different patients. Besides differences in the phenotype of a given tumor including intratumor variability, due to any fluctuation, epigenetic factors or others, like age, gender, diurnal changes, health status and many more, TEX number and composition will vary. What are reasonable criteria for selecting the panel?

There is strong evidence that CIC are responsible for radiation and chemoresistance as well as for metastasis formation [146, 147]. CIC are a small subpopulations of cells and migrating tumor cells may even be a subpopulation of CIC [148]. Nonetheless, it is highly recommended to search for the markers of this minority of tumor cells. Fortunately, there is evidence that CIC secrete far

higher amount of exosomes than non-CIC and that these exosomes are enriched in the serum [149]. For starting, I recommend to use exosomes from CIC-enriched subpopulations. CIC can be enriched by repeated passage as spheroids or holoclones. Both spheroid growth and holoclone formation are natural features of adult stem cells and as such independent of tumor transformation [150, 151], which increases the likeliness to select for stem-cell-like markers rather than for "mere" oncogenes. I recommend a pool of exosomes from 10 or more tumor lines, as far as available. For preevaluation exosomes should be highly purified by sucrose density gradient centrifugation [152], even if not feasibly in large screenings due to limits in serum availability and/or ultracentrifuge capacity.

For the selection of CIC protein markers, exosomes should be analyzed by MS using exosomes of non-transformed lines as controls. Differentially expressed proteins can be verified by ELISA [153] or flow cytometry, where latex beads can be coated with antibody in advance or latex beads are loaded with exosomes and marker expression is evaluated by incubation with antibodies after blocking free binding sites of the latex beads [135, 154]. The choice of the method depends on which screening procedure one is aiming for. For miRNA microarray screening, recently a thorough comparative evaluation of mRNA preparation was published [155], which should be taken into account as in dependence of the exosome source minor differences may lead to a pronounced loss of miRNA. Microarray data are controlled by qRT-PCR. For the miRNA, the prescreening procedure should be repeated with exosomes of at least two pools of patient versus healthy person serum. One might include exosomes from an additional serum pool from patients with chronic inflammatory disease of the relevant organ, like serum exosomes from patients with colitis ulcerosa for colon carcinoma TEX or chronic pancreatitis for PaCa patients' TEX. In our hands the results with exosomes from pancreatic cancer lines and serum from patients with PaCa were strikingly overlapping. Excluding those miRNA that were abundant in healthy donors serum exosomes, the 10 miRNA with the highest copy number were identical and the first 50 miRNA showed only slight variations in ranking [140]. A similar proteomic analysis with exosomes from patient sera is recommendable. By shortage of patient material and the comparably high cost, we did not perform a second proteomic analysis with exosomes from pooled patients' sera. Furthermore, the number of CIC protein markers is more restricted, Vesiclepedia, a compendium for extracellular vesicles [156], is continuously updated and there is large consensus on CIC markers in most cancer types [157], whereas with the exception of few miRNA known to be engaged, e.g. in mesenchymal transition, several reviews report on great variability between different studies even in concern about

the same tumor entity [158–160]. This may in part be fostered by the still growing number of newly described miRNA. With rapid progress in NGS, this caution may be outdated in a few years.

Having selected the protein marker panel, we suggest concomitant evaluation of the corresponding antibodies. This requires less material and is easier to translate in a clinical setting. We used flow cytometry; ELISA may work as well. However, two points need thorough control. Some antibodies may exhibit unspecific binding, when used in a panel. We experienced this with an anti-CD24 IgM antibody that exhibited no binding to exosomes from healthy donors. when tested by itself, but did so, when being concomitantly tested with additional CIC marker-specific antibodies, where the IgM nature of this antibody may contribute to unspecific binding. The second concern is about individual CIC marker expression in disease states other than cancer. This can be evaluated in the initial proteomic analysis, but also can be performed at this step with two or three pools of sera from patients with non-malignant disease. As exosomes from 0.5 ml serum are mostly sufficient for evaluating the CIC protein markers, the required amount of serum is low enough to proceed with a repetition of individual serum exosomes and individual marker antibodies to exclude an unsuited marker, where required. Last to mention, when using flow cytometry, we suggest indirect staining of latex bead-coated exosomes with unlabeled primary antibodies followed by dye-labeled secondary antibodies, as the staining intensity of exosomes loaded on latex beads is comparably low. On the other hand, distinct to most tumor cells, latex beads exhibit negligible autofluorescence. This allows FITC-labeled secondary antibodies and also high fluorescence intensity amplification without undue fluorescence intensity of the negative control.

Having finished prescreening, where a similar protocol was published for serum-free miRNA in PaCa [161] one should proceed with a comparably small number of samples split in a training and a validation set to define the AUC with a 95 % confidence interval and the receiver-operating characteristic (ROC). We used for both the miRNA evaluated by qRT-PCR and the CIC protein markers evaluated by flow cytometry a weighted score in as far as we took the mean RQ values for the selected miRNAs and a fluorescence index, which was defined as the % stained beads × the mean fluorescence intensity. According to our experience, there was a slight, but statistically significant gain in positive samples, without any increase in false positives. Finally, statistical evaluation revealed a higher sensitivity with a minimal loss in specificity, when taking into account both samples with a positive RQ or a positive FI [140]. The improved specificity relies mostly on CIC protein markers being already detected at early disease stages and RQ values increasing, though not reaching statistical significance, with tumor volume and metastatic settlement.

There are two additional concerns: (1) Due to some common features between chronic inflammation and cancer, patients with chronic inflammation are at risk to become false positive [162]. Should this be the case, we recommend to take an additional screening panel building on an inflammatory profile to exclude false positives; (2) We reached 100 % sensitivity and 93 % specificity testing PaCa patients' serum exosomes against exosomes from healthy donors and patients with non-malignant disease, but only 80 % specificity including exosomes from patients with non-PaCa malignancies located in the pancreatic gland [140]. Though this group was small and cancer entities at distinct organs were not included, one can expect this result taking into account that most of the CIC markers and the majority of oncomiR and metastomiR are not restricted to one tumor type, e.g. CD44 is a tumor stem cell marker in many solid organ derived cancers and leukemia [163] and miR-21 is recovered in different tumor entities [164]. I have no convincing solution to this problem. From a practical point of view, I would argue that for screening of people at risk for a given cancer as well as for a follow-up, it is in the first instance important to define a cancer-related exosome profile, which according to the patient's history may guide toward the expected type of cancer. Alternatively, one can include a third protein marker and/or miRNA panel which takes hold of organ-specific markers. I consider this not as a necessity in routine screening. It may be beneficial in rare cases where the origin of the primary tumor is unknown.

Taken together, serum (or other body fluid) exosomes appear as a most promising diagnostic tool. Exosomes from 1 ml of serum is sufficient for screening a panel of protein and miRNA markers, and mostly allows for an additional repetition. Exosome preparation, as far as not commercially available kits are used, takes some time. Flow cytometry and qRT-PCR are standard methods and not very time-consuming. Compared to most screening procedure, serum exosome analysis is comparably cheap. Besides easy accessibility, additional major advantages rely on (1) the enriched presence of CIC exosomes in the serum, although the underlying mechanisms remain to be clarified and (2) recovery and (strong) enrichment of CIC protein markers. This likely is due to all CIC markers described so far being located in internalization prone membrane domains or to be additionally engaged in endosome assembly [165]. (3) Exosomes selectively recruit miRNA, which also results in a strong enrichment. Though this prohibits a straightforward translation from cellular toward exosomal miRNA profiles, available studies argue for an enrichment of onco- and metastomiR in exosomes [166]. With progress in this field, part of the suggested prescreening can possibly be omitted.

Finally, I want to apologize for not going into detail on the TEX protein and miRNA marker profiles. There have been close to 200 reviews on exosomes and close to 100 reviews on TEX in

2014, which include extensive literature searches. In addition, Vesiclepedia [156] is continuously updated. Thus, I kindly ask the reader to look for the latest literature on her/his special request. Taken together and irrespective of the need for further defining miRNA regulation in cancer and TEX, data provided so far give hope that a combined evaluation of TEX protein and miRNA markers might soon be the method of choice for a first diagnostic screen including patients at risk as well as for follow-up studies.

6 Exosomes, Target Selection, and Exosome Uptake

Exosomes are the most potent intercellular communicators. This discovery has revolutionized many aspects of biological sciences and is expected to bring a major breakthrough in therapy, including cancer. The power of exosomes relies on their ubiquitous presence, their particular protein profile, their equipment with mRNA and miRNA and their most efficient transfer in target cells. Information on the latter aspect, the uptake by target cells and the exosomes' target cell selection are two prerequisites for clinical translation.

Exosomes are taken up by target cells. Thus, exosomal mouse mRNA and miRNA was recovered in human cells after co-culture [25] and luciferin-loaded exosomes induced bioluminescence in luciferase expressing cells [167], which implies merging of the exosomal cytosol with the target cell cytoplasm through membrane fusion either at the plasma membrane or after uptake. The most common method for detecting exosomes uptake uses lipophilic fluorescent dyes, like PKH67, PKH26, rhodamine B, DiI, and DiD. Alternatively, chemical compounds, like CFSE and CFDA can be used, which become fluorescent in the cytoplasm after esterification [168, 169]. To differentiate between binding and uptake, the target cell can be stripped by acid treatment or trypsin [170]. The latter procedures confirm exosome internalization. Exosome uptake can also be visualized by fusing an abundant exosomal protein with a fluorescent tag. All these methods have some drawbacks, like dye leakiness or exosome clumping or altered tagged protein configuration. Nonetheless there is overwhelming evidence that exosomes are, indeed, taken up. However, the route of uptake is still disputed.

There is evidence for clathrin-mediated endocytosis, GEM-supported endocytosis, phagocytosis, macropinocytosis and membrane fusion, which possibly are not mutually exclusive [168, 171–175]. The mode of exosome uptake depends in part on the protein signature of the exosomes and the target cell as well as on membrane subdomains of the target cell that may change with the target cells activation state. Accordingly, blocking antibodies and reagents that modulate targeted proteins or disturb membrane

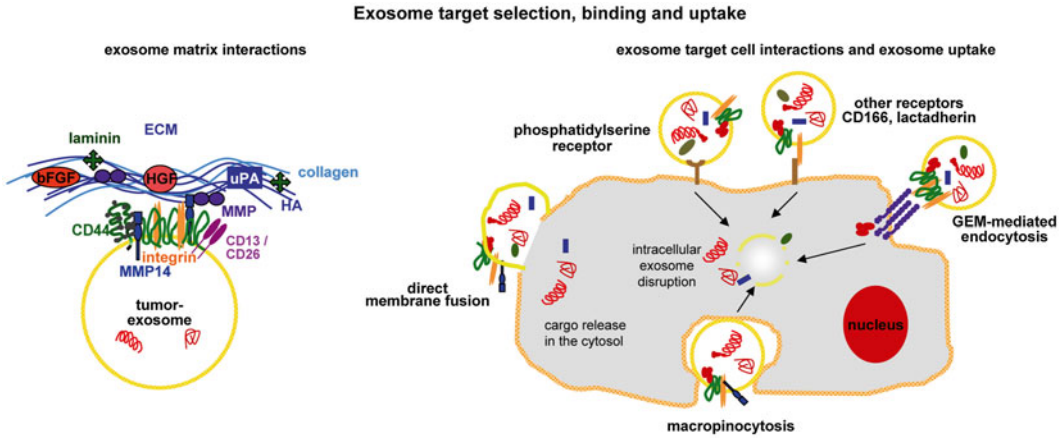


Fig. 3 Exosome binding and uptake. Exosomes bind via their adhesion molecules to the extracellular matrix, e.g. via CD44 to HA and via integrins to fibronectin, laminin and collagen. Exosomes also bind to target cells or fuse with the target cell membrane or are taken up by macropinocytosis or are endocytosed in glycolipid-enriched membrane microdomains. The mode of the intracellular exosome membrane disruption is not yet fully explored

microdomains are commonly used to obtain hints toward the underlying mechanism. To give a few examples: treatment with chemicals providing evidence for endocytosis have been heparin that targets heparansulfate proteoglycans and asialofetuin, which targets Galectin 5, or cytochalasinB and D and latrunculin that destroy the actin cytoskeleton [167, 176–179]. Caveolin- and clathrin-dependent uptake was suggested by dynamin and dynamin-2 inhibitors [176, 177]; the efficacy of cholesterol depletion by methyl-β-cyclodextrin, filipin or other chemicals argues for lipid raft-mediated endocytosis [35, 44, 167] (Fig. 3). Antibody blocking studies provided evidence for the engagement of tetraspanins, likely via tetraspanin-associated transmembrane molecules, e.g. integrins that bind to ICAMs on target cells [35, 180, 181]. In fact, the engagement of integrins, like CD11a binding to ICAM-1 [182] has been amply demonstrated though mostly without evaluating the association with tetraspanins. The T-cell receptor complex in association with accessory molecules binds to MHC complexes on DC exosomes [183, 184]. Blocking studies by chemicals and/or antibodies can be confirmed by a transient knockdown of the molecule suggested to be engaged in exosome uptake. Finally, there is also evidence for membrane fusion, which involves SNARES, Rab and Sec1-related proteins, which can be probed by fluorescent lipid quenching [52]. Though plasma membrane fusion may not play a major role, fusion with endosomal membranes is expected, where the acidic pH of endosomes will be supportive [55, 185].

As an alternative to antibody blocking, chemicals or siRNA downregulation, we suggest as a method of choice for identifying exosomal receptors and target cell ligands a mutual pulldown of target membrane lysates by sepharose-coupled exosome membranes and exosome membrane lysate passage over sepharose-coupled target cell membranes followed by SDS-PAGE and proteome analysis. Under mild lysis conditions this procedure allows, in addition, to uncover besides the direct target molecule, associated protein complexes. In our hands this procedure pointed toward an engagement of GEM-located membrane molecules as well as fission and scission engaged internalization complexes [35].

There are additional hints that the selectivity of exosome uptake is guided or, at least, facilitated by the engagement of protein complexes at the exosomes and the target cell. In fact, only defined tetraspanin–integrin complexes are taken up by selected target cells and exosome uptake proceeds via binding to internalization prone microdomains [35]. The constitutively high expression of tetraspanins, located in TEM and associated with a multitude of molecules [33, 34], favors our suggestion. Furthermore, the requirement of two signals to initiate target cell activation is a common phenomenon. Thus, T-cell activation requires engagement of the T-cell receptor and accessory molecules that interact with MHC and costimulatory molecules on DC [186]. Such a requirement for a double hit could well contribute to the selectivity of exosome uptake. However, the question is by no means solved.

Even less is known on the fate of uptaken exosomes [187]. There is some evidence in DC that exosomes may use different passenger molecules for the intracellular traffic toward early endosomes for fusion and/or integration into MVB, where colocalization with LAMP1 is poor, but colocalization with the tetraspanin CD81 and CathepsinD, one of the central proteases in MHCII-MVB is pronounced [184, 188]. For epithelial cells, itineration was described [189], but a systemic analysis of the trafficking of exosomes in host cells is still missing.

Taken together, there is strong evidence that target cells take up exosomes. Exosomal receptors and target cell ligands are poorly defined. We propose the engagement of complexes on both sites, which could contribute to the selectivity of exosome uptake. A precise answer to this question is most essential for exosome-based therapy. The processing of uptaken exosomes also is poorly defined, hints so far being mostly derived from cells of the hematopoietic system. Again, detailed studies are needed to unravel, how uptaken exosomes modulate or reprogram their targets, which involves exosomal proteins, including signaling molecules, mRNA, and miRNA [4, 190].

7 Target Modulation by TEX

Exosome components are function-competent [191, 192]. I will give an overview on the engagement of TEX in central features of tumor progression, which includes modulation of the matrix, the surrounding stroma and the premetastatic niche, angiogenesis and EMT, the latter likely relying on a crosstalk between CIC TEX and Non-CIC.

7.1 Matrix Modulation by Exosomes

Exosomes are rich in proteases, which are active and modulate the exosomes protein profile, but as well the extracellular matrix (ECM) and target cells. Besides others, MMP2, 7, 9, 14, ADAM (A disintegrin and metalloproteinase) 10, 15, 17, ADAMTS1, 13 and several dipeptidases have been detected in TEX [69, 193].

TEX proteases can modulate the TEX protein profile. This was described for L1 and CD44 shedding by ADAM10 and for EpCAM, CD46, TNFR1 by unknown metalloproteinases, where proteolytic activity can be regulated through the association with tetraspanins [194, 195].

Tetraspanin-associated proteases in TEX also modulate the ECM. Tetraspanins associate with integrins, major ligands for ECM proteins as well as peptidases, ADAMs, MMPs and uPAR, where particularly the association with MMP14 plays an important role in pericellular lysis [196, 197]. The collagenolytic and laminin-degrading activity of exosomes facilitates angiogenesis and metastasis which includes modulation of the matrix in the pre-metastatic organ [198, 199]. Degradation of aggrecan by glioma TEX ADAMTS1, 4 and 5 increases glioma cell invasiveness [200]. CD151 and Tspan8 complexes with α6β4, MMP14 or TACE in TEX are particularly important in collagen IV and laminin degradation [70, 72, 201]. TEX Cathepsin B also contributes to matrix remodeling [202].

The ECM is not only a structural element, but a storage of bioactive compounds and an essential component in tissue repair as well as in the crosstalk between tumor cells and the stroma [203]. Thus, modulation of the ECM by exosomal proteases also accounts for cytokine/chemokine and protease liberation and generation of cleavage products that promote motility, angiogenesis, and stroma cell activation [69, 204].

Taken together, the modulation of the ECM by TEX proteases creates a path for migrating cells, favors a tumor growth-promoting microenvironment, angiogenesis, and premetastatic niche establishment.

7.2 TEX-Initiated Target Cell Modulation

TEX-initiated signal transduction can be promoted by TEX binding and uptake and TEX uptake-induced target cell modulation may frequently represent the combined result of protein

transfer-initiated signal transduction, transferred mRNA translation, and mRNA silencing by miRNA. Precise answers are additionally handicapped by the still incomplete knowledge on miRNA and the only starting awareness of other forms of non-coding RNA. In concern about miRNA, most miRNA have multiple targets, which can hamper associating changes in the exosome target mRNA profile with the exosomal miRNA profile and/or signal transduction molecules or transcription factors transferred by the exosome or becoming activated by exosomal proteins. To give a simple example, we noted that high expression and transfer of miR-494 and miR-542-3p correlated with downregulation of cadherin-17, which is a target of these two miRNA. We also noted upregulation of MMP2 and MMP9 upon target cell cocultivation with TEX or these two miRNAs. Cadherin-17 is known to repress MMP2 and MMP9 transcription. Thus, the exosomal miRNA initiated a circle with upregulation of metastasis-promoting genes by downregulation of a metastasis suppressing mRNA [205]. We do not know, whether additional transcription factors in the exosomes directly promoted MMP transcription. Thus, looking for exosome-induced signal transduction is like scratching an iceberg with a pin. This should by no means be discouraging, instead increase awareness of the present limits and, accordingly, putting results in place. In silico network analyses and extended NGS will wave these concerns. At present, it is recommendable to start with miRNA microanalysis and proteomic of a given exosome population and to correlate modulation of the target cell by mRNA microarray with the miRNA and the protein profile of the exosomes. Alternatively, one may start with an analysis of exosome-induced target cell modulation and go back to search, which TEX component(s) can account for the observed effects. These studies already revealed convincing evidence that TEX interfere with immune response induction, are engaged in modulating the tumor stroma, support angiogenesis, promote preparation of a premetastatic niche, and may contribute to EMT.

7.2.1 TEX and the Immune System

DC exosomes are one of the best explored examples of exosome binding-initiated target cell activation. DC exosomes can replace DC in immune response induction and exosome-based therapy was first explored in the context of DC exosomes as a cancer vaccine [206]. DC exosomes are particularly rich in CD9 and CD81, MHC I and II, where MHC molecules colocalize and associate with tetraspanins during vesicle formation [207]. DC also take up exosomes secreted by other cells, including TEX, which they internalize and process for presentation, CD9 and CD81 also being important for uptake by DC [182]. Thus, DC use exosomes as a source of antigen and produce exosomes that suffice for

T-cell activation, both features expanding the operational range of DC [208].

Hope that DC exosomes are a promising means for immunotherapy [209], was dampened by TEX interfering with immune response induction (reviewed in [210–212]).

However, TEX also can support immune response induction and effector cell activity, one of the well explored mechanisms being based on the abundant recovery of HSP in TEX, which function as a danger signal in the immune system provoking NK and cytotoxic T-cell activation. Stress forces exosomal HSP release, which strengthens DC maturation, helper T cell, CTL, and NK activity [213–215]. Increased immunogenicity of TEX from heat-stressed tumor cells is further promoted by chemokines, which attract DC and T cells such that tumor growth becomes retarded [216, 217]. Finally, TEX can be a strong immunogen. Tumor antigens, which are non-immunogenic when presented by tumor cells, induce a potent Th, CTL and B-cell response and led to a decrease in regulatory T cells, when presented by TEX [218]. There is evidence that this is a sequel of TEX being more efficiently taken up and processed by DC than tumor lysate [184].

Taken together, TEX intensely communicate with all elements of the immune system. Depending on the TEX composition and the activation status of the immune system, they are immunosuppressive or support immune response induction and effector cell activity. Based on the knowledge gathered so far, TEX-loaded DC vaccination appears as a promising option in cancer immunotherapy [219, 220].

7.2.2 TEX and Angiogenesis

Angiogenesis is one of the hallmarks of cancer and first evidence for the engagement of TEX has already been presented in 1998, when it was noted that exosomes delivered by a Tspan8-expressing pancreatic tumor line induced a consumption coagulopathy [221], which was demonstrated to require the uptake of exosomes expressing a Tspan8–CD49d complex by endothelial cells, where the exosomes promoted maturation of progenitors and proliferation of mature EC [180, 222]. Meanwhile, intense efforts have been taken to elaborate the contribution of TEX in angiogenesis.

TEX contain TNFα, IL1β, TGFβ, and TNFR1, which recruit EC progenitors and stimulate EC by paracrine signaling [122, 223]. EGFR-positive TEX taken up by EC elicit EGFR-dependent responses including activation of the MAPK and Akt pathway and VEGFR2 expression [224]. Exosomal transfer of Notch-ligand-delta-like-4 inhibits Notch signaling and increases angiogenesis [217]. In chronic myeloid leukemia (CML) TEX-induced angiogenic activity is Src-dependent [225]. EC proliferation can also be promoted by the transfer of exosomal cell cycle-related mRNA, which was demonstrated for colorectal cancer TEX [226]; glioblastoma TEX-induced angiogenesis relies on the transfer of

exosomal proteins and mRNA [94]; renal cell CIC TEX carry proangiogenic mRNA and miRNA, which trigger the angiogenic switch [199]; TEX from a metastatic melanoma line educate BM progenitors via c-Met toward a provasculogenic phenotype, where an important contribution of exosomal miRNA is suggested [227]; leukemia TEX contain miR-92a that is transferred into EC and downregulates CD49e, which selectively increases migration and tube formation [228].

7.2.3 TEX, the Tumor Stroma, and the Premetastatic Niche

Besides on the ECM, hematopoietic, and endothelial cells, TEX exert a strong impact on tumor stroma as well as stroma of (pre) metastatic organs. Early reports on the information transfer via exosomes showed for embryonic stem cells (ESC)-exosomes that they transfer messages into hematopoietic progenitor cells to promote survival and expression of early pluripotency markers, where exosomal proteins, mRNA, and miRNA were active contributors [88].

TEX uptake induces changes in recipient non-tumor cells that suffice to drive tumor growth as described for tissue transglutaminase and fibronectin [229] or high level c-Met uptake by BMC, which leads to their re-education to support premetastatic niche formation for melanoma cells, where in melanoma patients, too, circulating BM-derived cells express c-Met [227]. TEX also transport apoptosis inhibitory proteins [230] and present TGFβ, which drives differentiation of fibroblast toward myofibroblasts, supporting tumor growth, angiogenesis, and metastasis formation [231]. Breast cancer TEX convert adipose-tissue-derived mesenchymal SC into myofibroblasts with increased expression of α-SMA, SDF1, VEGF, CCL5, TGFβ, TGFβRI and II, accompanied by SMAD pathway activation [232]. Lung cancer TEX uptake stimulates the tumor stroma to secret IL8, VEGF, LIF, oncostatin, and MMPs, which support tumor growth [233].

TEX also promote the generation of a premetastatic niche. We explored that TEX together with a soluble tumor matrix supported recruitment of hematopoietic progenitors from the BM as well as activation of stroma cells and leukocytes in premetastatic lymph nodes such that a non-metastatic tumor line settled and formed metastases [234]. Similar findings were reported by Hood et al. for melanoma TEX [223]. A complex of exosomal HSP90 with MMP2, tissue plasminogen activator, and annexin II promotes plasmin activation facilitating tumor cell motility [235]. Reprogramming of the BM niche by AML TEX is promoted by the transfer of IGF-IR mRNA, which supports stroma cell proliferation. By the transfer of miR-150 from AML TEX into hematopoietic progenitors expression of CXCR4 becomes reduced, which impairs progenitor cell migration [236].

While TEX modulate host cells to allow for recruitment and growth of tumor cells [12, 227, 234], exosomes from

non-transformed cells also affect tumor cells. This was first described for platelet-derived exosomes, which via the transfer of the αIIb integrin chain to lung cancer cells, stimulate the MAPK pathway, increase expression of MT1-MMP, cyclin D2 and angiogenic factors and enhance adhesion to fibrinogen and EC [237]. Activated T-cell exosomes can promote tumor invasion via FAS signaling, where FasL+ exosomes initiate activation of the ERK and NFκB pathway in melanoma cells with subsequent upregulation of MMP expression [238]. Fibroblast exosomes promote breast cancer motility via Wnt planar polarity signaling [239]. BM stroma exosomes from patients with multiple myeloma force multiple myeloma progression, these exosomes showing a lower content of tumor suppressor miR-15a, but higher levels of oncogenic proteins, cytokines, and adhesion molecules than healthy donor BM stroma exosomes [240]. Also tumor-associated macrophages secrete exosomes with high miR-223 that causes nuclear accumulation of β-catenin [241].

Finally, two points should be mentioned, (1) the cargo of TEX can vary depending on their route of generation and (2) the same TEX may induce different responses in distinct target cells. Thus, overexpression of CD9 or CD82 promotes secretion of exosomes that contain β-catenin, thereby reducing its cellular content and impairing Wnt signaling. The reduction in β-catenin proceeds via tetraspanin-associated E-cadherin and is ESCRT-independent [36]. Besides indicating that the cargo of exosomes differs depending on ESCRT- or tetraspanin-initiated internalization, this study demonstrates that by depletion of inhibitors or stimulators exosomes can opposingly affect signal transduction [242]. TEX-promoted tumor growth may also vary for individual tumors. A deficit in Rab27a leads to reduced TEX production, which affected growth of a tumor line that required recruitment of neutrophils, but not of another neutrophil-independent line [243].

7.2.4 CIC TEX, Non-CIC, and EMT

TEX, mostly CIC TEX, also spread information into other tumor cells via protein, mRNA, and miRNA transfer [88, 244].

One of the first evidences to support that TEX-uptake plays a critical role in autocrine stimulation of tumor growth revealed that the intercellular transfer of the oncogenic receptor EGFRvIII via TEX to glioma cells, lacking this receptor, causes transformation of indolent glioma cells [78] and reprograms growth factor pathways in EC [94]. Other oncogenes, like Ras, Myc, SV40T also induce signaling and gene expression [81, 91, 228], where, e.g. exosomal amphiregulin, an EGFR ligand, increased tumor invasiveness fivefold compared to the recombinant protein, indicating that the transfer of amphiregulin is accompanied by additional messages delivered via the TEX [245]. Exosomes from virus transfected cells also transfer viral miRNA [246], e.g. EBV transfected nasopharyngeal carcinoma transfer viral BART miRNA [247].

In lung cancer TEX miR-21 and miR-29a act as a TLR ligand, functioning as agonist and leading to NFκB activation and IL6 and TNFα secretion, which promote metastasis [248]. Hepatocellular carcinoma TEX contain a set of highly enriched miRNA that are not detected in the donor cell, including miR-584, miR-517c and others, one of the potential targets being identified as TGFβ-activated kinase 1, which activates JNK and MAPK pathway and NFκB. In cocultures, these exosomal miRNA promoted anchorage-independent growth and apoptosis resistance [249]. Apoptosis resistance can also rely on the transfer of MDR1 [250], which is enriched in TEX [251].

Finally, EMT, a fundamental process during embryology is transiently refreshed by tumor cells and suggested to allow for liberation from the primary tumor mass and to support the first steps of the metastatic cascade [252, 253]. It is accompanied by downregulation of E-cadherin, upregulation of vimentin, tissue factor and, notably, significant changes in the exosome profile [239, 254, 255], including TGFβ, TNFα, IL6, TSG101, Akt, ILK1, β-catenin, hepatoma-derived growth factor, casein kinase II, annexinA2, α3 integrin, caveolin, and MMPs [256–260]. These and additional factors were recovered in exosomes after oncogenic H-Ras-induced EMT. The authors point out that they demonstrated for the first time the strong impact of oncogenic transformation on the protein content of exosomes and that these exosomes likely can induce EMT in recipient cells [255]. Qualitative changes in the proteome of exosomes accompanying EMT were also reported for SSC-derived A431 upon transit into the mesenchymal state. At least 30 proteins were only recovered in "EMT" exosomes, with most pronounced changes in proteins involved in proliferation, communication, junction formation, ECM adhesion, and exosome generation [261]. Also, latent membrane protein 1 of EBV is found in TEX and modulates together with HIF1α EMT marker expression in recipient cells [262]. We reported that exosomes from CIC-enriched colon cancer lines can induce EMT in the CIC-depleted population, which is accompanied by upregulation of N-cadherin, downregulation of E-cadherin, and strong induction of Notch [263]. Finally, the transfer of miR-409 from stroma exosomes induces EMT in adjacent prostate cancer cells [264].

In brief, TEX can pave the way for tumor growth and progression by shielding tumor cells from an immune attack, by modulating the tumor and the host acellular matrix to promote tumor cell migration, by supporting tumor cell survival via angiogenesis induction, drug expulsion, and the transfer of drug resistance genes. TEX also affect the host stroma to provide a tumor growth and progression favorable milieu, where TEX-conditioned stroma cell can give a feedback toward EMT induction. Finally, CIC exosomes can confer CIC features toward non-CIC including EMT (Fig. 4). It is obvious that tumor cells adapt the generation

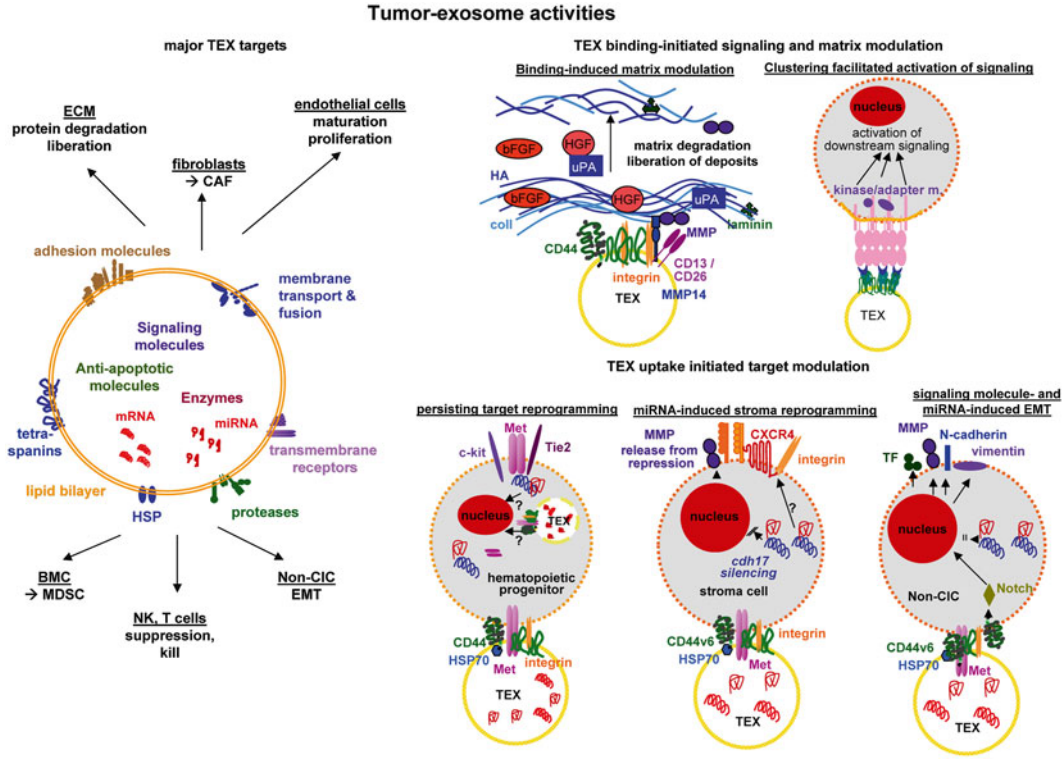


Fig. 4 Exosome functions. Major target of tumor exosomes (TEX) are the ECM, stroma cells, endothelial cells, BMC, and neighboring tumor cells. ECM binding promotes matrix degradation by exosomal proteins, which is accompanied by liberation of incorporated cytokines, chemokines, and (pro)proteases. Clustering of membrane receptor by exosomal protein complexes facilitates target cell activation, well known for TNFRI and CD95 clustering. Receptor activation may well cooperate with uptaken exosomal proteins, mRNA, and particularly miRNA. Uptaken exosomes were described to drive hematopoietic stem cells into maturation, myeloid progenitors into myeloid-derived suppressor cells, endothelial progenitors cells into maturation and expansion and to support adhesion molecule, protease and cytokine expression in stroma cells and to initiate epithelial–mesenchymal transition (EMT) in Non-CIC

of TEX according to their instantaneous need and also that TEX components work in concert. Still being at the descriptive level, technical resources and the hope for a new wave of therapeutics will push rapid knowledge progression.

8 Outlook: Exosomes as Therapeutics

Exosomes are optimal therapeutics. They are natural non-synthetic and non-viral products, are small and flexible, which allows them to cross biological membranes. By a lipid bilayer they protect their cargo from degradation. Being easily transfected, they appear as ideal and possibly most potent "drug" delivery system. In fact, unmodified exosomes can display native therapeutic activity [265]

as demonstrated, besides others, for DC exosomes that induce a T-cell response [220] or MSC exosomes that support heart reperfusion or neurite outgrowth [266, 267]. Instead, the activity of TEX, very efficient in promoting tumor growth and progression at all critical points from tumorigenesis until metastatic growth, urges for hindrance. Nonetheless, attempts to translate experimental studies on the power of exosomes into therapeutic settings, should be aware of "preparatory work" still to do.

8.1 Impeding TEX Delivery

One of the most straight forward approaches is affinity plasmapheresis, eliminating TEX carrying a specific receptor like HER2 (Aethlon ADAPT™) [268]. The efficiency of this approach is burdened by the ongoing delivery of TEX. As an alternative, blocking of uptake by annexinV or Diannexin [78, 224] or heparin, which competes with heparansulfate proteoglycans [269, 270] are discussed. Antibody blocking may in most instances display too low specificity. Nonetheless, in a rat model, where exosomes expressing the tetraspanin Tspan8 induced disseminated intravascular coagulation, blocking exosomes by a Tspan8-specific antibody completely prevented undue angiogenesis, although primary tumor growth was not impaired [201, 222]. Blocking with a bispecific antibody that targets with one arm a constitutive exosome component, preferentially a tetraspanin due to the engagement in target selection [35], and with the second arm a tetraspanin-associated adhesion molecule or a tumor marker, has to our knowledge not yet been explored.

Targeting the machinery for exosome generation, as described for the ESRCT machinery [271], rab27 and rab35 [272, 273] might well be considered as an additional support. Trials to interfere with exosome generation to avoid drug expulsion by changing the pH follow the same line [55].

One exception should be mentioned, where TEX are collected rather than being eliminated. TEX might be a superior antigen source for DC loading, particularly when immunogenic tumor antigens are unknown. Thus, DC pulsed with TEX of an AML line or a murine myeloid tumor line provoke a strong anti-leukemia response [184, 274].

8.2 Exosome Tailoring and Delivery

In concern about generating therapeutic exosomes, it is of invaluable advantage that exosomes are easy to manipulate [265, 275]. This can be achieved by transfection of the donor cell, mostly explored for mRNA and miRNA [276–279] or by directly loading exosomes, preferentially by electroporation [279, 280]. For protein or peptide loading, generating a fusion protein with an abundant natural exosome cargo like lactadherin or Lamp2 appears advantageous [276, 280–282]. Furthermore, utilizing the native mechanisms for packaging RNA can significantly increase the cargo load. This was achieved by incorporating RNA zipcodes in the 3′ untranslated region, that direct RNA localization in the cell [283].

There are multiple cargos to be discussed for "therapeutic" exosomes, like a drug that efficiency is hampered by poor solubility, but is efficiently delivered by exosomes, e.g. curcumin [284, 285], proteins that promote pro-apoptotic signaling or prevent activation of anti-apoptotic signaling cascades [286, 287] or oncosuppressive or metastasis suppressive miRNA like let-7 [288] or miR-34 [289, 290] or miR-26a, which induces cell cycle arrest via targeting cyclins D2 and E2 and exerted a dramatic protective effect in a mouse hepatoma model [291]. Various forms of miRNA inhibitors [276, 292], e.g. anti-miR-146 [78], anti-mi-R21 [293] also revealed promising results. Of central importance will be the elimination of CIC themselves, where the engagement of TEX in Wnt and Notch signaling may provide a point of attack [109, 294–296]. The delivery of miRNA and anti-miRNA is frequently facilitated by concomitantly equipping the exosomes with a selective targeting advice.

The targeted delivery of exosomes is still of major concern and possibly requires adapting for defined tumor entities. Virus-derived proteins/peptides have been reported to facilitate delivery toward B cells or neural cells [281, 297]. Ligands for tumor cell marker proteins have also been successfully applied as, e.g. the GE11 peptide, which binds the EGFR [276]. Finally, uptake and proper site delivery within the target cell can be facilitated by equipping exosomes with cell-penetrating peptides, short cationic or amphipathic peptides that induce fusion of cellular membranes [280, 298].

Taken together, though exosomes are ideal therapeutics, large-scale exosome collection for therapeutic use is expensive and time-consuming and availability can become a limiting factor. Thus, great efforts are undertaken to generate "exosome mimetics" [299, 300] (Fig. 5).

9 Conclusion

The discovery that the organism avails on a mobile ubiquitously present instruction system, exosomes, has revolutionized many aspects of biological sciences, where the small size and stability guarantees accessibility and availability. As in many other respects, tumors as "foreign bodies" in the organism have adopted this system to reeducate the host. Thus, knowing the system and its mode of action might bring a major breakthrough in oncology. We might be very close to a safe and highly specific non- or minimally invasive diagnostic tool. A broad range therapeutic use awaits further exploration of TEX targeting receptors and their ligands, to avoid side effects and, in concern about target cell reprogramming a precise knowledge on signaling molecules and miRNA targets. These are not insurmountable hurdles, but answering these questions will take some time.

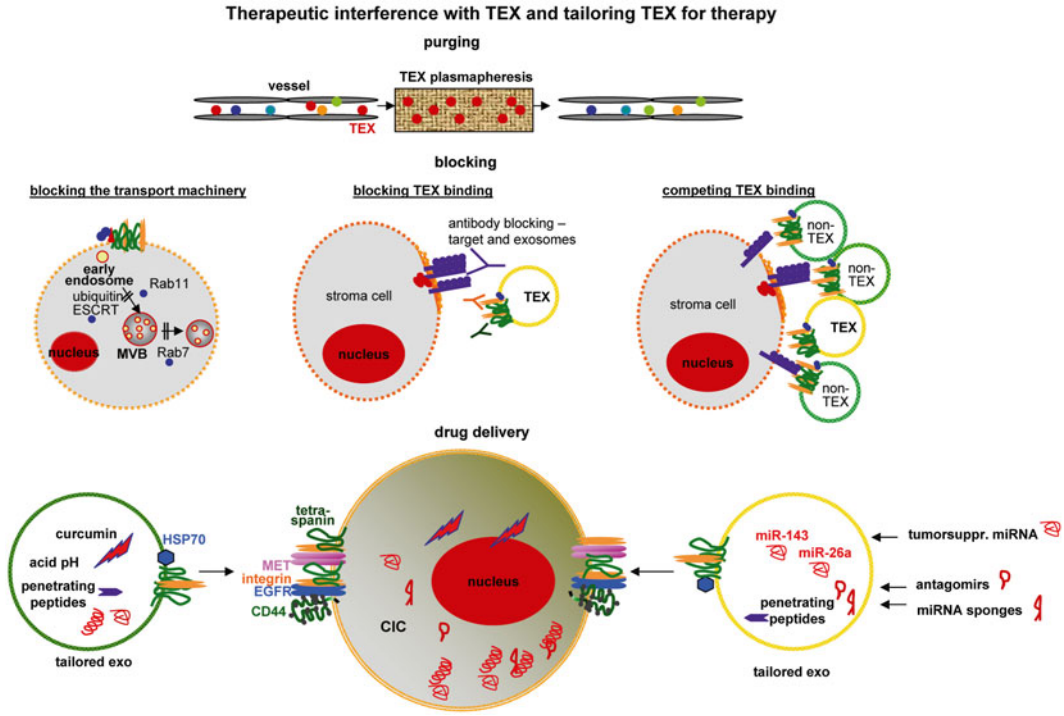


Fig. 5 Therapeutic interference with TEX. Two therapeutic strategies are discussed, elimination of blocking of TEX and interfering with artificial/tailored exosomes. Elimination of TEX can be approached by plasmapheresis, blocking by interfering with exosome generation by destruction of the lipid bilayer, the ESCRT machinery, Rabs (e.g. by shRNA), antibody blocking at the target cell or the exosome level or by competition with non-TEX exosomes that are equipped with the TEX target structures. Active therapeutic interference relies on the transfer of cytotoxic drugs, like curcumin, an excess of oncomiR or metastomiR suppressing miRNA or blocking of oncomiR/metastomiR by various forms of anti-miR. Therapeutic delivery can be facilitated by special targeting advices and/or penetrating peptides

Acknowledgement

Grant support: This investigation was supported by the German Cancer Aid (grant No: 109144) and the Wilhelm-Sander Stiftung.

References

1. Maryáš J, Faktor J, Dvořáková M et al (2014) Proteomics in investigation of cancer metastasis: functional and clinical consequences and methodological challenges. Proteomics 14: 426–440
2. Hong IS, Lee HY, Nam JS (2014) Cancer stem cells: the 'achilles heel' of chemo-resistant tumors. Recent Pat Anticancer Drug Discov 10:2–22
3. Colak S, Medema JP (2014) Cancer stem cells--important players in tumor therapy resistance. FEBS J 281:4779–4791
4. György B, Szabó TG, Pásztói M et al (2011) Membrane vesicles, current state-of-the-art: emerging role of extracellular vesicles. Cell Mol Life Sci 68:2667–2688
5. O'Loughlin AJ, Woffindale CA, Wood MJ (2012) Exosomes and the emerging field of

exosome-based gene therapy. Curr Gene Ther 12:262–274

6. Lee Y, El Andaloussi S, Wood MJ (2012) Exosomes and microvesicles: extracellular vesicles for genetic information transfer and gene therapy. Hum Mol Genet 21(R1): R125–R134
7. Corrado C, Raimondo S, Chiesi A et al (2013) Exosomes as intercellular signaling organelles involved in health and disease: basic science and clinical applications. Int J Mol Sci 14:5338–5366
8. Simons M, Raposo G (2009) Exosomes--vesicular carriers for intercellular communication. Curr Opin Cell Biol 21:575–581
9. Vlassov AV, Magdaleno S, Setterquist R et al (2012) Exosomes: current knowledge of their composition, biological functions, and diagnostic and therapeutic potentials. Biochim Biophys Acta 1820:940–948
10. Mathivanan S, Ji H, Simpson RJ (2010) Exosomes: extracellular organelles important in intercellular communication. J Proteomics 73:1907–1920
11. Kharaziha P, Ceder S, Li Q et al (2012) Tumor cell-derived exosomes: a message in a bottle. Biochim Biophys Acta 1826:103–111
12. Lee TH, D'Asti E, Magnus N et al (2011) Microvesicles as mediators of intercellular communication in cancer--the emerging science of cellular 'debris'. Semin Immunopathol 33:455–467
13. Simpson RJ, Lim JW, Moritz RL et al (2009) Exosomes: proteomic insights and diagnostic potential. Expert Rev Proteomics 6:267–283
14. Rak J (2010) Microparticles in cancer. Semin Thromb Hemost 36:888–906
15. Record M, Subra C, Silvente-Poirot S et al (2011) Exosomes as intercellular signalosomes and pharmacological effectors. Biochem Pharmacol 81:1171–1182
16. Martins VR, Dias MS, Hainaut P (2013) Tumor-cell-derived microvesicles as carriers of molecular information in cancer. Curr Opin Oncol 25:66–75
17. Lässer C (2012) Exosomal RNA as biomarkers and the therapeutic potential of exosome vectors. Expert Opin Biol Ther 12(Suppl 1): S189–S197
18. Raposo G, Stoorvogel W (2013) Extracellular vesicles: exosomes, microvesicles, and friends. J Cell Biol 200:373–383
19. Pap E, Pállinger E, Pásztói M et al (2009) Highlights of a new type of intercellular communication: microvesicle-based information transfer. Inflamm Res 58:1–8
20. Sotelo JR, Porter KR (1959) An electron microscope study of the rat ovum. J Biophys Biochem Cytol 5:327–342
21. Stahl PD, Barbieri MA (2002) Multivesicular bodies and multivesicular endosomes: the "ins and outs" of endosomal traffic. Sci STKE 2002(141):PE32
22. Stoorvogel W, Kleijmeer MJ, Geuze HJ et al (2002) The biogenesis and functions of exosomes. Traffic 3:321–330
23. Pan BT, Johnstone RM (1983) Fate of the transferrin receptor during maturation of sheep reticulocytes in vitro: selective externalization of the receptor. Cell 33:967–977
24. Denzer K, Kleijmeer MJ, Heijnen HF et al (2000) Exosome: from internal vesicle of the multivesicular body to intercellular signaling device. J Cell Sci 113:3365–3374
25. Valadi H, Ekström K, Bossios A et al (2007) Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells. Nat Cell Biol 9:654–659
26. Henne WM, Buchkovich NJ, Emr SD (2011) The ESCRT pathway. Dev Cell 21:77–91
27. Katzmann DJ, Odorizzi G, Emr SD (2002) Receptor downregulation and multivesicular-body sorting. Nat Rev Mol Cell Biol 3:893–905
28. Colombo M, Moita C, van Niel G et al (2013) Analysis of ESCRT functions in exosome biogenesis, composition and secretion highlights the heterogeneity of extracellular vesicles. J Cell Sci 126(Pt 24):5553–5565
29. Baietti MF, Zhang Z, Mortier E et al (2012) Syndecan-syntenin-ALIX regulates the biogenesis of exosomes. Nat Cell Biol 14: 677–685
30. Schuh AL, Audhya A (2014) The ESCRT machinery: from the plasma membrane to endosomes and back again. Crit Rev Biochem Mol Biol 49:242–261
31. Buschow SI, Nolte-'t Hoen EN, van Niel G et al (2009) MHC II in dendritic cells is targeted to lysosomes or T cell-induced exosomes via distinct multivesicular body pathways. Traffic 10:1528–1542
32. vanNiel G, Charrin S, Simoes S et al (2011) The tetraspanin CD63 regulates ESCRT-independent and-dependent endosomal sorting during melanogenesis. Dev Cell 21: 708–721

33. Hemler ME (2005) Tetraspanin functions and associated microdomains. Nat Rev Mol Cell Biol 6:801–811
34. Zöller M (2009) Tetraspanins: push and pull in suppressing and promoting metastasis. Nat Rev Cancer 9:40–55
35. Rana S, Yue S, Stadel D et al (2012) Toward tailored exosomes: the exosomal tetraspanin web contributes to target cell selection. Int J Biochem Cell Biol 44:1574–1584
36. Chairoungdua A, Smith DL, Pochard P et al (2010) Exosome release of beta-catenin: a novel mechanism that antagonizes Wnt signaling. J Cell Biol 190:1079–1091
37. Trajkovic K, Hsu C, Chiantia S et al (2008) Ceramide triggers budding of exosome vesicles into multivesicular endosomes. Science 319:1244–1247
38. Kajimoto T, Okada T, Miya S et al (2013) Ongoing activation of sphingosine 1-phosphate receptors mediates maturation of exosomal multi-vesicular endosomes. Nat Commun 4:2712
39. Laulagnier K, Grand D, Dujardin A et al (2004) PLD2 is enriched on exosomes and its activity is correlated to the release of exosomes. FEBS Lett 572:11–14
40. Kirkegaard T, Roth AG, Petersen NH et al (2010) Hsp70 stabilizes lysosomes and reverts Niemann-Pick disease-associated lysosomal pathology. Nature 463:549–553
41. Stenmark H (2009) Rab GTPases as coordinators of vesicle traffic. Nat Rev Mol Cell Biol 10:513–525
42. Barr F, Lambright DG (2010) Rab GEFs and GAPs. Curr Opin Cell Biol 22:461–470
43. Sonnichsen B, De Renzis S, Nielsen E et al (2000) Distinct membrane domains on endosomes in the recycling pathway visualized by multicolor imaging of Rab4, Rab5, and Rab11. J Cell Biol 149:901–914
44. Huotari J, Helenius A (2011) Endosome maturation. EMBO J 30:3481–3500
45. Sato M, Sato K, Liou W et al (2008) Regulation of endocytic recycling by C. elegans Rab35 and its regulator RME-4, a coated-pit protein. EMBO J 27:1183–1196
46. Hsu C, Morohashi Y, Yoshimura S et al (2010) Regulation of exosome secretion by Rab35 and its GTPase-activating proteins TBC1D10A-C. J Cell Biol 189:223–232
47. Kelly EE, Horgan CP, Goud B et al (2012) The Rab family of proteins: 25 years on. Biochem Soc Trans 40:1337–1347
48. Ducharme NA, Ham AJ, Lapierre LA et al (2011) Rab11-FIP2 influences multiple components of the endosomal system in polarized MDCK cells. Cell Logist 1:57–68
49. Hoshino D, Kirkbride KC, Costello K et al (2013) Exosome secretion is enhanced by invadopodia and drives invasive behavior. Cell Rep 5:1159–1168
50. Martin-Cofreces NB, Baixauli F, Sanchez-Madrid F (2014) Immune synapse: conductor of orchestrated organelle movement. Trends Cell Biol 24:61–72
51. Pfeffer SR (2010) Two Rabs for exosome release. Nat Cell Biol 12:3–4
52. Jahn R, Scheller RH (2006) SNAREs – engines for membrane fusion. Nat Rev Mol Cell Biol 7:631–643
53. Südhof TC, Rizo J (2011) Synaptic vesicle exocytosis. Cold Spring Harb Perspect Biol 3:12
54. Merendino AM, Bucchieri F, Campanella C et al (2010) Hsp60 is actively secreted by human tumor cells. PLoS One 5:e9247
55. Parolini I, Federici C, Raggi C et al (2009) Microenvironmental pH is a key factor for exosome traffic in tumor cells. J Biol Chem 284:34211–34222
56. Subra C, Laulagnier K, Perret B et al (2007) Exosome lipidomics unravels lipid sorting at the level of multivesicular bodies. Biochimie 89:205–212
57. Subra C, Grand D, Laulagnier K et al (2010) Exosomes account for vesicle-mediated transcellular transport of activatable phospholipases and prostaglandins. J Lipid Res 51: 2105–2120
58. Ramstedt B, Slotte JP (2002) Membrane properties of sphingomyelins. FEBS Lett 531:33–37
59. Allen TM, Austin GA, Chonn A et al (1991) Uptake of liposomes by cultured mouse bone marrow macrophages: influence of liposome composition and size. Biochim Biophys Acta 1061:56–64
60. Chernomordik LV, Kozlov MM (2003) Protein-lipid interplay in fusion and fission of biological membranes. Annu Rev Biochem 72:175–207
61. Costanzo M, Baryshnikova A, Bellay J et al (2010) The genetic landscape of a cell. Science 327:425–431
62. Muller G, Schneider M, Biemer-Daub G et al (2011) Microvesicles released from rat adipocytes and harboring glycosylphosphatidylinositol-anchored proteins transfer RNA stimulating lipid synthesis. Cell Signal 23: 1207–1223
63. Record M, Carayon K, Poirot M et al (2014) Exosomes as new vesicular lipid transporters involved in cell-cell communication and various pathophysiologies. Biochim Biophys Acta 1841:108–120

64. Raimondo F, Morosi L, Chinello C et al (2011) Advances in membranous vesicle and exosome proteomics improving biological understanding and biomarker discovery. Proteomics 11:709–720
65. Mathivanan S, Simpson RJ (2011) ExoCarta: a compendium of exosomal proteins and RNA. Available from: http://exocarta.org/index.html
66. Pols MS, Klumperman J (2009) Trafficking and function of the tetraspanin CD63. Exp Cell Res 315:1584–1592
67. Choi DS, Kim DK, Kim YK et al (2013) Proteomics, transcriptomics and lipidomics of exosomes and ectosomes. Proteomics 13: 1554–1571
68. Xie Y, Bai O, Zhang H et al (2010) Membrane-bound HSP70-engineered myeloma cell-derived exosomes stimulate more efficient CD8(+) CTL- and NK-mediated antitumour immunity than exosomes released from heat-shocked tumour cells expressing cytoplasmic HSP70. J Cell Mol Med 14:2655–2666
69. Shimoda M, Khokha R (2013) Proteolytic factors in exosomes. Proteomics 13: 1624–1636
70. Mu W, Rana S, Zöller M (2013) Host matrix modulation by TEX promotes motility and invasiveness. Neoplasia 15:875–887
71. Henderson MC, Azorsa DO (2012) The genomic and proteomic content of cancer cell-derived exosomes. Front Oncol 2:38
72. Yue S, Mu W, Erb U et al (2015) The tetraspanins CD151 and Tspan8 are essential exosome components for the crosstalk between cancer initiating cells and their surrounding. Oncotarget 6:2366
73. Choi DS, Yang JS, Choi EJ et al (2012) The protein interaction network of extracellular vesicles derived from human colorectal cancer cells. J Proteome Res 11:1144–1151
74. Sahu R, Kaushik S, Clement CC et al (2011) Microautophagy of cytosolic proteins by late endosomes. Dev Cell 20:131–139
75. Julich H, Willms A, Lukacs-Kornek V et al (2014) Extracellular vesicle profiling and their use as potential disease specific biomarker. Front Immunol 5:413
76. Andre F, Schartz NE, Movassagh M et al (2002) Malignant effusions and immunogenic tumour-derived exosomes. Lancet 360:295–305
77. Runz S, Keller S, Rupp C et al (2007) Malignant ascites-derived exosomes of ovarian carcinoma patients contain CD24 and EpCAM. Gynecol Oncol 107:563–571
78. Al-Nedawi K, Meehan B, Micallef J et al (2008) Intercellular transfer of the oncogenic receptor EGFRvIII by microvesicles derived from tumour cells. Nat Cell Biol 10:619–624
79. Corcoran C, Rani S, O'Brien K et al (2012) Docetaxel-resistance in prostate cancer: evaluating associated phenotypic changes and potential for resistance transfer via exosomes. PLoS One 7:e50999
80. Park JA, Sharif AS, Tschumperlin DJ et al (2012) Tissue factor-bearing exosome secretion from human mechanically stimulated bronchial epithelial cells in vitro and in vivo. J Allergy Clin Immunol 130:1375–1380
81. Demory Beckler M, Higginbotham JN, Franklin JL et al (2013) Proteomic analysis of exosomes from mutant KRAS colon cancer cells identifies intercellular transfer of mutant KRAS. Mol Cell Proteomics 12:343–355
82. Ji H, Greening DW, Barnes TW et al (2013) Proteome profiling of exosomes derived from human primary and metastatic colorectal cancer cells reveal differential expression of key metastatic factors and signal transduction components. Proteomics 13:1672–1680
83. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233
84. Calin GA, Croce CM (2006) MicroRNA signatures in human cancers. Nat Rev Cancer 6:857–866
85. Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9:102–114
86. Lim LP, Lau NC, Garrett-Engele P et al (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433:769–773
87. Tzimagiorgis G, Michailidou EZ, Kritis A et al (2011) Recovering circulating extracellular or cell-free RNA from bodily fluids. Cancer Epidemiol 35:580–589
88. Ratajczak J, Miekus K, Kucia M et al (2006) Embryonic stem cell-derived microvesicles reprogram hematopoietic progenitors: evidence for horizontal transfer of mRNA and protein delivery. Leukemia 20:847–856
89. Turchinovich A, Weiz L, Langheinz A et al (2011) Characterization of extracellular circulating microRNA. Nucleic Acids Res 39:7223–7233
90. Gibbings DJ, Ciaudo C, Erhardt M et al (2009) Multivesicular bodies associate with components of miRNA effector complexes and modulate miRNA activity. Nat Cell Biol 11:1143–1149

91. Balaj L, Lessard R, Dai L et al (2011) Tumour microvesicles contain retrotransposon elements and amplified oncogene sequences. Nat Commun 2:180
92. Salido-Guadarrama I, Romero-Cordoba S, Peralta-Zaragoza O et al (2014) MicroRNAs transported by exosomes in body fluids as mediators of intercellular communication in cancer. Onco Targets Ther 7:1327–1338
93. Chen X, Liang H, Zhang J et al (2012) Horizontal transfer of microRNAs: molecular mechanisms and clinical applications. Protein Cell 3:28–37
94. Skog J, Wurdinger T, van Rijn S et al (2008) Glioblastoma microvesicles transport RNA and proteins that promote tumour growth and provide diagnostic biomarkers. Nat Cell Biol 10:1470–1476
95. Kim VN, Han J, Siomi MC (2009) Biogenesis of small RNAs in animals. Nat Rev Mol Cell Biol 10:126–139
96. Han J, Lee Y, Yeom KH et al (2004) The Drosha-DGCR8 complex in primary microRNA processing. Genes Dev 18: 3016–3027
97. Suzuki HI, Miyazono K (2011) Emerging complexity of microRNA generation cascades. J Biochem 149:15–25
98. Lee YS, Pressman S, Andress AP et al (2009) Silencing by small RNAs is linked to endosomal trafficking. Nat Cell Biol 11:1150–1156
99. Chen TS, Lai RC, Lee MM et al (2010) Mesenchymal stem cell secretes microparticles enriched in pre-microRNAs. Nucleic Acids Res 38:215–224
100. Li L, Zhu D, Huang L et al (2012) Argonaute 2 complexes selectively protect the circulating microRNAs in cell-secreted microvesicles. PLoS One 7:e46957
101. Doench JG, Sharp PA (2004) Specificity of microRNA target selection in translational repression. Genes Dev 18:504–511
102. Pinto R, De Summa S, Petriella D et al (2014) The value of new high-throughput technologies for diagnosis and prognosis in solid tumors. Cancer Biomark 14:103–117
103. Braicu C, Tomuleasa C, Monroig P et al (2015) Exosomes as divine messengers: are they the Hermes of modern molecular oncology? Cell Death Differ 22:34–45
104. Sato-Kuwabara Y, Melo SA, Soares FA et al (2015) The fusion of two worlds: non-coding RNAs and extracellular vesicles--diagnostic and therapeutic implications (Review). Int J Oncol 46:17–27
105. Alečković M, Kang Y (2014) Regulation of cancer metastasis by cell-free miRNAs. Biochim Biophys Acta 1855:24–42
106. Garg M (2015) Targeting microRNAs in epithelial-to-mesenchymal transition-induced cancer stem cells: therapeutic approaches in cancer. Expert Opin Ther Targets 19:285
107. Garofalo M, Croce CM (2015) Role of microRNAs in maintaining cancer stem cells. Adv Drug Deliv Rev 81C:53–61
108. Mimeault M, Batra SK (2014) Molecular biomarkers of cancer stem/progenitor cells associated with progression, metastases, and treatment resistance of aggressive cancers. Cancer Epidemiol Biomarkers Prev 23: 234–254
109. Katoh M (2013) Therapeutics targeting angiogenesis: genetics and epigenetics, extracellular miRNAs and signaling networks (Review). Int J Mol Med 32:763–777
110. Park SM, Gaur AB, Lengyel E et al (2008) The miR-200 family determines the epithelial phenotype of cancer cells by targeting the E-cadherin repressors ZEB1 and ZEB2. Genes Dev 22:894–907
111. Joglekar MV, Patil D, Joglekar VM et al (2009) The miR-30 family microRNAs confer epithelial phenotype to human pancreatic cells. Islets 1:137–147
112. Ji Q, Hao X, Zhang M et al (2009) MicroRNA miR-34 inhibits human pancreatic cancer tumor-initiating cells. PLoS One 4:e6816
113. Misawa A, Katayama R, Koike S et al (2010) AP-1-Dependent miR-21 expression contributes to chemoresistance in cancer stem cell-like SP cells. Oncol Res 19:23–33
114. Khew-Goodall Y, Goodall GJ (2010) Myc-modulated miR-9 makes more metastases. Nat Cell Biol 12:209–211
115. Liu C, Kelnar K, Liu B et al (2011) The microRNA miR-34a inhibits prostate cancer stem cells and metastasis by directly repressing CD44. Nat Med 17:211–215
116. Wu ZS, Wu Q, Wang CQ et al (2011) miR-340 inhibition of breast cancer cell migration and invasion through targeting of oncoprotein c-Met. Cancer 117:2842–2852
117. Nabhan JF, Hu R, Oh RS et al (2012) Formation and release of arrestin domain-containing protein 1-mediated microvesicles (ARMMs) at plasma membrane by recruitment of TSG101 protein. Proc Natl Acad Sci U S A 109:4146–4151
118. Shen B, Wu N, Yang JM et al (2011) Protein targeting to exosomes/microvesicles by plasma membrane anchors. J Biol Chem 286:14383–14395
119. Fang Y, Wu N, Gan X et al (2007) Higher-order oligomerization targets plasma membrane proteins and HIV gag to exosomes. PLoS Biol 5:e158

120. Hurley JH, Emr SD (2006) The ESCRT complexes: structure and mechanism of a membrane-trafficking network. Annu Rev Biophys Biomol Struct 35:277–298
121. Rana S, Claas C, Kretz CC et al (2011) Activation-induced internalization differs for the tetraspanins CD9 and Tspan8: impact on tumor cell motility. Int J Biochem Cell Biol 43:106–119
122. Pant S, Hilton H, Burczynski ME (2012) The multifaceted exosome: biogenesis, role in normal and aberrant cellular function, and frontiers for pharmacological and biomarker opportunities. Biochem Pharmacol 83: 1484–1494
123. Villarroya-Beltri C, Gutierrez-Vazquez C, Sanchez-Cabo F et al (2013) Sumoylated hnRNPA2B1 controls the sorting of miRNAs into exosomes through binding to specific motifs. Nat Commun 4:2980
124. Munro TP, Magee RJ, Kidd GJ et al (1999) Mutational analysis of a heterogeneous nuclear ribonucleoprotein A2 response element for RNA trafficking. J Biol Chem 274: 34389–34395
125. Hollas H, Aukrust I, Grimmer S et al (2006) Annexin A2 recognises a specific region in the 3′-UTR of its cognate messenger RNA. Biochim Biophys Acta 1763:1325–1334
126. Thery C, Amigorena S, Raposo G et al (2006) Isolation and characterization of exosomes from cell culture supernatants and biological fluids. Curr Protoc Cell Biol 3:22
127. Taylor DD, Gercel-Taylor C, Lyons KS et al (2003) T-cell apoptosis and suppression of T-cell receptor/CD3-zeta by Fas ligand-containing membrane vesicles shed from ovarian tumors. Clin Cancer Res 9:5113–5119
128. Lai RC, Arslan F, Lee MM et al (2010) Exosome secreted by MSC reduces myocardial ischemia/reperfusion injury. Stem Cell Res 4:214–222
129. Petersen KE, Manangon E, Hood JL et al (2014) A review of exosome separation techniques and characterization of B16-F10 mouse melanoma exosomes with AF4-UV-MALS-DLS-TEM. Anal Bioanal Chem 406:7855–7866
130. Brownlee Z, Lynn KD, Thorpe PE et al (2014) A novel "salting-out" procedure for the isolation of tumor-derived exosomes. J Immunol Methods 407:120–126
131. Clayton A, Court J, Navabi H et al (2001) Analysis of antigen presenting cell derived exosomes, based on immuno- magnetic isolation and flowcytometry. J Immunol Methods 247:163–174
132. Chen C, Skog J, Hsu CH et al (2010) Microfluidic isolation and transcriptome analysis of serum microvesicles. Lab Chip 10:505–511
133. Navabi H, Croston D, Hobot J et al (2005) Preparation of human ovarian cancer ascites-derived exosomes for a clinical trial. Blood Cells Mol Dis 35:149–152
134. Lamparski HG, Metha-Damani A, Yao JY et al (2002) Production and characterization of clinical grade exosomes derived from dendritic cells. J Immunol Methods 270: 211–226
135. Orozco AF, Lewis DE (2010) Flow cytometric analysis of circulating microparticles in plasma. Cytometry A 77:502–514
136. Yurkovetsky Z, Skates S, Lomakin A et al (2010) Development of a multimarker assay for early detection of ovarian cancer. J Clin Oncol 28:2159–2166
137. Deatherage BL, Cookson BT (2012) Membrane vesicle release in bacteria, eukaryotes, and archaea: a conserved yet underappreciated aspect of microbial life. Infect Immun 80:1948–1957
138. Mariotto AB, Noone AM, Howlader N et al (2014) Cancer survival: an overview of measures, uses, and interpretation. J Natl Cancer Inst Monogr 2014:145–186
139. Paulson AS, Tran Cao HS, Tempero MA et al (2013) Therapeutic advances in pancreatic cancer. Gastroenterology 144:1316–1326
140. Madhavan B, Yue S, Galli U et al (2015) Combined evaluation of a panel of protein and miRNA serum-exosome biomarkers for pancreatic cancer diagnosis increases sensitivity and specificity. Int J Cancer 136:2616
141. Logozzi M, DeMilito A, Lugini L et al (2009) High levels of exosomes expressing CD63 and caveolin-1 in plasma of melanoma patients. PLoS One 4:e5219
142. Weber JA, Baxter DH, Zhang S et al (2010) The microRNA spectrum in 12 body fluids. Clin Chem 56:1733–1741
143. Mitchell PS, Parkin RK, Krih EM et al (2008) Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A 105:10513–10518
144. Jia S, Zocco D, Samuels ML et al (2014) Emerging technologies in extracellular vesicle-based molecular diagnostics. Expert Rev Mol Diagn 14:307–321
145. Sun Y, Liu J (2014) Potential of cancer cell-derived exosomes in clinical application: a review of recent research advances. Clin Ther 36:863–872

146. Di C, Zhao Y (2015) Multiple drug resistance due to resistance to stem cells and stem cell treatment progress in cancer (Review). Exp Ther Med 9:289–293
147. Elshamy WM, Duhé RJ (2013) Overview: cellular plasticity, cancer stem cells and metastasis. Cancer Lett 341:2–8
148. Tirino V, Desiderio V, Paino F et al (2013) Cancer stem cells in solid tumors: an overview and new approaches for their isolation and characterization. FASEB J 27:13–24
149. Wang H, Rana S, Giese N et al (2013) Tspan8, CD44v6 and alpha6beta4 are biomarkers of migrating pancreatic cancer-initiating cells. Int J Cancer 133:416–426
150. Marcelo CL, Peramo A, Ambati A et al (2012) Characterization of a unique technique for culturing primary adult human epithelial progenitor/"stem cells". BMC Dermatol 12:8
151. Papini S, Cecchetti D, Campani D et al (2003) Isolation and clonal analysis of human epidermal keratinocyte stem cells in long-term culture. Stem Cells 21:481–494
152. Tauro BJ, Greening DW, Mathias RA et al (2012) Comparison of ultracentrifugation, density gradient separation, and immunoaffinity capture methods for isolating human colon cancer cell line LIM1863-derived exosomes. Methods 56:293–304
153. Chen CL, Lai YF, Tang P et al (2012) Comparative and targeted proteomic analyses of urinary microparticles from bladder cancer and hernia patients. J Proteome Res 11: 5611–5629
154. Kim G, Yoo CE, Kim M et al (2012) Noble polymeric surface conjugated with zwitterionic moieties and antibodies for the isolation of exosomes from human serum. Bioconjug Chem 23:2114–2120
155. Lässer C, Eldh M, Lötvall J (2012) Isolation and characterization of RNA-containing exosomes. J Vis Exp 59:e3037
156. Kalra H, Simpson RJ, Ji H et al (2012) Vesiclepedia: a compendium for extracellular vesicles with continuous community annotation. PLoS Biol 10:e1001450
157. Skvortsov S, Debbage P, Skvortsova I (2014) Proteomics of cancer stem cells. Int J Radiat Biol 90:653–658
158. Wang Y, Gao X, Wei F et al (2014) Diagnostic and prognostic value of circulating miR-21 for cancer: a systematic review and meta-analysis. Gene 533:389–397
159. Zhang QH, Sun HM, Zheng RZ et al (2013) Meta-analysis of microRNA-183 family expression in human cancer studies comparing cancer tissues with noncancerous tissues. Gene 527:26–32
160. Fu X, Han Y, Wu Y et al (2011) Prognostic role of microRNA-21 in various carcinomas: a systematic review and meta-analysis. Eur J Clin Invest 41:1245–1253
161. Liu R, Chen X, Du Y et al (2012) Serum microRNA expression profile as a biomarker in the diagnosis and prognosis of pancreatic cancer. Clin Chem 58:610–618
162. Altevogt P, Bretz NP, Ridinger J et al (2014) Novel insights into exosome-induced, tumor-associated inflammation and immunomodulation. Semin Cancer Biol 28:51–57
163. Zöller M (2011) CD44: can a cancer-initiating cell profit from an abundantly expressed molecule? Nat Rev Cancer 11: 254–267
164. Wang W, Li J, Zhu W et al (2014) MicroRNA-21 and the clinical outcomes of various carcinomas: a systematic review and meta-analysis. BMC Cancer 14:819
165. Thuma F, Zöller M (2014) Outsmart TEX to steal the cancer initiating cell its niche. Semin Cancer Biol 28:39–50
166. Azmi AS, Bao B, Sarkar FH (2013) Exosomes in cancer development, metastasis, and drug resistance: a comprehensive review. Cancer Metastasis Rev 32:623–642
167. Montecalvo A, Larregina AT, Shufesky WJ et al (2012) Mechanism of transfer of functional microRNAs between mouse dendritic cells via exosomes. Blood 119:756–766
168. Tian T, Wang Y, Wang H et al (2010) Visualizing of the cellular uptake and intracellular trafficking of exosomes by live-cell microscopy. J Cell Biochem 111:488–496
169. van der Pol E, Hoekstra AG, Sturk A et al (2010) Optical and non-optical methods for detection and characterization of microparticles and exosomes. J Thromb Haemost 8:2596–2607
170. Feng D, Zhao WL, Ye YY et al (2010) Cellular internalization of exosomes occurs through phagocytosis. Traffic 11:675–687
171. Mulcahy LA, Pink RC, Carter DR (2014) Routes and mechanisms of extracellular vesicle uptake. J Extracell Vesicles 4:3
172. Taylor MJ, Lampe M, Merrifield CJ (2012) A feedback loop between dynamin and actin recruitment during clathrin-mediated endocytosis. PLoS Biol 10:e1001302
173. Parton RG, Simons K (2007) The multiple faces of caveolae. Nat Rev Mol Cell Biol 8: 185–194
174. Kerr MC, Teasdale RD (2009) Defining macropinocytosis. Traffic 10:364–371
175. Chernomordik LV, Kozlov MM (2008) Mechanics of membrane fusion. Nat Struct Mol Biol 15:675–683

176. Barrès C, Blanc L, Bette-Bobillo P et al (2010) Galectin-5 is bound onto the surface of rat reticulocyte exosomes and modulates vesicle uptake by macrophages. Blood 115: 696–705
177. Svensson KJ, Christianson HC, Wittrup A et al (2013) Exosome uptake depends on ERK1/2-heat shock protein 27 signalling and lipid raft-mediated endocytosis negatively regulated by caveolin-1. J Biol Chem 288: 17713–17724
178. Franzen CA, Simms PE, Van Huis AF et al (2014) Characterization of uptake and internalization of exosomes by bladder cancer cells. Biomed Res Int 2014:619829
179. Escola JM, Kleijmeer MJ, Stoorvogel W et al (1998) Selective enrichment of tetraspan proteins on the internal vesicles of multivesicular endosomes and on exosomes secreted by human B-lymphocytes. J Biol Chem 273: 20121–20127
180. Nazarenko I, Rana S, Baumann A et al (2010) Cell surface tetraspanin Tspan8 contributes to molecular pathways of exosome-induced endothelial cell activation. Cancer Res 70: 1668–1678
181. Zech D, Rana S, Büchler MW et al (2012) Tumor-exosomes and leukocyte activation: an ambivalent crosstalk. Cell Commun Signal 10:37
182. Morelli AE, Larregina AT, Shufesky WJ et al (2004) Endocytosis, intracellular sorting, and processing of exosomes by dendritic cells. Blood 104:3257–3266
183. Nolte-'t Hoen EN, Buschow SI, Anderton SM et al (2009) Activated T cells recruit exosomes secreted by dendritic cells via LFA-1. Blood 113:1977–1981
184. Gu X, Erb U, Büchler MW et al (2015) Improved vaccine efficacy of tumor exosome compared to tumor lysate loaded dendritic cells in mice. Int J Cancer 136:E74–E84
185. Valapala M, Vishwanatha JK (2011) Lipid raft endocytosis and exosomal transport facilitate extracellular trafficking of annexin A2. J Biol Chem 286:30911–30925
186. Van Gool SW, Vandenberghe P, de Boer M et al (1996) CD80, CD86 and CD40 provide accessory signals in a multiple-step T-cell activation model. Immunol Rev 153:47–83
187. Prydz K, Tveit H, Vedeler A et al (2013) Arrivals and departures at the plasma membrane: direct and indirect transport routes. Cell Tissue Res 352:5–20
188. Vega-Ramos J, Villadangos JA (2013) Consequences of direct and indirect activation of dendritic cells on antigen presentation: functional implications and clinical considerations. Mol Immunol 55:175–178
189. Lakkaraju A, Rodriguez-Boulan E (2008) Itinerant exosomes: emerging roles in cell and tissue polarity. Trends Cell Biol 18:199–209
190. Stanley S (2014) Biological nanoparticles and their influence on organisms. Curr Opin Biotechnol 28:69–74
191. Mathivanan S, Lim JW, Tauro BJ et al (2010) Proteomic analysis of A33-immunoaffinity-purified exosomes released from the human colon tumor cell line LIM1215 reveals a tissue-specific protein signature. Mol Cell Proteomics 9:197–208
192. Kosaka N, Iguchi H, Yoshioka Y et al (2010) Secretory mechanisms and intercellular transfer of microRNAs in living cells. J Biol Chem 285:17442–17452
193. Sevenich L, Joyce JA (2014) Pericellular proteolysis in cancer. Genes Dev 28:2331–2347
194. Arduise C, Abache T, Li L et al (2008) Tetraspanins regulate ADAM10-mediated cleavage of TNF-alpha and epidermal growth factor. J Immunol 181:7002–7013
195. Gutiérrez-López MD, Gilsanz A, Yáñez-Mó M et al (2011) The sheddase activity of ADAM17/TACE is regulated by the tetraspanin CD9. Cell Mol Life Sci 68: 3275–3292
196. Le Naour F, André M, Boucheix C et al (2006) Membrane microdomains and proteomics: lessons from tetraspanin microdomains and comparison with lipid rafts. Proteomics 6:6447–6454
197. Yanez-Mo M, Barreiro O, Gonzalo P et al (2008) MT1-MMP collagenolytic activity is regulated through association with tetraspanin CD151 in primary endothelial cells. Blood 112:3217–3226
198. Hendrix A, Westbroek W, Bracke M et al (2010) An ex(o)citing machinery for invasive tumor growth. Cancer Res 70:9533–9537
199. Grange C, Tapparo M, Collino F et al (2011) Microvesicles released from human renal cancer stem cells stimulate angiogenesis and formation of lung premetastatic niche. Cancer Res 71:5346–5356
200. Lo Cicero A, Majkowska I, Nagase H et al (2012) Microvesicles shed by oligodendroglioma cells and rheumatoid synovial fibroblasts contain aggrecanase activity. Matrix Biol 31:229–233
201. Ngora H, Galli UM, Miyazaki K et al (2012) Membrane-bound and exosomal metastasis-associated C4.4A promotes migration by associating with the α(6)β(4) integrin and MT1-MMP. Neoplasia 14:95–107

202. Lages E, Ipas H, Guttin A et al (2012) MicroRNAs: molecular features and role in cancer. Front Biosci 17:2508–2540
203. Sangaletti S, Colombo MP (2008) Matricellular proteins at the crossroad of inflammation and cancer. Cancer Lett 267: 245–253
204. Clayton A, Mitchell JP, Court J et al (2007) Human tumor-derived exosomes selectively impair lymphocyte responses to interleukin-2. Cancer Res 67:7458–7466
205. Rana S, Malinowska K, Zöller M (2013) Exosomal tumor microRNA modulates premetastatic organ cells. Neoplasia 15:281–295
206. Delcayre A, Shu H, Le Pecq JB (2005) Dendritic cell-derived exosomes in cancer immunotherapy: exploiting nature's antigen delivery pathway. Expert Rev Anticancer Ther 5:537–547
207. Artavanis-Tsakonas K, Kasperkovitz PV, Papa E et al (2011) The tetraspanin CD82 is specifically recruited to fungal and bacterial phagosomes prior to acidification. Infect Immun 79:1098–1106
208. Tumne A, Prasad VS, Chen Y et al (2009) Noncytotoxic suppression of human immunodeficiency virus type 1 transcription by exosomes secreted from CD8+ T cells. J Virol 83:4354–4364
209. Tan A, De La Peña H, Seifalian AM (2010) The application of exosomes as a nanoscale cancer vaccine. Int J Nanomedicine 5: 889–900
210. Taylor DD, Gercel-Taylor C (2011) Exosomes/microvesicles: mediators of cancer-associated immunosuppressive microenvironments. Semin Immunopathol 33: 441–454
211. Filipazzi P, Bürdek M, Villa A et al (2012) Recent advances on the role of tumor exosomes in immunosuppression and disease progression. Semin Cancer Biol 22:342–349
212. Whiteside TL (2014) Immune modulation of T-cell and NK (natural killer) cell activities by TEXs (tumour-derived exosomes). Biochem Soc Trans 41:245–251
213. Khalil AA, Kabapy NF, Deraz SF et al (2011) Heat shock proteins in oncology: diagnostic biomarkers or therapeutic targets? Biochim Biophys Acta 1816:89–104
214. Elsner L, Muppala V, Gehrmann M et al (2007) The heat shock protein HSP70 promotes mouse NK cell activity against tumors that express inducible NKG2D ligands. J Immunol 179:5523–5533
215. Dai S, Wan T, Wang B et al (2005) More efficient induction of HLA-A*0201-restricted and carcinoembryonic antigen (CEA)-specific CTL response by immunization with exosomes prepared from heat-stressed CEA-positive tumor cells. Clin Cancer Res 11: 7554–7563
216. Hurwitz MD, Kaur P, Nagaraja GM et al (2010) Radiation therapy induces circulating serum Hsp72 in patients with prostate cancer. Radiother Oncol 95:350–358
217. Chen T, Guo J, Yang M et al (2011) Chemokine-containing exosomes are released from heat-stressed tumor cells via lipid raft-dependent pathway and act as efficient tumor vaccine. J Immunol 186:2219–2228
218. Zeelenberg IS, van Maren WW, Boissonnas A et al (2011) Antigen localization controls T cell-mediated tumor immunity. J Immunol 187:1281–1288
219. Robbins PD, Morelli AE (2014) Regulation of immune responses by extracellular vesicles. Nat Rev Immunol 14:195–208
220. Pitt JM, Charrier M, Viaud S et al (2014) Dendritic cell-derived exosomes as immunotherapies in the fight against cancer. J Immunol 193:1006–1011
221. Claas C, Seiter S, Claas A et al (1998) Association between the rat homologue of CO-029, a metastasis-associated tetraspanin molecule and consumption coagulopathy. J Cell Biol 141:267–280
222. Gesierich S, Berezovskiy I, Ryschich E et al (2006) Systemic induction of the angiogenesis switch by the tetraspanin D6.1A/CO-029. Cancer Res 66:7083–7094
223. Hood JL, San Roman S, Wickline SA (2011) Exosomes released by melanoma cells prepare sentinel lymph nodes for tumor metastasis. Cancer Res 71:3792–3801
224. Al-Nedawi K, Meehan B, Rak J (2009) Microvesicles: messengers and mediators of tumor progression. Cell Cycle 8:2014–2018
225. Mineo M, Garfield SH, Taverna S et al (2012) Exosomes released by K562 chronic myeloid leukemia cells promote angiogenesis in a Src-dependent fashion. Angiogenesis 15:33–45
226. Hong BS, Cho JH, Kim H et al (2009) Colorectal cancer cell-derived microvesicles are enriched in cell cycle-related mRNAs that promote proliferation of endothelial cells. BMC Genomics 10:556
227. Peinado H, Alečković M, Lavotshkin S et al (2012) Melanoma exosomes educate bone marrow progenitor cells toward a pro-metastatic phenotype through MET. Nat Med 18:883–891
228. Verweij FJ, Middeldorp JM, Pegtel DM (2012) Intracellular signaling controlled by

the endosomal-exosomal pathway. Commun Integr Biol 5:88–93

229. Antonyak MA, Li B, Boroughs LK (2011) Cancer cell-derived microvesicles induce transformation by transferring tissue transglutaminase and fibronectin to recipient cells. Proc Natl Acad Sci U S A 108:4852–4857
230. Khan S, Jutzy JM, Aspe JR et al (2011) Survivin is released from cancer cells via exosomes. Apoptosis 16:1–12
231. Webber J, Steadman R, Mason MD et al (2010) Cancer exosomes trigger fibroblast to myofibroblast differentiation. Cancer Res 70:9621–9630
232. Cho JA, Park H, Lim EH et al (2012) Exosomes from breast cancer cells can convert adipose tissue-derived mesenchymal stem cells into myofibroblast-like cells. Int J Oncol 40:130–138
233. Wysoczynski M, Ratajczak MZ (2009) Lung cancer secreted microvesicles: underappreciated modulators of microenvironment in expanding tumors. Int J Cancer 125: 1595–1603
234. Jung T, Castellana D, Klingbeil P et al (2009) CD44v6 dependence of premetastatic niche preparation by exosomes. Neoplasia 11: 1093–1105
235. McCready J, Sims JD, Chan D et al (2010) Secretion of extracellular hsp90alpha via exosomes increases cancer cell motility: a role for plasminogen activation. BMC Cancer 10:294
236. Huan J, Hornick NI, Shurtleff MJ et al (2013) RNA trafficking by acute myelogenous leukemia exosomes. Cancer Res 73: 918–929
237. Janowska-Wieczorek A, Wysoczynski M, Kijowski J et al (2005) Microvesicles derived from activated platelets induce metastasis and angiogenesis in lung cancer. Int J Cancer 113:752–760
238. Cai Z, Yang F, Yu L et al (2012) Activated T cell exosomes promote tumor invasion via Fas signaling pathway. J Immunol 188: 5954–5961
239. Luga V, Zhang L, Viloria-Petit AM et al (2012) Exosomes mediate stromal mobilization of autocrine Wnt-PCP signaling in breast cancer cell migration. Cell 151: 1542–1556
240. Umezu T, Ohyashiki K, Kuroda M et al (2013) Leukemia cell to endothelial cell communication via exosomal miRNAs. Oncogene 32:2747–2755
241. Fabbri M, Paone A, Calore F et al (2012) MicroRNAs bind to Toll-like receptors to induce prometastatic inflammatory response. Proc Natl Acad Sci U S A 109:E2110–E2116
242. Hupalowska A, Miaczynska M (2012) The new faces of endocytosis in signaling. Traffic 13:9–18
243. Bobrie A, Krumeich S, Reyal F et al (2012) Rab27a supports exosome-dependent and -independent mechanisms that modify the tumor microenvironment and can promote tumor progression. Cancer Res 72: 4920–4930
244. Pan Q, Ramakrishnaiah V, Henry S et al (2012) Hepatic cell-to-cell transmission of small silencing RNA can extend the therapeutic reach of RNA interference (RNAi). Gut 61:1330–1339
245. Higginbotham JN, Demory Beckler M, Gephart JD et al (2011) Amphiregulin exosomes increase cancer cell invasion. Curr Biol 21:779–786
246. Meckes DG Jr, Shair KH, Marquitz AR et al (2010) Human tumor virus utilizes exosomes for intercellular communication. Proc Natl Acad Sci U S A 107:20370–20375
247. Gourzones C, Gelin A, Bombik I et al (2010) Extra-cellular release and blood diffusion of BART viral micro-RNAs produced by EBV-infected nasopharyngeal carcinoma cells. Virol J 7:271
248. Yang M, Chen J, Su F et al (2011) Microvesicles secreted by macrophages shuttle invasion-potentiating microRNAs into breast cancer cells. Mol Cancer 10:117
249. Zhang Y, Liu D, Chen X et al (2010) Secreted monocytic miR-150 enhances targeted endothelial cell migration. Mol Cell 39:133–144
250. Roninson IB (1987) Molecular mechanism of multidrug resistance in tumor cells. Clin Physiol Biochem 5:140–151
251. Bebawy M, Combes V, Lee E et al (2009) Membrane microparticles mediate transfer of P-glycoprotein to drug sensitive cancer cells. Leukemia 23:1643–1649
252. Puisieux A, Brabletz T, Caramel J (2014) Oncogenic roles of EMT-inducing transcription factors. Nat Cell Biol 16:488–494
253. Vella LJ (2014) The emerging role of exosomes in epithelial-mesenchymal-transition in cancer. Front Oncol 4:361
254. Garnier D, Magnus N, Lee TH et al (2012) Cancer cells induced to express mesenchymal phenotype release exosome-like extracellular vesicles carrying tissue factor. J Biol Chem 287:43565–43572
255. Tauro BJ, Mathias RA, Greening DW et al (2013) Oncogenic H-ras reprograms Madin-Darby canine kidney (MDCK) cell-derived exosomal proteins following epithelial-mesenchymal transition. Mol Cell Proteomics 12:2148–2159

256. Jeppesen DK, Nawrocki A, Jensen SG et al (2014) Quantitative proteomics of fractionated membrane and lumen exosome proteins from isogenic metastatic and nonmetastatic bladder cancer cells reveal differential expression of EMT factors. Proteomics 14:699–712
257. Ramteke A, Ting H, Agarwal C et al (2015) Exosomes secreted under hypoxia enhance invasiveness and stemness of prostate cancer cells by targeting adherens junction molecules. Mol Carcinog 54:554
258. Bijnsdorp IV, Geldof AA, Lavaei M et al. (2013) Exosomal ITGA3 interferes with noncancerous prostate cell functions and is increased in urine exosomes of metastatic prostate cancer patients. J Extracell Vesicles 2. doi: 10.3402/jev.v2i0.22097
259. Mathias RA, Chen YS, Wang B et al (2010) Extracellular remodelling during oncogenic Ras-induced epithelial-mesenchymal transition facilitates MDCK cell migration. J Proteome Res 9:1007–1019
260. Atay S, Banskota S, Crow J et al (2014) Oncogenic KIT-containing exosomes increase gastrointestinal stromal tumor cell invasion. Proc Natl Acad Sci U S A 111:711–716
261. Garnier D, Magnus N, Meehan B et al (2013) Qualitative changes in the proteome of extracellular vesicles accompanying cancer cell transition to mesenchymal state. Exp Cell Res 319:2747–2757
262. Aga M, Bradley JM, Wanchu R et al (2014) Differential effects of caveolin-1 and -2 knockdown on aqueous outflow and altered extracellular matrix turnover in caveolin-silenced trabecular meshwork cells. Invest Ophthalmol Vis Sci 55:5497–5509
263. Philip R, Heiler S, Mu W et al (2015) Claudin-7 promotes the epithelial-mesenchymal transition in human colorectal cancer. Oncotarget 6:2046
264. Josson S, Gururajan M, Hu P et al (2014) miR-409-3p/-5p promotes tumorigenesis, epithelial-to-mesenchymal transition, and bone metastasis of human prostate cancer. Clin Cancer Res 20:4636–4646
265. Marcus ME, Leonard JN (2013) FedExosomes: engineering therapeutic biological nanoparticles that truly deliver. Pharmaceuticals (Basel) 6:659–680
266. Xin H, Li Y, Buller B et al (2012) Exosome-mediated transfer of miR-133b from multipotent mesenchymal stromal cells to neural cells contributes to neurite outgrowth. Stem Cells 30:1556–1564
267. Arslan F, Lai RC, Smeets MB et al (2013) Mesenchymal stem cell-derived exosomes increase ATP levels, decrease oxidative stress and activate PI3K/Akt pathway to enhance myocardial viability and prevent adverse remodeling after myocardial ischemia/reperfusion injury. Stem Cell Res 10:301–312
268. Marleau AM, Chen CS, Joyce JA et al (2012) Exosome removal as a therapeutic adjuvant in cancer. J Transl Med 10:134
269. Christianson HC, Svensson KJ, van Kuppevelt TH et al (2013) Cancer cell exosomes depend on cell-surface heparan sulfate proteoglycans for their internalization and functional activity. Proc Natl Acad Sci U S A 110:17380–17385
270. Atai NA, Balaj L, van Veen H et al (2013) Heparin blocks transfer of extracellular vesicles between donor and recipient cells. J Neurooncol 115:343–351
271. Fabian MR, Sonenberg N, Filipowicz W (2010) Regulation of mRNA translation and stability by microRNAs. Annu Rev Biochem 79:351–379
272. Chen VY, Posada MM, Blazer LL et al (2006) The role of the VPS4A-exosome pathway in the intrinsic egress route of a DNA-binding anticancer drug. Pharm Res 23:1687–1695
273. Li J, Sherman-Baust CA, Tsai-Turton M et al (2009) Claudin-containing exosomes in the peripheral circulation of women with ovarian cancer. BMC Cancer 9:244
274. Shen C, Hao SG, Zhao CX et al (2011) Antileukaemia immunity: effect of exosomes against NB4 acute promyelocytic leukaemia cells. J Int Med Res 39:740–747
275. Johnsen KB, Gudbergsson JM, Skov MN (2014) A comprehensive overview of exosomes as drug delivery vehicles - endogenous nanocarriers for targeted cancer therapy. Biochim Biophys Acta 1846:75–87
276. Ohno S, Takanashi M, Sudo K et al (2013) Systemically injected exosomes targeted to EGFR deliver antitumor microRNA to breast cancer cells. Mol Ther 21:185–191
277. Kosaka N, Iguchi H, Yoshioka Y et al (2012) Competitive interactions of cancer cells and normal cells via secretory microRNAs. J Biol Chem 287:1397–1405
278. Mizrak A, Bolukbasi MF, Ozdener GB et al (2013) Genetically engineered microvesicles carrying suicide mRNA/protein inhibit schwannoma tumor growth. Mol Ther 21: 101–108
279. Wahlgren J, De L, Karlson T et al (2012) Plasma exosomes can deliver exogenous short

interfering RNA to monocytes and lymphocytes. Nucleic Acids Res 40:e130

280. El-Andaloussi S, Lee Y, Lakhal-Littleton S et al (2012) Exosome-mediated delivery of siRNA in vitro and in vivo. Nat Protoc 7: 2112–2126
281. Alvarez-Erviti L, Seow Y, Yin H et al (2011) Delivery of siRNA to the mouse brain by systemic injection of targeted exosomes. Nat Biotechnol 29:341–345
282. Hartman ZC, Wei J, Glass OK et al (2011) Increasing vaccine potency through exosome antigen targeting. Vaccine 29:9361–9367
283. Bolukbasi MF, Mizrak A, Ozdener GB et al (2012) miR-1289 and "Zipcode"-like Sequence Enrich mRNAs in Microvesicles. Mol Ther Nucleic Acids 1:e10
284. de Medina P, Paillasse MR, Segala G et al (2010) Identification and pharmacological characterization of cholesterol-5,6-epoxide hydrolase as a target for tamoxifen and AEBS ligands. Proc Natl Acad Sci U S A 107: 13520–13525
285. Sun D, Zhuang X, Xiang X et al (2010) A novel nanoparticle drug delivery system: the anti-inflammatory activity of curcumin is enhanced when encapsulated in exosomes. Mol Ther 18:1606–1614
286. Ristorcelli E, Beraud E, Verrando P et al (2008) Human tumor nanoparticles induce apoptosis of pancreatic cancer cells. FASEB J 22:3358–3369
287. Engelman JA (2009) Targeting PI3K signalling in cancer: opportunities, challenges and limitations. Nat Rev Cancer 9:550–562
288. Esquela-Kerscher A, Trang P, Wiggins JF et al (2008) The let-7 microRNA reduces tumor growth in mouse models of lung cancer. Cell Cycle 7:759–764
289. Kelnar K, Peltier HJ, Leatherbury N et al (2014) Quantification of therapeutic miRNA mimics in whole blood from nonhuman primates. Anal Chem 86:1534–1542
290. Bader AG (2012) miR-34 - a microRNA replacement therapy is headed to the clinic. Front Genet 3:120
291. Kota J, Chivukula RR, O'Donnell KA et al (2009) Therapeutic microRNA delivery suppresses tumorigenesis in a murine liver cancer model. Cell 137:1005–1017
292. Ajit SK (2012) Circulating microRNAs as biomarkers, therapeutic targets, and signaling molecules. Sensors (Basel) 12:3359–3369
293. Si ML, Zhu S, Wu H et al (2007) miR-21-mediated tumor growth. Oncogene 26: 2799–2803
294. Zhang L, Wrana JL (2014) The emerging role of exosomes in Wnt secretion and transport. Curr Opin Genet Dev 27:14–19
295. Gangoda L, Boukouris S, Liem M et al (2015) Extracellular vesicles including exosomes are mediators of signal transduction: are they protective or pathogenic? Proteomics 15:260–271
296. Wolfson B, Eades G, Zhou Q (2014) Roles of microRNA-140 in stem cell-associated early stage breast cancer. World J Stem Cells 6:591–597
297. Ruiss R, Jochum S, Mocikat R et al (2011) EBV-gp350 confers B-cell tropism to tailored exosomes and is a neo-antigen in normal and malignant B cells--a new option for the treatment of B-CLL. PLoS One 6:e25294
298. Rajendran L, Knölker HJ, Simons K (2010) Subcellular targeting strategies for drug design and delivery. Nat Rev Drug Discov 9:29–42
299. Kooijmans SA, Vader P, van Dommelen SM et al (2012) Exosome mimetics: a novel class of drug delivery systems. Int J Nanomedicine 7:1525–1541
300. van der Meel R, Fens MH, Vader P et al (2014) Extracellular vesicles as drug delivery systems: lessons from the liposome field. J Control Release 195:72–85

Chapter 8

MicroRNA (miRNA) Profiling

Lu Gao and Feng Jiang

Abstract

MicroRNAs (miRNAs) are small, highly conserved noncoding RNA molecules involved in the regulation of gene expression. Since each miRNA regulates the expression of hundreds of target mRNAs, miRNAs could function as master coordinators, efficiently regulating fundamental cellular processes, including proliferation, apoptosis, and development. Furthermore, miRNAs may provide useful diagnostic and therapeutic targets in a variety of diseases. However, miRNA expression profiling is essential for the investigation of the biological functions and clinical applications of miRNAs. Therefore, in this chapter, we review and discuss commonly used techniques for miRNAs profiling, as well as their advantages and restrictions.

Key words miRNA, Microarray, DDPCR

1 Introduction

MicroRNAs (miRNAs) are small endogenous noncoding RNAs, consisting of 19–24 nucleotides in length [1]. Since the discovery of the first miRNA (lin-4) in *C. elegans*, thousands of miRNAs have been identified by experimental or computational approaches in a variety of species (XX). miRNAs have important roles in regulating protein coding genes' functions by binding to the 3′-UTR sequences. The discovery of miRNAs and their biological functions could be one of the most exciting scientific breakthroughs in the last decade. For example, although miRNAs comprise up to 5 % of animal sequences, they can regulate approximately 30 % of protein coding genes, thus being the most abundant classes of regulators. Furthermore, given the important biological roles, moRNAs may have oncogenic functions in the development and progression of tumorigenesis and, could be used as biomarkers for malignancies. However, the investigation of the biological functions and clinical applications of miRNAs will be based on the development of miRNA expression profiling method. Indeed, miRNA profiling has helped to identify and detect miRNAs that regulate a range of processes, including organismal development

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_8,

and different diseases. In addition, the ability to effectively profile miRNAs could lead to the discoveries of disease- or tissue-specific miRNA biomarkers, our deep understanding of how miRNAs regulate cell differentiation and function. Therefore, miRNA expression profiling is crucial for the investigation of the biological functions and clinical applications of miRNAs. Several major profiling approaches for identification and validation of miRNAs are discussed below [2–6].

2 Profiling Approaches

2.1 cDNA Library-Based Platforms

Lee and Ambros [1] first proposed cDNA library-based platforms for searching miRNAs through discovering *lin-4* and *let-7* of *C. elegans*. Briefly, the cDNA was cloned and sequenced. The cloned sequences were homologically compared in the species genome database with NCBI Blast by using the related software. The secondary structure of homologous genomic sequence was predictably analyzed using the program mfold. A small-molecule RNA with a hairpin structure was detected by Northern blot. Although this method holds great promise, there are several challenges to overcome. These include the low abundance of miRNA expression, and its specific expression in different tissues and different stages of development. Furthermore, the degradation products of endogenous mRNA and other noncoding RNA have a certain interference effect.

2.2 Computation for the Prediction of miRNAs

Alternatively, computational methods for the prediction of miRNAs have gained popularity. Currently, two computer analytic tools are commonly used to support the approaches. The first one is called MiRscan [7, 8]. It produced an initial set of candidates by scanning the genome of *C. elegans* with a sliding-window of 110 nt. The regions were folded and filtered according to more permissive structural criteria. Potential homologues were sought in *C. briggsae* sequences and only conserved hairpins were retained, yielding a total of ~36,000 candidates. The second one is miRseeker [4] that represents the first attempt to identify conserved stem-loops due to selection, and not as an artifact of considering genomes that are not sufficiently distant. One can align the non-annotated intergenic and intronic sequences of the genomes of *D. melanogaster* and *D. pseudoobscura*. Both tools have been successfully identified a large number of miRNA genes and confirmed by the experiments. Furthermore, some researchers have combined high-throughput experimental methods with computational procedures in order to identify a wider range of miRNAs [9]. However, the computed-generated data need to be vigorously and reliably validated by conventional and gold stand experimental approaches.

2.3 Quantitative Reverse Transcription PCR-Based Methods

Because mature miRNAs are very small, they require appropriate small size primers for the quantification. It was challenging to use qPCR for the analysis of miRNAs. However, successful real-time RT-PCR technologies are recently developed to amplify and quantify both the precursor and mature microRNA [10]. One major approach relies on reverse transcription from miRNA to cDNA, followed by qPCR with real-time monitoring of reaction product accumulation. An appealing aspect of this approach is the ease of incorporation into the workflow for laboratories that are familiar with real-time PCR. In order to scale this approach for miRNA profiling, reactions are carried out in a highly parallel, high-throughput form. Basically, qRT-PCR methods designed for miRNAs include SYBR green and TaqMan assays. Several manufacturers offer SYBR green detection for small RNA species. Generally, this method including Qiagen miScript and WaferGen system rely on polyadenylation of small RNAs, followed by a reverse transcription using an oligo-dT primer with tag. This tag sequence is then used as a universal reverse primer site SYBR-green detection [11]. Qigen maintains specificity for small RNA species using a proprietary Hi-Spec buffer, which inhibits the reverse transcription of longer coding and noncoding RNAs. Exiqon miRNA PCR. Exiqon's microRNA qPCR system combines the speed of a Universal RT reaction with the sensitivity and specificity of LNA™-enhanced PCR primers, and based on SYBR green reagents [12]. Because of ribose modifications, locked nucleic acids increase the acidity of Watson-Crick binding and specificity of primers allowing for similar primer Tms with short sequences.

The TaqMan-probe method is designed to detect and accurately quantify mature miRNAs using real-time PCR system [10]. The principle of the TaqMan™ microRNA assays is similar to conventional TaqMan™ RT-PCR ones. A major difference is the use of a novel target-specific stem-loop reverse transcription primer during the RT reaction, which address the challenge of the short length of mature miRNA. The primer extends the 3′ end of the target to produce a template that can be used in standard TaqMan® Assay-based real-time PCR. Also, the stem-loop structure in the tail of the primer confers a key advantage to these assays: specific detection of the mature, biologically active miRNA. Moreover, TaqMan technology can detect mature miRNAs that differ by as little as one nucleotide.

Since mature miRNA exerts its activity by binding to the 3′ untranslated region of mRNA, quantification of the active, mature miRNA, rather than the inactive, premiRNA, is generally preceded. Pre-miRNA exists as a stable hairpin of approximately 70 nts in length [13]. To amplify the pre-miRNA, forward and reverse primers were designed to anneal to the stem portion of the hairpin. Isoforms present another issue that needs to be carefully considered when designing quantified miRNA. Numerous miRNAs exists

as isoforms of identical mature and precursor sequences. Using SYBR green detection, it is often not possible for the PCR primers designed to the hairpin to discriminate among various isoforms. However, TaqMan™ minor groove binding (MGB) probes can be used to detect a family of different isoforms [14]. Sequences of the primers and TaqMan™ MGB probes for the analysis of the miRNA might be found in the website [15].

2.4 Microarray-Based Techniques for Quantification of miRNAs

Microarrays have been widely used to profile large numbers of mRNAs [16, 17]. cDNA microarrays are an increasingly popular technology to profile miRNAs [18], which includes synthesis of cDNA, labeling the product with fluorophore followed by dissociation and hybridization to complementary probes immobilized on a surface. It is practical to profile miRNA expression using real-time PCR in 384-well reaction plates. Gene expression profiling using real-time PCR has better sensitivity, which translates into smaller sample size. However, a disadvantage of real-time PCR profiling of gene expression is how to efficiently and accurately transfer small volumes of liquid into 384-well plates. Furthermore, some challenges also exist in microarray primer design.

The major commercial hybridized-based platforms, such as Affymetrix, Agilent, Exiqon (miRNA only), and Illumina BeadChip, have all been demonstrated to provide similar data quality [19–21]. The commercial micro-assays have typically software for extracting probe intensities form hybridization images as well as preprocessing of the data, including background connection and normalization. In addition, ones need to acquire their own tools for the downstream data management and analysis by using bioinformatics tools. Table 1 shows comparison of some common sued platforms for various qualities, providing some information to help decide which platform might be chosen for certain purposes.

Table 1
The commonly used microarray platforms for profiling miRNAs

	Affymetrix	Agilent	Illumina
Amount of DNA requested for service	please inquire	100 ng total RNA	200 ng total RNA
1- or 2-color	1-color	1-color	1-color
Probe size	25-mer	60-mer	50-mer
Species for which catalog arrays are available[a]	Human, mouse, rat, canine, monkey	Human, mouse, rat, custom printing	Human, mouse, custom printing
Array formats	Mini-array (format 400) 4 identical sense probes for each miRNA	8 × 15 K	12-sample BeadChip

[a]All species are represented on one array

Table 2
Next-generation sequencing platforms

	Roche's 454 sequencing	Illumina/Solexa	ABI SOLiD
Sequencing Chemistry	Pyrosequencing	Polymerase-based sequence-by-synthesis	Ligation-based sequencing
Amplification	Emulsion PCR	Bridge amplification	Emulsion PCR
Paired-end (PED) separation	3 kb	200–500 bp	3 kb
Mb per run	100 Mb	1300 Mb	3000 Mb
Time per PED run	<0.5 day	4 days	5 days
Read length (update)	100–400 bp	15–200 bp	15, 35, and 50 bp

2.5 Deep Sequencing/Next-Generation Sequencing

qRT-PCR and hybridization-based microarray platforms have been used to identify cancer-associated miRNA aberrations [22]. Yet these technologies only measure relative abundant and known miRNA sequences, and have limited capacity in identifying novel miRNAs whose aberrations are associated with cancer. Next-generation deep sequencing has emerged as a powerful tool for global miRNA analysis. DNA sequencing was first reported by Sanger [23], providing a tool to decipher genes. However, low throughput and high cost stalled its use for deciphering the human genome. A more cost-effective sequencing technology was developed by 545 Life Sciences [24]. Since then, several next-generation sequencing (NGS) platforms, such as Illumina Genome Analyzer (Illumina, Inc., San Diego, CA, USA) and SOLiD™ (Life Technologies Corporation, Carlsbad, CA, USA) have been developed. The newly develop NGS platforms have been used to various fields of biological and medical research, including measuring expression levels of known miRNAs and detecting unknown miRNAs as shown in Table 2. Deep sequencing processes millions of independent sequencing events, allows providing billions of nucleotide information within a single experiment. Furthermore, deep sequencing system enables comprehensive analyses of large amounts of sequence data, resulting in dramatically accelerated research compared to traditional labor-intensive efforts and is a powerful approach to determine accurate encoded-information from nucleotide fragments [25]. Therefore, its advantages over the current techniques include pooling of samples for high-throughput purposes, a wide detectable expression range, analyzing expression of all annotated miRNAs, and detecting novel miRNAs [26].

Hu et al. [27] used Solexa sequencing to evaluate miRNA profiling in serum of patients with stages I to IIIa NSCLC. Levels of four serum-based miRNAs (miR-486, miR-30d, miR-1, and miR-499) were significantly associated with overall survival. Using SOLiD transcriptome sequencing of miRNAs in peripheral blood of lung cancer patients, Keller et al. [28] identified 32 annotated and seven unknown miRNAs that were altered in the blood specimens of cancer patients. We recently used next-generation deep sequencing to comprehensively characterize miRNA profiles in eight lung tumor tissues consisting of two major types of NSCLC. We successfully identified 896 known miRNAs and 14 novel miRNAs, of which 24 miRNAs displayed dysregulation with fold change ≥4.5 in either stage I ACs or SCCs or both relative to normal tissues [29]. In comparison with NGS platforms, microarray only covers known genes and probe design is based to the reference sequence. Therefore microarray is able to detect the concentration of known sequence fragments. Microarray may have better accuracy and precision than NGS, which is based on PCR character and sequencing by synthesis (SBS) technology. Using NGS, all fragments can be detected without reference sequence, and the fragment sequence is well presented. In addition, during building in sequencing library, PCR amplification increases a relatively sensitivity for detection, but following the imbalance of amplification, it is lacking in quantitative accuracy. Therefore, NGS can discover new and small fragments without tedious probe design. However, for detection of gene expression levels, we should choose microarray analysis in my studies.

454 deep sequencing system from Roche was one of the first NGS platforms on the market, launching in 2005. The system uses emulsion PCR (emPCR) to clonally amplify the fragments that are then sequenced via sequencing-by-synthesis (SBS) technology [30]. Differing from 454 deep sequencing system, the Illumina sequencing is a base-by-base sequencing technology using a reversible terminator-based method, enabling detection of single base that is incorporated into growing DNA strands complementary to the template [31]. Since this technology reads out one base at a time, the main error mode is substitution rather than insertion or deletion. However, Applied Biosystems' SOLiD sequencing technology is based on ligation of oligonucleotides. 16 different dinucleotides are encoded with four fluorescent color dyes, each dye encoding four dinucleotides. SOLiD performs double interrogation of each base by combining the four-dye encoding scheme with a sequencing assay for every base in samples [32].

2.6 Droplet Digital PCR (ddPCR)

Quantitative polymerase chain reaction (qPCR) is one of the most commonly used techniques that can estimate expression levels of miRNAs in clinical specimens [22, 33–35]. However, qPCR has two major challenges for the assessment of plasma miRNAs [36, 37].

First, qPCR is an indirect and labor-consuming approach to analyze miRNAs, as it relies on an increase in fluorescence signal that is proportional to the polymerase reaction product, and uses the cycle threshold (C_T) as a metric. C_T values for miRNA targets are referenced to endogenous small RNA controls across samples and used for normalization. This can become problematic, because expression levels of the endogenous controls and their transcripts may differ between samples [36, 38]. Furthermore, numerous endogenous genes have been evaluated for determination of target miRNAs, including U6, U6B, 18S rRNA, 5S RNA, RNU38B, and RNU43; yet none has been widely accepted as a standard control [22, 33]. These problems can be partially solved through the use of an exogenous "spike-in" control, which, however, does not account for any template-specific effect or bias introduced through primer design. Moreover, to estimate the absolute abundance of a given miRNA, data must be compared to a previously generated standard curve of the same template with identical primers and conditions. However, the additional manipulations are labor intensive, and extreme care should be taken when measuring the reference samples and comparing the references and experimental standard curves [36]. Second, the sensitivity of qPCR for the detection of low copy number of genes is not high enough, as it only resolves ~1.5-fold changes of nucleic acids [37]. Given that a proportion of the cancer-associated miRNAs is derived from primary tumor and could be "diluted" in a background of normal miRNAs [39–41], the miRNAs presenting at low levels in plasma could be undetectable by qPCR.

ddPCR is a direct method for quantitatively measuring nucleic acids [42–49], as it depends on limiting partition of the PCR volume, where a positive result of a large number of microreactions indicates the presence of a single molecule in a given reaction. The number of positive reactions, together with Poisson's distribution, can be used to produce a straight and high-confidence measurement of the original target concentration [47]. Therefore, ddPCR does not require the reliance on rate-based measurements (C_T values), endogenous controls, and the use of calibration curves. Furthermore, previous studies targeting low copy number of nucleic acids have demonstrated that ddPCR has a high degree of sensitivity and precision than does qPCR [50–52]. We recently investigated the efficacy of using ddPCR for quantitative detection of two miRNAs (miRs-21-5p and 335-3p) in artificially seeded samples, RNA of cancer cells, and clinical plasma samples. miRs-21-5p and 335-3p were chosen, because our previous studies [22, 33–35] showed that miR-21-5p displayed a high expression level, whereas miR-335-3p had an endogenously low level in plasma. We then used ddPCR to quantify copy number of plasma miR-21-5p and miR-335-3p in 36 lung cancer patients and 38 controls. ddPCR showed a high degree of linearity and quantitative correlation

($R^2 = 0.96–0.99$) of measuring the miRNAs in a dynamic range from one to 10,000 copies/μl of input with high reproducibility. qPCR exhibited a dynamic range from 100 to 1×10^7 copies/μl of input. ddPCR had a higher sensitivity to detect copy number of the miRNAs compared with qPCR (one vs. 100 copies/μl, $P < 0.05$). In plasma, ddPCR could detect copy number of both miR-21-5p and miR-335-3p, whereas qPCR was only able to assess miR-21-5p. Quantification of the plasma miRNAs by ddPCR provided 71.8 % sensitivity and 80.6 % specificity in distinguishing lung cancer patients from cancer-free subjects. Therefore, as ddPCR becomes more established, it might be a robust tool for quantitative assessment of miRNA copy number in cancer diagnosis.

3 In Summary

Each platform has the advantages and disadvantages. In addition to carefully selecting the appropriate one for each research or clinical applications regarding efficacy and cost, we also need to pay attention on intra- and interlaboratory reproducibility, developing method standardization, establishing guidelines for sample collection and preparation. For instance, one of the major technical challenges in applying the techniques in clinical settings is how to standardize protocols for miRNA extraction from biological specimens such as serum or plasma, and to normalizing measured values and controls. Nevertheless, the rapid advance in the development of the sophistication of miRNA profiling tools provides the technical capabilities required for function analysis of miRNAs and miRNA biomarker discovery and validation. Future use of the techniques will dramatically deep understanding of biological function of miRNAs, and develop the small molecules as important diagnostic and therapeutic targets for various human diseases.

Acknowledgements

This work was supported in part by NCI R01CA161837, VA merit Award I01 CX000512, LUNGevity/Upstage Foundation Early Detection Award, and Award from the Geaton and JoAnn DeCesaris Family Foundation (F. J.).

References

1. Lee RC, Ambros V (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. Science 294:862–864
2. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T (2002) Identification of tissue-specific microRNAs from mouse. Curr Biol 12:735–739
3. Sunkar R, Girke T, Jain PK, Zhu JK (2005) Cloning and characterization of microRNAs from rice. Plant Cell 17:1397–1411

4. Lai EC, Tomancak P, Williams RW, Rubin GM (2003) Computational identification of Drosophila microRNA genes. Genome Biol 4:R42
5. Jones-Rhoades MW, Bartel DP (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. Mol Cell 14:787–799
6. Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J (2003) Computational and experimental identification of *C. elegans* microRNAs. Mol Cell 11: 1253–1263
7. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP (2003) The microRNAs of *Caenorhabditis elegans*. Genes Dev 17:991–1008
8. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP (2003) Vertebrate microRNA genes. Science 299:1540
9. Mendes ND, Freitas AT, Sagot MF (2009) Current tools for the identification of miRNA genes and their targets. Nucleic Acids Res 37:2419–2433
10. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, Barbisin M, Xu NL, Mahuvakar VR, Andersen MR, Lao KQ, Livak KJ, Guegler KJ (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33, e179
11. Yu Z, Zhu Y, Zhang Y, Li J, Fang Q, Xi J, Yao B (2011) Nanoliter droplet array for microRNA detection based on enzymatic stem-loop probes ligation and SYBR Green real-time PCR. Talanta 85:1760–1765
12. Rajwanshi VK, Hakansson AE, Sorensen MD, Pitsch S, Singh SK, Kumar R, Nielsen P, Wengel J (2000) The eight stereoisomers of LNA (locked nucleic acid): A remarkable family of strong RNA binding molecules. We acknowledge the Danish Natural Science Research Council, the Danish Technical Research Council, and Exiqon A/S for financial support. Ms Britta M. Dahl is thanked for oligonucleotide synthesis, Dr. Carl E. Olsen for MALDI-MS analysis, and Ms. Karen Jorgensen for recording CD spectra. Angew Chem Int Ed Engl 39:1656–1659
13. Lin SL, Chang D, Ying SY (2005) Asymmetry of intronic pre-miRNA structures in functional RISC assembly. Gene 356:32–38
14. Zhou S, Liu M, An W, Liang X, Yu W, Piao F (2014) A new method for analyzing the Duffy blood group genotype by TaqMan minor groove binding probes. J Clin Lab Anal 29(3):203–207
15. Jiang J, Lee EJ, Gusev Y, Schmittgen TD (2005) Real-time expression profiling of microRNA precursors in human cancer cell lines. Nucleic Acids Res 33:5394–5403
16. Liu CG, Calin GA, Volinia S, Croce CM (2008) MicroRNA expression profiling using microarrays. Nat Protoc 3:563–578
17. Xu G, Chen J, Pan Q, Huang K, Pan J, Zhang W, Yu F, Zhou T, Wang Y (2014) Long noncoding RNA expression profiles of lung adenocarcinoma ascertained by microarray analysis. PLoS One 9, e104044
18. Liu CG, Calin GA, Meloon B, Gamliel N, Sevignani C, Ferracin M, Dumitru CD, Shimizu M, Zupo S, Dono M, Alder H, Bullrich F, Negrini M, Croce CM (2004) An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues. Proc Natl Acad Sci U S A 101:9740–9744
19. Dee S, Getts RC (2012) MicroRNA expression analysis using the Affymetrix Platform. Methods Mol Biol 822:117–129
20. D'Andrade PN, Fulmer-Smentek S (2012) Agilent microRNA microarray profiling system. Methods Mol Biol 822:85–102
21. Tsao J, Yau P, Winegarden N (2010) Profiling microRNA expression with the Illumina BeadChip platform. Methods Mol Biol 632: 73–86
22. Shen J, Jiang F (2012) Applications of MicroRNAs in the diagnosis and prognosis of lung cancer. Expert Opin Med Diagn 6:197–207
23. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687–695
24. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

25. Git A, Dvinge H, Salmon-Divon M, Osborne M, Kutter C, Hadfield J, Bertone P, Caldas C (2010) Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. RNA 16: 991–1006
26. Schee K, Lorenz S, Worren MM, Gunther CC, Holden M, Hovig E, Fodstad O, Meza-Zepeda LA, Flatmark K (2013) Deep sequencing the MicroRNA transcriptome in colorectal cancer. PLoS One 8, e66165
27. Hu Z, Chen X, Zhao Y, Tian T, Jin G, Shu Y, Chen Y, Xu L, Zen K, Zhang C, Shen H (2010) Serum microRNA signatures identified in a genome-wide serum microRNA expression profiling predict survival of non-small-cell lung cancer. J Clin Oncol 28:1721–1726
28. Keller A, Backes C, Leidinger P, Kefer N, Boisguerin V, Barbacioru C, Vogel B, Matzas M, Huwer H, Katus HA, Stahler C, Meder B, Meese E (2011) Next-generation sequencing identifies novel microRNAs in peripheral blood of lung cancer patients. Mol Biosyst 7:3187–3199
29. Ma J, Mannoor K, Gao L, Tan A, Guarnera MA, Zhan M, Shetty A, Stass SA, Xing L, Jiang F (2014) Characterization of microRNA transcriptome in lung cancer by next-generation deep sequencing. Mol Oncol 8:1208–1219
30. Ekblom R, Slate J, Horsburgh GJ, Birkhead T, Burke T (2012) Comparison between normalised and unnormalised 454-sequencing libraries for small-scale RNA-Seq studies. Comp Funct Genomics 2012:281693
31. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16: 545–552
32. Li S, Wang H, Qi Y, Tu J, Bai Y, Tian T, Huang N, Wang Y, Xiong F, Lu Z, Xiao Z (2011) Assessment of nanomaterial cytotoxicity with SOLiD sequencing-based microRNA expression profiling. Biomaterials 32:9021–9030
33. Shen J, Stass SA, Jiang F (2013) MicroRNAs as potential biomarkers in human solid tumors. Cancer Lett 329:125–136
34. Shen J, Liu Z, Todd NW, Zhang H, Liao J, Yu L, Guarnera MA, Li R, Cai L, Zhan M, Jiang F (2011) Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microRNA biomarkers. BMC Cancer 11:374
35. Shen J, Todd NW, Zhang H, Yu L, Lingxiao X, Mei Y, Guarnera M, Liao J, Chou A, Lu CL, Jiang Z, Fang H, Katz RL, Jiang F (2011) Plasma microRNAs as potential biomarkers for non-small-cell lung cancer. Lab Invest 91: 579–587
36. Dodd DW, Gagnon KT, Corey DR (2013) Digital quantitation of potential therapeutic target RNAs. Nucleic Acid Ther 23:188–194
37. Whale AS, Huggett JF, Cowen S, Speirs V, Shaw J, Ellison S, Foy CA, Scott DJ (2012) Comparison of microfluidic digital PCR and conventional quantitative PCR for measuring copy number variation. Nucleic Acids Res 40, e82
38. Warren L, Bryder D, Weissman IL, Quake SR (2006) Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. Proc Natl Acad Sci U S A 103: 17807–17812
39. Chen X, Ba Y, Ma L, Cai X, Yin Y, Wang K, Guo J, Zhang Y, Chen J, Guo X, Li Q, Li X, Wang W, Wang J, Jiang X, Xiang Y, Xu C, Zheng P, Zhang J, Li R, Zhang H, Shang X, Gong T, Ning G, Zen K, Zhang CY (2008) Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res 18:997–1006
40. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova-Agadjanyan EL, Peterson A, Noteboom J, O'Briant KC, Allen A, Lin DW, Urban N, Drescher CW, Knudsen BS, Stirewalt DL, Gentleman R, Vessella RL, Nelson PS, Martin DB, Tewari M (2008) Circulating microRNAs as stable blood-based markers for cancer detection. Proc Natl Acad Sci U S A 105:10513–10518
41. Pritchard CC, Kroh E, Wood B, Arroyo JD, Dougherty KJ, Miyaji MM, Tait JF, Tewari M (2012) Blood cell origin of circulating microRNAs: a cautionary note for cancer biomarker studies. Cancer Prev Res (Phila) 5:492–497
42. Bhat S, Herrmann J, Armishaw P, Corbisier P, Emslie KR (2009) Single molecule detection in nanofluidic digital array enables accurate measurement of DNA copy number. Anal Bioanal Chem 394:457–467
43. Kiss MM, Ortoleva-Donnelly L, Beer NR, Warner J, Bailey CG, Colston BW, Rothberg JM, Link DR, Leamon JH (2008) High-throughput quantitative polymerase chain reaction in picoliter droplets. Anal Chem 80: 8975–8981
44. Kreutz JE, Munson T, Huynh T, Shen F, Du W, Ismagilov RF (2011) Theoretical design and analysis of multivolume digital assays with wide dynamic range validated experimentally with microfluidic digital PCR. Anal Chem 83:8158–8168
45. Pinheiro LB, Coleman VA, Hindson CM, Herrmann J, Hindson BJ, Bhat S, Emslie KR (2012) Evaluation of a droplet digital polymerase chain reaction format for DNA copy number quantification. Anal Chem 84:1003–1011

46. Pohl G, Shih Ie M (2004) Principle and applications of digital PCR. Expert Rev Mol Diagn 4:41–47
47. Vogelstein B, Kinzler KW (1999) Digital PCR. Proc Natl Acad Sci U S A 96:9236–9241
48. Hayden RT, Gu Z, Ingersoll J, Abdul-Ali D, Shi L, Pounds S, Caliendo AM (2013) Comparison of droplet digital PCR to real-time PCR for quantitative detection of cytomegalovirus. J Clin Microbiol 51:540–546
49. Day E, Dear PH, McCaughan F (2013) Digital PCR strategies in the development and analysis of molecular biomarkers for personalized medicine. Methods 59:101–107
50. Diehl F, Diaz LA Jr (2007) Digital quantification of mutant DNA in cancer patients. Curr Opin Oncol 19:36–42
51. Hindson BJ, Ness KD, Masquelier DA, Belgrader P, Heredia NJ, Makarewicz AJ, Bright IJ, Lucero MY, Hiddessen AL, Legler TC, Kitano TK, Hodel MR, Petersen JF, Wyatt PW, Steenblock ER, Shah PH, Bousse LJ, Troup CB, Mellen JC, Wittmann DK, Erndt NG, Cauley TH, Koehler RT, So AP, Dube S, Rose KA, Montesclaros L, Wang S, Stumbo DP, Hodges SP, Romine S, Milanovich FP, White HE, Regan JF, Karlin-Neumann GA, Hindson CM, Saxonov S, Colston BW (2011) High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. Anal Chem 83:8604–8610
52. Sanders R, Huggett JF, Bushell CA, Cowen S, Scott DJ, Foy CA (2011) Evaluation of digital PCR for absolute DNA quantification. Anal Chem 83:6474–6484

Chapter 9

Cancer Gene Profiling for Response Prediction

B. Michael Ghadimi and Peter Jo

Abstract

The revolution of genomic technologies, including gene expression profiling, high-resolution mapping of genomic imbalances, and next-generation sequencing, allows us to establish molecular portraits of cancer cells with unprecedented accuracy. This generates hope and justifies anticipation that disease diagnosis, prognosis, and the choice of treatment will be adapted to the individual needs of patients based on molecular evidence.

Preoperative treatment strategies are now recommended for a variety of human cancers. Unfortunately, the response of individual tumors to a preoperative treatment is not uniform, and ranges from complete regression to resistance. This poses a considerable clinical dilemma, as patients with a priori resistant tumors could either be spared exposure to radiation or DNA-damaging drugs, i.e., could be referred to primary surgery, or dose-intensified protocols could be pursued. Because the response of an individual tumor as well as therapy-induced side effects represent the major limiting factors of current treatment strategies, identifying molecular markers of response or for treatment toxicity has become exceedingly important.

However, complex phenotypes such as tumor responsiveness to multimodal treatments probably do not depend on the expression levels of just one or a few genes and proteins. Therefore, methods that allow comprehensive interrogation of genetic pathways and networks hold great promise in delivering such tumor-specific signatures, since expression levels of thousands of genes can be monitored simultaneously. Over the past few years, microarray technology has emerged as a central tool in addressing pertinent clinical questions, the answers to which are critical for the realization of a personalized genomic medicine, in which patients will be treated based on the biology of their tumor and their genetic profile (Quackenbush, N Engl J Med 354:2463–72, 2006; Jensen et al., Curr Opin Oncol 18:374–380, 2006; Bol and Ebner, Pharmacogenomics 7:227–235, 2006; Nevins and Potti, Nat Rev Genet 8:601–609, 2007).

Key words Gene expression profiling, Microarrays, Rectal cancer, Preoperative chemoradiotherapy, Response prediction, Personalized medicine

1 Introduction

The major advantage of microarray technology over other techniques that study expression levels of genes is that tens of thousands of genes can be studied simultaneously in one single experiment. It has been shown that gene expression profiles of cancer cell lines correlate with drug activity [1–3] or radiosensitivity [4].

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_9, © Springer Science+Business Media New York 2016

It has also been demonstrated that gene expression signatures predicting sensitivity to chemotherapeutic drugs in vitro can also accurately predict clinical response in patients treated with these drugs in vivo [5]. In analogy to these model systems, gene expression signatures have been identified that predict the response of breast cancers to preoperative chemotherapy [6], of esophageal carcinomas to preoperative chemoradiotherapy [7], or colon cancers to postoperative chemotherapy [8]. In addition, prognostic signatures have been established for patients with pancreatic ductal adenocarcinoma [9], breast cancers [10–13], and non-small-cell lung cancer [14, 15], leading to the initiation of multicenter trials to test the clinical effectiveness of these gene sets [16, 17].

We recently demonstrated that gene expression profiling might be useful for predicting the response of locally advanced rectal cancers to preoperative chemoradiotherapy [18]. These results led us to initiate prospective profiling of tumor samples from patients enrolled in the CAO/ARO/AIO-04 and TransValid A trial of the German Rectal Cancer Study Group, which is integrated into a Clinical Research Unit (KFO 179) funded by the German Research Foundation (DFG). In this framework, a multicenter biobank could be established, which meets the current standards and requirements of modern biobanks to guarantee reliable results based on quality assured biospecimen [19]. Large-scale tumor sample analyses are generally necessary to identify or validate molecular markers that could help to individualize patient treatment by response prediction. Beside the tissue that has been retrieved and stored based on rigid standardized operating procedures (SOPs) high quality clinical data are necessary. To provide an excellent platform for translational research tissue acquisition and clinical data collection are best performed within clinical trials [20]. This again requires a well-developed information technology (IT) infrastructure that is of increasing importance if patient numbers are high and combinations of clinical data, biomaterial, and molecular results are warranted.

This chapter is focused on preparing tumor samples which were taken preoperatively from patients with locally advanced rectal cancer for assessing the response prediction with the Agilent Technologies platform (Agilent, Santa Clara, CA), using the Low Input Quick Amp Labeling one-color- and the Human Gene Expression 4x44K v2 Microarray Kit. The Agilent One-color Microarray-based Gene Expression Analysis uses cyanine 3-labeled targets to measure gene expression in experimental samples. Figure 1 is a standard workflow for sample preparation and array hybridization design. Total RNA was isolated from RNAlater tumor samples. Gene expression profiling was performed using 200 ng of total RNA.

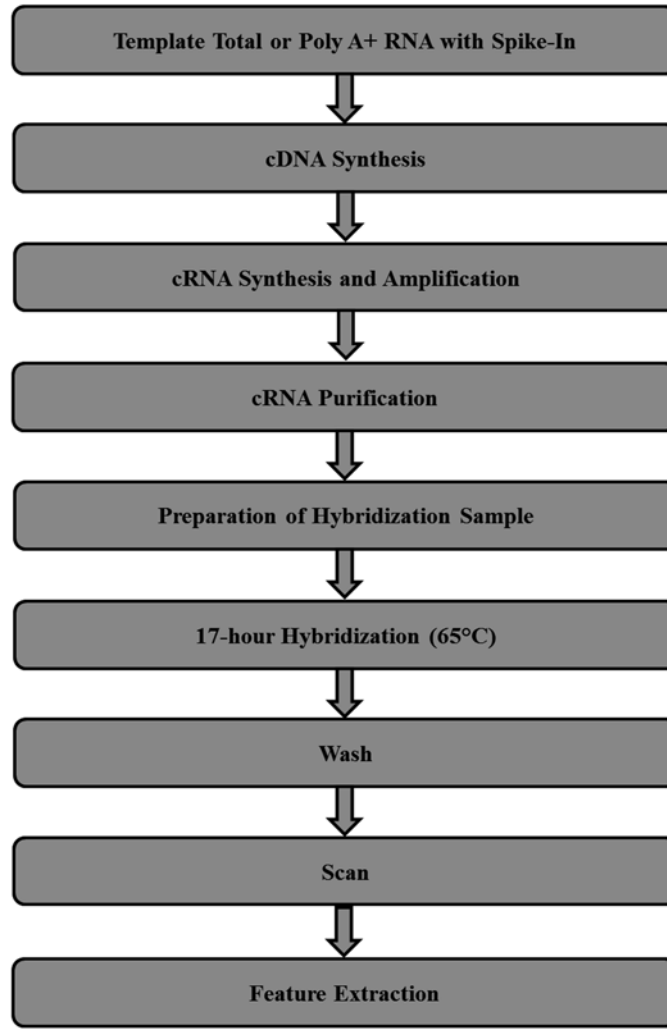


Fig. 1 Workflow for sample preparation and array processing

2 Materials

2.1 Sample Accrual and Storage

1. RNAlater (Qiagen, Hilden, Germany). Make aliquots of 1 ml in polypropylene tubes and store at room temperature.

2.2 RNA Isolation

1. TRIzol (Life Technologies, Carlsbad, CA). Cover bottle with aluminum foil.
2. Anatomical forceps that can be sterilized. You need one forceps per sample.
3. Tissue Lyser with Stainless Steel Beads, 5 mm (Qiagen, Hilden, Germany).

4. Chloroform (Sigma Aldrich, St. Louis, MO).
5. GlycoBlue 20 μg/μl (Life Technologies, Carlsbad, CA).
6. Isopropyl alcohol (Carl Roth, Karlsruhe, Germany).
7. 200-proof ethyl alcohol (Warner-Graham, Cockeysville, MD).
8. RNase-, DNase-free water (ATCC, Manassas, VA).
9. Recommended: Spectrophotometer (Nanodrop, Rockland, DE).
10. Recommended: Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA).
11. RNA-, DNA-, RNase-, DNase-free sterile, cotton-plugged pipette tips.
12. RNA-, DNA-, RNase-, DNase-free microcentrifuge tubes.
13. Always wear gloves!

2.3 RNA Amplification and Hybridization

1. Low Input Quick Amp Labeling Kit Cy3 OneColor (Agilent, Santa Clara, CA).
2. RNA Spike-In Kit, onecolor (Agilent, Santa Clara, CA).
3. 2× HIRPM Hybridization Buffer (Agilent, Santa Clara, CA).
4. 2x Hybridization Buffer (Agilent, Santa Clara, CA).
5. 25× Fragmentation Buffer (Agilent, Santa Clara, CA).
6. 10× Blocking Agent (Agilent, Santa Clara, CA).
7. RNase-free water/MilliQ water (ATCC, Manassas, VA).
8. RNeasy® Mini Kit (Qiagen, Hilden, Germany).
9. 20× SSPE (AppliChem, Darmstadt, Germany).
10. 20 % N-Lauroylsarcosine (Sigma Aldrich, St. Louis, MO).
11. Triton X 102 (Agilent, Santa Clara, CA).
12. Acetonitrile (AppliChem, Darmstadt, Germany).

2.4 Equipment

1. Vortex (IKA, Staufen, Germany).
2. Agilent 2100 BioAnalyzer (Agilent, Santa Clara, CA).
3. Centrifuge (Eppendorf, Hamburg, Germany).
4. RNase-free disposable (pipettes, tips 1.5 ml tubes, 2 ml tubes).
5. Heating Block (Eppendorf, Hamburg, Germany).
6. Thermocycler (Eppendorf, Hamburg, Germany).
7. Hybridization Chamber (Agilent, Santa Clara, CA).
8. Backings (Agilent, Santa Clara, CA).
9. Hybridization oven G2545A (Agilent, Santa Clara, CA).
10. Staining dishes (Agilent, Santa Clara, CA).
11. NanoDrop ND1000 (Thermo Scientific, Waltham, MA).
12. DNA Microarray Scanner G2505B (Agilent, Santa Clara, CA).

3 Methods

The setup of gene expression microarray experiments largely depends on two factors: the amount of RNA of a given sample, and the microarray platform. Depending on the samples that are used, RNA amplification may be required. Some microarray manufacturers like Agilent Technologies have already included an RNA amplification step in their protocol. If RNA amplification is necessary to obtain sufficient amounts of RNA (e.g., for repeat hybridizations), there are many companies that provide special kits.

We have experience with Agilent Technologies (Santa Clara, CA, USA), recently with the Low Input Quick Amp Labeling and the Human Gene Expression 4x44K v2 Microarray Kit. Since the biopsies of the rectal cancer tumor are usually small and the isolated total RNA yield low, we exceeded with the abovementioned platform good results. The Low Input Quick Amp Labeling Kit, One-Color generates fluorescent cRNA (complimentary RNA) with a sample input RNA range between 10 and 200 ng of total RNA or a minimum of 5 ng of poly A+ RNA for one-color processing. The method uses T7 RNA Polymerase Blend (red cap), which simultaneously amplifies target material and incorporates Cyanine 3-CTP. Amplification is typically at least a 100-fold from total RNA to cRNA with the use of this kit. All samples were hybridized to the oligonucleotide-based Whole Human Genome Microarrays. Slides were washed, scanned, and analyzed using a G2565BA scanner and finally Feature Extraction software applied.

It should be noted that some kits generate sense RNA, while others generate anti-sense RNA. This is not a factor when hybridizing to cDNA arrays, but must be taken into consideration for oligonucleotide arrays. Additionally, some kits enable two rounds of amplification for higher yield. The decision which kit to use is therefore based on the design of the microarray platform, since (spotted) microarrays can represent single- or double-stranded sequences.

Another very important aspect to consider is if one-color or dual-color hybridizations should be performed. Both techniques are accepted in the microarray field for use with specific platforms, and each has advantages and disadvantages which are discussed elsewhere [21, 22]. As mentioned above, we performed recently one-color hybridizations.

Finally, because of the dynamic nature of this technology, it should be noted that this protocol is optimized for those microarrays that we purchased. Other commercially available microarrays obviously require different protocols. The general considerations outlined here and the protocol for collecting tissue samples and RNA isolation, however, hold true for those too.

3.1 Ascertainment of Tumor Biopsies and Storage

The time interval between sampling and storage is very important because even partial degradation impairs microarray analyses. For gene expression profiling, we therefore strongly recommend to take tissue samples directly in the operating room, or in an endoscopic unit. The samples should be immediately stored in a RNA stabilization reagent or frozen directly in liquid nitrogen. We and others have good experience with RNAlater. The advantage of RNA stabilization reagents is that they are ready for use, and can be stored in cups or tubes at room temperature for months, especially in hospitals, where liquid nitrogen is not available.

According to established Standard Operating Procedures (SOPs) pretherapeutic tumor biopsies from patients with rectal cancer were collected prospectively and stored in RNAlater (Qiagen, Hilden, Germany), the first 24 h at room temperature and for long-term storage at −20 °C until further reprocessing.

3.2 Quality Assurance

Histopathological assessment by a board-certified pathologist was performed for each biopsy to exclude biosamples containing less than 50 % of tumor cells. RNA isolation was performed using TRIZOL (Invitrogen, Carlsbad, CA) as described in Subheading 3.2. RNA samples with an RNA Integrity Number (RIN) of <5 were excluded. The asservation, storage, and quality assurance of biospecimen should be taken under stringent SOPs to avoid artificial results and to achieve reliable gene expression analyses. All samples were first labeled and then stored at −80 °C before hybridization. Hybridizations were randomized, and controls were included to minimize potential experimental artifacts. All samples were hybridized to the same batch of the oligonucleotide-based Whole Human Genome Microarray.

3.3 RNA Isolation

The single most critical factor for a successful microarray experiment is the RNA, i.e., its purity and integrity. There are many different protocols available for RNA isolation. Since we have not only focused on the cellular transcriptome, but also on genomic and proteomic analysis, we have been primarily using TRIzol. The isolation protocol described here is based on the manufacturer's recommendation with minor modifications according to our experience. In our hands, we have been able to successfully isolate sufficient amounts of RNA from cancer biopsies with weights ranging from 5 to 150 mg.

1. Thaw tumor samples that have been stored in RNAlater, and using a sterile forceps, transfer the tissue immediately into a 2 ml polypropylene tube containing 1 ml of the TRIzol reagent with additionally 1 Stainless Steal Beads, 5 mm (*see* **Note 1**).
2. Thoroughly homogenize samples to disrupt cells and dissolve components with the Tissue Lyser, which usually takes approximately 3–5 min at an amplitude of 50 Hz, and incubate

for 5 min at room temperature to dissociate nucleoprotein complexes.

3. Remove the Beads and add 200 µl chloroform, tightly cap tubes, and shake vigorously for 30 s.
4. Allow phase separation for 15 min on ice, and centrifuge at 12,000 × *g* for 15 min at 4 °C (phase separation).
5. Transfer very carefully only the upper aqueous phase (colorless, ca. 600 µl), containing mostly RNA, to a new 1.5 ml polypropylene tube (*see* **Note 2**).
6. Add 1 µl GlycoBlue and mix briefly (*see* **Note 3**).
7. Add 500 µl isopropyl alcohol to precipitate the RNA.
8. Vortex tube and incubate for at least 1 h at –20 °C (*see* **Note 4**).
9. Centrifuge at 12,000 × *g* for 30 min at 4 °C.
10. Remove the supernatant (*see* **Note 5**), and add 1 ml of 75 % ethanol to wash off residuals of TRIzol.
11. Break up the pellet by pipetting up and down and vortex a few seconds.
12. Wash the pellet for 10 min at room temperature on a rotator (*see* **Note 6**).
13. Centrifuge at 10,000 × *g* for 15 min at 4 °C to pellet the RNA.
14. Remove the supernatant, and add 1 ml of 75 % ethanol.
15. Break up the pellet by pipetting up and down, and transfer it to a new RNase-free 1.5 ml microcentrifuge tube.
16. Vortex, and wash the pellet for 10 min at room temperature on a rotator.
17. Centrifuge at 7500 × *g* for 15 min at 4 °C.
18. Carefully remove supernatant, and briefly air-dry the pellet at room temperature (*see* **Note 7**).
19. Resuspend the pellet in 20–100 µl RNase-, DNase-free water (*see* **Note 8**), and incubate at 65 °C for 5 min on a shaking thermomixer.
20. Cool down the sample on ice, and determine the quantity, purity, and integrity of your RNA (*see* **Note 9**).
21. Store RNA at –80 °C.

4 Procedure

Before you begin:

1. Determine the integrity of the input RNA with the Agilent BioAnalyzer (*see* **Note 9**).
2. Vortex all thawed reagents.

4.1 Preparation of the Spike-In Mix

1. Mix the stock solution (stored at −80 °C) vigorously on a vortex mixer.
2. Heat at 37 °C for 5 min, mix on a vortex mixer, and centrifuge briefly.
3. Dilute the stock solution according to the below:

Starting amount of RNA		Serial dilution				
Total RNA (ng)	Max volume of RNA (µl)	First	Second	Third	Fourth	*Spike-in mix volume to be used* (µl)
25	1.5	1:20	1:25	1:20	1:4	2
50	1.5	1:20	1:25	1:20	1:2	2
100	1.5	1:20	1:25	1:20		2
200	1.5	1:20	1:25	1:10		2

4.1.1 Example for Preparation of the Spike-In Mix Dilutions for 200 ng RNA Input

1. Mix the thawed Spike-In Mix concentrate vigorously on a vortex mixer.
2. Heat at 37 °C for 5 min.
3. Mix the Spike-In concentrate again on a vortex mixer.
4. Spin briefly in a centrifuge.
5. Add 2 µl of the Spike-In Mix concentrate to 38 µl of Dilution Buffer provided in the kit (1:20).
6. Mix thoroughly on a vortex mixer and spin down briefly. This tube contains the First Dilution.
7. Add 2 µl of First Dilution to 48 µl of Dilution Buffer for the Second Dilution (1:25).
8. Mix thoroughly on a vortex mixer and spin down briefly. This tube contains the Second Dilution.
9. Depending on the number of samples for microarray analysis add:

   2 µl of Second Dilution to 18 µl of Dilution Buffer or

   4 µl of Second Dilution to 36 µl of Dilution Buffer

   for the Third Dilution (1:10).
10. Mix thoroughly on a vortex mixer and spin down briefly. This tube contains the Third Dilution (5000-fold dilution).
11. Add 2 µl of Third Dilution to 200 ng of total RNA and continue with the preparation of the labeling reaction.

4.1.2 Storage

1. Store stock solutions at −80 °C for up to 1 year.
2. Store the first dilution at −80 °C for up to 2 months and freeze/thaw up to eight times.
3. Discard the second and third dilution tubes.

4.2 Prepare the Labeling Reaction

1. Add 25–200 ng of total RNA to an RNase-free tube.
2. Add water to get the max RNA volume indicated in the table below (1.5 µl).
3. If the input RNA has a greater volume than the max RNA volume you have to concentrate the sample in a speed vac.
4. Add 2 µl of diluted Spike-In Mix to each tube. Each tube now contains a total volume of 3.5 µl.
5. Add 0.8 µl of T7 Promoter Primer and 1.0 µl of nuclease-free water to get a total volume of 5.3 µl.

Total RNA input (ng)	Max RNA volume (µl)	Fourth dilution of spike-mix (µl)	T7 primer (µl)	Water (µl)	Total volume (µl)
25	1.5	2 (Fourth Dilution)	0.8	1.0	5.3
50	1.5	2 (Fourth Dilution)	0.8	1.0	5.3
100	1.5	2 (Third Dilution)	0.8	1.0	5.3
200	1.5	2 (Third Dilution)	0.8	1.0	5.3

6. Denature primer and template at 65 °C for 10 min in a thermocycler.
7. Place reactions on ice for 5 min.
8. Prewarm the 5× first-strand buffer at 80 °C for 3–4 min. Keep at RT.

4.3 Prepare the cDNA Master Mix

The cDNA master mix is prepared at RT. Thaw all reagents at RT. Be sure to use the 10 mM dNTP mix. Add an extra half-reaction to your Master Mix.

Component	Volume per rx (µl)
5× First-strand buffer	2.0
0.1 M DTT	1.0
10 mM dNTP mix	0.5
AffinityScript RNase Block Mix	1.2
Total volume	*4.7*

1. Add 4.7 µl of cDNA Master Mix to each sample tube and mix by pipetting up and down. Each tube now contains a total volume of 10 µl.
2. Incubate samples in a thermocycler at 40 °C for 2 h.

3. Incubate samples in a thermocycler at 70 °C for 15 min. Incubation at 70 °C inactivates the AffinityScript enzyme.
4. Place reactions on ice for 5 min.

Stopping Point: If you do not immediately continue to the next step, store the samples at –80 °C.

4.4 Prepare the Transcription Master Mix

Immediately prior to use, gently mix the components listed in the table below in the order indicated for the Transcription Master Mix by pipetting at room temperature.

The T7 RNA polymerase blend is a blend of enzymes. Keep the T7 RNA polymerase on ice and add to the Transcription Master Mix just before use.

Component	Volume per rx (μl)
Water	0.75
5× Transcription buffer	3.2
0.1 M DTT	0.6
NTP mix	1.0
T7 RNA Polymerase	0.21
Cyanine 3-CTP	0.24
Total volume	*6.0*

1. Add 6.0 μl of Transcription Master Mix to each sample and mix by pipetting. Each tube now contains a total volume of 16 μL.
2. Incubate samples in a thermocycler at 40 °C for 2 h.

Stopping Point: If you do not immediately continue to the next step, store the samples at –80 °C.

4.5 Purify the Labeled aRNA

Qiagen's RNeasy Mini Kit is used to purify the amplified RNA samples.

1. Add 84 μl water to your samples for a total volume of 100 μl.
2. Add 350 μl of buffer RLT and mix well by pipetting.
3. Add 250 μl of 100 % ethanol to precipitate the RNA. Mix directly by pipetting up and down four to six times. Do not centrifuge!
4. Transfer your sample to an RNeasy mini column and centrifuge at 12,000 × *g* at 4 °C for 30 s at 13,000 rpm. Discard the flow-through and collection tube.
5. Transfer the column to a new collection tube.

6. Add 500 µl of buffer RPE (containing ethanol) to the column and centrifuge at 4 °C for 30 s at 13,000 rpm. Discard flow-through.
7. Add another 500 µl of buffer RPE to the column and centrifuge at 4 °C for 30 s at 13,000 rpm. Discard flow-through and collection tube.
8. Transfer the column to a new collection tube and centrifuge for an additional minute at 13,000 rpm. Discard the collection tube.
9. Transfer the column to a new 1.5 ml tube and elute the aRNA by adding 30 µl RNase-free water directly onto the membrane of the column. Wait for 1 min, and then centrifuge at RT for 1 min at 13,000 rpm.

4.6 Quantify the Labeled aRNA

1. Quantify the labeled aRNA using the NanoDrop ND1000.
2. Determine the yield and specific activity of each reaction as follows:

   Use the concentration of cRNA (ng/µl) to determine the µg cRNA yield:

$$(\text{Conc. of cRNA}) \times 30\ \mu\text{l}\ (\text{elution volume}) / 1000 = \mu\text{g of cRNA}$$

   Use the concentration of cRNA (ng/µl) and cyanine 3 (pmol/µl) to determine the specific activity:

$$(\text{Conc. of Cy3}) / (\text{Concentration of cRNA}) \times 1000 = \text{pmol Cy3 per } \mu\text{g cRNA}$$

   If the yield is < 1.65 µg per array and the specific activity is < 6.0 pmol/µg do not proceed to the hybridization step. Repeat cRNA preparation.

4.7 Prepare the 10× Blocking Agent

1. Add 500 µl RNase-free water to the vial containing lyophilized 10× blocking agent supplied with the Agilent Gene Expression Hybridization Kit, or add 1250 µl RNase-free water to the vial containing lyophilized large volume 10x blocking agent.
2. Mix by gently vortexing. If the pellet does not go into solution completely, heat the mix for 4–5 min at 37 °C.
3. Centrifuge for 5–10 s to drive down material adhering to the tube wall or cap.
4. Storage: 10× blocking agent can be stored at –20 °C for up to 2 months. After thawing, vortex and centrifuge again before use.

4.8 Prepare Hybridization Samples

1. Set the hybridization oven already to 65 °C.
2. For each microarray, add the following components to a 1.5 ml tube.
   Fragmentation mix 4x44K microarrays

Component	Volume/Mass
Cyanine 3-labeled cRNA	1.65 µg
10× blocking agent	11 µl
RNase-free water	Bring volume to 52.8 µl
25× fragmentation buffer	2.2 µl
Total volume	*55*µl

3. Incubate at 60 °C for exactly 30 min to fragment the RNA.
4. Add buffer to stop the fragmentation reaction:

   2× GEx hybridization buffer HIRPM to the 4x44K array formats.

   Hybridization mix

Component	4x44K
cRNA from fragmentation mix	55 µl
2× GEx hybridization buffer HI-RPM	11 l

5. Mix by careful pipetting.
6. Spin for 1 min at RT at full speed. Use immediately.

4.9 Prepare the Hybridization Assembly (See Note 10)

1. Load a clean backing (gasket slide) into the Agilent SureHyb chamber base with the label facing up.
2. Slowly dispense the volume of hybridization sample onto the backing. Do not allow the hybridization solution to touch the walls of the backing.
   Hybridization sample

Component	4x44K
Volume prepared	110 µl
Hybridization sample volume	100

3. Slowly place an array "active side" down onto the backing, so that the "Agilent"-labeled barcode is facing down and the numeric barcode is facing up.
4. Place the SureHyb chamber cover onto the sandwiched slides and hand-tighten the clamp onto the chamber.
5. Vertically rotate the assembled chamber to wet the backing and assess the mobility of the bubbles. If necessary, tap the assembly on a hard surface to move stationary bubbles.
6. Place the slide chamber in the hybridization oven set to 65 °C. Be sure to balance the loaded hybridization chambers on the rack.

7. Set the rotor to rotate at 4 rpm when using the 2× hybridization buffer or 10 rpm when using the 2× GEx hybridization buffer HIRPM.
8. Hybridize at 65 °C for 17 h.

4.10 Wash the Microarray Slides (See Notes 11 and 12)

1. Prepare the following wash solutions: Prepare 500 ml per four slides.
2. Wash solution 1: 6× SSPE, 0.005 % N-Lauroylsarcosine, Triton X-102.

Component	Volume	Volume
Water	700 ml	350 ml
20× SSPE	300 ml	150 ml
20 % N-Lauroylsarcosine	250 µl	125 µl
Triton X-102	250 µl	125 µl
Total volume	*1000*ml	*500*l

3. Wash solution 2: 0.06× SSPE, 0.005 % N-Lauroylsarcosine, Triton X102.
4. Prepare 250 ml per four slides.

Component	Volume	Volume
Water	997 ml	249 ml
20× SSPE	3 ml	750 µl
20 % N-Lauroylsarcosine	250 µl	62.5 µl
Triton X-102	250 µl	125 µl
Total volume	*1000*ml	*250*ml

5. Prepare four staining dishes for washing the slides:
6. To the first staining dish add 250 ml of wash solution 1 at RT.
7. To the second staining dish add also 250 ml of wash solution 1 and a slide rack and a magnetic stir bar. Place the dish on a magnetic stir plate.
8. To the third staining dish add 250 ml wash solution 2 at RT and a magnetic stir bar. Place the dish on a magnetic stir plate.
9. To the fourth staining dish add 250 ml acetonitrile. Cover the dish with a glass plate.
10. Remove one hybridization chamber from the oven and determine if all bubbles rotate freely.
11. Place the chamber assembly on a flat surface and loosen the clamp.

12. Slide off the clamp assembly and remove the chamber cover.
13. Remove the sandwiched slides (array + backing) from the chamber base by grabbing the slides from their ends and transfer the sandwich quickly to the first staining dish and submerge them in wash solution 1.
14. With the sandwiched slides completely submerged in the wash solution, pry the two slides apart from the barcode end. Do this by slipping one end of the tweezers between the slides and then gently turn the tweezers to separate the slides.
15. Let the backing drop to the bottom of the staining dish.
16. Remove the slide quickly to the slide rack in the second staining dish. Minimize exposure of the slide to air. Touch only the barcoded portion of the slide or its edges.
17. Repeat this procedure for the next slides. Do not wash more than four slides in one staining dish.
18. After all the slides have been collected in the staining dish containing the Wash Solution 1, turn on the magnetic stir plate adjusted to 300 rpm. Wash slides for 5 min at RT.
19. Transfer the slide rack quickly to the third staining dish with wash solution 2. Wash slides for exactly 5 min at 37 °C.
20. Remove the glass cover from the fourth staining dish and transfer the slide rack to the staining dish containing acetonitrile.
21. After 5–10 s slowly remove the slide rack from the staining dish. It should take 5–10 s to remove the slide rack.
22. Use fresh wash buffer for each new wash group.
23. Scan slides.
24. Wash all dishes, racks, and stir bars with MilliQ water.

4.11 Scanning

- Scan slides within 24 h, preferably directly after the washes. Use settings according to the recommendations in the scanner manual (*see* **Note 13**).

5 Notes

1. TRIzol is toxic and should be handled under a fume hood. We recommend weighing the tissue samples at this point, since it might give a rough estimate on the amount of RNA to expect.
2. Three phases should be visible: a lower red phenol–chloroform phase (proteins), an interphase (DNA), and a colorless upper aqueous phase (RNA). Be careful not to disturb the interphase when removing the upper phase; it is better to loose some RNA than risking contamination with DNA. If you wish to subsequently isolate DNA and proteins as well, you need to

keep the phenol–chloroform phase and the interphase (please read the manufacturer's protocol).

3. Glycogen serves as an inert co-precipitant and increases nucleic acid recovery. It also helps to visualize the RNA pellet after precipitation, and does not inhibit further reactions.
4. One may wish to stop at this point and leave the tubes over night at –20 °C.
5. Depending on the size of the tissue samples, this pellet might be very small and difficult to see (*see* **Note 3**).
6. You can stop at this point, and store your sample at –20 °C (months).
7. Do not vacuum-dry the RNA pellet, and do not air-dry completely; otherwise its solubility will be decreased.
8. The volume of RNase-, DNase-free water to be added is strongly influenced by the amplification protocol that you wish to use.
9. To evaluate RNA quantity and purity, perform spectrophotometric readings at wavelengths of 260 nm and 280 nm. Depending on the quality of the RNA, expect 260/280 ratios between 1.9 and 2.1. For subsequent microarray experiments, we strongly recommend analyzing your samples with Agilent's Bioanalyzer or a similar technique. A spectrophotometer does not provide information about the RNA integrity, and even RNA with a perfect 260/280 ratio can be degraded. Even though a Bioanalyzer is not capable of determining the percentage of full-length mRNA it is in our opinion more reliable than a conventional denaturing agarose gel and requires a smaller amount of RNA.
10. Hybridize the arrays in an ozone free room. Wear nitrile gloves during working with arrays.
11. Wash the arrays in an ozone free room. Wear nitrile gloves during working with arrays.
12. Alternatively you can purchase the Gene Expression Wash Buffer Kit, Agilent Technologies, containing Gene Expression Wash Buffer 1 and 2, and Triton X-102 (10 %) and wash the slides according to the manufacturer's instructions.
13. We recommend to document each step and to archival storage the documentation by filling a form.

Acknowledgements

The authors would like to thank PD Dr. Jochen Gaedcke, PD Dr. Marian Grade, Dr. Gabriela Salinas-Riester, and Mr. Chan Rong Lai for their advice. This work was supported by the Deutsche Forschungsgemeinschaft (KFO 179).

References

1. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, Pergamenschikov A, Lee JC, Lashkari D, Shalon D, Myers TG, Weinstein JN, Botstein D, Brown PO (2000) Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 24:227–35
2. Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT, Scudiero DA, Eisen MB, Sausville EA, Pommier Y, Botstein D, Brown PO, Weinstein JN (2000) A gene expression database for the molecular pharmacology of cancer. Nat Genet 24:236–44
3. Mariadason JM, Arango D, Shi Q, Wilson AJ, Corner GA, Nicholas C, Aranes MJ, Lesser M, Schwartz EL, Augenlicht LH (2003) Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. Cancer Res 63:8791–812
4. Torres-Roca JF, Eschrich S, Zhao H, Bloom G, Sung J, McCarthy S, Cantor AB, Scuto A, Li C, Zhang S, Jove R, Yeatman T (2005) Prediction of radiation sensitivity using a gene expression classifier. Cancer Res 65:7169–76
5. Potti A, Dressman HK, Bild A, Riedel RF, Chan G, Sayer R, Cragun J, Cottrill H, Kelley MJ, Petersen R, Harpole D, Marks J, Berchuck A, Ginsburg GS, Febbo P, Lancaster J, Nevins JR (2006) Genomic signatures to guide the use of chemotherapeutics. Nat Med 12:1294–300
6. Lønning PE, Knappskog S, Staalesen V, Chrisanthar R, Lillehaug JR (2007) Breast cancer prognostication and prediction in the post-genomic era. Ann Oncol 18:1293–306
7. Luthra R, Wu TT, Luthra MG, Izzo J, Lopez-Alvarez E, Zhang L, Bailey J, Lee JH, Bresalier R, Rashid A, Swisher SG, Ajani JA (2006) Gene expression profiling of localized esophageal carcinomas: association with pathologic response to preoperative chemoradiation. J Clin Oncol 24:259–67
8. Del Rio M, Molina F, Bascoul-Mollevi C, Copois V, Bibeau F, Chalbos P, Bareil C, Kramar A, Salvetat N, Fraslon C, Conseiller E, Granci V, Leblanc B, Pau B, Martineau P, Ychou M (2007) Gene expression signature in advanced colorectal cancer patients select drugs and response for the use of leucovorin, fluorouracil, and irinotecan. J Clin Oncol 25:773–80
9. Haider S, Wang J, Nagano A, Desai A, Arumugam P, Dumartin L, Fitzgibbon J, Hagemann T, Marshall JF, Kocher HM, Crnogorac-Jurcevic T, Scarpa A, Lemoine NR, Chelala C (2014) A multi-gene signature predicts outcome in patients with pancreatic ductal adenocarcinoma. Genome Med 6(12):105
10. van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415:530–6
11. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009
12. Buyse M, Loi S, van't Veer L, Viale G, Delorenzi M, Glas AM, d'Assignies MS, Bergh J, Lidereau R, Ellis P, Harris A, Bogaerts J, Therasse P, Floore A, Amakrane M, Piette F, Rutgers E, Sotiriou C, Cardoso F, Piccart MJ (2006) TRANSBIG Consortium. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 98:1183–92
13. Drukker CA, Bueno-de-Mesquita JM, Retèl VP, van Harten WH, van Tinteren H, Wesseling J, Roumen RM, Knauer M, van't Veer LJ, Sonke GS, Rutgers EJ, van de Vijver MJ, Linn SC (2013) A prospective evaluation of a breast cancer prognosis signature in the observational RASTER study. Int J Cancer 133:929–36
14. Potti A, Mukherjee S, Petersen R, Dressman HK, Bild A, Koontz J, Kratzke R, Watson MA, Kelley M, Ginsburg GS, West M, Harpole DH Jr, Nevins JR (2006) A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N Engl J Med 355:570–80
15. Botling J, Edlund K, Lohr M, Hellwig B, Holmberg L, Lambe M, Berglund A, Ekman S, Bergqvist M, Pontén F, König A, Fernandes O, Karlsson M, Helenius G, Karlsson C, Rahnenführer J, Hengstler JG, Micke P (2013) Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. Clin Cancer Res 19(1):194–204
16. Bogaerts J, Cardoso F, Buyse M, Braga S, Loi S, Harrison JA, Bines J, Mook S, Decker N, Ravdin P, Therasse P, Rutgers E, van't Veer LJ, Piccart M, TRANSBIG Consortium (2006) Gene signature evaluation as a prognostic tool:

challenges in the design of the MINDACT trial. Nat Clin Pract Oncol 3:540–51

17. Anguiano A, Potti A (2007) Genomic signatures individualize therapeutic decisions in non-small-cell lung cancer. Expert Rev Mol Diagn 7:837–44
18. Ghadimi BM, Grade M, Difilippantonio MJ, Varma S, Simon R, Montagna C, Füzesi L, Langer C, Becker H, Liersch T, Ried T (2005) Effectiveness of gene expression profiling for response prediction of rectal adenocarcinomas to preoperative chemoradiotherapy. J Clin Oncol 23:1826–38
19. Dangl A, Demiroglu SY, Gaedcke J, Helbing K, Jo P, Rakebrandt F, Rienhoff O, Sax U (2010) The IT-infrastructure of a biobank for an academic medical center. Stud Health Technol Inform 160:1334–8
20. Hall JA, Brown R (2013) Developing translational research infrastructure and capabilities associated with cancer clinical trials. Expert Rev Mol Med 15, e11
21. de Reyniès A, Geromin D, Cayuela JM, Petel F, Dessen P, Sigaux F, Rickman DS (2006) Comparison of the latest commercial short and long oligonucleotide microarray technologies. BMC Genomics 7:51
22. Patterson TA, Lobenhofer EK, Fulmer-Smentek SB, Collins PJ, Chu TM, Bao W, Fang H, Kawasaki ES, Hager J, Tikhonova IR, Walker SJ, Zhang L, Hurban P, de Longueville F, Fuscoe JC, Tong W, Shi L, Wolfinger RD (2006) Performance comparison of one-color and two-color platforms within the microarray quality control (MAQC) project. Nat Biotechnol 24:1140–50

Chapter 10

Target Gene Discovery for Novel Therapeutic Agents in Cancer Treatment

Sanjay Tiwari, Ole Ammerpohl, and Holger Kalthoff

Abstract

Target identification of novel therapeutic drugs is pivotal for the establishment of (1) new anticancer regiments, (2) to control side effects of the drugs, and (3) to identify appropriate combinations with established drugs.

Here, we describe several in vitro assays applicable to characterize different characteristics of tumor cells. Furthermore, we present a protocol for establishing a reporter gene system for in vivo imaging, allowing for the study of drug effects in small animal models.

Key words Apoptosis, Cell cycle, Tumor invasion, FACS, In vivo imaging, Fluorescence

1 Introduction

Worldwide a large number of studies and projects are in progress to identify and isolate new anticancer drugs from plants, animals, or microorganisms from all parts of the world, from the rain forests to the oceans. Identified by high-throughput assays every year, numerous new potential drugs await further investigation to determine their mode of action. This kind of translation approach also covers the evaluation of putative side effects and the establishment of combination therapies.

Cancer progression, its spreading and prognosis is characterized by different parameters such as cell proliferation, cell motility, cell viability, or angiogenesis. Unfortunately, there is no simple assay available addressing all these biologic processes. Thus, to characterize a putative new drug, several assays have to be performed. In this chapter, we present various tests which can be very useful to identify the mode of action of new drugs and the signaling pathway by which they mediate their activity. Enhanced with this knowledge it is also possible to further delineate putative target genes. This is especially true, when gene expression profiling or DNA methylation data showing the status of specific genes involved

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_10,

in different pathways can be combined with functional assays. For the molecular assays we would like to refer here to the other chapters of this book.

Numerous known anticancer drugs induce cell cycle arrest, often followed by induction of apoptosis. Since this affects primarily rapidly dividing cells (like tumor cells), drugs affecting the cell cycle and inducing apoptosis are useful for many fast growing cancers. To test the ability of an unknown drug to influence the cell cycle, a standard ex vivo assay is the staining of treated tumor cells and untreated control cells with propidium iodide (PI) followed by analysis with a cytometer [1, 2].

If only the cell viability or the total number of cells is of interest, other less apparatus-dependent methods and high-throughput suitable tests like the MTT test [3–6] (measures the activity of mitochondrial enzymes and therefore cell viability) or the crystal violet assay (determines an optical density corresponding to the number of viable cells) are available [7].

Besides the increased cell division another important aspect of many malignant tumors is their ability to spread and to induce metastasis. To spread to a new location, a tumor cell has to leave its current tissue. For this purpose it must gain the ability to destroy or digest the connective tissue or extracellular matrix. There are many tests available investigating different aspects of tumor cell spreading. Here, we present an easy and non-complex assay which determents the tumor cells' ability to digest organic material [8, 9].

Last but not least, we present some information for the in vivo imaging of xenotransplanted animals. In contrast to the in vitro assays discussed above, this approach offers the possibility to investigate the effect of a new drug and to identify mechanisms and pathways involved in an in vivo situation. Here, side effects of the drug and additional contributions of other organs and cells (metabolism of other organs, effect of the immune system) to the tumor growth or metastasis can be studied. Since optical imaging is the technique of choice for preclinical molecular imaging due to its versatility with utilizing multispectral probes, high resolution, fast imaging times and user-friendly imaging systems which are cheaper compared to other modalities, the review will be limited to optical imaging.

Optical imaging utilizes a reporter system based on bioluminescent or fluorescent proteins. Fluorescence refers to a physical property of molecules that are able to absorb light of a specified wavelength and to subsequently emit light of a longer wavelength. The transmission of incident photons through biologic materials is determined by the proportion that are reflected, scattered, or absorbed [5]. Bioluminescence imaging is based on the expression of the luciferase enzyme, which catalyzes the conversion of luciferin, adenosine triphosphate, and oxygen to oxyluciferin, adenosine monophosphate, pyrophosphate, carbon dioxide, and visible

light. The major drawback of optical imaging is the limited depth sensitivity since light is absorbed by chromophores such as hemoglobin, melanin or water and scattered by cellular components, especially membranes. Optical imaging in the near infrared (NIR) and mid-infrared emitted light (650–900 nm) is receiving greater attention since absorption and scattering is at a minima and several imaging probes have been developed which emit in the NIR have recently been developed [10, 11]. Although the poor tissue penetration restrict optical imaging applications in the clinic, for preclinical studies it is extensively used for monitoring the expression of transgenes, tumor growth and metastasis, infections and gene therapy assessments. Optical reporters can be built into animals and linked to target genes and cells. When cloned into promoter/enhancer sequences or engineered into fusion proteins, imaging reporters enable the interrogation of signal transduction pathways and protein–protein interactions in live cells and animals [12]. The in vivo visualization of specific molecular pathways enables the identification of key targets for the development of novel therapeutic agents in cancer treatment. Furthermore, the evaluation of treatment efficiency and optimization of new drugs in reporter models can also lead to the further elucidation and characterization of these pathways, thus improving identification for novel target drugs. Following the identification of a target gene, gene ablation using siRNA is often utilized to determine if the target gene is a "drugable" target. The effect of gene ablation can be performed in vivo utilizing a reporter gene fused to the target gene of interest.

2 Materials

2.1 Cell Culture and Cell Preparation

1. Cell culture media (appropriate for the cell line investigated, e.g. Dulbecco's Modified Eagle's Medium (DMEM); Ham's F12 or RPMI-1640) supplemented with 10 % fetal calf serum (FCS) (Pan-Biotech, Aidenbach, Germany). Depending on the cell line used for the experiments, additional supplements might be mandatory. Information about cell line prerequisites is available from ATCC, DSMZ, ECACC, or the specific literature.
2. Phosphate buffered saline (PBS): sodium hydrogen phosphate (Na_2HPO_4, 10 mM), sodium dihydrogen phosphate (NaH_2PO_4, 2 mM), sodium chloride (130 mM), pH 7.4.
3. Solution of trypsin (0.25 %) and ethylenediamine tetraacetic acid (EDTA, 1 mM) in phosphate buffered saline (PBS).

All solution used for cell culturing must be sterile; either by autoclaving or sterile filtration.

2.2 Cell Cycle Investigations by FACS (PI Staining)

1. Propidium Iodide Stain: Prepare a 0.1 % (1 mg/ml) Stock solution of propidium iodide in PBS. Propidium iodide is a mutagen and carcinogen. It should be handled with care (see instruction of the manufacturer).
2. Ethanol (absolute).
3. EDTA (1 mM) in PBS.
4. RNase A (20 mg/ml). RNase A solution can be purchased from several providers (e.g. Invitrogen) or prepared by dissolving RNase A in PBS to get a 20 mg/ml solution.

2.3 Assays

2.3.1 Cell Viability Assay (MTT-Assay)

1. MTT solution: Dissolve 50 mg MTT powder (3-(4,5-Dimethyl-2-thiazolyl)-2,5-diphenyl-2H-tetrazolium bromide; Sigma-Aldrich, Munich, Germany) in 100 ml RPMI-1640 (without any other additives like FCS) to get a 0.05 % solution.

2.3.2 Cell Assay (Crystal Violet Stain)

1. Solution of 0.5 % Hexamethylpararosaniline chloride (also known as Crystal violet, Methyl violet 10B or Gentian violet) in 20 % methanol and aqua dest (Carl Roth, Karlsruhe, Germany). Crystal violet is a putative mutagen and carcinogen. It should be handled with care (*see* instruction of the manufacturer).

2.3.3 An Easy Tumor Cell Invasion Assay

1. Trypan blue: Prepare a 0.2 % solution of trypan blue in PBS. Alternatively, dilute a commercially available 0.4 % trypan blue solution (Invitrogen, Karlsruhe, Germany) 1:2 with PBS. If photometric measurement is intended, a lysis buffer is required. We recommend the PolyATtract® GTC Extraction Buffer available from Promega (Promega, Mannheim, Germany).

2.4 Stable Transfection

1. Accutase (PAA Laboratories GmbH).
2. 1 % low melting soft agarose (Sigma).
3. 2× DMEM + 20 % FBS.
4. G418 200 mg/ml stock solution (Carl Roth GmbH).
5. LipofectAmine Plus (Invitrogen).
6. Cloning cylinder (Fisher).

3 Methods

3.1 Cell Cycle Investigations by FACS (PI Staining)

3.1.1 Preparing Cells for FACS Analysis

1. Non-adherent cells
   (a) When investigating non-adherent cells, transfer cells including media into a falcon tube.
2. Adherent cells
   (a) When using adherent cells remove media including non-adherent (save it if these cells are to be analyzed also).

(b) Add prewarmed (37 °C) PBS to the cells carefully not to detach the cells until cells are covered.

(c) Remove PBS by decantation or aspiration.

(d) Cover cell layer with trypsin solution.

(e) Store cells at 37 °C in a cell culture incubator until cells have detached. Detaching can be supported by several careful thumps of the culture vessel against the open hand every 5 min during the incubation. Depending on the cell line, the cells usually have detached after 10–20 min.

(f) Stop trypsination by adding one volume cell culture medium including 10 % FCS. Transfer cell suspension into a suitable centrifuge tube.

(g) When non-adherent cells should be included into analysis, add the medium including the floating cells stored from **step** (**a**).

3. Spin at 500 × *g* (approximately 1400 rpm at standard benchtop centrifuges) for 5 min in a centrifuge to pellet the cells.
4. Remove medium carefully by decantation or aspiration not to disturb the cell pellet.
5. Resuspend cell pellet carefully in at least 10 volumes (15–30 ml) PBS.
6. Spin at 500 × *g* (approximately 1400 rpm at standard benchtop centrifuges) for 5 min in a centrifuge to pellet the cells.
7. Resuspend cell pellet carefully in 1–5 ml PBS.
8. Perform cell count using a hematocytometer. Numerous hematocytometers are available (e.g. according to Fuchs-Rosenthal, Thoma, Bürker or Neubauer). If you are not familiar with the system available please refer to the manufacture's manual of your chamber.
9. When not continuing directly with Subheading 3.1.2 store the remaining cells on ice (this time should be kept as short as possible).

3.1.2 Stain Cells for FACS Analysis

1. Spin the remaining cells not used for counting at 500 × *g* (approximately 1400 rpm at standard benchtop centrifuges) for 5 min in a centrifuge to pellet the cells.
2. Add PBS to a final concentration of 2–5 × 10^6 cells/ml and resuspend cell pellet.
3. Transfer 1 ml cell suspension into a new centrifuge tube (e.g. 15 ml conical tube).
4. By vortexing the cell suspension very carefully, add ethanol (abs.) dropwise, 1 ml.
5. Incubate for 30 min at room temperature.

6. Spin the suspension at 500×*g* (approximately 1400 rpm at standard benchtop centrifuges) for 5 min in a centrifuge to pellet the fixed cells.
7. Remove the supernatant by decantation, thumb the tube upside down against an absorbent paper to remove any remaining alcohol (do not lose the pellet!).
8. Add 500 µl PBS containing 1 mM EDTA.
9. Add 2.5 µl RNase A solution to a final concentration of 50 µg/ml and resuspend pellet carefully by flicking against the tube.
10. Transfer cells into an appropriate FACS tube.
11. Spin at 500×*g* (approximately 1400 rpm at standard benchtop centrifuges) for 5 min in a centrifuge.
12. Remove the supernatant by decantation; thumb the tube upside down against an absorbent paper to remove any remaining liquid.
13. Resuspend pellet in 500 µl PBS containing 200 µg/ml propidium iodide.
14. Incubate at room temperature for 30 min.
15. Incubate for at least 30 min at 4 °C in the dark (can be stored overnight if desired).

3.1.3 Run FACS Analysis

A wide range of FACS equipment is available on the market. Thus, we cannot give a detailed description that is appropriate for every cytometer available. Instead, we will present some general considerations and explain some steps exemplary as to do when using a FACScan™ or FACSCalibur™ running CellQuest™ software (Becton Dickenson, Heidelberg, Germany) or WinMDI (http://facs.scripps.edu/software.html; freeware).

1. In contrast to, e.g. DAPI staining which requires a UV-laser, for PI staining a standard blue laser (488 nm) is sufficient. For detection, use the orange fluorescence channel (FL2). This is available for most cytometers.
2. For data acquisition, we recommend to create the following plots:
   (a) Forward scatter (FSC) versus side scatter (SSC) either as dot plot or density plot: This plot provides insights into the quality and integrity of your cells. Even if this plot is not essential for cell cycle analysis, it would provide additional information which might be interesting for putative trouble shooting.
   (b) FL2-A (FL2-area) versus FL2-W (FL2-width) either as dot plot or density plot: This plot allows the discrimination of single cells and cell aggregates which might alternate the results. A cell in the G2-phase has the double

DNA amount than in the G1-phase. Thus, when a cell in G2 passes the laser beam, it will provide a stronger fluorescence signal than a cell in G1. But two cells in G1 clumping together and passing the beam at the same time will be difficult to distinguish from a single cell in G2. But the analysis of the photomultiplier signal allows a discrimination of these cases. This will be accomplished by this plot. By adjusting the amplifiers (FL2-H, FL2-A, and FL2-W), cells in G1-phase should be located around channel 200 (linear presentation). Signals or events generated by cell aggregates should be excluded from the further analysis by gating the signals generated from single cells. Figure 1 provides the typical view of a cell cycle analysis. In some cases, you might prefer to define a region which contains the G1-phase and acquire a fixed number of cells in this region. In this case, the histograms have similar heights. However, this might be critical when some of your samples show a strong arrest in a cell cycle phase other than the G1-phase.

(c) FL2-A histogram plot: This histogram displays the different phases of the cell cycle. Only signals from single cells are shown in this histogram (they have been gated and selected in the FL2-A versus FL2-W plot). The statistic function of the FACS software allows a determination of the cell numbers in G0/G1-, S-, or G2-phase. When information about apoptotic cells is of interest, the subG1-fraction can be included into the gate in the FL2-A versus FL2-W plot. The subG1-fraction includes apoptotic cells. However, also cellular fragments including DNA other than from apoptotic origin (e.g. mechanically damaged cells or aneuploid cells) might contribute to the subG1-signals and this might cause misleading results. When a solely apoptosis assay is desired, we recommend other assays like Annexin V-PI staining (detection of phosphatidylserine at the cell surface) or TUNNEL staining.

3.2 Cell Viability Assay (MTT-Assay)

The viability assay can be performed in numerous formats from 96-well plates to T75 cell culture flasks. Here, we will describe an assay suitable for 12-well plates. According to the individual conditions, the assay can easily be scaled up or down. The assay presented here works with adherent cells only (Fig. 2).

1. Seed cells for the experiments. For most experiments the cells should not have reached confluence at the time point of performing the experiment. As a guideline, seed 2×10^5 cells per well (12-well plate) 24 h before you perform the MTT assay. When an additional treatment of the cells with drugs is intended, adjust the time.

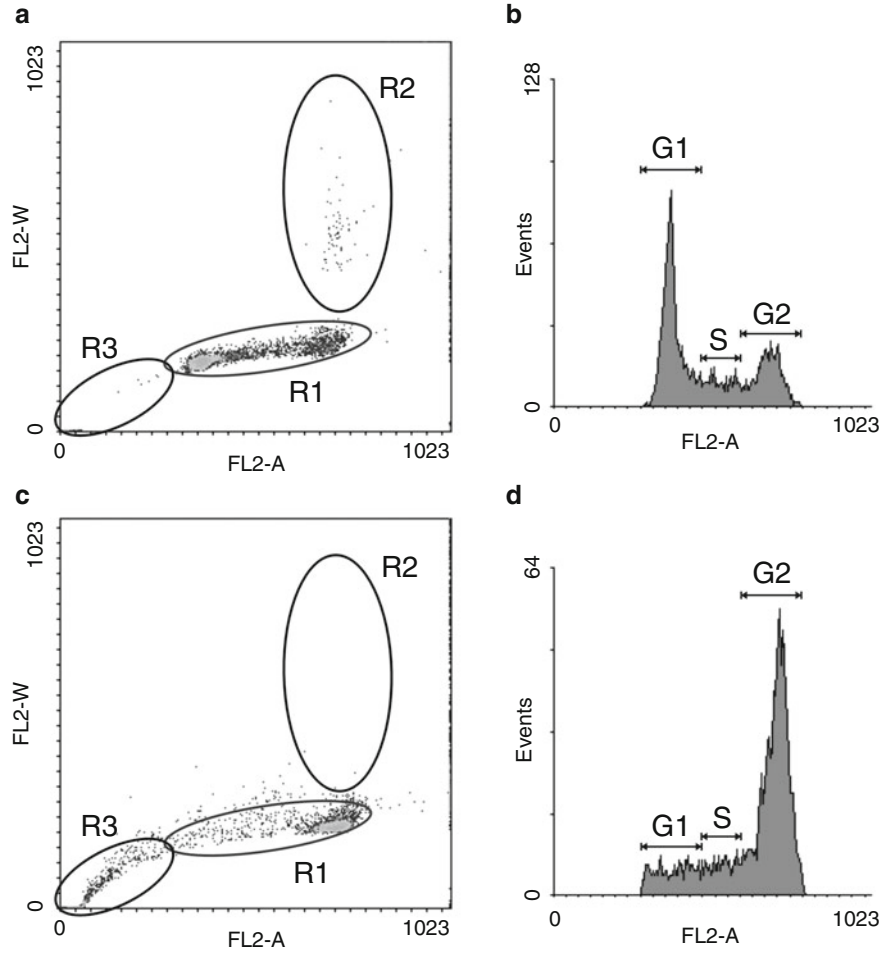


Fig. 1 Cell cycle analysis of a pancreatic cancer cell line. *Left column* (**a**; **c**): Density plots, *right column* (**b**; **d**): histograms. The *upper row* (**a**; **b**) demonstrates an average result of a cell cycle analysis from a pancreatic cancer cell line growing under standard cell culture conditions. (**a**) Signals obtained from single cells are located in region R1. This region is pivotal for the cell cycle analysis and should be gated for the further analysis. Cells in region R2 show cell aggregates (duplets), which causes artificial results and therefore should be excluded from the analysis. Signals in region R3 (subG1) represent apoptotic cells, apoptotic bodies, or otherwise broken cells. (**b**; **c**) Histogram of the cells gated in region R1 of the density plot. The *first peak* corresponds to the G0/G1-phase, while the second peak represents cells in the G2-phase. The valley between the peaks displays cells in the S-phase of the cell cycle. The *lower row* (**c**; **d**) shows a cell population treated with a strong inhibitor of the cell cycle, leading to a cell cycle arrest in the G2-phase. While signals obtained from cells in G1- and S-phase are substantially reduced, the peak corresponding to the G2-phase is significantly increased. In parallel, the subG1-fraction (R3) is increased indicating increased cell death

2. Start treatment of the cells at the time point applicable (e.g. 12 h before starting MTT assay).
3. Remove the medium from the cells by aspiration carefully not to disturb the cells.

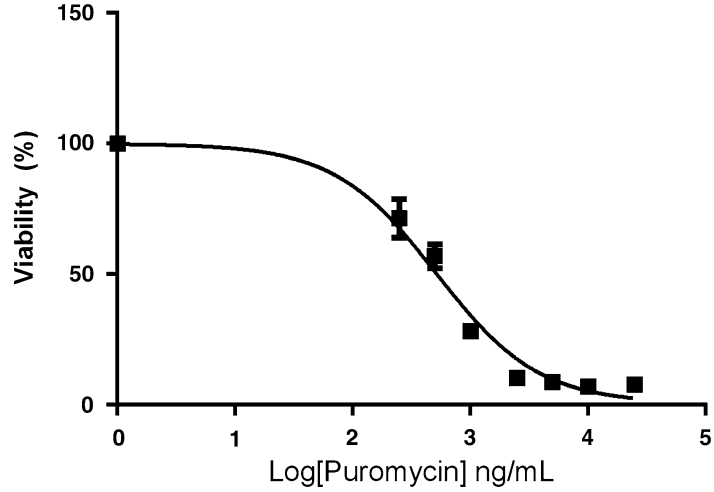


Fig. 2 Dose–response curve of A431 cell line treated with Puromycin. The IC_{50} is the concentration of the drug that induces a response half way between the maximal (*Top*) response and the maximally inhibited (*Bottom*) response. The IC_{50} was calculated using GraphPad Prism 6 software as follows: three replicate values are entered, in an XY data table, in the Y subcolumns as the percentage response of non-treated control. The percentage response can be easily calculated in excel using the following formula: =(A1/A$1)*100, where column A1 is the spectrophotometric value for the untreated control (100 % viability). Dragging the cell down will calculate the percent viability from the absorbance value of the next dose entered in column A2. From the data table, click "Analyze," under the XY analyses, choose nonlinear regression (curve fit), choose the panel of equations "Dose-response—Inhibition" and then choose the equation "*log*(*inhibitor*) vs. *response*" (*three parameter*). Under "Constrain," enter top to a constant equal to 100 and bottom to 0 and click OK. An analyses chart is depicted which indicates the IC_{50} values. For comparing different drugs, if the drug treatment does not define a minimum response, consider constraining the bottom to a shared value. The program then finds the best-fit parameters with greater confidence. To determine another value other than IC_{50}, for example IC_{90}, under choose, Dose-response—Special choose log(agonist) vs. response—Find ECanything. Constrain F to $F = 10$, the concentration that brings the curve down by 90 %, to 10 %

4. Add MTT solution. The cells should be covered completely. As a guideline, add 500 µl MTT solution per well (12-well plate). Add the solution not directly to the cells but to the border of the wells to prevent detachment of the cells.
5. Incubate at 37 °C in a cell culture incubator.
6. Check staining of the cells every 10 min. Depending on the cell viability dark violet crystals become visible. The incubation time should be adjusted so the wells do not become completely deep violet as the measurement become difficult in this case. Usually an incubation time between 30 and 60 min is sufficient.

7. Remove MTT solution completely from the cells by aspiration carefully not to disturb the cells.
8. Add 300 μl 2-propanol (abs.) to the wells.
9. Incubate on a shaker at 200 rpm at room temperature until all crystals are completely dissolved (usually 20 min).
10. Transfer 50–150 μl of the alcoholic solution (depending on its density) into a well of a new 96-well plate.
11. Measure optical density of the solution at 570 nm (with reference at 650 nm) in an optical plate reader.
12. When including appropriate controls (e.g. untreated cells), the effect of a specific treatment on the cell viability can be determined.

3.3 Cell Assay (Crystal Violet Stain)

This is an alternative assay to stain viable cells and to get an overview about viable cells in a culture vessel. This cell assay can be performed in numerous formats from 96-well plates to T75 cell culture flasks. Here, we will describe an assay suitable for 24-well plates. The assay presented here works with adherent cells only.

1. Seed cells for the experiments. For most experiments the cells should not have reached confluence at the time point of running the crystal violet stain. As a guideline, seed $0.5–1 \times 10^5$ cells per well (24-well plate) 24 h before you perform the crystal violet stain. When an additional treatment of the cells with drugs is intended, adjust the time.
2. Start treatment of the cells at the time point applicable (e.g. 12 h before crystal violet stain).
3. Remove the medium from the cells by aspiration carefully not to scratch the cells.
4. Add 100 μl of crystal violet solution (24-well plate). The cells should be covered completely. Add the solution not directly to the cells but to the border of the wells to prevent detachment of the cells.
5. Incubate for 15–30 min with slight shaking.
6. Aspirate the crystal violet solution completely.
7. Wash 3–4 times with aqua dest. by adding the water to the border of the wells to prevent detachment of the cells followed by aspiration.
8. Let wells dry at room temperature completely.
9. Add 400 μl methanol to each well.
10. Shake at room temperature until the dye is solved completely.
11. Transfer 100 μl of the alcoholic solution into a well of a new 96-well plate.

12. Determine optical density at 590 nm with an optical plate reader.
13. When including appropriate controls (e.g. untreated cells), the effect of a specific treatment on the amount of cells can be estimated.

3.4 An Easy Tumor Cell Invasion Assay

1. Seed 3×10^5 cells, preferentially a fibroblast cell line (e.g. normal diploid human skin fibroblasts), in 1 ml complete cell culture medium into 1 well of a 24-well cell culture vessel.
2. Let cells grow until they have reached confluence.
3. Remove the medium from the cells by aspiration carefully not to disturb the cells.
4. Add 1 ml PBS to each well. Add PBS not directly to the cells but to the border of the wells to prevent detachment of the cells.
5. Remove PBS from the cells by aspiration carefully not to disturb the cells.
6. Add 400 µl DMSO carefully to the cells to prevent cell detachment.
7. Incubate for 1 h at room temperature.
8. Remove the DMSO from the cells by aspiration carefully not to disturb the cells.
9. Add 1 ml PBS to each well. Add PBS not directly to the cells but to the border of the wells to prevent detachment of the cells.
10. Remove PBS from the cells by aspiration carefully not to scratch the cells.
11. Repeat **steps 9** and **10** two times.
12. Seed 2×10^4 invasive growing cells in 1 ml complete cell culture medium on top of the DMSO-treated cells.
13. Add drugs to the medium where applicable.
14. Incubate in a cell culture incubator for 24–48 h.
15. Remove the medium from the cells by aspiration carefully not to disturb the cells.
16. Add 1 ml PBS to each well. Add PBS not directly to the cells but to the border of the wells to prevent detachment of the cells.
17. Remove PBS from the cells by aspiration carefully not to disturb the cells.
18. Add PBS-trypan blue solution to the cells carefully. Prevent disturbing the cells.
19. Incubate for 20 min at room temperature.

20. Remove the staining solution from the cells by aspiration. Do not disturb the cells.
21. Add 1 ml PBS carefully to each well.
22. Remove PBS from the cells by aspiration. Do not disturb the cells.
23. Repeat **steps 21** and **22** two times.
24. Only the DMSO-fixed cells simulating the organic layer to be destroyed and invaded by the tumor cells become stained by trypan blue. Cell staining can be measured by microscopy or photometry:

    Microscopy: For microscopic analysis, take pictures of numerous areas of the culture vessel and measure stained or unstained areas using your adequate software.

    Photometry:

    (a) For photometric analysis, add 300 μl lysis buffer (e.g. PolyATtract GTC extraction buffer) to each well.

    (b) Incubate with shaking at room temperature for 30 min.

    (c) Depending on the viscosity of the solution, it might be necessary to shear the chromatin. This can be done either by ultrasound or by aspiration of the solution through a 20- or 24-gauge needle several times.

    (d) Transfer 100 μl of this solution into a well of a new 96-well plate and determine optical density at 590 nm.

    (e) A high optical density corresponds to a low tumor cell invasion. A sample without any tumor cell addition might be useful to determine the maximal optical density (no invasion at all). A sample without trypan blue staining will present the minimal density (maximal invasion, matrix completely destroyed).

3.5 Establishment of a Stable Transfected Cell Line with a Reporter Gene

3.5.1 Overview

A reporter gene with a strong viral vector permits not only longitudinal monitoring of cell trafficking and/or tumor growth but also monitoring of cell survival following engraftment into the mouse since constitutive expression of the reporter gene occurs only when the cell is alive. It is necessary to stably express the reporter gene in order to visualize all implanted cells, get an accurate read out of tumor growth and to ensure that the reporter expression is not diluted upon cell division.

As shown in Fig. 3, a molecular construct (plasmid) is obtained commercially that contains the reporter gene (DsREd2) and its promoter. The plasmid also contains a drug selection marker (for, e.g. neomycin-resistant gene, Neo). The plasmid vector can be used directly to transfect the cells; their entry into cells are facilitated by cationic lipid-based transfecting agents or by electroporation. After transfection, cells are subjected to a selection procedure

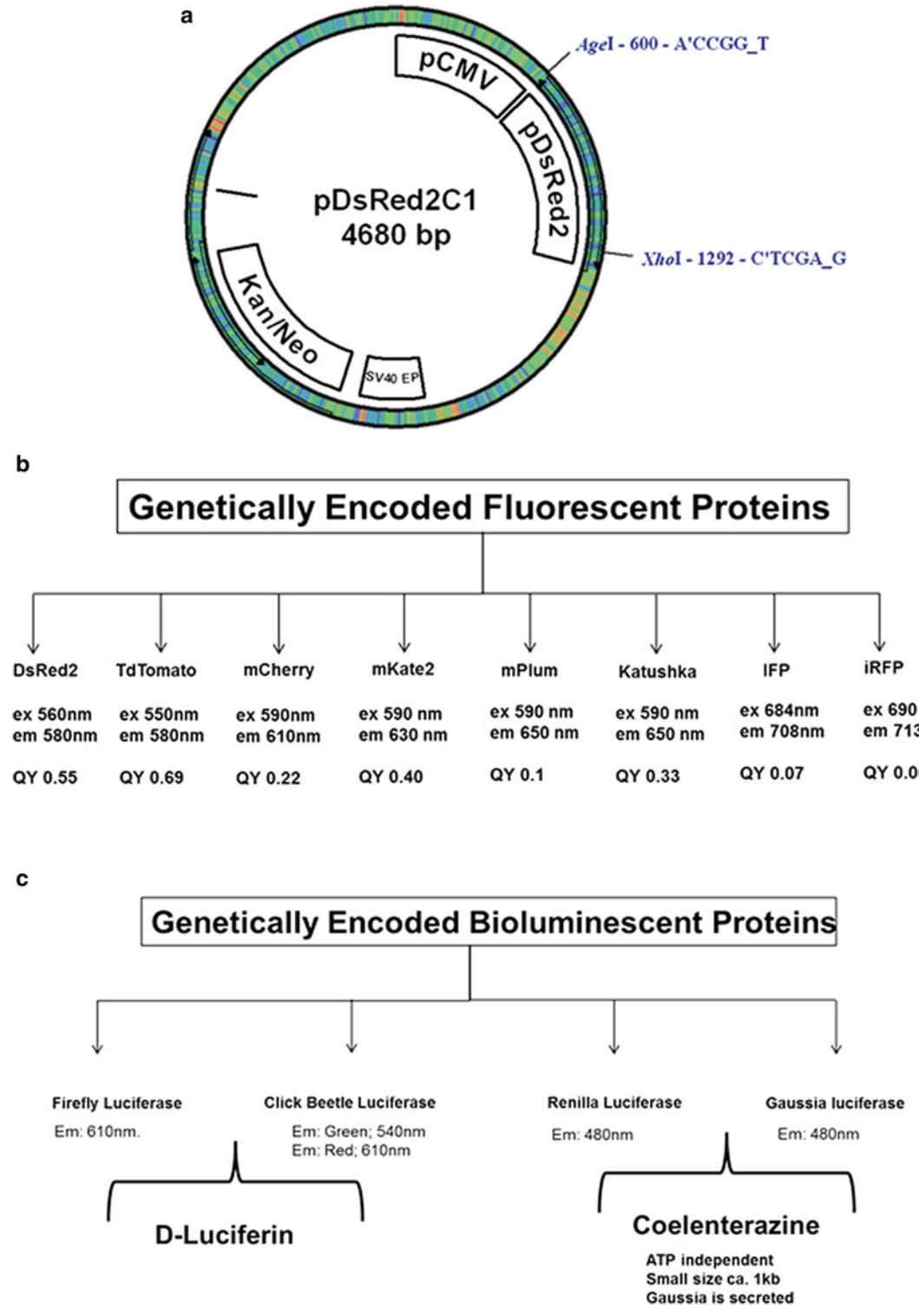


Fig. 3 (**a**) Representative fluorescent reporter gene plasmid is shown with the reporter gene (DsRED2, Clontech) under the control of a viral vector (CMV). A drug selection marker, Neo (neomycin resistant gene) with its promoter is also present. Kanamycin (Kan)-resistant gene is present for selection in bacteria. (**b**). Depiction of some commonly used fluorescent reporter genes which emit in the near infrared. Listed are the maximal excitation and emission wavelength for each reporter gene and their respective quantum yield. (**c**) Depiction of commonly used bioluminescent reporter genes. Click Beetle luciferase is available as a green-emitting and a red-emitting luciferase

in which only cells that have integrated the reporter gene in the genome will be selected. This is achieved by incubation of cells with a suitable antibiotic drug, cells not expressing the antibiotic-resistant gene (e.g. Neo) will be killed by the drug Neomycin. Clones of surviving cells will be collected and amplified for further characterization. The level of reporter gene expression may vary among the clones; therefore, the clone that stably expresses the highest level of the reporter gene is chosen.

3.5.2 Titrating G418 (Neomycin) to Establish a Kill Curve

Since each cell line has a different sensitivity to G418, you should determine the optimal concentration of drug for selection.

1. Split confluent cells 1:5 in 10 ml DMEM + 10 % FBS media.
2. Transfer 0.5 mL cell suspension into 24-well plate containing 500 μl of (media + G418). Use a G418 range starting at 50 μg/ml with the highest concentration at 1 mg/ml.
3. Use the lowest concentration of drug that begins to give massive cell death in 3 days and kills all the cells within 2 weeks.

3.5.3 Transfection and Drug Selection

1. Grow cells to ~80 % confluence in complete medium and transfect your plasmid with appropriate method, for example LipofectAmine Plus (Invitrogen). Include a mock transfection control which contains only the transfection reagent but no DNA.
2. After 24–48 h of transfection, cells are split to 1:10, 1:20, or 1:50 into (2) 15 cm plates containing 25 ml of DMEM + 10 % FBS + appropriate concentration of G418.
3. Observe cell growth in every 2–3 days and change medium with G418 every week or more if necessary. After 2–4 weeks, isolated colonies should begin to appear. At this time point, the mock transfected control cells should be dead and you can proceed with the cloning.

3.5.4 Isolation of Drug-Resistant Clones

This can be done in one of two ways:

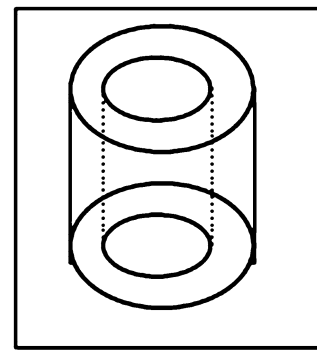
A. (adapted from Dario Neri's lab, Institute of Pharmaceutical Sciences ETHZ)

1. Take 4–6 (or more) 96-well plates and fill them with 80 μl of G418-supplemented medium. Detach cells using Accutase (gives no clumps!) and resuspend an aliquot of cells in a 15 ml falcon tube. Then make 1 or 2 serial dilutions (e.g. one in ten) and determine the cell concentration with the cytometer. Check your dilutions until you can only count 1–5 cells in the cytometer.
2. Prepare a sufficient amount of a 3 cells/20 μl (150 cells/ml) dilution

3. Let the cells grow in the wells for about 2 weeks. About 1 week after plating, you can also add 100 μl of fresh G418-medium to each well to reestablish high G418 concentration (it may get broken down with time) and prevent contamination.
4. Once "big" colonies (about 1/6 of the well diameter) are visible, also by eye when viewing the wells from the bottom of the plate, and the color of the medium starts to change (typically after about 2 weeks), colonies can be screened for expression by observing the plate under a fluorescent microscope.
5. Choose the highest expressing clones to transfer them to a 24-well plate. Let them grow and determine the protein-expression levels again once they have reached a suitable cell number. At the level of 24- or 6-well plates, it is also possible to make a first cryotube of each selected clone.
6. Grow several best expressing candidates up to the 75 cm^2 level.

B. (adapted from The Morimoto Lab, Northwestern University).

1. Melt 1 % low melting agarose by microwave and incubate in a water bath at 37 °C until it has cooled down below 37 °C. Mix 10 ml of 1 % agarose and 10 ml of 2× DMEM + 20 % FBS and pour into 15 cm plate. Leave the plate at RT for less than 30 min. Place the plate into a CO_2 incubator.
2. Mark large, healthy and well separated colonies and put colony separator (*see* figure) directly on surface of soft-agarose. Apply gentle pressure to top of separator to prevent movement.



3. Add 100 μl of accutase, pipeting several times gently to penetrate the soft-agarose surrounding the isolated colony and incubate at RT.
4. Place accutase-treated colony into one well of a 48-well plate containing 1 ml of DMEM + 10 % FBS + G418.
5. Split clones that reach ~80 % confluence into one 12-well plate and one 6-well plate. Use one plate for checking protein expression/induction. Check the protein expression of each

clone by its fluorescent intensity under a fluorescent microscope.

6. Split only expression-positive clones to 10 cm plate and store in liquid N_2.
7. For sub-cloning, re-plate about 100 cells per plate. For example, immediately after splitting, take 10–100 μl of culture and re-plate in 10 cm plate.
8. Repeat **steps 5–11** for up to six colonies to check for protein expression.

3.6 Image Analyses

Four- to six-week-old SCID beige mice were inoculated orthotopically with PancTuI cells, stably expressing DsRED2 fluorescent protein. The Berthold LB983 NightOwl optical imager (EG&G Berthold, Bad Wildbad, Germany) was used to monitor tumor growth by detection of DsRed2 fluorescence. The imager contains a Peltier cooled backlit CCD camera (2184 × 1472 pixels) housed within a light-tight enclosure. The excitation source is a ring light used for epi-illumination, mounted 12 cm above the mice. For excitation a 550 nm (10 nm) filter, for emission a 605 nm (55 nm) was used. Exposure time was 2 s. Using the WinLight 32 software (Berthold), fluorescent signals from the images were calculated by selecting a circular region of interest around the materials and integrating the signal from that area. Signals were expressed in ph/s. In addition, color-enhanced overlays of fluorescent images on photographic images were created using the WinLight software (Fig. 4).

3.7 Imaging NF-kB Activation and Apoptosis Simultaneously In Vivo

One approach to monitoring transcription factor activity is to clone the responsive element of the promoter upstream of a reporter gene. Expression of the reporter gene will then occur upon transcription factor activation and binding to the responsive element. However, in a longitudinal assay with drug treatment, cell

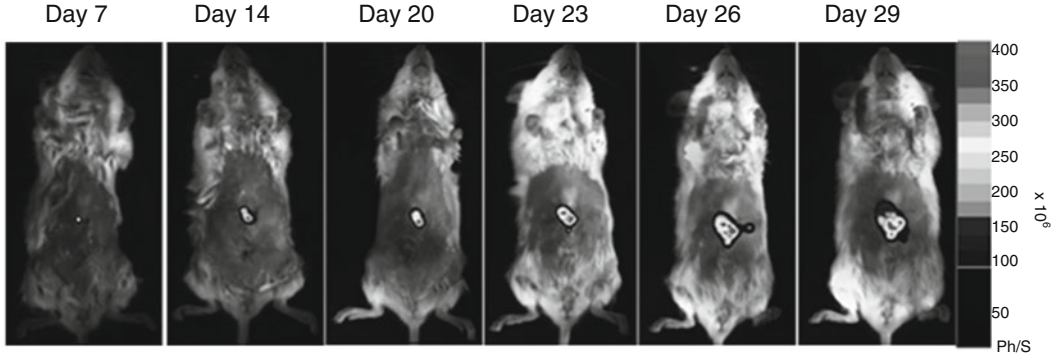


Fig. 4 Tumor growth rate of pancreatic adenocarcinoma cell line, PancTu1. Orthotopic implantation of 0.5×10^6 stably transfected DsRED2 cells in SCID mice was performed. DsREd2 fluorescence emission was collected at regular intervals (Tiwari et al., unpublished data)

numbers expressing the transgene will likely be changing due to, for example, treatment-induced cell death. In order to compensate for the decrease in cell number, a second reporter gene is necessary to report on cell number and viability. This reporter gene should be driven by a promoter that is constitutively active and signal will only be emitted from viable cells due to the requirement for ATP. Constitutively active promoters used in mammalian cells include the simian virus 40 early promoter, cytomegalovirus immediate-early promoter (CMV), human ubiquitin C promoter, human elongation factor 1a promoter, mouse phosphoglycerate kinase 1 promoter, and chicken β-actin promoter coupled with CMV early enhancer. A ratio of the transcription factor-induced signal and the constitutively active signal will then accurately indicate induction of transcription factor activity.

This approach was recently adopted by Mezzanotte et al. [13] by introducing three transgenes into a cell utilizing two different vectors. One vector encoding the constitutively active human phosphoglycerate kinase (PGK) promoter driving the expression of the CBG99 green-shifted luciferase ($\lambda_{max} = 540$ nm) was transduced and used to monitor tumor size. A second vector was transduced which encoded bidirectional promoters in which one promoter was cloned upstream and in opposite orientation to the other promoter, driving concurrent expression of two transgenes. An insulator sequence was placed between the two promoters to maintain independence of the individual domains from their surroundings. This bicistronic gene expression system encoded the red shifted luciferase PpyRE9 ($\lambda_{max} = 620$ nm) under the control of the NF-kB promoter responsive elements and the Gaussia luciferase ($\lambda_{max} = 480$ nm) under the control of the PGK promoter. Gaussia luciferase unlike firefly luciferase requires coelenterazine as a substrate and is exported to the transmembrane. This latter property allows for the in vitro selection of cells stably transduced with the bicistronic vector by utilizing an antibody directed against Gaussia and performing flow cytometric sorting. The induction of NF-kB activity in response to drug treatment was determined by spectral unmixing of the red and green light and determining the red/green signal ratio (Fig. 5).

Since bioluminescence systems depend on the delivery of a specific substrate to the reporter expressing cells, it is possible to modify this substrate to keep it "silent" until activated by a protease. Such a system has been established which enables imaging of caspase-3/7 activity in cells undergoing apoptosis. In this system, the DEVD tetrapeptide sequence recognized by caspase-3 or -7 is incorporated into aminoluciferin to create a firefly luciferase prosubstrate [14]. Upon activation of caspase-3 or -7, the DEVD peptide is cleaved and the aminoluciferin is now available as a substrate for luciferase to generate measurable luminescent signal. In the study described above by Mezzanotte et al., this assay was utilized to link NF-kB promoter activity to induction of caspase-mediated cellular apoptosis.

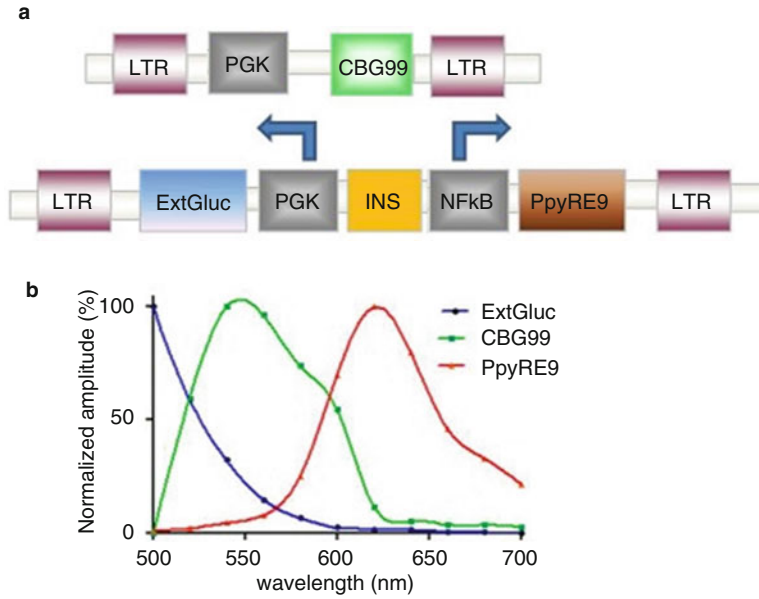


Fig. 5 Generation of a triple-color reporter gene strategy. (**a**) Schematic representation of the lentivirus constructs used for generating expression of three reporter genes in a single cell. (**b**) Graph representing the emission spectrum of the three luciferases used for the triple-color strategy. Gaussia Luciferase (ExtGluc) in *blue*, Green-emitting luciferase (CBG99) in *green*, and red-emitting luciferase (PpyRE9) in *red*. Taken from Mezzanotte L, An N, Mol IM, Löwik CWGM, Kaijzel EL (2014) A New Multicolor Bioluminescence Imaging Platform to Investigate NF-kB Activity and Apoptosis in Human Breast Cancer Cells. PLoS One 9(1): e85550. doi:10.1371/journal.pone.0085550

4 Note

1. For FACS analyses (Subheading 3.1) cells need to be in a single cell suspension as much as possible. Vortexing pellets while adding solutions is important to minimize cell clumping. Fixation in ice-cold alcohol may also help.

   If there is obvious clumping, pass cells through nylon mesh before analyses on the FACS and vortex each sample just prior to running it through the FACs. If the coefficient of variation (CV) of the G1 peak is high (i.e. 10 % or more), estimation of the fraction of cells in different phases of the cell cycle is difficult. A broad CV may be due to inadequate fixation procedures or concentration of stain that is either too low or too high, leading, respectively, to non-stoichiometric or nonspecific binding.

2. MTT enters the cell by endocytosis and is converted by dehydrogenases to insoluble purple formazan crystals which are read spectrophotometrically after organic solvents have dissolved the crystal. The formation of insoluble formazan

(Subheading 3.2) has been proven to be unstable at room temperature. Therefore, it is recommended to standardize the time before the plates are read in the spectrophotometer and also to store the plates at 4 °C if several plates need to be read at the same time. The test is less effective if cells have been cultured in the same media that has supported growth for a few day, which leads to underestimation of control and untreated samples.

3. The invasion assay (Subheading 3.4) relies on the invasion and/or digestion of the fibroblast layer by tumor cells which lead to areas with mostly unstained cells. The dimensions of these unstained regions can be estimated semi-quantitatively and correspond to the invasive potential of the tumor cells. Therefore it is paramount that the confluent fibroblast layer are not detached during the procedure of adding PBS or seeding of invasive cells. Particular care needs to be taken following DMSO treatment since the detachment of the fibroblast layer of cells can occur very easily.
4. In the establishment of a stably transfected cell line (Subheading 3.5), circular or linearized plasmid can be used for transfection. It is recommended to use linearized plasmid and to cut the plasmid at a unique site which does not disrupt the coding sequence of the reporter gene or drug resistance. Stable transfection with a circular plasmid may result in recombination into the genome at a random position within the plasmid and can lead to antibiotic-resistant clones that do not express the reporter gene. A screening of the transfection method and reagents (e.g. Nucleofection, GT porator) may be initially required to obtain high transfection efficiency for the cell line of interest. If the method of picking colonies described in Subheading 3.5.4 is not successful, a simpler method is to wait until sizeable colonies appear and then mark the position of individual colonies on the underside of the dish with a marker pen with the help of an inverted microscope. Then, in the sterile tissue culture hood, wash the plate with a little trypsin and then add fresh trypsin just enough to cover the plate. To obtain individual colonies, use a 1 ml pipetteman to suck up the colony using the marking on the underside of the plate to determine the localization of that colony.
5. A disadvantage with optical imaging is the problem of depth sensitivity. Thus, for two-dimensional planar imaging, as described in Subheading 3.6, the emission signal detected is dependent on the proximity of the reporter gene to the surface. In cases where tumors grow close to the skin, the tumor may be perceived as being larger than those growing deeper inside the animal. Hoffman and colleagues have demonstrated that for tumors less than 1500 mm^3, measurement of the

fluorescent tumor area detected by planar optical imaging correlated strongly with the tumor volume measured by MRI [15]. Therefore, measurement of fluorescent tumor area may be a better indication of tumor growth and drug efficacy.

References

1. Ammerpohl O et al (2007) Complementary effects of HDAC inhibitor 4-PB on gap junction communication and cellular export mechanisms support restoration of chemosensitivity of PDAC cells. Br J Cancer 96(1):73–81
2. Orntoft TF, Petersen SE, Wolf H (1988) Dual-parameter flow cytometry of transitional cell carcinomas. Quantitation of DNA content and binding of carbohydrate ligands in cellular subpopulations. Cancer 61(5):963–970
3. Asklund T et al (2004) Histone deacetylase inhibitor 4-phenylbutyrate modulates glial fibrillary acidic protein and connexin 43 expression, and enhances gap-junction communication, in human glioblastoma cells. Eur J Cancer 40(7):1073–1081
4. Svechnikova I, Ammerpohl O, Ekstrom TJ (2007) p21waf1/Cip1 partially mediates apoptosis in hepatocellular carcinoma cells. Biochem Biophys Res Commun 354(2):466–471
5. Ammerpohl O et al (2004) HDACi phenylbutyrate increases bystander killing of HSV-tk transfected glioma cells. Biochem Biophys Res Commun 324(1):8–14
6. Appelskog IB et al (2004) Histone deacetylase inhibitor 4-phenylbutyrate suppresses GAPDH mRNA expression in glioma cells. Int J Oncol 24(6):1419–1425
7. Tolboom TC, Huizinga TW (2007) In vitro matrigel fibroblast invasion assay. Methods Mol Med 135:413–421
8. Casey RC et al (2003) Establishment of an in vitro assay to measure the invasion of ovarian carcinoma cells through mesothelial cell monolayers. Clin Exp Metastasis 20(4):343–356
9. Trauzold A et al (2005) CD95 and TRAF2 promote invasiveness of pancreatic cancer cells. FASEB J 19(6):620–622
10. Frangioni JV (2003) In vivo near-infrared fluorescence imaging. Curr Opin Chem Biol 7(5):626–634
11. Weissleder R et al (1999) In vivo imaging of tumors with protease-activated near-infrared fluorescent probes. Nat Biotechnol 17(4):375–378
12. Villalobos V, Naik S, Piwnica-Worms D (2007) Current state of imaging protein-protein interactions in vivo with genetically encoded reporters. Annu Rev Biomed Eng 9:321–349
13. Mezzanotte L et al (2014) A new multicolor bioluminescence imaging platform to investigate NF-kappaB activity and apoptosis in human breast cancer cells. PLoS One 9(1), e85550
14. O'Brien MA et al (2005) Homogeneous, bioluminescent protease assays: caspase-3 as a model. J Biomol Screen 10(2):137–148
15. Bouvet M, Spernyak J, Katz MH, Mazurchuk RV, Takimoto S, Bernacki R, Rustum YM, Moossa AR, Hoffman RM (2005) High correlation of whole-body red fluorescent protein imaging and magnetic resonance imaging on an orthotopic model of pancreatic cancer. Cancer Res 65(21):9829–9833

Chapter 11

Application of Proteomics in Cancer Biomarker Discovery: GeLC-MS/MS

Pedro R. Cutillas and Tatjana Crnogorac-Jurcevic

Abstract

Proteomic approaches are being increasingly applied to study multiple facets of healthy and diseased processes. In particular, the application of global proteome profiling in the field of oncology is already starting to shape the diagnostic, prognostic, monitoring, and therapeutic approaches. At the heart of such approaches lies a quest for clinically relevant biomarkers, particularly arising from global analyses of body fluids, as, in major part, they represent easily accessible and noninvasive matrices. A detailed protocol of one of the popular approaches for global proteome profiling, SDS-PAGE-liquid chromatography-tandem mass spectrometry or GeLC-MS/MS, and its application for biomarker discovery in urine is provided here.

Key words Proteomics, GeLC-MS/MS, Biomarker discovery, Body fluids

1 Introduction

Biomarkers comprise any measurable biological indicators of a normal biological or pathological process or a pharmacological response to a therapeutic intervention [1]. With the advent of global proteomics profiling techniques, unbiased biomarker discovery based on changes in protein expression levels between analysed healthy and diseased (cancer) samples became possible. While the concentration of any candidate biomarker is the highest in cancer tissues, body fluids, in particular proximal fluids that bathe the diseased tissue/organ, while being the "second" best, are still rich sources of potentially useful biomarkers.

Most commonly used body fluids are blood and urine. Blood is the most sampled but also the most complex proteome [2, 3], with currently over 10,500 proteins described (http://www.plasmaproteomedatabase.org). It has an enormous dynamic range spanning over 12 orders of magnitude, and with 22 abundant proteins representing almost 99 % of its protein content, leaves only a small compartment as a source of potentially useful biomarkers [4, 5].

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_11,

In addition, both plasma and serum proteomes are very sensitive to sample handling, processing, and analytical procedures.

Urine presents an attractive alternative; in fact, the first bona fide protein biomarker still utilised today was described in urine in 1848 by Henry Bence-Jones [6]. Healthy urine sample comprises at least 2000-3000 proteins (depending on the peptide identification criteria) [7, 8], of which at least a third are of a systemic origin [9]. Urine is much less complex than serum or plasma, with more stable proteome; however, it contains low protein content and large amounts of salts and other interfering compounds [10, 11]. Nevertheless, with robust experimental design, standardised collection and storage, as well as sample processing protocols in place (for more detailed review *see* [12]), combined with sensitive global proteome profiling and subsequent thorough validation of candidate biomarkers, both body fluids could prove to be a source of multitude of useful candidate biomarkers, and therefore of immense clinical benefit.

In urine, biomarker discovery may be achieved using direct liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis of unfractionated protein or peptide samples [13, 14]. However, single LC-MS/MS may not always offer the depth of analysis required to detect low abundant proteins present in complex protein mixtures, although recent advances in MS instrumentation make this less of a problem. Therefore, in order to increase the coverage of the analysed proteomes, several methods for fractionating proteomes prior to LC-MS/MS have been developed and applied to the analysis of the urinary proteome. One of such methods for global proteome profiling widely used for biomarker discovery is sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) prefractionation followed by LC-MS/MS or GeLC-MS/MS. In GeLC-MS/MS, complex protein mixtures are first separated by SDS-PAGE, and gel sections excised and digested prior to subsequent analysis by nanocapillary LC-MS/MS. In-gel digestion of proteins compatible with MS analysis was developed in the 1990s [15], and the technique have undergone some optimisation with the publication of more robust protocols [16]. The abundances of protein-derived peptides may be assessed by LC-MS/MS using the spectral peptide counts (i.e., the times a peptide derived from a given protein was sequenced in the LC-MS/MS run), which roughly correlate with protein amounts in the original sample and thus provide semi-quantitative readout of protein amounts. Contrary to the spectral counting approach, when combined with label-free quantification based on ion currents, GeLC-MS/MS can be used to quantify proteins and their sites of modification with very good accuracy [17, 18]. This chapter describes the materials and detailed steps in GeLC-MS/MS procedure, an outline of which is illustrated in Fig. 1.

The appeal of the technique lies in combining a technically straightforward 1D electrophoretic protein separation which

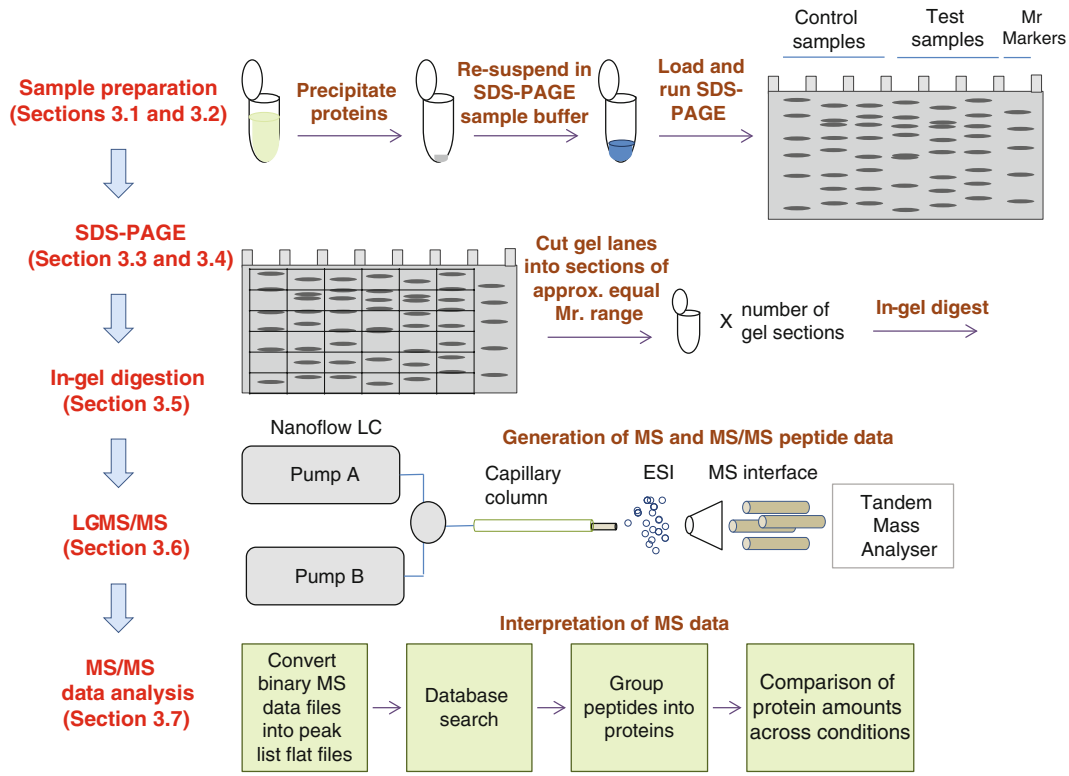


Fig. 1 Generalized workflow for the analysis of urinary proteins by GeLC-MS/MS. The analytical steps represented diagrammatically in the figure are explained in detail in the named sections of the main text

allows representation of membranous and large proteins that were traditionally not captured in 2D PAGE, with analytical sensitivity of protein sequencing, and has allowed identification of diverse core proteomes, including these of both freshly frozen and formalin-fixed tissues [19], cell lines [20], and gastroduodenal and pancreatic fluids [21]. Our laboratories have used GeLC-MS/MS to investigate protein expression in the urine of renal Fanconi syndrome patients [22], to quantify phosphoproteins regulated in cancer cells [17], to profile the proteomes of mouse tissues [18], to discover proteins that interact with PI3K [23], and more recently to detect biomarkers in urine specimens collected from patients with pancreatic adenocarcinoma [24].

2 Materials

Solutions for electrophoresis should be prepared using deionized water to achieve a resistivity of 18 megaohms-cm. Reagents for in-gel digestion, liquid chromatography, and mass spectrometry should be of LC-MS grade.

2.1 Protein Precipitation Reagents

1. LC-MS-grade methanol (Optigrade, LGC, Teddington, UK).
2. LC-MS-grade chloroform (Optigrade, LGC, Teddington, UK).

2.2 Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis

1. Sample Buffer (Laemmli Buffer): 40 mM Tric–HCl pH 6.8, 2 % (w/v) SDS, 4 % (v/v) glycerol, 2 mM DTT, 0.01 % (w/v) bromophenol blue.
2. Pre-casted 10 % acrylamide SDS-PAGE gels.
3. Running buffer: 0.025 M Tris–HCl pH 8.3, 0.192 M glycine, 0.1 % SDS.

2.3 Colloidal Coomassie Blue Staining

1. Fixing solution: 50 % (v/v) ethanol, 2 % (v/v) phosphoric acid (from an 85 % stock dissolve 2.35 ml in 97.6 ml ddH_2O).
2. Staining solution: 34 % (v/v) methanol, 17 % (w/v) $(NH_4)_2SO_4$, 3 % (v/v) phosphoric acid (from a 85 % stock dissolve 3.53 ml in 96.5 ml ddH_2O).
3. Solid Coomassie Blue G-250 (BioRad, Hemel Hempstead, UK).
4. ddH_2O.

2.4 In-Gel Digestion

1. Acetonitrile (ACN Optigrade, LGC, Teddington, UK).
2. Formic acid (99 % pure, LC-MS grade, Fisher Scientific, UK).
3. Trifluoroacetic acid (Optigrade, LGC, Teddington, UK).
4. Ammonium bicarbonate (Sigma).
5. Sequence grade trypsin (Promega).

2.5 Liquid Chromatography-Tandem Mass Spectrometry

1. Mobile phase A: 0.1 % (v/v) formic acid (Optigrade LC-MS grade) in water (LC-MS grade).
2. Mobile phase B: 0.1 % (v/v) formic acid (Optigrade LC-MS grade) in ACN (LC-MS grade).
3. Quality control standards: enolase protein digest (Enolase mass-prep standard, Waters, Manchester, UK).
4. Resuspension solution: 0.1 % formic acid in LC-MS grade water.

2.6 Equipment

1. Bench-top centrifuge.
2. Vortex.
3. Gel electrophoresis equipment.
4. Vacuum concentrator (Speed-Vac).
5. High-resolution LC-MS/MS system equipped with nanoflow gradient pumps and nanoelectrospray ion source.

3 Methods

The methods described below, and illustrated diagrammatically in Fig. 1, describe a generalized workflow for GeLC-MS/MS analysis of urinary proteins and, with minor modifications, should work in laboratories with a range of different analytical and bioinformatics platforms. The methods are also applicable to the analysis of proteins present in other biofluids and cell lysates.

3.1 Sample Preparation

1. Thaw 1 ml urine samples on ice.
2. Centrifuge defrosted samples at maximal r.p.m. using a bench-top centrifuge for 5 min.
3. Remove 400 μl of supernatant into a fresh low bind tube.

3.2 Chloroform Precipitation of Proteins (See Note 1)

1. Add 400 μl methanol and 100 μl chloroform to the 400 μl centrifuged urine sample.
2. Centrifuge on a bench-top centrifuge at maximal r.p.m. for 1 min. Protein should precipitate at the interface. Pull upper phase.
3. Add 300 μl methanol, vortex, and centrifuge on a bench-top centrifuge at maximal r.p.m. for 2 min.
4. Protein will have precipitated at the bottom of the tube. Discard supernatant and allow pellet to dry by placing the opened tube in a fume hood.

3.3 SDS-PAGE

1. Add 50 μl of SDS-PAGE sample buffer to the dried pellet, vortex and incubate at 90 °C for 2 min. Vortex again and let solution cool down before proceeding.
2. Load sample in a pre-cast 10 % SDS-PAGE gel.
3. Load at least one lane with protein markers.
4. Apply a fixed current to the SDS-PAGE gel and run the electrophoretic system until the solvent front reaches the bottom of the gel. The time required for separation depends on the voltage applied and the percentage of acrylamide used. A voltage of 100 V applied to a 10 % separation gel requires about 1 h of separation.

3.4 Visualization of SDS-PAGE Protein Bands

This protocol is adapted from Neuhoff et al. [25].

1. Remove the SDS-PAGE gel from the electrophoretic system place and place it in a plastic container with sufficient fixing solution to cover gel. Incubate from 3 h to overnight.
2. Wash the gel by discarding the fixing solution and adding sufficient ddH_2O to cover gel pieces. Leave gel shaking gently in a laboratory rocker for 30 min. Repeat this step a total of three times, replacing ddH_2O with fresh one each time.

3. Remove ddH_2O and add sufficient staining solution to cover gel. Incubate with shaking for 1 h.
4. Add 0.5 g/l crushed solid Coomassie Blue G-250 to the gel in staining solution. Leave shaking gently for 1–2 days or until bands of the required intensity are visible on the gel. Seal container to avoid evaporation.
5. Rinse gels in ddH_2O and scan (*see* **Note 2**).

3.5 In-Gel Tryptic Digestion

1. Cut gel lanes into sections of approximate equal molecular weight (*see* **Note 3**).

   The following is an adaptation of the Shevchenco in-gel digestion protocol [21]. In each case add sufficient solution to cover gel pieces.
2. Destain, wash, and dehydrate gel pieces by: (a) washing three times in 50 % CAN; and (b) washing once with 200 µl of 100 % ACN (aspirate as much liquid as possible after final wash as this will facilitate drying in **step 3**). Wait 3–4 min between washes with occasional agitation.
3. Dry gel pieces in a speed-vac for 20–30 min.
4. Reduce cysteines by adding 10 mM DTT solution; incubate for 45 min at 50 °C.
5. Alkylate reduces cysteines by removing DTT and adding 50 mM IAM solution; incubate for 60 min at room temperature in the dark.
6. Remove IAM solution wash gel pieces twice with 50 % ACN.
7. Dry gel pieces in a speed-vac for 40–45 min or until all ACN has evaporated and gel pieces are dry.
8. Add enough trypsin solution to cover the dried gel pieces, wait for 5 min to allow gel pieces to absorb, some of the solution. Add more trypsin solution as required so that gels are saturated with trypsin. Incubate overnight at 37 °C.
9. On the following morning centrifuge tubes (1 min at maximal r.p.m) to spin liquid down from lids.
10. Extract peptides from gel pieces by adding 100 µl extraction solution to the tubes. Transfer supernatant to a new 1.5 ml low-bind microfuge tube. Add sufficient extraction solution to cover gel pieces, vortex, spin down (1 min at maximal r.p.m), and transfer supernatant to the same microfuge tube used to collect first elution. Repeat this extraction step twice.
11. Concentrate samples by drying eluted peptides in a speed-vac (*see* **Note 4**).
12. Store dry peptide extract at –20 °C until ready to be analysed by LC-MS/MS.

3.6 LC-MS/MS

1. Resuspend dried peptides in 12 μl of 0.1 % TFA.
2. Load 4 μl of dissolved peptides into an LC-MS/MS system equipped with nanoflow gradient pumps and a nano-electrospray ion source. Run a LC gradient separation on a reversed phase nanoflow column using solvents A and B as mobile phases. A suitable gradient consists of ramping the percentage of B from 2 to 35 % in 60 min. As peptides elute from the column these will be ionized by nanoflow ESI and sampled in the mass spectrometer (*see* **Note 5**).

3.7 Protein Identification and Data Analysis

This is a generic protocol for identifying peptides and proteins from tandem mass spectrometry data. Several software packages, search engines and databases can be used for this purpose and although the following lists the tools used in our laboratory for this purpose to illustrate the approach, other software may also be used.

1. Convert mass spectrometry raw data into peak lists in the Mascot Generic Format (.mgf) using a peak piking software such as Mascot Distiller (*see* **Note 6**).
2. Submit the .mgf files to the Mascot Search Engine. Select the latest version of Swissprot protein database restricted to human entries. Choose carbamidomethyl of cysteines as fixed modification and oxidation of methionines and pyro-glutamate from glutamine as variable modifications. Adjust the search parameters to match the mass accuracy of the mass spectrometer used for analysis.
3. Parse data from Mascot result files into lists of peptide spectral matches (PSMs). Adjusts the thresholds of tolerance for accepting significant peptides matches so that false discovery rate is less than 1 %. This can be done by comparing results of searches against a decoy database (*see* **Note 7**).
4. Combine PSMs into protein groups (*see* **Note 8**). A simple comparison of protein expression across conditions can be done by comparing the number of PSMs for given proteins between control and test samples (*see* **Note 9**).

4 Notes

1. The concentration of protein in urine is low and this biofluid contains salts that interfere with electrophoresis. Therefore, a protein enrichment step that removes salts is recommended before SDS-PAGE. In addition to chloroform precipitation, other solvents, such as TCA and acetone, have been used to precipitate urinary proteins. Alternatives to precipitation include ultrafiltration and solid-phase extraction.

2. If the gel appears blue, and for applications where a well-defined visualization of protein bands is required, further destaining can be performed in 35 % MeOH/7.5 % acetic acid.
3. The number of gel sections to which a lane may be cut depends on the desired analytical depth. The larger the number of gel sections cut, the larger the number of proteins that will be identified, at the expense of mass spectrometry time (and expense) required for completing the experiment.
4. Depending on the type of speed-vac used this will probably be an overnight drying step.
5. Several types of mass spectrometers and liquid chromatographs can be used for this purpose. Those based on Orbitrap or high resolution Q-TOF are preferred because their high mass accuracy allows unambiguous protein identification based on tandem mass spectrometry data. This protocol assumes that the LC-MS/MS system will have been tuned and calibrated and quality control samples will have been used to assess the performance of the system before running samples.
6. Alternative tools to Mascot Distiller include vendor specific software such as Proteome Discoverer and MassLynx for ThermoFisher and Waters instruments, respectively. Open-source software for this purpose include MSconvert (http://proteowizard.sourceforge.net/tools.shtml) and MaxQuant (http://www.maxquant.org).
7. The Mascot Search engine performs this step automatically if the option is selected. Alternatives to the Mascot Search Engine include Open MS, Protein Prospector, MaxQuant, Sequest, ProteoWizard, Proteome Discoverer among others.
8. This laboratory uses a script written in-house for this purpose but commercial software (such as Scaffold) may be used instead.
9. Comparing PSMs across samples, the so-called spectral count approach, provides a qualitative means to analyse LC-MS/MS data. More accurate methods for protein quantification include isotope labeling methods introduced chemically or metabolically and label-free methods using peak intensities as the quantitative readout.

References

1. Anonymous (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin Pharmacol Ther 69:89–95
2. Anderson NL, Anderson NG (2002) The human plasma proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 1:845–867
3. Omenn GS, States DJ, Adamski M et al (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analyti-

cal groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics 5:3226–3245

4. Hortin GL, Sviridov D (2010) The dynamic range problem in the analysis of the plasma proteome. J Proteomics 73:629–636
5. Tirumalai RS, Chan KC, Prieto DA et al (2003) Characterization of the low molecular weight human serum proteome. Mol Cell Proteomics 2:1096–1103
6. Bence Jones H (1848) On a new substance occurring in the urine of a patient with molli-ties ossium. Phil Trans R Soc 138:55–62
7. Kentsis A, Monigatti F, Dorff K et al (2009) Urine proteomics for profiling of human disease using high accuracy mass spectrometry. Proteomics Clin Appl 3:1052–1061
8. Santucci L, Candiano G, Petretto A et al (2014) From hundreds to thousands: widening the normal human Urinome (1). J Proteomics 112C:53–62
9. Jia L, Zhang L, Shao C et al (2009) An attempt to understand kidney's protein handling function by comparing plasma and urine proteomes. PLoS One 4:e5146
10. Schaub S, Wilkins J, Weiler T et al (2004) Urine protein profiling with surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry. Kidney Int 65:323–332
11. Thongboonkerd V (2007) Practical points in urinary proteomics. J Proteome Res 6: 3881–3890
12. Alrawashdeh WaC-J T (2011) Biomarker discovery in biological fluids. Springer
13. Cutillas PR, Norden AG, Cramer R et al (2003) Detection and analysis of urinary peptides by on-line liquid chromatography and mass spectrometry: application to patients with renal Fanconi syndrome. Clin Sci 104: 483–490
14. Quintana LF, Campistol JM, Alcolea MP et al (2009) Application of label-free quantitative peptidomics for the identification of urinary biomarkers of kidney chronic allograft dysfunction. Mol Cell Proteomics 8:1658–1673
15. Shevchenko A, Wilm M, Vorm O et al (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Anal Chem 68:850–858
16. Shevchenko A, Tomas H, Havlis J et al (2006) In-gel digestion for mass spectrometric characterization of proteins and proteomes. Nat Protoc 1:2856–2860
17. Cutillas PR, Geering B, Waterfield MD et al (2005) Quantification of gel-separated proteins and their phosphorylation sites by LC-MS using unlabeled internal standards: analysis of phosphoprotein dynamics in a B cell lymphoma cell line. Mol Cell Proteomics 4:1038–1051
18. Cutillas PR, Vanhaesebroeck B (2007) Quantitative profile of five murine core proteomes using label-free functional proteomics. Mol Cell Proteomics 6:1560–1573
19. Tanca A, Pagnozzi D, Burrai GP et al (2012) Comparability of differential proteomics data generated from paired archival fresh-frozen and formalin-fixed samples by GeLC-MS/MS and spectral counting. J Proteomics 77: 561–576
20. Schirle M, Heurtier MA, Kuster B (2003) Profiling core proteomes of human cell lines by one-dimensional PAGE and liquid chromatography-tandem mass spectrometry. Mol Cell Proteomics 2:1297–1305
21. Paulo JA, Kadiyala V, Banks PA et al (2012) Mass spectrometry-based (GeLC-MS/MS) comparative proteomic analysis of endoscopically (ePFT) collected pancreatic and gastroduodenal fluids. Clin Transl Gastroenterol 3:e14
22. Cutillas PR, Chalkley RJ, Hansen KC et al (2004) The urinary proteome in Fanconi syndrome implies specificity in the reabsorption of proteins by renal proximal tubule cells. Am J Physiol Renal Physiol 287:F353–F364
23. Beltran L, Chaussade C, Vanhaesebroeck B et al (2011) Calpain interacts with class IA phosphoinositide 3-kinases regulating their stability and signaling activity. Proc Natl Acad Sci U S A 108:16217–16222
24. Radon TP, Massat NJ, Jones R et al (2015) Identification of a three-biomarker panel in urine for early detection of pancreatic adenocarcinoma. Clin Cancer Res (in press)
25. Neuhoff V, Stamm R, Pardowitz I et al (1990) Essential problems in quantification of proteins following colloidal staining with coomassie brilliant blue dyes in polyacrylamide gels, and their solution. Electrophoresis 11:101–117

Chapter 12

Meta-analysis of Cancer Gene Profiling Data

Janine Roy, Christof Winter, and Michael Schroeder

Abstract

The simultaneous measurement of thousands of genes gives the opportunity to personalize and improve cancer therapy. In addition, the integration of meta-data such as protein-protein interaction (PPI) information into the analyses helps in the identification and prioritization of genes from these screens.

Here, we describe a computational approach that identifies genes prognostic for outcome by combining gene profiling data from any source with a network of known relationships between genes.

Key words Network-based, Outcome prediction, Gene expression, PageRank, Cancer biomarker

1 Introduction

Outcome prediction tries to define the future state of a patient based on its current disease state.

The main goal of gene expression-based cancer outcome prediction is to improve the diagnosis and the treatment of cancer through more accurate disease classification and patient stratification.

Outcome-based cancer research spans from discovery research to validation and into clinical utility including:

1. Identification of cancer biomarkers and therapeutic targets
2. Elucidation of the mechanisms of cancer pathways
3. Validation of therapeutic targets and cancer biomarkers
4. Clinical classification and stratification

Although thousands of genes can be profiled simultaneously by high-density oligonucleotide arrays (DNA microarrays) or next-generation sequencing (NGS), it is often desirable for signatures used for outcome prediction to comprise only few genes.

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_12,

To this end, genes need to be ranked according to their influence on the patient's outcome. However, standard techniques like fold change, *t*-test, and other prioritization methods are strongly influenced by noise in the data.

The problem with the noisy expression data can be addressed by meta-analysis using data integration. This general principle was summarized recently [1] and applied to a breast cancer data set resulting in an improvement of 8 % in prediction accuracy [2].

Other studies followed these lines and showed that network information efficiently improves outcome prediction and reduces noise in microarray [3–6] or sequencing experiments [7–9]. The basic principle of all network-based analyses is the usage of functional association networks. Edges in these networks are based on the integration of different sources such as high-throughput protein interaction experiments (like yeast-2-hybrid screens), evidence of physical binding extracted from literature as well as co-expression networks built from gene expression data, or pathway databases. Furthermore, gene annotations from Gene Ontology [10] or MeSH [11] can be used to build associations between genes. Although such a network covers only a fraction of the real interactome, integration of gene expression data and network data has been shown to dramatically improve the outcome prediction of diseases [2, 12, 13].

Here we present one meta-approach, called NetRank [3], combining interaction data and gene expression data for cancer outcome prediction.

It has been shown that this approach is able to systematically improve outcome prediction using gene expression data and gene or protein association information [12].

2 Material

1. R packages affy, rma, e1071, genefilter, MLInterfaces, siggenes, Matrix, stats, LIBSVM (*see* **Note 1**).
2. Network represented as symmetric adjacency matrix, such that $wij = wji = 1$ if genes i and j are connected, and $wij = wji = 0$ otherwise (*see* **Note 2** and Subheading 3.1).

   Variable: W
3. As we use a Monte Carlo cross-validation schema, a list is needed indicating which patient belongs to the train or test set (*see* **Note 3**).

   For NetRanks internal validation, the train set is again split (Fig. 1 and Subheading 3.3).

   Variable: t.sets

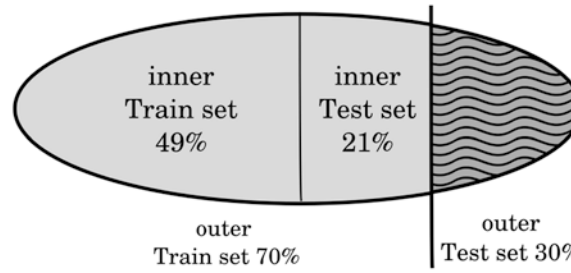


Fig. 1 To obtain an unbiased estimation of the performance of the algorithm, the dataset is initially split into train and test set. NetRank needs an internal validation for finding the best damping factor for the dataset; therefore the train set is again split into an "inner" train and test set

4. Mapping table of DNA-microarray features (Probesets) or NGS identifiers to network features (e.g., Entrezids) (*see* **Note 4**).

   Variable: f2e
5. An expression set containing the gene expression values (*see* **Note 5**).

   Variable: eset
6. A binary patient outcome variable.

   Variable: Prognosis

3 Methods

3.1 Meta-data

The expression "Meta data" covers a wide range of information. This can be different types of PPI information as used in HPRD [14], Transfac [15], STRING [16], Hprint [17], as well as gene annotations from Gene Ontology [10] or MeSH [11].

Transfac is a manually curated database, providing data on eukaryotic transcription factors, their genomic binding sites, and regulated genes.

HPRD, the Human Protein Reference Database, is manually created and integrates information from domain architecture, posttranslational modifications, interaction networks, and disease association for each known human protein.

Hprint contains predicted physical and functional interactions by integrating various data sources like text mining, genetic relationships, evolutionary information, as well as domain profiles.

STRING is a database of known and predicted protein interactions. Interactions included are either direct (physical) or indirect (functional) associations and are derived from different sources: high-throughput and/or co-expression experiments, text mining, as well as downloaded from existing databases.

In several studies it was shown that Transfac is most suited for cancer outcome prediction despite the smaller network size compared to other networks [3, 12]. This implies that regulatory information as provided by Transfac is particularly beneficial in outcome prediction. The majority of cancers arise due to mutations in regulatory elements governing cell cycle and cell survival; therefore the efficiency of Transfac in cancer outcome prediction is biologically reasonable.

The starting point of the analyses explained below is a dataset after filtering, consisting of a gene expression matrix with features (rows) and patient samples (columns). The following steps creating this matrix are exemplified using Affymetrix gene expression data, but any gene expression data from high-density oligonucleotide arrays or from NGS data can be used as well.

3.2 PreSteps

1. Affymetrix raw probe-level intensity (CEL) files need to be background-corrected, normalized, and summarized (here RMA is used) (*see* **Note 6**).

```
ab = ReadAffy(verbose=T)
eset = rma(ab, verbose=TRUE)
```

2. To remove genes with low expression, probe sets with a mean expression below 6 on the $\log_2$ scale are discarded. In addition, genes are filtered out whose expression shows only a small variance between patients. They are not informative and therefore cannot discriminate between patient outcome groups. In our example, this includes probe sets with a standard deviation below 0.5 on the log2 scale (*see* **Note 7**).

```
f1 = pOverA(0.1, 6)
f2<- function(x) (IQR(x) > 0.5)
ff = filterfun(f1,f2)
selected = genefilter(eset, ff)
eset = eset[selected,]
```

3. We kept for each gene only the probe set with the highest mean expression over all patients. A high correlation between probe sets reporting for the same gene was generally found (*see* **Note 8**).
4. Add a binary variable called Prognosis to the eset representing the outcome of the patients (*see* **Note 9**).
5. Support vector machines (SVM) are powerful supervised machine learning algorithms for classification problems. Here, an SVM is used to classify tumors samples into poor or good prognosis groups based on the expression levels of selected genes (*see* **Note 10**).

```
svmLearn <- function(dataset, tr.set.dataset) {
svm.m = mlSVM(dataset, tr.set.dataset)
conf.matrix = confuMat(svm.m)
```

```
correctly.classified = correct.class.rate(conf.
    matrix)
return(correctly.classified)
}
mlSVM <- function(dataset, tr.set.dataset, …) {
MLearn(Prognosis ~ ., dataset, svmI, trainInd=tr.
    set.dataset, …)
}
```

6. To create training sets, with N as the number of samples and No and Ni as the number of outer and inner training sets the following code is used. Ocv.size, icv.size represent the number of patients in the outer- and inner validation sets, respectively. The function called balanced.sample secures a balanced splitting of the samples (*see* **Note 3** and Fig. 1).

```
for (rep.o in 1:No) {
tr.plus.icv.set = balanced.sample(eset, (N - ocv.
    size))
tr.plus.icv.eset = eset[, tr.plus.icv.set]
ocv.set = (1:N)[-c(tr.plus.icv.set)]
for (rep.i in 1:Ni) {
tr.set = balanced.sample(tr.plus.icv.eset, (N - icv.
    size - ocv.size))
tr.set = match(sampleNames(tr.plus.icv.eset[, tr.
    set]), sampleNames(eset))
icv.set = (1:N)[-c(tr.set, ocv.set)]
tr.sets[rep.o, rep.i][[1]] = list(tr.set=tr.set,
    icv.set=icv.set,
ocv.set=ocv.set)
    }
}

balanced.sample <- function(eset, n) {
cat("Sampling training set size", n, "out of", N,
    "\n")
all.good = which(eset$Prognosis == "good")
all.poor = which(eset$Prognosis == "poor")
if (N%%2 == 0) {# N is even
add = sample(c(-0.25, 0.25), 1)
sample.good = sample(all.good, round(add+length(all.
    good)/N*n), replace=FALSE)
sample.poor = sample(all.poor, round(-add+length(all.
    poor)/N*n), replace=FALSE)
}
else {# N is odd
sample.good = sample(all.good, round(length(all.
    good)/N*n), replace=FALSE)
```

```
sample.poor = sample(all.poor, round(length(all.
    poor)/N*n), replace=FALSE)
}
training.set = sample(c(sample.good, sample.poor)) #
    indices only
return(training.set)
}
```

7. To create the sparse Matrix M from the meta network data an existing matrix M is used and transformed either via

   ```
   W = Matrix(M, sparse=TRUE)
   ```

   or directly from a data frame (edge.table), with 2 columns (node1 and node2) representing the network connections (*see* **Note 11**).

   ```
   W = xtabs(~ node1 + node2, data = edge.table, sparse
       = T)
   diag(W) = rep.int(0,dim(W)[1])
   ```

3.3 NetRank Work Flow

Given a fixed signature size n and a feature selection method (*see* **Note 12**), the following steps should be repeated several times (Fig. 2, figure adapted from [12]). This assures an unbiased estimate of the performance of the algorithm (*see* **Note 13**).

1. The starting point is a dataset after filtering, consisting of a gene expression matrix with features (rows) and patient samples (columns). Genes (features) that did not pass filtering are included with their initial values set to zero (*see* **Note 14**).
2. The data are randomly split into training and test sets.
3. Scores are defined for the initialization of each node in the network. These scores can either be the fold change between outcome groups, the p-values of the Students t-test, as well as the Pearson or Spearman correlation between gene expression and survival time (*see* **Note 15**).
4. Using the training set data only, features are ranked according to a feature selection method (here NetRank).
5. The top-ranked n features are selected. These features become the signature. The gene signature created on the training set is used to train a classifier on the sample outcome of the training set.
6. Afterwards the trained classifier is used to predict the outcome of the patients in the test set.
7. Finally, the predicted outcome is compared with the true outcome. The accuracy is defined by the fraction of patients predicted correctly.

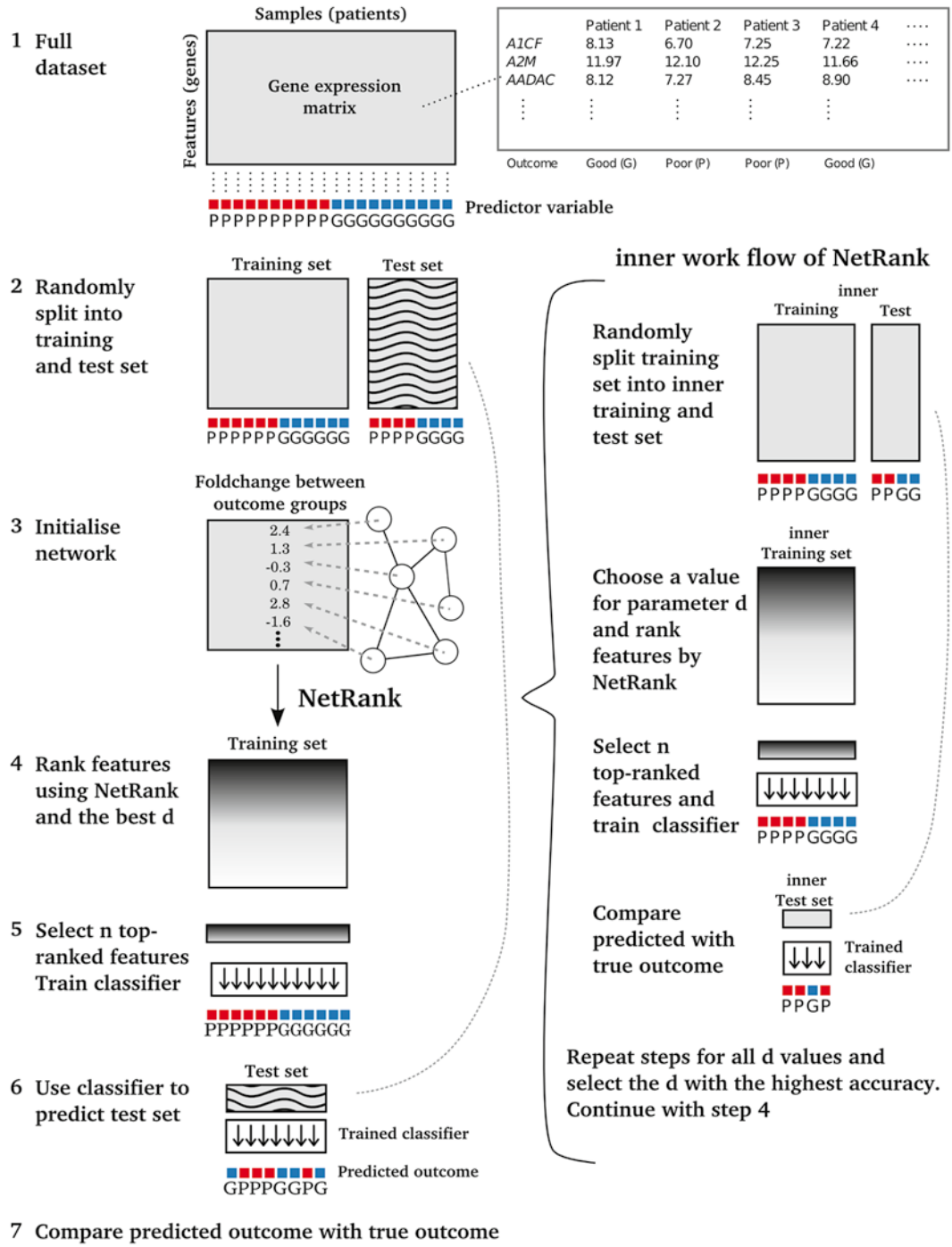


Fig. 2 The workflow of NetRank. The figure is adapted from [3]

For NetRank, additional steps are taken between **steps 3** and **4**, to identify the best damping factor for that dataset. Details are explained below.

The overall classification accuracy is the average of all repeated work flow accuracies.

3.4 NetRank

For ranking of genes, NetRank combines the expression level of a gene with the outcome variable of the patient by using a network of known gene-gene relationships. The ranking might be computed by eigenvalue decomposition (*see* for details ref. 3) or iteratively. Here, we follow the notation and implementation of Morrison et al. [18] that defines the ranking of gene j at the nth iteration as follows:

$$r_j^n = (1-d)c_j + d\sum_{i=1}^{N} \frac{w_{ij} r_i^{n-1}}{\deg_i}$$

W in $RNxN$ is a symmetric adjacency matrix for the network. c is a vector of coefficients representing the difference between the outcome groups. The damping factor d in $(0,1)$ is a fixed parameter regulating the influence of the network on the rank of a page. No influence of the network and full influence of the gene expression data is achieved with $d=0$, whereas setting $d=1$ corresponds to the opposite.

To achieve a full equality of each gene in the voting procedure, the result is scaled by $1/\deg_i$ in the summation. The final rank is obtained via the sum of $1-d$ and d times the votes given by other genes.

```
NetRank <- function(W, c, d) {
# W is a symmetric adjacency matrix representing the
   network
# c is a vector representing the difference between out-
   come groups, created in step 2
# d describes the influence of the network on the rank of
   a page (see Note 16)
# Starts NetRank algorithm with
c = abs(c)
norm.c = c/max(c)
# Calculates degrees
degrees = colSums(W)
ind = which(degrees == 0)
degrees[ind] = 1
# Initialization with 1/degrees
D1 = as(diag(1/degrees), "sparseMatrix")
# Constructing identity matrix
eye = as(diag(1, nrow(W)), "sparseMatrix")
```

```
# Adding degrees
A = eye - d*(t(W) %*% D1)
b = (1-d) * norm.c
# Solve linear system
r = solve(A, b)
return(r) # contains the rank scores, the higher the
   better
}
Special cases:
if (d==0)
return(c)
if (d==1) {
degrees = colSums(W)
return(degrees)
}
```

4 Notes

1. Libraries are available through Bioconductor.

   ```
   Source("http://bioconductor.org/biocLite.R")
   biocLite("XXX")
   ```

2. There exists a wide range of gene or protein interaction information. For human-based data networks like Transfac, HPRD or STRING are well-known and maintained data sources. Table 1 shows an overview of possible meta-data sources. Transfac has been proven to be most beneficial regarding outcome prediction.

Table 1
Overview of meta-data sources for human gene expression data analysis

Meta data	Type of information	Source
HPRD [14]	Physical protein interaction	www.hprd.org
Transfac [15]	Gene regulation	www.gene-regulation.com
STRING [16]	Predicted physical and functional protein-protein interaction	www.string-db.org
Hprint [17]	Predicted physical and functional protein-protein interaction	http://print-db.org
Reactome [22]	Pathway	www.reactome.org
GO [10]	Gene annotation	http://geneontology.org
MeSH [11]	Gene annotation	www.nlm.nih.gov/mesh

3. The splitting should be balanced such that the numbers of poor and good samples in the test set are either equal or differ by at most one. This ensures no over representation of one of the outcome groups in the training set.
4. Mapping can easily be done using bioconductor's annotation libraries. These libraries are easily accessible via (shown for Affymetrix chip hgu133plus2):

```
library(hgu133plus2)
Entrez = unlist(mget(featureNames(eset), hgu-
    133plus2ENTREZID, ifnotfound = NA)))
f2e = data.frame(PROBESET = featureNames(eset),
    ENTREZID = Entrez)
```

5. Data used in this pipeline can be either obtained via DNA-microarrays or next-generation sequencing. The generation of an expression set is explained in Subheading 3.2.
6. There exist a wide range of preprocessing methods, e.g., MAS5 [19], dChip [20], and RMA [21]. These methods differ in several aspects: (1) assessing background intensity, (2) normalization of expression levels across samples, and (3) summarization of probes to probe sets. RMA is a widely used method and applied in this work flow.
7. Values given are just examples for data created by Affymetrix DNA-Microarrays. Change accordingly to your experimental setup. For sequencing data a cutoff based on the amount of reads or FPKMs is advisable. In our experience, this filtering step removes between half and two thirds of all protein-coding genes.
8. This step is not needed when NGS data is used. It is important that each gene in the network is represented by only one value. For sequencing this could be raw counts or FPKM values, where individual transcript FPKM values for a gene are summed up to one value.
9. Make sure the order of the Prognosis vector for the patient outcome is the same order as in the expression set. Indicate a positive outcome as "good" and a negative as "poor." Add via eset$Prognosis = as.factor(prognosis) or if you have more clinical data directly before the normalization in Subheading 3.1, **step 1**, via:

```
clinical.adf = new('AnnotatedDataFrame', data=
    clinical)
phenoData(ab) = clinical.adf
```

10. Any other machine learning method is suitable, e.g., k-nearest neighbor, and random forest. A wide range of different methods are implemented in the Bioconductor package "MLearn."
11. Make sure that the matrix contains all genes measured on the DNA-Microarray.

12. Any sophisticated method can be used for ranking. Here, we introduce NetRank, but any other network ranking method can be applied in this work flow.
13. For NetRank the work flow was repeated between 500 and 1000 times, depending on the amount of patients in a dataset.
14. This serves to keep the network structure intact.
15. Make sure that the order of the vector created matches the order of the features in the matrix W.
16. For speed up, the matrixes for each damping factor can be precomputed.

Acknowledgement

We kindly acknowledge funding from EU, DFG, BMWi (PPI-Marker, OpenScienceLink, SigSax, GeneCloud).

References

1. Ideker T, Dutkowsky J, Hood L et al (2011) Boosting signal-to-noise in complex biology: prior knowledge is power. Cell 144:860–863
2. Chuang HY, Lee E, Liu YT et al (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140
3. Winter C, Kristiansen G, Kersting S et al (2012) Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS Comput Biol 8(5):e1002511
4. Johannes M, Brase JC, Fröhlich H et al (2010) Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics 26(17):2136–2144
5. Fortney K, Kotlyar M, Jurisica I (2010) Inferring the functions of longevity genes with modular subnetwork biomarkers of Caenorhabditis elegans aging. Genome Biol 11(2):R13
6. Chowdhury SA, Koyutürk M (2010) Identification of coordinately dysregulated subnetworks in complex phenotypes. Pac Symp Biocomput
7. Akula N, Baranova A, Seto D et al (2011) A network-based approach to prioritize results from genome-wide association studies. PLoS One 6(9):e24220
8. Davis NA, Crowe JE, Pajewski NM et al (2010) Surfing a genetic association interaction network to identify modulators of antibody response to smallpox vaccine. Genes Immun 11(8):630–636
9. Vandin F, Upfal E, Raphael BJ (2011) Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol 18(3): 507–522
10. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29
11. Rogers FB (1963) Communications to the editor. Bull Med Libr Assoc 51(1):114–116
12. Roy J, Winter C, Isik Z et al (2012) Network information improves cancer outcome prediction. Brief Bioinform 15(4):612–625
13. Cun Y, Fröhlich H (2013) Network and data integration for biomarker signature discovery via network smoothed T-statistics. PLoS One 8(9), e73074
14. Prasad TSK, Goel R, Kandasamy K et al (2009) Human protein reference database–2009 update. Nucleic Acids Res 37(Database issue): D767–D772
15. Matys V, Kel-Margoulis OV, Fricke E et al (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res 34(Database issue):D108–D110
16. Franceschini A, Szklarczyk D, Frankild S et al (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and

integration. Nucleic Acids Res 41(Database issue):D808–D815

17. Elefsinioti A, Saraç ÖS, Hegele A et al (2011) Large-scale de novo prediction of physical protein-protein association. Mol Cell Proteomics 10(11):M111.010629
18. Morrison JL, Breitling R, Higham DJ et al (2005) GeneRank: using search engine technology for the analysis of microarray experiments. BMC Bioinformatics 6:233
19. Affymetrix (2001) Microarray suite user's guide, 5th edn. Affymetrics
20. Li C, Wong WH (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci U S A 98:31–36
21. Irizarry RA, Hobbs B, Collin F et al (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264
22. Croft D, Mundo AF, Haw R et al (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42(Database issue): D472–D477

Chapter 13

Reproducible, Scalable Fusion Gene Detection from RNA-Seq

Vladan Arsenijevic and Brandi N. Davis-Dusenbery

Abstract

Chromosomal rearrangements resulting in the creation of novel gene products, termed fusion genes, have been identified as driving events in the development of multiple types of cancer. As these gene products typically do not exist in normal cells, they represent valuable prognostic and therapeutic targets. Advances in next-generation sequencing and computational approaches have greatly improved our ability to detect and identify fusion genes. Nevertheless, these approaches require significant computational resources. Here we describe an approach which leverages cloud computing technologies to perform fusion gene detection from RNA sequencing data at any scale. We additionally highlight methods to enhance reproducibility of bioinformatics analyses which may be applied to any next-generation sequencing experiment.

Key words RNA-Seq, Cloud, Fusion, Cancer, Reproducible, Genomics, Next-generation sequencing

1 Introduction

The chromosomal translocation now known to result in the BCR-ABL1 fusion gene was identified by Peter Nowell and David Hungerford in 1960 when, upon microscopic examination of white blood cells from a patient with chronic myelogenous leukemia, they noted that chromosome 22 was unusually short while the length of chromosome 9 was extended [1, 2]. Coined the Philadelphia chromosome, this rearrangement leads to the fusion of BCR and ABL1 genes and has been identified to play an important role in multiple types of carcinoma [3]. It is now clear that chromosomal alterations, including fusion genes, contribute to the initiation and progression of nearly all neoplastic subtypes [4]. While initially driven by advances in cytogenetic techniques, the advent of next-generation sequencing has transformed our ability to identify and detect fusion genes.

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_13,

1.1 The Role of Fusion Genes in Cancer

Fusion genes can result from rearrangement of DNA sequences located on different chromosomes (interchromosomal) or a shuffling of sequences located on the same chromosome (intrachromosomal). The BCR-ABL1 fusion represents an example of interchromosomal rearrangement, while the prostate cancer-associated TMPRSS2-ERG occurs as a result of an intrachromosomal translocation [5]. Moreover, two types of fusion can be distinguished: ones that alter regulatory regions and thus change gene expression without affecting the gene product, and those that give rise to a new gene product which is independent of either partner. In the latter type, the breakpoints occur within the coding regions of one or both genes, while in the former case breakpoints are found in the noncoding regions of both partners.

Abnormal gene expression or novel gene products caused by chromosomal translocation is frequently observed in cancer cells. The observation that specific rearrangements are associated with specific cancer types suggests that some gene fusions may represent "cancer-driving" events. For example, the fusion gene TMPRSS2-ERG occurs in approximately 50 % of prostate cancers [6]. While the mechanisms of fusion gene-driven carcinogenesis has been explored for a small number of recurrent fusion genes, the function of most fusion products remains an area of active investigation. An increased occurrence of chromosomal rearrangements has been found in regions with copy number transitions and high-level DNA amplifications. This observation suggests that fusion events may contribute to a selective advantage of cells harboring DNA amplifications and deletions [7]. The molecular mechanisms that lead to an increased rate of fusion gene production in carcinogenesis remain poorly understood. However, several lines of evidence suggest that altered DNA repair processes following double-strand breaks may contribute to the increased rate of chromosomal translocation [8].

1.2 Impact of Next-Generation Sequencing on Fusion Gene Identification and Detection

For decades, understanding the role of fusion genes in cancer development was hindered by a technical bottleneck in the detection process. However, advances in next-generation sequencing, including novel sequencing techniques and bioinformatic detection algorithms, have revealed that cancer-driving fusion genes occur more commonly and in a wider range of tumors than previously appreciated [9]. Indeed, the number of fusion genes deposited into the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer has grown at a nearly exponential rate since its creation in 2004 [10] (*see* Fig. 1). Chromosomal translocations and resulting fusion genes can be detected by sequencing DNA (using either whole-genome sequencing or targeted sequencing technologies). However, in tumor samples with highly rearranged genomes many alterations may be present, but only a fraction might alter transcribed gene products. RNA-Sequencing

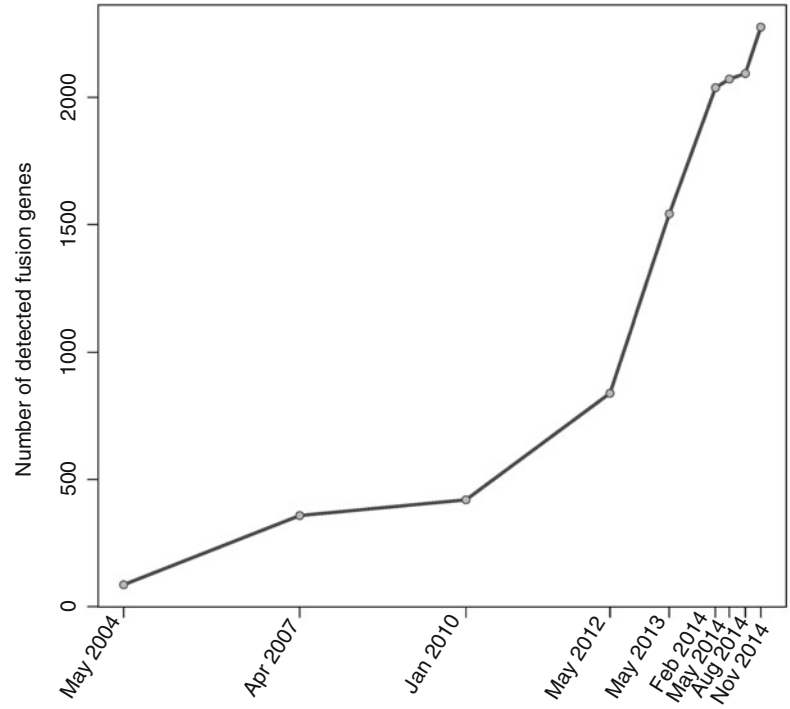


Fig. 1 Number of reported fusion genes over time. The number of unique fusion gene products deposited into the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer has grown steadily since its creation in 2004

(RNA-Seq) has the advantage of detecting only those fusion events that produce transcripts. Additionally, RNA-Seq allows the expression level of these transcripts to be directly evaluated. As a result, the majority of new fusion transcripts have been identified through RNA-Sequencing strategies [11].

1.3 Generalized Approach for Fusion Gene Detection from RNA-Seq

In contrast to array-based or targeted approaches, RNA-Seq allows quantitation of gene expression levels without a priori knowledge of the transcriptome. Many RNA-Seq analysis approaches begin by mapping the short reads produced during the library sequencing procedure to a reference genome or transcriptome. In situations where a reference genome is not available, de novo transcriptome assembly can be performed [12, 13]. While some reads will map directly to the genome, in many situations a read may span one or more exon-exon junctions. Indeed, in the case of the human genome, upwards of 30 % of 100 bp sequencing reads are expected to span two or more exons [14]. Multiple approaches have been developed to address the challenge of aligning reads to non-contiguous regions of the genome. While some tools implement a two-step approach to first discover exon junctions and then use these junctions to guide final alignment, other tools use existing

annotations to inform alignment. In some cases, accuracy is further improved by prioritizing alignments with consistent read pair mapping. A full discussion of split-read aligners is beyond the scope of this chapter and the reader is directed to a recent review and comparison of 26 alignment methods for further information [15]. Importantly, researchers have adapted and extended methods that allow alignment across exon-exon junctions resulting from gene fusions.

Just as in the case of analyzing gene expression using RNA-Seq, typically the first step of detecting gene fusions is the mapping of sequenced reads to a reference genome. Reads that do not map to the reference may span multiple transcripts or large distances. To examine this possibility, a second alignment step is performed in which reads are split into smaller units or (depending on the algorithm) treated as single-end reads. If reads or read pairs can be mapped to two different chromosomes or regions of the same chromosome, a candidate gene fusion is suggested. Because many reads may satisfy these criteria, fusion detection algorithms also take into account a number of other conditions to refine the selected candidates. The breakpoints of the fusion gene are then determined based on the transcript reference databases such as RefSeq [16] or Ensembl [17]. Some fusion detection methods discard any candidates not containing a breakpoint within an annotated gene. As a result, the choice of transcript reference can strongly influence the outcome of a fusion gene detection analysis.

Candidate fusion genes are detected when reads map uniquely to transcripts from different loci. There are two types of evidence that can support a gene fusion event. Encompassing reads are those whose both ends/mates are mapped to different chimeric segment, i.e., different chromosomes, that jointly make a fusion. This type of evidence requires paired-end sequencing to be performed. On the other hand, spanning reads can be detected with unpaired sequencing reads and occur when a single read contains sequences from both fusion partners (*see* Fig. 2). These two types of reads are handled differently by different tools. Some methods use the sum of encompassing reads and spanning reads, while others may weight spanning reads more heavily. In most cases, the number of reads supporting a particular fusion event is taken into consideration. This de facto introduces a filter that removes weakly expressed fusions from the identified candidates; e.g., junctions supported by just one read are removed. This filtering step leads to a list of potential fusion transcripts. To find the exact junction sequence, most algorithms search for spanning reads within the list of unmapped reads and then map them to the virtual reference made by the pairs of genes representing the candidate fusion genes. Finally, candidates are ranked by quality scores, thus leaving space for further filtering of false positives.

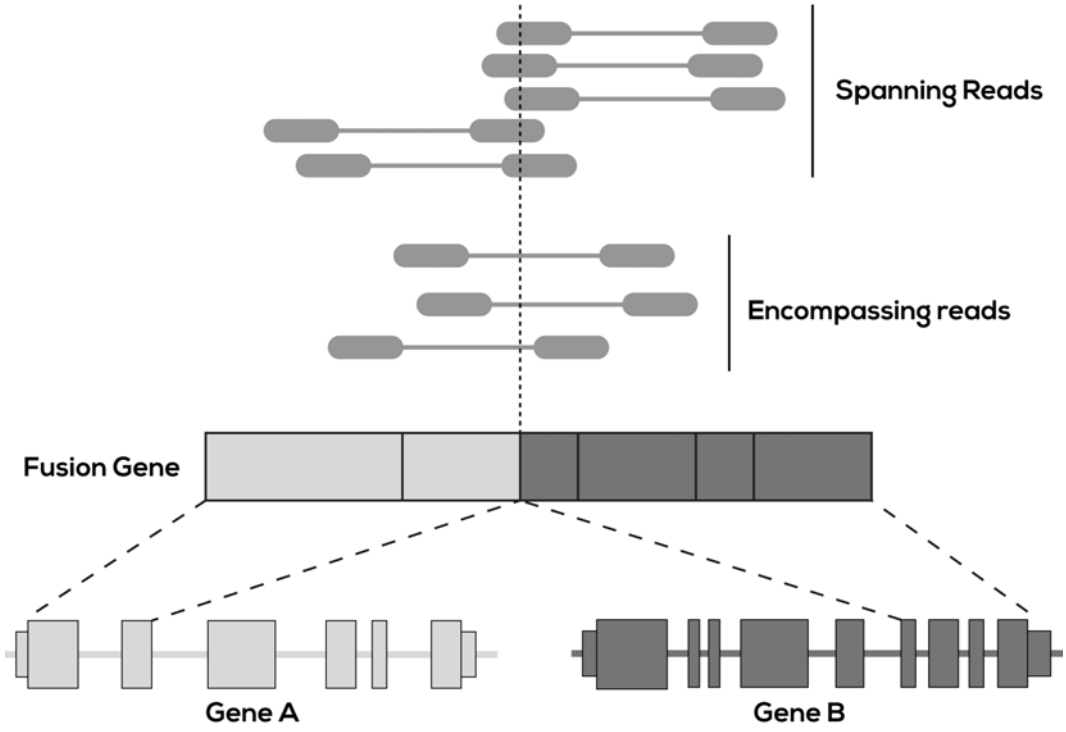


Fig. 2 Schematic of RNA-Seq reads supporting a fusion event. When paired-end sequencing is performed, two different configurations of reads can provide evidence for a fusion gene. Spanning reads directly cover the junction site. Read pairs where individual mates map on either side of the junction are termed encompassing reads

To date, dozens of fusion detection methods have been developed (*see* Table 1). An important challenge for researchers seeking to identify novel, or detect known, fusion genes is to evaluate which of these tools is best suited to their specific application. This decision is informed by determining the acceptable number of true positives, false positives, and false negatives, as well as the computational requirements and time needed to perform the analysis. For example, an investigator wishing to identify novel fusions may prefer a larger number of detections which can then be experimentally evaluated, while an investigator seeking to determine if a known fusion is present in a clinical sample may have much less tolerance for false positives even if some true positives are missed. It should be emphasized that computational methods for fusion detection from RNA-Seq are still in their relative infancy. As the number of validated fusions continues to grow and training sets for false and positive fusion detection improve, so too will the algorithms for fusion detection.

1.4 Considerations for Scalability and Reproducibility

As the cost of generating RNA-Seq data continues to decrease, investigators have the ability to perform fusion detection across hundreds or thousands of samples. Although individual tools vary in terms of processor and memory requirements, most tools require substantial computational resources. In particular, the initial steps

Table 1
Computational tools for fusion detection from RNA-Seq data. Numerous tools have been developed to facilitate detection of fusion genes. Links to source code are provided

Application	Tool name	Source code	Year	References
Detection	STAR	https://code.google.com/p/rna-star/	2012	[18]
Detection/identification	Bellerophontes	http://eda.polito.it/bellerophontes/	2012	[19]
Detection/identification	BreakFusion	http://bioinformatics.mdanderson.org/main/BreakFusion	2012	[20]
Detection/identification	ChimeraScan	http://chimerascan.googlecode.com	2011	[21]
Detection/identification	deFuse	http://sourceforge.net/projects/defuse/	2011	[22]
Detection/identification	Dissect	http://dissect-trans.sourceforge.net/Home	2012	[23]
Detection/identification	FusionCatcher	https://code.google.com/p/fusioncatcher/	2014	[24]
Detection/identification	FusionFinder	http://code.google.com/p/fusioncatcher/	2012	[25]
Detection/identification	FusionHunter	http://bioen-compbio.bioen.illinois.edu/FusionHunter/	2011	[26]
Detection/identification	FusionMap	http://www.arrayserver.com/wiki/index.php?title=FusionMap	2011	[27]
Detection/identification	FusionQ	https://sites.google.com/site/fusionq1/home/	2013	[28]
Detection/identification	FusionSeq	http://archive.gersteinlab.org/proj/rnaseq/fusionseq/	2010	[29]
Detection/identification	JAFFA	https://code.google.com/p/jaffa-project/	2015	[30]
Detection/identification	MOJO	https://github.com/cband/MOJO	2014	[31]
Detection/identification	shortFuse	https://bitbucket.org/mckinsel/shortfuse/overview	2011	[32]

Detection/identification	SnowShoes-FTD	http://mayoresearch.mayo.edu/mayo/research/biostat/stand-alone-packages.cfm	2011	[33]
Detection/identification	SOAPfuse	http://soap.genomics.org.cn/soapfuse.html	2013	[34]
Detection/identification	SOAPfusion	http://soap.genomics.org.cn/SOAPfusion.html	2013	[35]
Detection/identification	TopHat-Fusion	http://tophat-fusion.sourceforge.net/	2011	[36]
Detection/identification	TRUP	https://github.com/ruping/TRUP	2015	[37]
Detection/identification—focus on viral events	ViralFusionSeq	http://sourceforge.net/projects/viralfusionseq/	2013	[38]
Detection/identification combined RNA and DNA	Comrad	https://code.google.com/p/comrad/	2011	[39]
Detection/identification combined RNA and DNA	nFuse	https://code.google.com/p/nfuse/	2012	[40]
Detection/identification/ visualization	FusionAnalyzer	http://www.ngsbicocca.org/html/fusion_analyser.html	2012	[41]
Filtering and prioritization	Chimera	http://www.bioconductor.org/packages/release/bioc/html/chimera.html	2014	[42]
Filtering and prioritization	Oncofuse	http://www.unav.es/genetica/oncofuse.html	2013	[43]
Filtering and prioritization	Pegasus	http://sourceforge.net/projects/pegasus-fus/	2014	[44]

of building the reference index and mapping full length reads against the reference genome are particularly resource intensive. Multiple CPU cores as well as a substantial amount of RAM (in many cases more than 20 GB) can be required to process a single sample. Extending this analysis to even tens of samples can present a significant challenge if computation is performed on local architecture. As an alternative, researchers are beginning to turn to dynamic compute networks ("the cloud") to perform massive computation on demand, at scale. Under this paradigm, a researcher can access thousands of high-powered machines immediately, and only for the amount of time that they need them. This can provide a significant cost and speed advantage when compared to building up local compute clusters that may be only partially utilized. The advantages of this approach become even more apparent as the cost of cloud computation and storage continues to decrease.

Researchers are faced with an increasing number of tools to perform fusion detection and the results of an analysis conducted with each of these tools are invariably different. Moreover, each tool has a wide range of possible parameters and inputs. Together, these realities make reproducing the results of an analysis extremely challenging. Methods of capturing the tools and parameters used during an analysis are required to promote transparency and ensure that results can be reproduced. The Common Workflow Language (CWL) represents one effort to address this problem [45]. This initiative seeks to describe bioinformatics tools and pipelines using a simple, yet fully reproducible, approach. CWL allows tools and pipelines to be described once and then run on any architecture (cloud or local). Additionally, CWL allows tools to be referenced in Docker containers which ensures that the same version of a tool, including all of its dependencies, can be run anywhere exactly as it was originally created [46]. The first open-source implementation of this specification is called Reproducible Analysis for Bioinformatics (Rabix) (*see* **Note 1**) [47]. The combination of reproducible workflows and on-demand computational resources will enable researchers to meet the demands of increasingly data-intensive analysis for fusion detection as well as other next-generation sequencing applications.

2 Materials

The computational analysis of fusion genes described below requires a modern personal computer equipped with a web browser (*see* **Note 2**). The input data files can be obtained from RNA-Seq libraries constructed with standard approaches including stranded or unstranded mRNA or total RNA approaches. While both single- and paired-end sequencing strategies can be used to identify fusion genes, paired-end sequencing approaches provide more accurate detection due to more efficient mapping.

3 Methods

To date dozens of computational approaches for fusion detection have been reported. Here we provide an approach to detect fusion genes using ChimeraScan. This approach requires paired-end sequencing reads. Readers wishing to perform fusion detection using single-end reads should refer to **Note 3**. We briefly review the particular tools used in this workflow before proceeding to its execution.

3.1 Alignment and Fusion Detection

Alignment and initial fusion detection is performed by ChimeraScan [21]. This tool incorporates Bowtie [48] to align paired-end reads to a combined genome-transcriptome reference, discover discordant reads, predict an optimal fusion breakpoint location, and detect fusions. ChimeraScan also includes an indexing program, ChimeraScan Index, which creates the combined index from genomic reference sequences (FASTA format) and custom transcriptome reference format (UCSC GenePred format). ChimeraScan conveniently outputs an HTML-formatted results table with links to detailed descriptions of identified fusion products.

3.2 Filtering

While ChimeraScan performs an initial filtering of detected fusions, additional filtering and prioritization are frequently beneficial. In particular, long repeat and noncanonical junctions are typically removed because these are often the result of mapping or library preparation artifacts. Low-abundance junctions located between highly expressed genes are also more likely to be false positives and so are removed. Although several tools are available to perform filtering of detected fusions (*see* **Note 4**) we suggest the use of Chimera [42]. Retaining only canonical junctions with a repeat less than 5 allows the number of false positive fusion gene detections to be significantly reduced.

3.3 Visualization

Visualizing fusion events across long chromosomal distances is facilitated by representing the genome as a circle. Each chromosome is shown as a segment in the outer ring and detected fusions are drawn as arcs (*see* Fig. 3). These plots are created by Circos [49].

3.4 Executing the Pipeline

A ready-to-run fusion detection pipeline incorporating the tools described above is available on the Seven Bridges Platform [50]. A pictorial representation of the pipeline is shown in Fig. 4. The pipeline has additionally been implemented in Rabix to allow its execution on other cloud platforms supporting CWL, or on local infrastructure. The steps to executing the pipeline are as follows:

1. Provide input files.
   (a) Reads (paired-end): The pipeline accepts one pair of paired-end RNA-Seq data in FASTQ format (plain text or

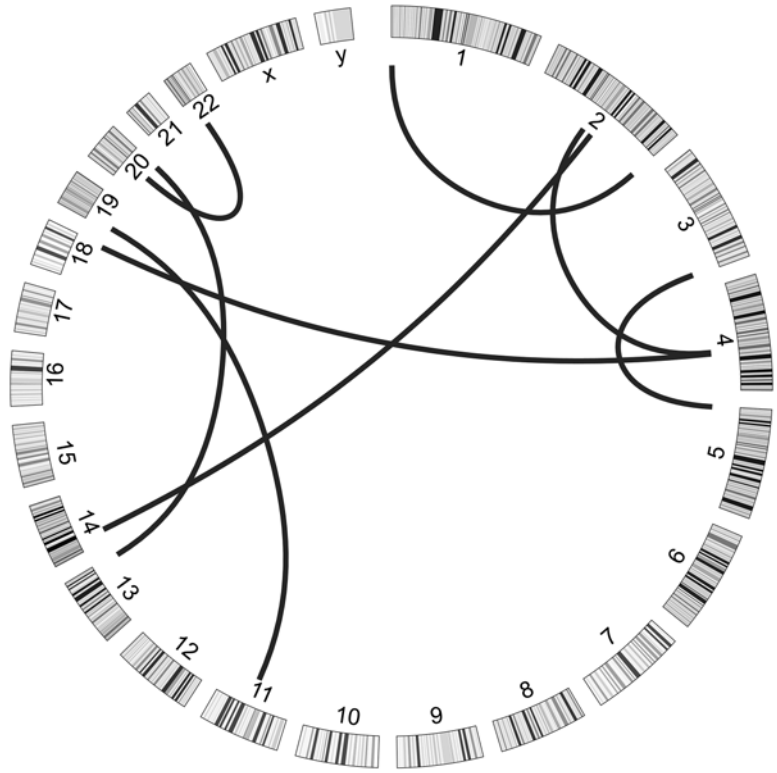


Fig. 3 Visualization of fusion genes on a Circos plot. Chromosomes are shown as segments in the outer ring and arcs are used to represent gene fusions across or within chromosomes

compressed files). If reads for samples are present in multiple files, they should be merged prior to alignment. The read pair metadata fields must be set as 1 and 2.

(b) Genome reference: FASTA file containing reference genome. For human samples we recommend the UCSC hg19 annotation. *See* **Note 5** for information about creating this file.

(c) Transcriptome reference: Transcriptome reference file containing all known transcripts in a custom format similar to BED. *See* **Note 6** for information about creating this file.

(d) False-positive chimeras (optional): List of known false-positive fusion detections from normal tissues. *See* **Note 7** for more information about this file.

2. Adjust tool parameters.

(a) ChimeraScan parameters can be adjusted to optimize alignment or accommodate stranded sequencing. In most cases the default values are appropriate. Considering only annotated genes and providing a file of known likely false positives can improve the positive detection rate.

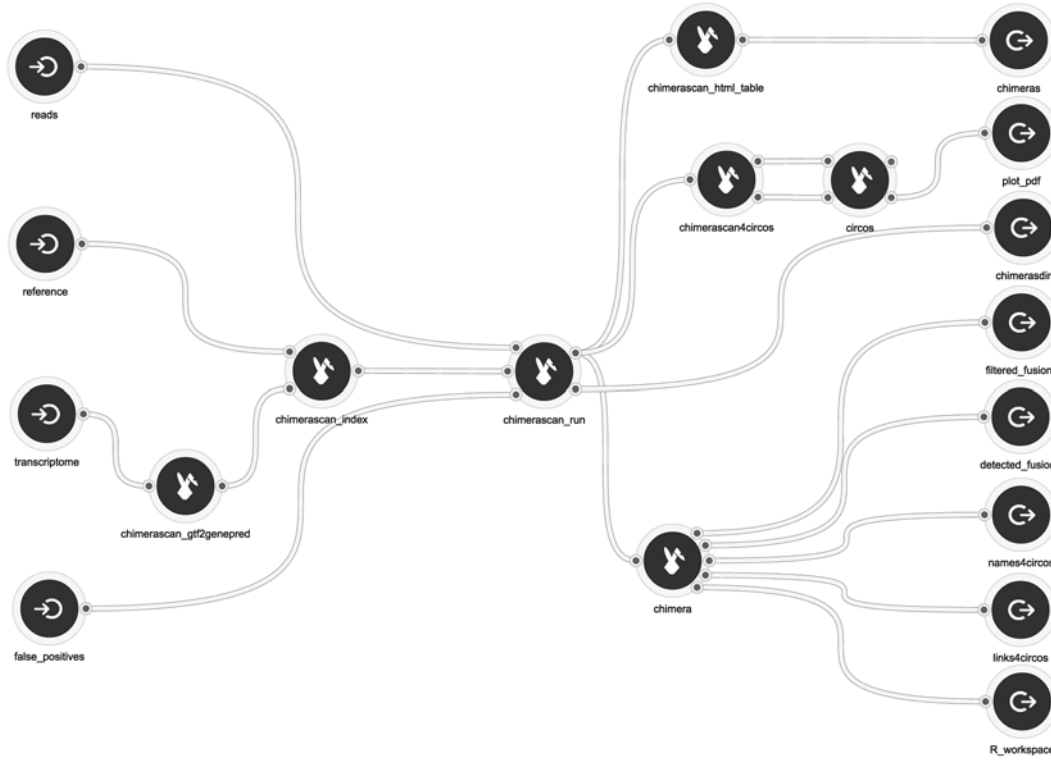


Fig. 4 Fusion detection pipeline. The computational workflow for fusion detection is shown. *Nodes with arrows* indicate data entering or exiting the workflow while other nodes represent computational steps performed on the data

(b) Chimera parameters should be adjusted to allow filtering of fusion candidates. These include:

- Selecting the minimal number of supporting reads (calculated as the sum of spanning and encompassing reads). Recommended (minimal) value: 2
- Selecting filtering type: Depending on the selection, a fusion is removed if (1) it has less supporting reads than a set value, (2) its name is not in the given list, (3) the intronic regions are included in the fusion, (4) the partner genes are not annotated, or (5) gene partners are the same. Recommended value: Filter unannotated gene partners.

3. Run the analysis

(a) Once input files and parameters are set, the pipeline can be executed. On the Seven Bridges Platform selecting "Run" sends the job to optimized Amazon Web Services compute instances. Typical RNA-Seq libraries (~5 GB zipped fastq files) are fully analyzed in 6–8 h. An e-mail is sent once the analysis is completed.

4. View the results
   (a) The following files are created upon completion of the pipeline:
      - The chimeras.bedpe file contains information about the chromosomal regions, transcript IDs, genes, and statistics for each identified fusion gene.
      - A sortable table of detected fusions is provided in as an HTML page with links to detailed descriptions of detected fusion genes.
      - An Index file is provided in an archive format. In the case of subsequent pipeline executions, this Index can be reused to decrease the overall run time.
      - A text file containing a list of fusions passing Chimera filters.
      - Circos plots for visual representation of fusion genes in pdf and svg format.

3.5 Optimization and Testing

The performance of a particular fusion detection workflow can be evaluated and optimized using data from samples with wet-lab-validated fusions as well as with synthetic positive data (reads with simulated fusions) and synthetic negative data (free of fusions). This analysis can assist the researcher in understanding the role of each parameter and aid in fine-tuning the workflow. Nevertheless, some members of the community argue that the imperfectness of simulated fusions in synthetic datasets may complicate algorithm optimization, and wet-lab validation of identified fusions remains the gold standard for evaluation of fusion detection tools and workflows.

Analysis on the cloud allows researchers with access to disparate computational resources to perform fusion detection at scale and on demand. Additionally, improved methods of capturing the precise tools and parameters (such as CWL) used during fusion detection will enable these analyses to be reproducible and shareable.

4 Notes

1. Directions for using Rabix for local execution of pipelines is described at https://github.com/rabix/rabix
2. Adaptation of the described approach to local compute architecture requires access to a compute cluster with multi-core processors and at least 20GB of RAM. Docker (www.docker.io) should be installed according to the directions on their website.
3. Detection of fusion genes from single read RNA Seq libraries can be performed using the Fusion Transcript Detection – STAR + Chimera pipeline available at https://igor.sbgenomics.

com/lab/pipeline/view/540dd2fad79f00766c174eb0/. Alignment and fusion detection with STAR are substantially faster than alignment with Bowtie (which is used by ChimeraScan). However, STAR does not allow identification of fusion gene orientation unless stranded RNA-Seq reads are supplied. Filtering of fusions detected by STAR is particularly important as many false positives may be detected.

4. Oncofuse [43] is a naive Bayesian classifier can alternatively be used to perform filtering and prioritization of detected fusion genes. This tool allows prioritization of fusions based on their oncogenic potential with the goal of identifying driver events. It is noted that while both Oncofuse and Chimera use the same read coverage prefilter, the outputs of these tools can vary substantially due to differences in the reference gene annotations used by each of these tools.
5. A genome reference can be obtained from UCSC using either the Seven Bridges FTP uploader or a wget command from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips.
6. The transcriptome reference file can be prepared using the ChimeraScan GTF to genePred tool on the Seven Bridges Platform or by following the directions available at https://code.google.com/p/chimerascan/wiki/Installation.
7. The authors of ChimeraScan provide a list of fusions detected in the Illumina BodyMap dataset. This dataset was generated from 16 normal human tissue types. While the fusions may not all represent false positives, as they were derived from normal tissues they are not likely to be implicated in disease. The file can be obtained at https://code.google.com/p/chimerascan/downloads/detail?name=hg19_bodymap_false_positive_chimeras.txt.gz&can=2&q=.

References

1. Nowell P, Hungerford D (1960) A minute chromosome in human chronic granulocytic leukemia [abstract]. Science 132:1497
2. Groffen J, Stephenson JR, Heisterkamp N et al (1984) Philadelphia chromosomal breakpoints are clustered within a limited region, bcr, on chromosome 22. Cell 36:93–99
3. Koretzky GA (2007) The legacy of the Philadelphia chromosome. J Clin Invest 117: 2030–2032
4. Mitelman F, Johansson B, Mertens F (2007) The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 7: 233–245
5. Tomlins SA, Laxman B, Varambally S et al (2008) Role of the TMPRSS2-ERG gene fusion in prostate cancer. Neoplasia 10:177–188
6. Tomlins SA, Rhodes DR, Perner S et al (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310:644–648
7. Edgren H, Murumagi A, Kangaspeska S et al (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 12:R6
8. Aplan PD (2006) Causes of oncogenic chromosomal translocation. Trends Genet 22: 46–55
9. Mitelman F, Johansson B, Mertens F (2004) Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer. Nat Genet 36:331–334
10. Mitelman database of chromosome aberrations and gene fusions in cancer. http://cgap.nci.

nih.gov/Chromosomes/Mitelman. Accessed 1 Feb 2015
11. Wang Q, Xia J, Jia P et al (2013) Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. Brief Bioinform 14:506–519
12. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12: 671–682
13. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63
14. Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36
15. Engström PG, Steijger T, Sipos B et al (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10: 1185–1191
16. Pruitt KD, Brown GR, Hiatt SM et al (2014) RefSeq: an update on mammalian reference sequences. Nucleic Acids Res 42:D756–D763
17. Hubbard T, Barker D, Birney E et al (2002) The Ensembl genome database project. Nucleic Acids Res 30:38–41
18. Dobin A, Davis CA, Schlesinger F et al (2012) STAR: ultrafast universal RNA-seq aligner. Bioinformatics. doi:10.1093/bioinformatics/bts635
19. Abate F, Acquaviva A, Paciello G et al (2012) Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. Bioinformatics 28:2114–2121
20. Chen K, Wallis JW, Kandoth C et al (2012) BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. Bioinformatics 28:1923–1924
21. Iyer MK, Chinnaiyan AM, Maher CA (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics 27:2903–2904
22. McPherson A, Hormozdiari F, Zayed A et al (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 7, e1001138
23. Yorukoglu D, Hach F, Swanson L et al (2012) Dissect: detection and characterization of novel structural alterations in transcribed sequences. Bioinformatics 28:i179–i187
24. Nicorici D, Satalan M, Edgren H et al (2014) FusionCatcher—a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv. doi: 10.1101/011650
25. Francis RW, Thompson-Wicking K, Carter KW et al (2012) FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. PLoS One 7, e39987
26. Li Y, Chien J, Smith DI, Ma J (2011) FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 27:1708–1710
27. Ge H, Liu K, Juan T et al (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics 27:1922–1928
28. Liu C, Ma J, Chang CJ, Zhou X (2013) FusionQ: a novel approach for gene fusion detection and quantification from paired-end RNA-Seq. BMC Bioinformatics 14:193
29. Sboner A, Habegger L, Pflueger D et al (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 11:R104
30. Davidson NM, Majewski IJ, Oshlack A (2015) JAFFA: high sensitivity transcriptome-focused fusion gene detection. Genome Med 7(1):43
31. Bandlamudi C, Lin P, Tian J et al (2014) Discovery and functional characterization of recurrent gene fusions from 7,470 primary tumor transcriptomes across 28 human cancers. ASHG 2014 meeting abstracts
32. Kinsella M, Harismendy O, Nakano M et al (2011) Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics 27:1068–1075
33. Asmann YW, Hossain A, Necela BM et al (2011) A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res 39, e100
34. Jia W, Qiu K, He M et al (2013) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol 14:R12
35. Wu J, Zhang W, Huang S et al (2013) SOAPfusion: a robust and effective computational fusion discovery tool for RNA-seq reads. Bioinformatics 29:2971–2978
36. Kim D, Salzberg SL (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12:R72
37. Fernandez-Cuesta L, Sun R, Menon R et al (2015) Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. Genome Biol 16:7

38. Li J-W, Wan R, Yu C-S et al (2013) ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics 29:649–651
39. McPherson A, Wu C, Hajirasouliha I et al (2011) Comrad: detection of expressed rearrangements by integrated analysis of RNA-Seq and low coverage genome sequence data. Bioinformatics 27:1481–1488
40. McPherson A, Wu C, Wyatt AW et al (2012) nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Res 22:2250–2261
41. Piazza R, Pirola A, Spinelli R et al (2012) FusionAnalyser: a new graphical, event-driven tool for fusion rearrangements discovery. Nucleic Acids Res 40, e123
42. Beccuti M, Carrara M, Cordero F et al (2014) Chimera: a Bioconductor package for secondary analysis of fusion products. Bioinformatics 30:3556–3557
43. Shugay M, Ortiz de Mendíbil I, Vizmanos JL, Novo FJ (2013) Oncofuse: a computational framework for the prediction of the oncogenic potential of gene fusions. Bioinformatics 29: 2539–2546
44. Abate F, Zairis S, Ficarra E et al (2014) Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. BMC Syst Biol 8:97
45. Common-workflow-language common-workflow-language/common-workflow-language. In: GitHub. https://github.com/common-workflow-language/common-workflow-language. Accessed 22 Feb 2015
46. Docker build, ship, and run any app, anywhere. https://www.docker.com/. Accessed 1 Aug 2014
47. rabix rabix/rabix. In: GitHub. https://github.com/rabix/rabix. Accessed 22 Feb 2015
48. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
49. Krzywinski M, Schein J, Birol I et al (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19:1639–1645
50. Arsenijevic V fusion transcript detection—ChimeraScan. https://igor.sbgenomics.com/lab/pipeline/view/540dd19dd79f00766c174ead/

Chapter 14

Transcriptome Sequencing for the Detection of Chimeric Transcripts

Hsueh-Ting Chu

Abstract

The occurrence of chimeric transcripts has been reported in many cancer cells and seen as potential biomarkers and therapeutic targets. Modern high-throughput sequencing technologies offer a way to investigate individual chimeric transcripts and the systematic information of associated gene expressions about underlying genome structural variations and genomic interactions. The detection methods of finding chimeric transcripts from massive amount of short read sequence data are discussed here. Both assembly-based and alignment-based methods are used for the investigation of chimeric transcripts.

Key words Transcriptome sequencing, Chimeric transcript, Fusion genes, Cancer genes, De novo assembly

1 Introduction

Genomic abnormalities including gene mutations and fusion genes were suggested to be causative events in tumorigenesis [1]. A fusion gene is a chimeric transcript formed from two different genes and is usually a result of a chromosome translocation event. For example, the BCR-ABL1 fusion gene was found in patients with chronic myelogenous leukemia (CML) in the early 1980s [2]. The fusion links the *abl1* gene on chromosome 9 and the *bcr* gene on chromosome 22. Besides, the BCAS4-BCAS3 fusion gene was found in breast cancer [3]. Originally, the *bcas4* gene is on chromosome 20 and the *bcas3* gene is on chromosome 17. So far more than 300 gene fusions have been identified in malignant neoplastic disorders [4]. These fusion genes have become potential biomarkers and therapeutic targets of different cancers. Therefore it is an essential topic to finding novel chimeric transcripts for the research of cancer diagnosis and treatment.

Transcriptome sequencing is one of next-generation sequencing (NGS) technologies to reveal a snapshot of RNA presence and

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7_14,

quantity from a genome at a given moment in time [5]. In contrast to microarray technology, transcriptome sequencing allows identification of individual transcripts, does not require a reference genome, and interferes with noisy fluorescence quantification. Therefore a great potential is recognized for learning the cause and function of cancer from transcriptome sequencing studies. Consequently, discovering fusion genes from RNA-Seq data has become one of the important topics for cancer research.

1.1 Transcriptome Sequencing

The expression of transcripts in a cell is dynamic. Next-generation sequencing is rapidly becoming the method of choice for the exploration of gene expressions. Furthermore, unlike hybridization-based detection, RNA-Seq allows genome-wide analysis of transcription at single-nucleotide resolution, including identification of including alternative splicing sites, posttranscriptional modifications, and fusion genes. As of 2015, the three most widely used NGS platforms for RNA-Seq are SOLiD and Ion Torrent and Illumina (*see* **Note 1**). Most of the RNA-Seq experiments use Illumina paired-end sequencing that allows users to sequence both ends of a fragment and generate high-fidelity reads. For the reason, most of published methods majorly deal with Illumina RNA-Seq data for finding chimeric transcript.

1.2 Chimeric Reads Versus Chimeric Transcripts

Not all chimeric reads in a RNA-Seq dataset are from chimeric transcripts. There are a lot of aberrant chimeric reads being produced during the sequencing process. Kircher et al. (2011) had addressed the problem as one of the major challenges on the analysis of Illumina sequencing data [6]. Such chimeric reads are called adapter-induced chimeras in comparison with transcription-induced chimeras which are the targets of fusion detection (*see* **Note 2**). It is still difficult to undoubtedly discriminate the two types of chimeras in RNA-Seq datasets. Therefore, most of current fusion-finding tools have to employ filtering criteria to get rid of false positive detections. Carrara et al. compared the performance of several state-of-the-art tools. All of the tools neglect some of transcription-induced chimeras and reach more than 20 % false discovery rate from their tests [7].

1.3 Alignment-Based Versus Assembly-Based Chimera Detection

Most of the published fusion-finding methods are alignment based and have similar pipelines which employ read-alignment tool, such as bowtie, BWA, BLAT, and BLAST. The basic steps of an alignment-based method begin with aligning reads to reference sequences and collect unmapped reads which are examined if they are chimeric. Such an approach relies heavily on the completeness of reference sequences (*see* **Note 3**).

De novo assembly of sequencing data is the method of reconstruct sequences without the aid of any reference. EBARDenovo is a paired-end RNA-Seq assembler [8] and firstly employs an

efficient chimera detection function to abrogate the effect of aberrant chimeric reads (*see* **Note 4**). The detected chimeric reads can be examined if they are from a chimeric transcript as well as the assembled contigs (Fig. 1).

2 Materials

2.1 RNA-Seq Test Datasets

To test the detection of chimeric transcripts, two RNA-Seq datasets with SRA Accession: SRX025827 and SRX025832 are examined. These datasets were obtained from the output of sequencing breast cancer cell lines (MCF-7 and KPL-4) on an Illumina Genome Analyzer II platform. Both of the datasets have paired-end reads of length 50 bp [9].

1. The MCF-7 dataset (SRX025827) includes 8,412,431 spots (841.2 M bases) in a run (SRR064286). There are three previously reported fusion genes, including BCAS4-BCAS3, ARFGEF2-SULF2, and RPS6KB1-VMP1.
2. The KPL-4 dataset (SRX025832) includes 6,800,166 spots (680 M bases) in a run (SRR064287). There are three previously reported fusion genes, including PPP1R12A-SEPT10, BSG0-NFIX, and NOTCH1-NUP214.

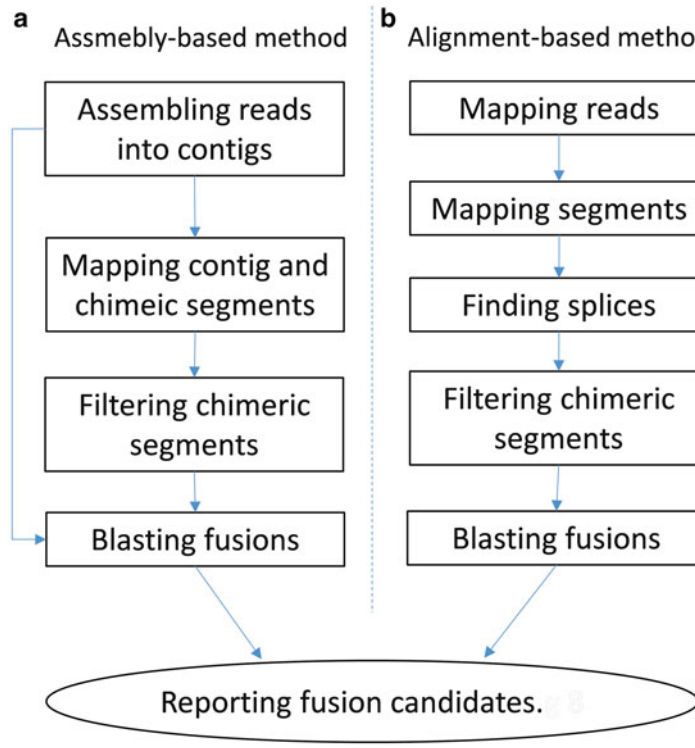


Fig. 1 The basic workflow for the detection of chimeric transcripts

2.2 Alignment-Based Tool

There are more than 15 alignment-based fusion-finding tools (*see* **Note 5**). Tophat-fusion was chosen as a representative alignment-based tool for finding chimeric transcripts from transcriptome sequencing data [10].

1. Install Tophat package. Tophat-fusion is a function of the Tophat package which has been included in the software repositories of different Linux distributions, such as Ubuntu (http://www.ubuntu.com/) or Linux Mint (http://www.linuxmint.com/). Linux users can install the Tophat package through Ubuntu Software Center or Mint Software Manager.
2. Install three required bioinformatics packages: blast, bowtie, and samtools. All are available through Ubuntu Software Center or Mint Software Manager as well.
3. Download and extract the Bowtie indexes of the human reference genome hg19 from ftp://ftp.cbcb.umd.edu/pub/data/bowtie_indexes/hg19.ebwt.zip.
4. Download and extract the Ensemble gene annotation file ensGene.txt from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/ensGene.txt.gz.
5. Download and extract the RefSeq gene annotation file refGene.txt from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/database/refGene.txt.gz.
6. Create the directory "~/blast/human_genomic" and download and extract the files human_genomic*.gz of blast database from ftp://ftp.ncbi.nlm.nih.gov/blast/db/.
7. Create the directory "~/blast/nt" and download and extract the files nt*.gz of blast database from ftp://ftp.ncbi.nlm.nih.gov/blast/db/.

2.3 Assembly-Based Tool

EBARDenovo is the de novo assembler with chimera detection [8]. It is coded in Microsoft .Net C# framework and 64-bits Windows is its native running platform. For Linux user, the program mono is required for running any .Net program. The Linux Mint (http://www.linuxmint.com/) is suggested to run EBARDenovo. Just through the Linux Mint Software Manager, a user can execute most of the .Net programs with mono (http://www.mono-project.com).

1. Install wine and mono package through Software Manager of the Linux Mint OS.
2. Download and extract the packages EBARDenovo and TranslocCheck from http://sourceforge.net/projects/ebardenovo/.
3. Download and extract the gmap-gsnap package from http://research-pub.gene.com/gmap/. In the software repositories

of Linux Mint, there is an old version of the gmap-gsnap package [11]. Since the gmap program had changed the format of reference database after 2014-03-28, the usage of the old versions are not suggested.

4. Install zlib-dev package through Software Manager of the Linux Mint OS. It's required for the compilation of the gmap program.
5. Run the following commands to install the gmap program:

```
./configure
make
sudo make install
```

6. Download the chromosome files *.gz from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/ and extract them into a folder and run the command

```
sudo gmap_build -d hg19 -k 15 *.fa
```

7. Install UGENE package through Software Manager of the Linux Mint OS [12]. UGENE is used to display the alignment of chimeric transcripts.

3 Methods

3.1 Fusion Finding by Tophat-Fusion

The execution of Tophat-fusion relies on three parts of reference databases including Bowtie indexes, Blast indexes and gene annotations (RefSeq and Ensemble). The hg19 Bowtie indexes should be put in the subfolder "~/hg19" and the Blast indexes should be put in the subfolder "~/blast." The gene annotation files ensGene.txt and refGene.txt should be put on the home directory.

1. Run the following command line of the Tophat:

```
$ tophat --fusion-search -o tophat_MCF7 -p 6 --keep-fasta-order --bowtie1 --no-coverage-search -r 0 --mate-std-dev 80 --max-intron-length 100000 --fusion-min-dist 100000 --fusion-anchor-length 13 --fusion-ignore-chromosomes chrM ~/hg19/hg19 SRR064286_1.fastq SRR064286_2.fastq
```

2. Run the following command of the Tophat-fusion-post program:

```
$ tophat-fusion-post -p 6 --num-fusion-reads 1 --num-fusion-pairs 2 --num-fusion-both 5 ~/hg19/hg19
```

3. The result of fusion finding is output into the folder ~/tophat-fusion_out. There is a text file "result.txt" which list the detected fusions, and an html file "result.html" which shows the read alignments at the splices of fragments.

3.2 Fusion Finding by EBARDenovo

De novo assembly of RNA-Seq data with EBARDenovo files does not need the configuration of any reference databases.

1. To run the EBARDenovo program, just list the input data files and the output filename after the parameter "-o" such as

```
$ mono EBARDenovo.exe SRR064286_1.fastq SRR064286_
   2.fastq -o ebar286.fa -Q
```

2. In general, EBARDenovo output three files: a contig file (.fa), a log file (.log), and a statistics file (.csv). The parameter "-Q" of EBARDenovo will cause the output of detected chimeras into the file with the postfix "-chimera.fa," e.g., ebar286-chimera.fa.
3. The assembled contigs can be checked whether they are chimeric through the online nucleotide blast service (*see* **Note 1**) http://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?PROGRAM=blastn.
4. The program gmap can detect translocations of assembled sequences caused by mis-assembly or being real chimeric transcripts. To check translocations of EBARDenovo outputs, run the gmap commands as

```
$ gmap -d hg19 --split-output=map_ebar_286 ebar286.
   fa
$ gmap -d hg19 --split-output=map_chix_286 ebar286-
   chimera.fa
```

5. The parameter "--split-output" of gmap let the outputs sent to individual files with file extensions: nomapping, uniq, mult, and transloc. After the above commands, the files map_ebar_286.transloc and map_chix_286.transloc store the candidates of chimeras.
6. Do the post-analysis of gmap results with the TranslocCheck programs. The program produces the alignment of reads cross the splices detected by the gmap program.

```
$ mono TranslocCheck.exe SRR064286_1.fastq
   SRR064286_2.fastq -c ebar286.fa -t map_ebar_
   ebar 286.transloc
$mono TranslocCheck.exe SRR064286_1.fastq
   SRR064286_2.fastq -c ebar286-chimera.fa -t map_
   chix_286.transloc -S
```

7. The TranslocCheck program aligns reads onto the translocated contigs and detected chimeric fragments and output a text file for the chimeric information and three ace files. The ace format is for the alignment of short reads and can be visually browsed by the UGENE program.

8. The command and major output files of each steps are as follows:

Commands	Outputs
EBARDenovo S_1.fastq S_2.fastq -o xxx.fa -Q	xxx.fa xxx-chimera.fa
gmap -d hg19 --split-output=map_xxx xxx.fa	map_xxx.transloc
gmap -d hg19 --split-output=map_ccc xxx-chimera.fa	map_ccc.transloc
TranslocCheck S_1.fastq S_2.fastq -c xxx.fa -t map_xxx.transloc	xxx.fa.txt xxx.fa_tloc.ace
TranslocCheck S_1.fastq S_2.fastq -c xxx-chimera.fa -t map_ccc.transloc -S	xxx-chimera.fa.txt xxx-chimera.fa_tloc.ace

3.3 Results

1. There are six fusion genes in the breast cancer cell lines MCF-7 and KPL-4 verified in the literature. Therefore, the outputs of Tophat-fusion and EBARDenovo were checked for the six chimeric transcripts. It shows that all of the six chimeric transcripts can be found by both Tophat-fusion and EBARDenovo. The findings by EBARDenovo are listed here (Table 1). Five of the fusions were found from the assembled contigs. Moreover, one of the fusions RPS6KB1-VMP1 is detected as aberrant chimera. It means that normal transcripts RPS6KB1 and VMP1 and the chimeric transcript RPS6KB1-VMP1 coexist in the cell transcriptome and the normal transcripts have much higher expression levels than the chimeric transcript.

Output	MCF-7	KPL-4
Assembled contigs	BCAS4-BCAS3 ARFGEF2-SULF2	NFIX-BSG, NOTCH1-NUP214 PPP1R12A-SEPT10
Detected chimeras	RPS6KB1-VMP1	

2. Excluding the six verified fusion genes, there are other translocated contigs from the EBARDenovo outputs. Because the de novo assembly by EBARDenovo does not refer to any reference sequences, there are still possible mis-assemblies caused by sequence repeats and homologous genes. In general, EBARDenovo outputs less artificial contigs than other state-of-the-art RNA-Seq assemblers because of the built-in chimera detection function.

Table 1
Fusion finding and de novo assembly programs for RNA-Seq analysis

Name	Source	References
Fusion finding algorithms		
BreakDancer	http://breakdancer.sourceforge.net/	Fan et al. [15]
FusionSeq	http://archive.gersteinlab.org/proj/rnaseq/fusionseq/	Sboner et al. [16]
MapSplice	http://www.netlab.uky.edu/p/bioinfo/MapSplice	Wang et al. [17]
Tophat-fusion	http://ccb.jhu.edu/software/tophat/fusion_index.html	Kim et al. [10]
deFuse	https://sourceforge.net/projects/soapfuse/	McPherson et al. [18]
FusionHunter	http://bioen-compbio.bioen.illinois.edu/FusionHunter/	Li et al. [19]
SnowShoes-FTD	http://bioinformaticstools.mayo.edu/research/snowshoes-ftd/	Yan et al. [20]
ChimeraScan	https://code.google.com/p/chimerascan/	Iyer et al. [21]
FusionMap	http://www.arrayserver.com/wiki/index.php?title=FusionMap	Ge et al. [14]
FusionFinder	http://bioinformatics.childhealthresearch.org.au/software/fusionfinder/	Francis et al. [22]
FusionAnalyser	http://www.ngsbicocca.org/html/fusionanalyser.html	Piazza et al. [23]
SAOPfusion	http://soap.genomics.org.cn/SOAPfusion.html	Wu et al. [24]
SAOPfuse	http://soap.genomics.org.cn/soapfuse.html	Jia et al. [25]
FusionCatcher	https://code.google.com/p/fusioncatcher/	Nicorici et al. [13]
ViralFusionSeq	https://sourceforge.net/projects/viralfusionseq/	Li et al. [26]
PRADA	http://bioinformatics.mdanderson.org/main/PRADA:Overview	Torres-Garcia et al. [27]
Chimera	http:///bioconductor.org/packages/release/bioc/html/chimera.html	Beccuti et al. [28]
TRUP	https://github.com/ruping/TRUP	Fernandez et al. [29]
De novo assembly algorithms		
EBARDenovo	https://sourceforge.net/projects/ebardenovo/	Chu et al. [8]
Trinity	http://trinityrnaseq.sourceforge.net/	Grabherr et al. [30]
Trans-ABySS	http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss	Robertson et al. [31]
Oases	https://www.ebi.ac.uk/~zerbino/oases/	Schulz et al. [32]

An updated list of RNA-Seq tools is maintained at http://fusion_gene.aiplab.net/index.php/Tools

3. EBARDenovo does not limit the minimal length of output contigs by default. Many assemblers only output contigs whose lengths are greater than 100 bp or 200 bp despite short contigs may contain extra information of gene expression profile. The parameter "-c 200" of EBARDenovo will limit the minimal contig length greater than or equal to 200 bp, or the outputs can be filtered by the small program ContigC200.exe after the assembly procedure.
4. One of the advantages of the proposed assembly-based method is the potential output of full-length chimeric transcripts which can be used to investigate the function of synthesized chimeric protein. The Blast result of the contig for the full-length chimeric transcript ARFGEF2-SULF2 is shown in Fig. 2.
5. Since it is difficult to discriminate transcription-induced chimeras from adapter-induced chimeras, wet laboratory is necessary to verify the detected results by any fusion finding method. The TranslocCheck program outputs ace files for the visualization of read alignment cross the splice junctions (Fig. 3).

4 Notes

1. The up-to-date (Jan 31, 2015) statistics of RNA-Seq experiments from the NCBI Sequence Read Archive http://www.ncbi.nlm.nih.gov/sra/ includes 138,491 experiments using Illumina platforms, 535 experiments using Ion Torrent platforms, and 7714 experiments using SOLiD platforms. More than 90 % of total RNA-Seq experiments were performed on Illumina platforms. Illumina currently have four series of sequencing systems: MiSeq, NextSeq, HiSeq, and HiSeq X. Illumina RNA-Seq data are used to find chimeric transcripts most frequently.
2. Illumina sequencing requires that the molecules to be determined are converted into special sequencing libraries. This is achieved by adding specific adapter sequences on both ends of fragments (called inserts). The majority of adapter dimers are removed by a gel excision step after library preparation. Sometimes, two or more short segments are combined into an insert. It is called adapter chimeras. Sequencing of these inserts produces aberrant chimeric reads.
3. Aberrant chimeric reads from adapter-induced amplicons influence the sensitivity of fusion finding methods. Accordingly, most of the fusion finding algorithms employ variant filtering

a Blast result for the left part of Contig26862

Homo sapiens sulfatase 2 (SULF2), transcript variant 3, mRNA
Sequence ID: ref|NM_001161841.1| **Length:** 4248 **Number of Matches:** 1

Range 1: 845 to 4190 GenBank Graphics ▼ Next Match ▲ Pre

Score	Expect	Identities	Gaps	Strand
6410 bits(3334)	0.0	3342/3346(99%)	0/3346(0%)	Plus/Minus

```
Query  1     TGCTGGTCATTGGCTGGGGTTTTGAACACTGTATGACAATACTTAAACTACAAAAttttt  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4190  TGCTGGTCATTGGCTGGGGTTTTGAACACTGTATGACAATACTTAAACTACAAAATTTTT  4131

Query  61    tttACAAATACAGTTTCATCAAAACTTGGGACATACATCAACTTCATTTCTTTTCAGTAC  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4130  TTTACAAATACAGTTTCATCAAAACTTGGGACATACATCAACTTCATTTCTTTTCAGTAC  4071

Query  121   CTTaaaaaaaaaaCATCAGTTCTGGGACATAACAAAGAAATACTAGGAGAAATGGTATCT  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4070  CTTAAAAAAAAAAACATCAGTTCTGGGACATAACAAAGAAATACTAGGAGAAATGGTATCT  4011

Query  181   GGACAGGAACAGAAATGTCCACAACTGCGAGGGATTTTCTTTTACACTGGCCACAGAGCG  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  4010  GGACAGGAACAGAAATGTCCACAACTGCGAGGGATTTTCTTTTACACTGGCCACAGAGCG  3951

Query  241   TTTATTGACACCACCACTCCTGAAAATTGGGATTTCTTATTAGGTTCCCCTAAAAGTTCC  300
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3950  TTTATTGACACCACCACTCCTGAAAATTGGGATTTCTTATTAGGTTCCCCTAAAAGTTCC  3891

Query  301   CATGTTGATTACATGTAAATAGTCACATATATACAATGAAGGCAGTTTCTTCAGAGGCAA  360
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3890  CATGTTGATTACATGTAAATAGTCACATATATACAATGAAGGCAGTTTCTTCAGAGGCAA  3831

Query  361   CCAGGGTTTATAGTGCTAGGTAAATGTCATCTCTTTTGTGCTACTGACTCATTGTCAAAC  420
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3830  CCAGGGTTTATAGTGCTAGGTAAATGTCATCTCTTTTGTGCTACTGACTCATTGTCAAAC  3771
```

b Blast result for the right part of Contig26862

Homo sapiens ADP-ribosylation factor guanine nucleotide-exchange factor 2 (brefeldin A-inl
chromosome 20
Sequence ID: ref|NG_011490.1| **Length:** 121956 **Number of Matches:** 1

Range 1: 5007 to 5273 GenBank Graphics ▼ Next Match ▲ Pre

Score	Expect	Identities	Gaps	Strand
514 bits(267)	4e-142	267/267(100%)	0/267(0%)	Plus/Minus

```
Query  1     CGAGCGCCACCTGGCAGGCCCTGCGCAGCTGGGAGTGCTGGGGCCGCTTCACCTCCTTGT  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5273  CGAGCGCCACCTGGCAGGCCCTGCGCAGCTGGGAGTGCTGGGGCCGCTTCACCTCCTTGT  5214

Query  61    CGGCTAGGATCTTCTCCAGGGCCCGGGACACGAACATGCTCTTGGTCTGGCTCTCCTGCA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5213  CGGCTAGGATCTTCTCCAGGGCCCGGGACACGAACATGCTCTTGGTCTGGCTCTCCTGCA  5154

Query  121   TggccccggccccggcgggggctgacggccccgcgggcgggcggcgTGAGCGAGATGGCGA  180
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5153  TGGCCCCGGCCCGGCGGGGGCTGACGGCCCCGCGGGCGGGCGGCGTGAGCGAGATGGCGA  5094

Query  181   GCTAGGCTGCGGGCCCGGCGTCCCGGCCGGCACCGGCCGCGTCCGTCCGCGCCGCCCCGT  240
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  5093  GCTAGGCTGCGGGCCCGGCGTCCCGGCCGGCACCGGCCGCGTCCGTCCGCGCCGCCCCGT  5034

Query  241   CAGGAAGCGACACCTCGGCCCCACGGC  267
             |||||||||||||||||||||||||||
Sbjct  5033  CAGGAAGCGACACCTCGGCCCCACGGC  5007
```

Fig. 2 Blast results of the assembled contig 26862 by EBARDenovo from the MCF-7 dataset (SRX025827)

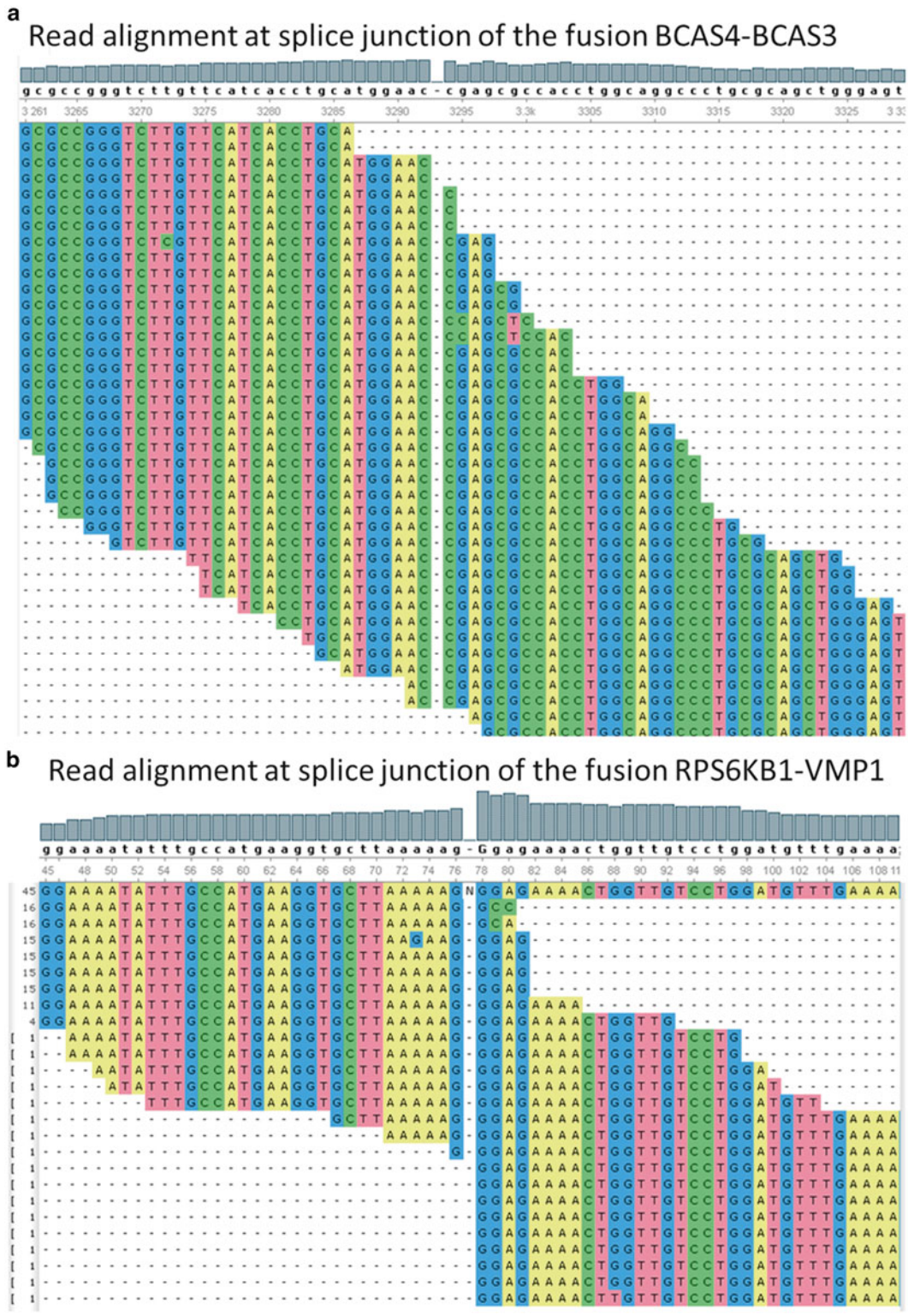


Fig. 3 Read alignments for the fusions BCAS4-BCAS3 and RPS6KB1-VMP1

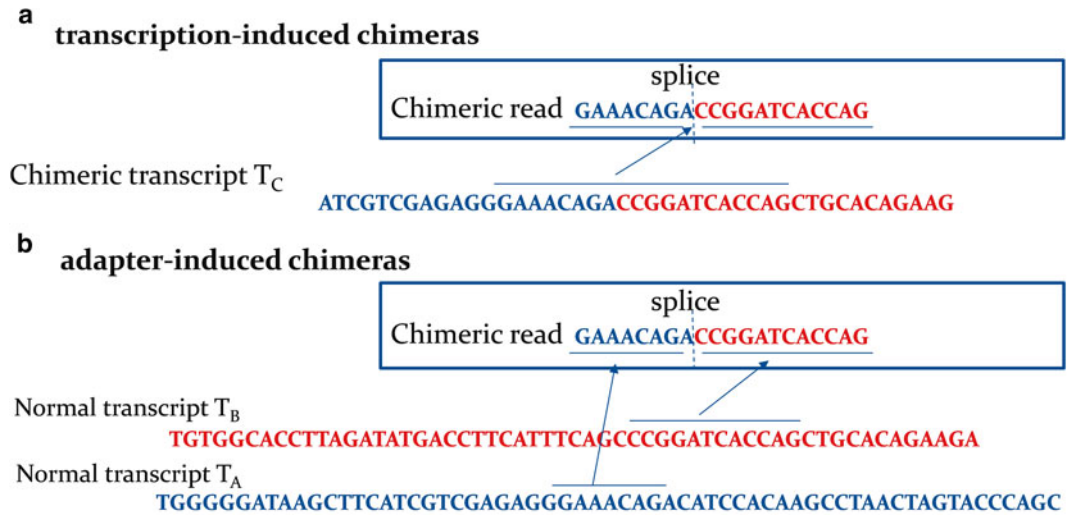


Fig. 4 Two types of chimeric reads: (**a**) transcription-induced chimeras. (**b**) Adapter-induced chimeras

criteria to get rid of false-positive detections. For example, FusionCatcher removes reads which align on ribosomal/transfer RNA and mitochondrial genes [13]; FusionMap has a white lists and a black list of gene names to filtering aberrant chimeras [14]. Different rules may reduce the sensitivity and specificity of fusion identification (Fig. 4).

4. The chimera detection test of the EBARDenovo assembler is performed on a candidate read which can be aligned into a contig and extend it. Assuming that there is a chimeric read R_c composed of two fragments from transcripts T_A and T_B, the correct assembly should always have higher read coverage than the contig containing the chimeric reads. In this case, the back-tracking from the other end of chimeric read R_c will find a set of correct reads from transcript T_B and the alignment of these reads is compared with the nucleotide bases of the current contig to determine whether the program can correctly expand the contig or cause a translocation of the two segments. The detected chimeric reads are not aligned into any contig and output to a chimera file later (Fig. 5).
5. The availability of a subset of RNA-Seq tools, including the many fusion finding programs and state-of-the-art de novo assembly programs, is shown in Table 1.

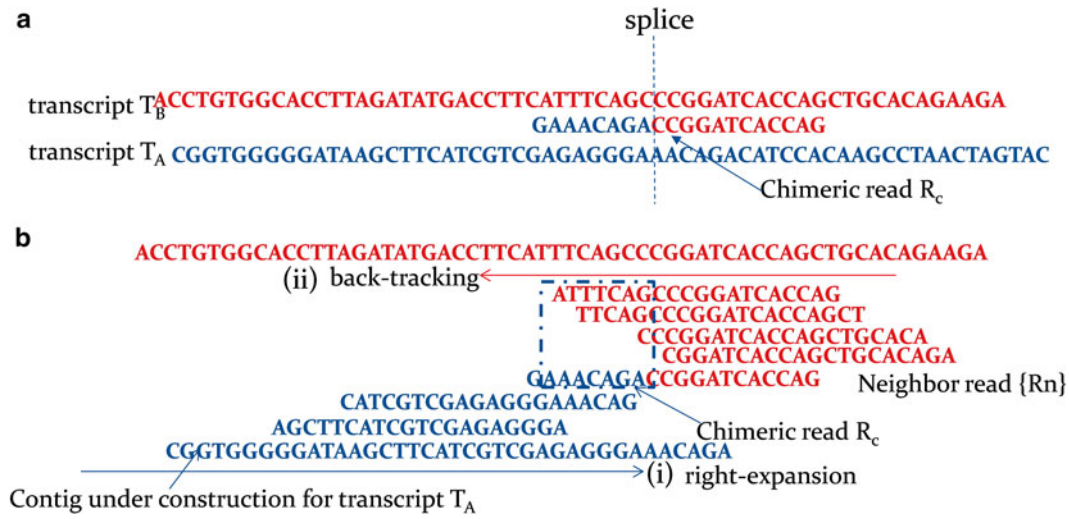


Fig. 5 Chimera detection of de novo assembly for removal of potential adapter-induced chimeras. The EBARDenovo program performs a back-tracking process if a lowly abundant read can be aligned to a constructing contig. The aligned reads from the back-tracking process are compared with the contig to detect potential chimeric junctions. In brief, a read is segregated if the alignment of its right neighbors forms different sequence from the alignment of its left neighbors

References

1. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM (2009) Transcriptome sequencing to detect gene fusions in cancer. Nature 458(7234):97–101. doi:10.1038/nature07638
2. de Klein A, van Kessel AG, Grosveld G, Bartram CR, Hagemeijer A, Bootsma D, Spurr NK, Heisterkamp N, Groffen J, Stephenson JR (1982) A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukaemia. Nature 300(5894):765–767
3. Barlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi OP, Kallioniemi A (2002) Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer. Genes Chromosomes Cancer 35(4):311–317. doi:10.1002/gcc.10121
4. Mitelman F, Johansson B, Mertens F (2007) The impact of translocations and gene fusions on cancer causation. Nat Rev Cancer 7(4):233–245. doi:10.1038/nrc2091
5. Meyerson M, Gabriel S, Getz G (2010) Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 11(10):685–696. doi:10.1038/nrg2841
6. Kircher M, Heyn P, Kelso J (2011) Addressing challenges in the production and analysis of illumina sequencing data. BMC Genomics 12:382. doi:10.1186/1471-2164-12-382
7. Carrara M, Beccuti M, Lazzarato F, Cavallo F, Cordero F, Donatelli S, Calogero RA (2013) State-of-the-art fusion-finder algorithms sensitivity and specificity. Biomed Res Int 2013:340620. doi:10.1155/2013/340620
8. Chu HT, Hsiao WW, Chen JC, Yeh TJ, Tsai MH, Lin H, Liu YW, Lee SA, Chen CC, Tsao TT, Kao CY (2013) EBARDenovo: highly accurate de novo assembly of RNA-Seq with efficient chimera-detection. Bioinformatics 29(8):1004–1010. doi:10.1093/bioinformatics/btt092
9. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, Rye IH, Nyberg S, Wolf M, Borresen-Dale AL, Kallioniemi O (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 12(1):R6. doi:10.1186/gb-2011-12-1-r6
10. Kim D, Salzberg SL (2011) TopHat-Fusion: an algorithm for discovery of novel fusion transcripts.

Genome Biol 12(8):R72. doi:10.1186/gb-2011-12-8-r72

11. Wu TD, Watanabe CK (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21(9):1859–1875. doi:10.1093/bioinformatics/bti310
12. Okonechnikov K, Golosova O, Fursov M (2012) Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics 28(8):1166–1167. doi:10.1093/bioinformatics/bts091
13. Nicorici D, Satalan M, Edgren H, Kangaspeska S, Murumagi A, Kallioniemi O, Virtanen S, Kilkku O (2014) FusionCatcher—a tool for finding somatic fusion genes in paired-end RNA-sequencing data. bioRxiv, Nov. 2014. doi:10.1101/011650
14. Ge H, Liu K, Juan T, Fang F, Newman M, Hoeck W (2011) FusionMap: detecting fusion genes from next-generation sequencing data at base-pair resolution. Bioinformatics 27(14):1922–1928. doi:10.1093/bioinformatics/btr310
15. Fan X, Abbott TE, Larson D, Chen K (2014) BreakDancer—identification of genomic structural variation from paired-end read mapping. Curr Protoc Bioinformatics 2014. doi:10.1002/0471250953.bi1506s45
16. Sboner A, Habegger L, Pflueger D, Terry S, Chen DZ, Rozowsky JS, Tewari AK, Kitabayashi N, Moss BJ, Chee MS, Demichelis F, Rubin MA, Gerstein MB (2010) FusionSeq: a modular framework for finding gene fusions by analyzing paired-end RNA-sequencing data. Genome Biol 11(10):R104. doi:10.1186/gb-2010-11-10-r104
17. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38(18), e178. doi:10.1093/nar/gkq622
18. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, Griffith M, Heravi Moussavi A, Senz J, Melnyk N, Pacheco M, Marra MA, Hirst M, Nielsen TO, Sahinalp SC, Huntsman D, Shah SP (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 7(5):e1001138. doi:10.1371/journal.pcbi.1001138
19. Li Y, Chien J, Smith DI, Ma J (2011) FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. Bioinformatics 27(12):1708–1710. doi:10.1093/bioinformatics/btr265
20. Asmann YW, Hossain A, Necela BM, Middha S, Kalari KR, Sun Z, Chai H-S, Williamson DW, Radisky D, Schroth GP, Kocher J-PA, Perez EA, Thompson EA (2011) A novel bioinformatics pipeline for identification and characterization of fusion transcripts in breast cancer and normal cell lines. Nucleic Acids Res. doi:10.1093/nar/gkr362
21. Iyer MK, Chinnaiyan AM, Maher CA (2011) ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics 27(20):2903–2904. doi:10.1093/bioinformatics/btr467
22. Francis RW, Thompson-Wicking K, Carter KW, Anderson D, Kees UR, Beesley AH (2012) FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. PLoS One 7(6), e39987. doi:10.1371/journal.pone.0039987
23. Piazza R, Pirola A, Spinelli R, Valletta S, Redaelli S, Magistroni V, Gambacorti-Passerini C (2012) FusionAnalyser: a new graphical, event-driven tool for fusion rearrangements discovery. Nucleic Acids Res 40(16), e123. doi:10.1093/nar/gks394
24. Wu J, Zhang W, Huang S, He Z, Cheng Y, Wang J, Lam TW, Peng Z, Yiu SM (2013) SOAPfusion: a robust and effective computational fusion discovery tool for RNA-seq reads. Bioinformatics 29(23):2971–2978. doi:10.1093/bioinformatics/btt522
25. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, Yu Y, Zhu D, Nickerson ML, Wan S, Liao X, Zhu X, Peng S, Li Y, Wang J, Guo G (2013) SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. Genome Biol 14(2):R12. doi:10.1186/gb-2013-14-2-r12
26. Li JW, Wan R, Yu CS, Co NN, Wong N, Chan TF (2013) ViralFusionSeq: accurately discover viral integration events and reconstruct fusion transcripts at single-base resolution. Bioinformatics 29(5):649–651. doi:10.1093/bioinformatics/btt011
27. Torres-Garcia W, Zheng S, Sivachenko A, Vegesna R, Wang Q, Yao R, Berger MF, Weinstein JN, Getz G, Verhaak RG (2014) PRADA: pipeline for RNA sequencing data analysis. Bioinformatics 30(15):2224–2226. doi:10.1093/bioinformatics/btu169
28. Beccuti M, Carrara M, Cordero F, Lazzarato F, Donatelli S, Nadalin F, Policriti A, Calogero RA (2014) Chimera: a Bioconductor package for secondary analysis of fusion products. Bioinformatics 30(24):3556–3557. doi:10.1093/bioinformatics/btu662
29. Fernandez-Cuesta L, Sun R, Menon R, George J, Lorenz S, Meza-Zepeda LA, Peifer M, Plenker D, Heuckmann JM, Leenders F,

Zander T, Dahmen I, Koker M, Schottle J, Ullrich RT, Altmuller J, Becker C, Nurnberg P, Seidel H, Bohm D, Goke F, Ansen S, Russell PA, Wright GM, Wainer Z, Solomon B, Petersen I, Clement JH, Sanger J, Brustugun OT, Helland A, Solberg S, Lund-Iversen M, Buettner R, Wolf J, Brambilla E, Vingron M, Perner S, Haas SA, Thomas RK (2015) Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. Genome Biol 16(1):7. doi:10.1186/s13059-014-0558-0

30. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29(7):644–652. doi:10.1038/nbt.1883
31. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I (2010) De novo assembly and analysis of RNA-seq data. Nat Methods 7(11):909–912. doi:10.1038/nmeth.1517
32. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28(8):1086–1092. doi:10.1093/bioinformatics/bts094

INDEX

Robert Grützmann and Christian Pilarsky (eds.), *Cancer Gene Profiling: Methods and Protocols*, Methods in Molecular Biology, vol. 1381, DOI 10.1007/978-1-4939-3204-7, © Springer Science+Business Media New York 2016

I

L

M

N

P

Q

R

S

T

V

W

X