

Methods in
Molecular Biology 1168

Springer Protocols

Ronald Trent *Editor*

Clinical Bioinformatics

Second Edition

 Humana Press

METHODS IN MOLECULAR BIOLOGY

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

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

Clinical Bioinformatics

Second Edition

Edited by

Ronald Trent

*Department of Medical Genomics, Royal Prince Alfred Hospital and
Sydney Medical School, University of Sydney, Camperdown, Australia*

 **Humana Press**

Editor

Ronald Trent
Department of Medical Genomics
Royal Prince Alfred Hospital and Sydney Medical School
University of Sydney
Camperdown, Australia

ISSN 1064-3745 ISSN 1940-6029 (electronic)
ISBN 978-1-4939-0846-2 ISBN 978-1-4939-0847-9 (eBook)
DOI 10.1007/978-1-4939-0847-9
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2014939636

© Springer Science+Business Media New York 2008, 2014

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Preface

Developments in *omics* are now impacting on patient care. They are being driven by increasingly more sophisticated analytic platforms allowing changes in diagnostic strategies from the traditional focus on a single or a small number of analytes (DNA, RNA, metabolites, proteins) to what might be possible when large numbers or all analytes are measured. These developments come at a price as the data sets generated are growing exponentially with megabytes now giving way to gigabytes or terabytes as routine outputs. Bioinformatics has arrived as a component of patient care as shown in somatic and germ-line DNA testing, and, more recently metabolic medicine through metabolomics. With genomics, samples can be sent to distant centralized analytic facilities leading to faster and cheaper DNA sequencing thereby shifting the focus even more to bioinformatics at the laboratory–patient interface. This interface is particularly relevant in patient care where understanding the clinical significance of data generated remains a significant roadblock. While a few years ago the focus was on *how to generate large data sets*, today the questions revolve around *what do the data mean?* This is a significant challenge for the medical testing laboratory particularly as the time taken for translation of novel findings from research to clinical practice shortens.

I would like to acknowledge the help of Carol Yeung in the preparation of this edition.

Sydney, Camperdown, Australia

Ronald Trent

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>ix</i>
1 From the Phenotype to the Genotype via Bioinformatics <i>Cali E. Willet and Claire M. Wade</i>	1
2 Production and Analytic Bioinformatics for Next-Generation DNA Sequencing <i>Richard James Nigel Allcock</i>	17
3 Analyzing the Metabolome <i>Francis G. Bowling and Mervyn Thomas</i>	31
4 Statistical Perspectives for Genome-Wide Association Studies (GWAS). <i>Jennifer H. Barrett, John C. Taylor, and Mark M. Iles</i>	47
5 Bioinformatics Challenges in Genome-Wide Association Studies (GWAS) <i>Rishika De, William S. Bush, and Jason H. Moore</i>	63
6 Studying Cancer Genomics Through Next-Generation DNA Sequencing and Bioinformatics <i>Maria A. Doyle, Jason Li, Ken Doig, Andrew Fellowes, and Stephen Q. Wong</i>	83
7 Using Bioinformatics Tools to Study the Role of microRNA in Cancer <i>Fabio Passetti, Natasha Andressa Nogueira Jorge, and Alan Durham</i>	99
8 Chromosome Microarrays in Diagnostic Testing: Interpreting the Genomic Data <i>Greg B. Peters and Mark D. Pertile</i>	117
9 Bioinformatics Approach to Understanding Interacting Pathways in Neuropsychiatric Disorders <i>Ali Alawieh, Zahraa Sabra, Amaly Nokkari, Atlal El-Assaad, Stefania Mondello, Fadi Zaraket, Bilal Fadlallah, and Firas H. Kobeissy</i>	157
10 Pathogen Genome Bioinformatics <i>Vitali Sintchenko and Michael P.V. Roper</i>	173
11 Setting Up Next-Generation Sequencing in the Medical Laboratory <i>Bing Yu</i>	195
12 Managing Incidental Findings in Exome Sequencing for Research. <i>Marcus J. Hinchcliffe</i>	207
13 Approaches for Classifying DNA Variants Found by Sanger Sequencing in a Medical Genetics Laboratory <i>Pak Leng Cheong and Melody Caramins</i>	227

14	Designing Algorithms for Determining Significance of DNA Missense Changes	251
	<i>Sivakumar Gowrisankar and Matthew S. Lebo</i>	
15	DNA Variant Databases: Current State and Future Directions	263
	<i>John-Paul Plazzer and Finlay Macrae</i>	
16	Natural Language Processing in Biomedicine: A Unified System Architecture Overview.	275
	<i>Son Doan, Mike Conway, Tu Minh Phuong, and Lucila Ohno-Machado</i>	
17	Candidate Gene Discovery and Prioritization in Rare Diseases.	295
	<i>Anil G. Jegga</i>	
18	Computer-Aided Drug Designing.	313
	<i>Mohini Gore and Neetin S. Desai</i>	
	<i>Index</i>	323

Contributors

- ALI ALAWIEH • *Department of Neurosciences, Medical University of South Carolina, Charleston, SC, USA*
- RICHARD JAMES NIGEL ALLCOCK • *School of Pathology and Laboratory Medicine, University of Western Australia, Nedlands, WA, Australia; Pathwest Laboratory Medicine WA, Department of Diagnostic Genomics, QEII Medical Centre, Nedlands, WA, Australia*
- JENNIFER H. BARRETT • *Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, St James's University Hospital, Leeds, UK*
- FRANCIS G. BOWLING • *Biochemical Diseases, Mater Children's Hospital, South Brisbane, QLD, Australia*
- WILLIAM S. BUSH • *Department of Biomedical Informatics, Center for Human Genetics Research, Vanderbilt University Medical School, Nashville, TN, USA*
- MELODY CARAMINS • *Genetics, Primary Health Care Group/SDS Pathology, North Ryde, NSW, Australia*
- PAK LENG CHEONG • *Department of Medical Genomics, Royal Prince Alfred Hospital, Camperdown, NSW, Australia*
- MIKE CONWAY • *Division of Behavioral Medicine, University of California, San Diego, La Jolla, CA, USA*
- RISHIKA DE • *Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA*
- NEETIN S. DESAI • *Department of Biotechnology and Bioinformatics, Padmashree Dr. D.Y. Patil University, Navi Mumbai, MS, India*
- SON DOAN • *Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA*
- KEN DOIG • *Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia*
- MARIA A. DOYLE • *Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia*
- ALAN DURHAM • *Department of Computer Science, Institute of Mathematics and Statistics, Universidade de São Paulo (USP), São Paulo, SP, Brazil*
- ATLAL EL-ASSAAD • *Department of Electrical and Computer Engineering, Faculty of Engineering and Architecture, American University of Beirut, Beirut, Lebanon*
- BILAL FADLALLAH • *Department of Electrical and Computer Engineering, College of Engineering, University of Florida, Gainesville, FL, USA*
- ANDREW FELLOWES • *Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia*
- MOHINI GORE • *Department of Biotechnology and Bioinformatics, Padmashree Dr. D.Y. Patil University, Navi Mumbai, MS, India*
- SIVAKUMAR GOWRISANKAR • *Partners HealthCare Center for Personalized Genetic Medicine, Cambridge, MA, USA*
- MARCUS J. HINCHCLIFFE • *Department of Medical Genomics B65L6, Royal Prince Alfred Hospital, Camperdown, NSW, Australia; Sydney Medical School, University of Sydney, Sydney, NSW, Australia*
- MARK M. ILES • *Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, St James's University Hospital, Leeds, UK*

- ANIL G. JEGGA • *Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA; Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA; Department of Computer Science, University of Cincinnati College of Engineering, Cincinnati, OH, USA*
- NATASHA ANDRESSA NOGUEIRA JORGE • *Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ, Brazil*
- FIRAS H. KOBEISSY • *Department of Biochemistry and Molecular Genetics, American University of Beirut, Beirut, Lebanon*
- MATTHEW S. LEBO • *Partners HealthCare Center for Personalized Genetic Medicine, Cambridge, MA, USA; Department of Pathology, Harvard Medical School, Boston, MA, USA*
- JASON LI • *Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia*
- FINLAY MACRAE • *Department of Colorectal Medicine and Genetics, Royal Melbourne Hospital, RMH, Parkville, VIC, Australia*
- STEFANIA MONDELLO • *Department of Neurosciences, University of Messina, Messina, Italy*
- JASON H. MOORE • *Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA*
- AMALY NOKKARI • *Department of Biochemistry and Molecular Genetics, American University of Beirut, Beirut, Lebanon*
- LUCILA OHNO-MACHADO • *Division of Biomedical Informatics, University of California, San Diego, La Jolla, CA, USA*
- FABIO PASSETTI • *Bioinformatics Unit, Clinical Research Coordination, Instituto Nacional de Câncer (INCA), Rio de Janeiro, RJ, Brazil*
- MARK D. PERTILE • *Victorian Clinical Genetics Services, Murdoch Childrens Research Institute, Parkville, VIC, Australia*
- GREG B. PETERS • *Sydney Genome Diagnostics, The Childrens Hospital at Westmead, Westmead, NSW, Australia*
- TU MINH PHUONG • *Department of Computer Science, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam*
- JOHN-PAUL PLAZZER • *Department of Colorectal Medicine and Genetics, Royal Melbourne Hospital, RMH, Parkville, VIC, Australia*
- MICHAEL P.V. ROPER • *Centre for Infectious Diseases and Microbiology – Public Health, Pathology West ICPMR, Westmead Hospital C24, Westmead, NSW, Australia; Sydney Medical School, University of Sydney, Sydney, NSW, Australia*
- ZAHRAA SABRA • *Department of Electrical and Computer Engineering, Faculty of Engineering and Architecture, American University of Beirut, Beirut, Lebanon*
- VITALI SINTCHENKO • *Centre for Infectious Diseases and Microbiology – Public Health, Pathology West ICPMR, Westmead Hospital C24, Westmead, NSW, Australia; Sydney Medical School, University of Sydney, Sydney, NSW, Australia*
- JOHN C. TAYLOR • *Section of Epidemiology and Biostatistics, Leeds Institute of Cancer and Pathology, University of Leeds, St James's University Hospital, Leeds, UK*
- MERVYN THOMAS • *Emphron Informatics Pty Ltd, Toowong, QLD, Australia*
- CLAIRE M. WADE • *Faculty of Veterinary Science, University of Sydney, Sydney, NSW, Australia*
- CALI E. WILLET • *Faculty of Veterinary Science, University of Sydney, Sydney, NSW, Australia*

STEPHEN Q. WONG • *Peter MacCallum Cancer Centre, East Melbourne, VIC, Australia*

BING YU • *Department of Medical Genomics, Royal Prince Alfred Hospital, Camperdown, NSW, Australia; Sydney Medical School, University of Sydney, Sydney, NSW, Australia*

FADI ZARAKET • *Department of Electrical and Computer Engineering, Faculty of Engineering and Architecture, American University of Beirut, Beirut, Lebanon*

Chapter 1

From the Phenotype to the Genotype via Bioinformatics

Cali E. Willet and Claire M. Wade

Abstract

Moving a project from the status of observing a trait of interest to identifying the underlying causal variant is a challenging task requiring a series of bioinformatics procedures and ideally the availability of a suitable reference genome sequence and its associated resources. We visit common practices for discovering the biology underlying observed traits in mammals.

Key words Association analysis, Bioinformatics, Candidate gene, Causal variant, Exome, Filtering, Gene mapping, Mutation detection, Sequencing, Whole-genome sequence

Abbreviations

bp	Base pair
CNV	Copy number variant
GC	Guanine–cytosine
GWAS	Genome-wide association study
Indel	Insertion deletion
Kb	Kilobase
Mb	Megabase
NGS	Next-generation sequencing
RAD	Restricted site-associated DNA
SNP	Single-nucleotide polymorphism

1 Introduction

Low-cost DNA sequencing has revolutionized our ability to locate and characterize the mutations that are responsible for many inherited disorders in mammals. The approaches taken to discover genetic alterations that underpin characteristics or disorders of interest by individual researchers are heavily influenced by access to technology and access to samples that reflect the phenotype of interest. In the ideal world, researchers would have at their disposal

unlimited samples, time, and money. Dealing with the limitations of the data is one of the primary challenges of science, but with developments in genomic tools and resources this task is becoming far easier.

In human medicine, we are driven to discover the mutations underlying observed phenotypes primarily by a desire to understand the biology of how the human body functions. For disorder phenotypes, this biological understanding facilitates the discovery of new interventions or improved treatment regimens. Screening for known disease mutations can enable parents to make informed reproductive decisions or prepare for potential disorders in their offspring. Similarly in animal species, identifying the mutations underlying traits of interest can enable targeted treatment options for disorder phenotypes, and the results of genetic tests can be utilized by animal breeders to enable better breeding decisions. The efficacy of applying this information within animal populations is affected by several factors including the complexity of the inheritance of the phenotypes and the quality of the test being applied.

In this chapter we outline the methods currently used for determining the genotypic changes that underlie observed phenotypes.

2 Materials

2.1 *Genetic Marker Discovery*

In most cases pinpointing the genetic basis of an observed phenotype begins with gene mapping, a process which identifies a location within the affected individual's genome that is significantly more likely to harbor the causal variant for the trait of interest than the remainder of the genome. Gene mapping is not possible without first having a means of distinguishing unique DNA signatures in individuals with and without the characteristic being assessed. This requires the discovery of genetic markers at known coordinates in the genome of the species under investigation. The most commonly assessed type of genetic marker in recent years has been the single-nucleotide polymorphism (SNP). These single-base differences in DNA sequence are the most abundant form of genetic variant, with approximately four to six million SNPs within mammalian genomes. Their high frequency and the ease of which they can be genotyped on modern high-throughput multiplexed systems make them ideal for gene mapping, enabling a higher density and thus statistical power than previously used restriction fragment length polymorphisms and microsatellites.

SNP markers are discovered through the comparison of DNA sequences of different individuals from the same species or from the comparison of the maternal and paternal chromosomes within the same individual. SNPs most commonly have two possible alleles, although triallelic SNPs do occur at low frequency.

Triallelic SNPs are less useful for mapping and so are usually discarded. Typically a reference genome (*see* Subheading 2.4) provides one allele, and the discovery of alternate alleles in the population is provided by the alignment of sequences from other individuals of the same species to the reference. The quality of the nucleotide calling in both the reference and the query sequence is of paramount importance to the expected validation rate of the markers discovered this way. Typical quality metrics interrogate the proximity of the marker to others nearby, assess the reference sequence for the presence of known repetitive elements which can lead to false-positive SNP calls, ensure that the alignment of the query sequence is uniquely placed on the reference, and assess the base calling quality assigned by the sequencing technology to the individual nucleotides in the aligned sequence reads. Similar methods can be used to discover other markers, such as microsatellites and small insertions or deletions (indels) although these are relatively less abundant and so provide lower resolution in the mapping process. Rare variants are not informative for gene mapping, so a wider sample of individuals are genotyped at potential markers to estimate population polymorphism rates and only those with minor allele frequency of >10 % are carried forward for inclusion in the gene mapping resource.

2.2 Genotyping Arrays

Genotyping arrays offer collections of genetic markers (most commonly these are SNPs) that are assessed simultaneously in a single individual. DNA probes complementary to the target sequence containing the SNP are physically anchored to the surface of microarrays at known coordinates. Single-stranded genomic DNA from the individual of interest is applied to the surface of the array. Hybridization or failure to hybridize is detected by imaging software and indicates the genotype at each SNP marker. Such biological devices are capable of rapidly assessing up to many millions of SNPs in a single experiment. Many competing commercial platforms for SNP genotyping are available, not only for human genomics but also for many animal species notably including the domestic dog, domestic cat, horse, cattle, and sheep. Most platforms offer sample multiplexing, with numerous individuals assayed at the same markers simultaneously. Genotyping accuracy is generally >99 % and genotyping rate >95 % depending on the quality of input DNA. The ability to quickly and accurately genotype multiple individuals at numerous genomic locations in a single experiment makes genotyping arrays a very affordable resource, with cost per sample currently less than a few hundred US dollars depending on platform, sample size and service provider.

2.3 Association Mapping Software

Association analysis or genome-wide association analysis (GWAS) (*see* Subheading 3.1) is a bioinformatics approach to gene mapping. Data generated from genotyping arrays or other genotyping methodologies are analyzed statistically to isolate a candidate

genomic region for the causal variant of interest. A number of freely available algorithms enable the conduct of association analyses including PLINK [1], Emma-X [2], and GCTA [3]. Extensive documentation for each tool can be found from the software websites.

2.4 Reference Genome Sequence

The post-Sanger sequencing technologies developed over the last decade, referred to as next-generation sequencing (NGS), have revolutionized our ability to sequence entire genomes for relatively small costs. Thanks to these technologies and improved bioinformatics strategies and algorithms enabling assembly of shorter sequence fragments, reference genome sequences are now available for over a thousand species, with this number increasing on a daily basis. Up-to-date information on sequenced genomes and access to related resources are available from the National Center for Biotechnology Information (www.ncbi.nlm.nih.gov/genome).

A reference genome comprises the entire collection of DNA of a single representative individual of a species. While all mammals have a maternal and paternal set of chromosomes, the reference assembly is presented as a single *reference* chromosome which may be made up of a combination of haplotypes from each parental set. Although the goal of an assembled reference genome is to portray the complete sequence along each chromosome, current technological limitations prevent this. Regions of the genome which are high in guanine and cytosine (GC) content are difficult to sequence, leaving gaps in the genome particularly around first exons of genes which are notoriously high in GC. Repetitive genomic features, representing as much as 50 % of the DNA content of mammalian genomes, are difficult to assemble accurately, creating many gaps in the genome represented as strings of *N* for unknown nucleotides.

The next or *third* generation of sequencers will ameliorate these limitations to some extent by sequencing single DNA molecules in a theoretically unbroken chain of sequence, in contrast to current platforms that first amplify fragmented libraries of genomic DNA. While current reports from Oxford Nanopore Technologies state 10,000 base pairs (bp) as the read length achieved, we are yet to see these targets attained in practice and base calling accuracy is currently much lower than next (current)-generation sequencers. It is also important to keep in mind that even if the reference individual's set of chromosomes could be entirely sequenced and seamlessly assembled, genomic variation at the nucleotide and structural levels means that there is no such thing as the perfect reference. Sequenced samples may contain DNA that is absent from the reference individual, meaning that those fragments will not be placed against the reference and vice versa.

Despite these limitations, a reference genome sequence for the species of interest is an indispensable resource in modern genetics. It serves as the foundation for genetic marker discovery, gene

mapping, primer design, comparative studies, and re-sequencing. *Re-sequencing* is the term given to sequencing the genome of an individual for which there exists a reference genome from the same species and using comparison of the sequenced samples and the reference genome to identify sequence variants. Re-sequencing is quickly becoming the go-to experimental approach for isolating the genetic basis for phenotypes of interest (discussed further in Subheading 3). For species which currently lack a reference genome, cross-species approaches can be used to variable success, generally with greater success when the study species has a closely related species with a good-quality reference genome.

The quality of a reference genome will impact its usefulness in genetic investigations. Quality is affected by many aspects but notably the level of fold coverage to which the individual was sequenced and the nature of the methods used to assemble the fragments. Typically, the best genomes are assembled using a combination of long reads to aid contig and scaffold creation and a large pool of shorter preferably paired-end reads to fill in the gaps and enhance sequence accuracy. This can be achieved using traditional approaches such as Sanger sequencing of bacterial artificial chromosomes in conjunction with NGS. Currently, a number of reference genome sequences are being released that have relied solely on NGS and these are at risk of being poorly assembled with a high frequency of gaps due to the increased challenges associated with assembling shorter fragments in the face of repetitive DNA sequence. In addition to continuity and coverage, the level of genome annotation also impacts the ease of downstream analysis.

Genome annotation refers to the informative tracks that can be placed upon a genome and include features such as the location and orientation of genes and their intron and exon boundaries, regulatory features, and stretches of sequence that show conservation across mammalian genomes. Genome annotation allows the researcher to quickly identify potentially important features within a region of the genome mapped by GWAS and often drives experimental design of the mutation detection phase of the project. Genomes are annotated through a combination of manual addition of known information as well as bioinformatics approaches including gene prediction software and comparative tools that apply information from other annotated genomes to suggest where important features may lie in the new assembly. The University of Santa Cruz genome browser is ideal for visualizing various genome annotation tracks (<http://genome.ucsc.edu/>).

2.5 Samples

Access to accurately phenotyped samples for the trait of interest is crucial to the success of a study. The numbers of case and control samples that must be evaluated to provide adequate statistical power in gene mapping are affected by the mode of inheritance and the effective sizes and lengths of linkage disequilibrium of the populations in which the trait occurs and the magnitudes of the observed effects.

Typically, recessively inherited conditions in populations with long linkage disequilibrium and low effective size such as within domestic dog breeds require the fewest samples. Strategies for determining power in association analysis are outlined in the literature [4, 5].

During the mutation detection stage, an increased number of cases and controls decrease the list of variants that segregate with the phenotype. Given our ability to sequence entire candidate regions or whole genomes, the numbers of variants discovered are numerous, and case-control filtering substantially increases our prospects of isolating the causal variant. Causal loci have been successfully identified from single-case studies (for example in refs. 6 and 7); however, both these investigations benefited from a strong set of candidate genes as well as genetic material from the unaffected parents.

The propensity of the trait to be accurately phenotyped also impacts the number of samples required. Traits which are difficult to measure, traits which may appear later in life thus leading to inappropriate designation as control rather than case, and those that are incompletely penetrant or epistatically modified, all require a greater number of samples than do projects investigating simple-to-measure and simply inherited phenotypes.

3 Methods

3.1 Association Mapping Using Genotyping Arrays

The highest chance of success in discovering the genetic change underlying a phenotype is experienced when the phenotype of interest is inherited in a simple Mendelian recessive pattern. Typically recessive deleterious mutations exist for long periods in populations because if the allele drifts to high frequency, natural selection will work to reduce the frequency again by removing homozygous affected individuals from the breeding population. If the mutant allele frequency falls to a low value, the allele can exist for long periods in the population, escaping detection because its expression relies upon the mating of two carrier individuals. Such alleles most frequently come to light when effective population sizes are reduced, when certain breeding animals are used widely in a population, or when related parents have offspring.

When the allele exists on a segment of DNA that is experiencing neutral evolution, small alterations such as SNPs and indels accumulate on the DNA and enable it to be readily distinguished as a unique genetic haplotype, or pattern, that can be readily mapped by association analysis. Depending upon the age of the allele, the specific haplotype on which it exists might span several hundreds of thousands of nucleotides in populations with modest effective population sizes. Such long segments of *linkage disequilibrium* (which refers to the physical connectivity of markers) reduce the numbers of markers that must be assessed in order to

map the trait of interest. The lengths of these segments are population specific, and having some idea of the expected length of linkage disequilibrium aids the researcher in selecting the appropriate marker density. Nowadays, this knowledge is gained through whole-genome assembly projects and commercial genotyping arrays are developed with the required density for most studies within the species. When greater resolution is required to map a trait, genotypes at additional loci can be inferred with reasonably high accuracy on the basis of haplotypes through an approach known as genotype *imputation* (reviewed in ref. 8).

Using association analysis (*see* also Chapters 4 and 5), unrelated individuals with the trait to be mapped are matched with counterparts without the trait of interest. The two groups of individuals are statistically compared at the large number of polymorphic sites assayed by genotyping arrays. The algorithms exploit linkage disequilibrium and variable allelic frequencies to identify candidate genomic regions. More sophisticated techniques can be used to either first correct the genomic region for unforeseen population structure using mixed-model methodology or make use of the regional allelic structure of haplotypes to discern better genetic differences between individuals under assessment. Detailed guidelines on using PLINK to map causal variants to a candidate genomic region using worked examples are available [9, 10].

Association analysis has enabled researchers to harness the inherent power of chromosomal recombination to identify stretches of the genome underlying phenotypes caused by common variants. However, disorders of a complex nature are recalcitrant to the use of association analysis, inspiring debate as to the relative importance of common and rare variants in the genetics of human complex genetic disorders [11].

3.2 Fine Mapping

Fine mapping is the method taken to reduce the size of a candidate genomic region by identifying associated markers that tightly flank the causal locus. Regions mapped by association analysis may contain hundreds of genes, hindering mutation detection. While gene databases can be used to refine the list of genes within the region to functional candidates, current knowledge about the function of all genes is incomplete and thus poses the risk of overlooking the gene which influences the trait of interest. Reducing the size of the candidate region and so the number of genes to explore is a good approach to avoiding over-filtering and false negatives.

To conduct fine mapping, additional markers are selected within the candidate region and genotyped in cases and controls. Custom genotyping arrays may be designed for this purpose, or samples may be genotyped using traditional molecular methods. Genotype imputation from whole-genome sequence can also be applied. The increased marker density may afford the ability to genotype the new markers in a reduced set of samples compared to

the original gene mapping analysis. Association analysis is then performed on this dense collection of markers and may reduce the size of the region by an order of magnitude, often from megabases (Mb) or hundreds of kilobases (Kb) down to tens of Kb.

3.3 Detecting Mutations Affecting Simply Inherited Phenotypes

NGS and high-throughput genomic technologies have seen a shift in the conduct of mutation detection. Nowadays, the most challenging aspect is not discovering polymorphisms but sifting through millions of variants to find that which causes the trait of interest. Identification of candidate genomic regions through GWAS can significantly reduce this task. However, prior knowledge about the condition or similar conditions, or some assumption about the nature of the causal variant, may guide researchers to select a particular method to detect mutations. Three main approaches are considered: (1) targeted sequencing of candidate genes, (2) whole-exome sequencing, and (3) whole-genome sequencing.

For candidate gene sequencing, the researcher makes assumptions about the nature of the genes involved. Whole-exome sequencing may be chosen either as a means of rapidly assaying coding variants since they are most likely to be functional or if the phenotype is one consistent with ablation or major disruption of a gene. Whole-genome sequencing makes no assumptions about the locus but has the attendant disadvantage of increased cost and bioinformatics involvement required. While each method can be employed without prior gene mapping, most commonly a candidate region has first been identified, particularly for candidate gene and whole-genome approaches.

3.3.1 Targeted (Candidate) Gene Sequencing

For many classes of disorders and phenotypes, candidate genes may be identifiable based on the biology of the trait and knowledge of gene function. Identification of candidate genes may be performed manually using a literature search or online databases such as Gene Cards (www.genecards.org/) which summarize current knowledge of the gene or Mouse Genome Browser (<http://gbrowse.informatics.jax.org/>) which describes the phenotypes of gene knockouts. This browser is a useful tool for extracting likely candidates from large genomic regions by first identifying the genomic region in mouse which is syntenic with the candidate region in the species of interest and then filtering genes in this region by those that have been found to affect a certain pathway or body system in mouse mutants. Often, the genes annotated onto the reference genome assembly within a mapped region can be quickly assessed for candidate status. For very large candidate regions or dense gene clusters harboring numerous genes, the list can be filtered bioinformatically by downloading the official gene names or accession codes for all genes within the region and searching this list against known mutation databases such as Online Mendelian Inheritance in Man or OMIM (www.omim.org/) and Online Mendelian

Inheritance in Animals or OMIA (<http://omia.angis.org.au/home/>). Candidate gene selection can be automated using tools like GeneSeeker [12] and Endeavour [13], which consider the biology of the phenotype in light of existing knowledge to rank genes according to those most likely to produce such a phenotype. One or a handful of the most likely candidate genes are then taken forward for sequencing.

Sequencing may target the coding and regulatory portions of the genes only or sequence the gene in its entirety including flanking sequence and introns. The former approach may be faster and more cost effective producing a more concise list of identified variants. The latter method has the advantage of enabling detection of noncoding variants which may influence the phenotype. The success of either approach depends on the researcher having selected the right candidate gene, and this is the main disadvantage of targeted sequencing over whole-exome or whole-genome approaches. To sequence the desired regions, researchers may design overlapping primer pairs and employ Sanger sequencing to produce a consensus sequence.

In recent years projects are more commonly utilizing NGS following the use of a sequence capture array. The reference genome sequence is used to design oligonucleotide probes complementary to the target regions, and these probes are anchored to the surface of a microarray which is hybridized with target DNA. The captured DNA is then eluted, sequenced, and aligned to the reference genome sequence, and differences from the reference are identified with variant detection software (*see* Subheading 3.3.3). While Sanger sequencing is more accurate at the individual nucleotide level than NGS, it is generally less cost efficient and may have diminished capacity to detect heterozygous indels than NGS.

3.3.2 Exome Sequencing

Exome sequencing obtains DNA sequence for all known exons within the reference genome sequence. This is typically achieved using commercially available or custom-designed sequence capture arrays which use complementary probes to capture the target DNA fragments for NGS [14, 15]. This approach has the benefit of obtaining a large amount of information from what is generally perceived as the most important fraction of the genome in a very short time frame and at low cost. The smaller dataset is an important consideration in studies with large numbers of sequenced samples, as whole-genome datasets and their analysis require incredibly large amounts of physical disk space and computing power.

The major deficiency of exome sequencing is that it can only detect mutations occurring in the actual regions targeted and is unable to accurately detect genomic insertions or rearrangements. If the causal variant happens to exist within a previously unidentified exon, within the non-coding region of the genome, or resides in the approximate 5 % of the targeted sequence that fails to be captured for sequencing it will fail to be recovered using exome sequencing. In these instances, time

and financial resources have been wasted and additional experiments are required to locate the genetic variant contributing to the trait. Despite these limitations exome sequencing has been a highly successful tool in mutation detection and diagnosis. An overview of the computational approaches to whole-exome analysis is provided in [16].

3.3.3 *Whole-Genome Sequencing*

Over the last decade, the cost per Mb of DNA sequence has fallen from >US \$5,000 to <10 cents. This has seen whole-genome sequencing emerge as the method of choice for many laboratories, both large and small. The affordability of massively parallel sequencing technologies in combination with a reference genome has made it possible to observe the DNA of individuals with very fine detail. Theoretically, every variant from SNPs through to large-scale rearrangements can be identified within each sequenced individual. In practice, this is limited by (1) the quality and content of the reference genome which the sequence is aligned to, (2) the amount and quality of sequence coverage obtained, and (3) the evenness of distribution of sequence coverage across the genome.

Another favorable aspect of whole-genome datasets is the unlimited life-span of the data. As the basic unit of biology, raw DNA sequence can be shared amongst research groups in a platform-independent manner. Existing sequence libraries can be added to over time as sequencing technology improves and current technical challenges are overcome. The ability to reuse whole-genome sequences across multiple projects helps to offset the relatively high initial cost of obtaining the data.

Given that each mammalian genome contains millions of variants, even within fine mapped candidate regions the number of potential candidate loci is impracticable without bioinformatics filtering. Even after filtering to those that segregate as expected between cases and controls, the remaining variants will number in the tens of thousands depending on the number of samples in the dataset. Researchers must choose to prioritize these segregating variants for investigation as candidates using one or a combination of methods. Mutations that reside within genes with a function likely to influence the trait are often considered first. Alternatively, bioinformatics software can be used to rank the variants according to their predicted severity on gene function.

There are likely to be a number of potentially devastating variants within such lists, and again some knowledge about the biology of the phenotype and molecular pathways involved is beneficial in choosing which variants to investigate further. It is important to keep in mind that prediction software emits both false positives and false negatives. By reducing the quantity of input variants, a well-mapped candidate region can ameliorate the temptation to over-filter results.

Experimental designs typically include sequenced case and control individuals, but researchers also have the option of sequencing case and control pools. While detailed information is lost at the

individual level, the benefit is a much reduced dataset and increased ease of identification of variants which segregate with the trait of interest. Many studies have successfully identified causal variants from pooled DNA sequences.

Given the trend towards mutation detection from whole-genome sequence data, we provide a general overview of the steps taken using this method.

1. Obtain DNA samples from cases and controls. DNA sourced from blood samples typically gives the best sequence results. However, other forms of DNA can be sequenced.
2. Submit either DNA samples or prepared DNA libraries for individuals or case and control pools to a sequencing center of choice. Larger laboratories often own NGS equipment and perform sequencing in-house. Libraries must be prepared following the recommendations of the sequencing platform used. Most commercial platforms have kits available.
3. DNA is sequenced on high-throughput sequencing machine. Most platforms offer paired-end sequencing, where the DNA fragment is sequenced from either end with an unsequenced gap in the center of approximately known size. Paired-end reads are currently 100–150 bp in length from either end of a fragment of around 500 bp. Mate pair sequencing obtains paired data from either end of fragments up to 10 Kb in length. These large insert sizes are beneficial to resolve structural variations. In either case, having sequence from two ends of a single molecule has greatly improved the ability of bioinformatics software to unambiguously place sequence reads in the face of repetitive sequence. Sequencers that produce single reads require longer read lengths to facilitate alignment.
4. Sequence is returned in fastq format, which contains the unique read identifier, the sequence read itself, as well as individual quality scores for each sequenced base. The quality scores are utilized by alignment and variant calling software. The data are often compressed, and most freely available alignment tools can process the data in this form.
5. Fastq files are aligned to the reference genome sequence. If paired, reads within the files must be retained in their original order, as read members of a pair appear on corresponding lines of the fastq files. The steps to perform alignment and the parameters applied vary depending on the alignment software used. Popular software include MAQ [17], BWA [18], Bowtie [19], SOAP2 [20], and Stampy [21]. A discussion of alignment software selection criteria is presented in [22]. For most experiments, performing the alignment with the default settings recommended by the software developers is sufficient. Some sequencing centers offer commercial alignment packages tailored to their instrument or may provide aligned data as well

as raw reads. The result of alignment is a sequence alignment map (SAM) file, which is a standardized format developed after the *Human Genome Project* to facilitate compatibility amongst datasets and analysis tools. This may be seen in the binary form (BAM).

6. Alignment processing is optional and may include removing PCR duplicates, which are evident as pairs of reads with identical outer mapping coordinates; trimming low-quality sequence or removing reads with consistently low base quality; and performing local realignment around indels. The primary alignment is conducted in a naïve fashion, one read pair at a time, frequently leading to mal-alignments and false SNP calls around indel sites. During local realignment, information from collections of adjacent and overlapping reads is considered simultaneously to resolve polymorphisms or poor-quality alignment. The Broad Institute Genome Analysis Toolkit (GATK, www.broadinstitute.org/gatk/) documents a best practice approach to alignment processing as well as variant calling.
7. Variant calling is performed either genome wide or within select regions of the genome. Many freely available tools exist to call variants. Popular software include GATK and SAMtools [23]. Most programs that call SNPs can also call indels, but different algorithms are required to identify other forms of variants such as copy number variations (CNVs), inversions, translocations, and structural rearrangements. Calling variants within the mapped candidate region only saves substantial computing time; however, for long-term data storage it is often beneficial to have genome-wide variants at hand for sequenced samples rather than the entire alignments, particularly if samples from one project are to be used as controls in another project.
8. Variant filtering for quality is an essential step. Calling software usually assigns a quality score to identified variants. These scores are based on many features including the number of reads that support the variant call, the quality of the individual base calls from the sequencing machine within these reads, the number of mismatches surrounding the variant, the allelic frequency (if multi-sample variant calling or pooled sequencing is carried out), and whether the variant is seen on both strands (strand bias can suggest a false-positive variant call triggered by repetitive sequence). Researchers can apply a hard filter when calling variants, so that only those above a certain quality threshold are reported. This greatly simplifies the bioinformatics involvement required. However, experience has shown that applying a hard filter may discard true variants. This is particularly relevant for samples sequenced to an average read depth of <math><10\times</math> (tenfold). In such cases, the standard thresholds may

be too stringent and bespoke filtering may be more effective. The complexities of variant calling and filtering with current algorithms are reviewed in [24].

9. Variant filtering for those that segregate with the phenotype is then applied to produce a list of candidate mutations. For simply inherited *recessive* traits, this is a straightforward approach of extracting all variants that are homozygous in all cases and either heterozygous or homozygous for the other allele in all controls. For *dominant* traits, the opposite is true. In both cases, it is essential to consider the level of sequence coverage at the variant site in all samples. Absent or inadequate quality sequence at a locus can cause a no call within one or a few samples, and this must be considered when filtering variants to avoid missing potential candidate loci affected by coverage issues. Identifying homozygous variants from pooled case data circumvents this issue. Coverage issues within individual samples are exacerbated when relatively low levels of sequence coverage are employed. However, even very high levels of sequence redundancy cannot overcome the difficulty in sequencing regions of high GC or sequence that has a propensity to form internal secondary structures, or in aligning regions of repetitive structure.
10. Selecting candidate causal variants: In most cases, the number of variants segregating within the candidate region are too numerous to allow investigation of each locus. Researchers must prioritize these variants for consideration as candidates. Variants within candidate genes are usually examined first although bioinformatics tools have made it possible to assess the likely functional status of very large lists of SNPs and indels. These include SIFT [25], PolyPhen-2 [26], ANNOVAR [27], SnpEff [28], and Ensembl Variant Effect Predictor (formerly SNP Effect Predictor) [29]. Coding mutations are categorized on their effect on protein sequence and structure, and non-coding variants within important genomic features such as splice sites, conserved sequence, microRNA, promoters, and various other regulatory elements are identified. The quality and quantity of genomic annotation are crucial to the extent to which these tools can identify important functional variants. While designed for human genomics, where an extremely well-annotated reference genome is available, these tools can also be applied to non-model animal datasets with varying levels of ease and accuracy. Candidate variants may also be excluded by consulting common variant databases such as dbSNP (www.ncbi.nlm.nih.gov/SNP/).
11. Investigating candidate causal variants: Plausible candidates are explored systematically by performing simple molecular genetic tests such as PCR, RFLP, or Sanger sequencing in a wider cohort of cases and controls. If the variant continues to segregate

with the phenotype, functional studies may be carried out to further support the role of the variant in the trait. A variant predicted to alter protein sequence can be tested by sequencing RNA. The impact of regulatory mutations can be assessed through gene expression array data from relevant tissues or the use of reporter gene constructs in cell culture. The impact of the variant on an organism level can be observed through induced mutant laboratory mouse models.

3.3.4 *Mutation Detection Without a Reference Genome Sequence*

While an ever-increasing number of species have reference genomes available, availability is not universal. The discovery of mutations in the absence of a reference genome is a far more complicated task that requires the application of advanced computational methods including de novo genome assembly. A competitive evaluation of current de novo methodologies is described in [30]. The task is made easier when a reference genome for a closely related species is available. Any close species reference assists the relative positioning of the assembled genomic fragments and allows the fragments from individuals with and without the phenotype of interest to be aligned and compared for genetic mapping.

If computational resources are limited, the quantity of sequence to be assembled can be substantively reduced by the application of technologies that limit genomic complexity such as restriction site-associated DNA (RAD) sequencing [31]. RAD sequencing involves first digesting the genomes to be compared with the same restriction enzyme. The resulting fragments for each sample are sized using electrophoresis, and a narrow size range is selected for genomic sequencing. The expectation is that most of the resulting fragments should be common for individuals of the same species, and when assembled, the fragments should align to common loci in the genome. This enables high-coverage sequencing of randomly ascertained loci at relatively low cost. The calling of variants within the fragments enables markers to be discovered which can be designed onto custom genotyping arrays, allowing gene mapping by association analysis to be carried out. Conserved sequence primers can also be used to amplify regions of the genome for sequencing variant discovery.

With ever-increasing computing power and concomitant bioinformatics advancements we are beginning to see a shift away from basic re-sequencing and alignment to reference-guided or reference-independent approaches. Reference-guided approaches, like that of using a closely related species to facilitate de novo genome assembly or cross-species alignment, may be adopted within species that have a reference genome assembly available. This method is useful in articulating structural rearrangements and CNVs that fail to be identified through straightforward alignment. It also allows identification of smaller variants present within tracts of sequence that may not be a part of the reference genome sequence, either

biologically or technically missing from the assembly. A reference-independent method for the detection of homozygous causal mutations from case and control pools was recently published [32]. Direct comparison of sequence reads was performed by analyzing the frequencies of substrings within reads. Developments of this and other reference-free bioinformatics techniques are likely over the coming years, not only benefiting research within species without a reference genome sequence but also offering a novel strategy to overcome the limitations associated with a reference genome.

4 Conclusion

Association-based gene mapping remains a powerful tool in mutation detection despite the ability to sequence entire genomes on an unprecedented scale. The technique complements whole-genome studies by reducing the amount of sequence to be searched for causal loci. We can dissect candidate regions using advanced computational filtering techniques to reduce the millions of genomic variants to a handful of loci most likely to impact phenotype. Given the rapid decline in sequencing costs and improvement in sequencing technologies, targeted re-sequencing of candidate genes and possibly exome sequencing are likely to be completely surpassed by whole-genome re-sequencing. As the third wave of sequencing technologies reach the market, many of the technical challenges of current whole-genome sequencing affecting regions that are difficult to sequence, assemble, align, or call variants within will be overcome, potentially making mutation detection as simple as reference-free single-molecule string comparison. In the meantime, GWAS followed by sequencing and bioinformatics ranking of candidate variants will remain a tried and true format for determining the genetic basis underlying phenotypes of interest.

References

1. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
2. Kang HM, Sul JH, Service SK et al (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348–354
3. Yang J, Lee SH, Goddard ME et al (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88:76–82
4. Purcell S, Cherny SS, Sham PC (2003) Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* 19:149–150
5. Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet* 10:381–391
6. Lupski JR, Reid JG, Gonzaga-Jauregui C et al (2010) Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. *N Engl J Med* 362:1181–1191
7. Hauswirth R, Haase B, Blatter M et al (2012) Mutations in MTF and PAX3 cause “splashed white” and other white spotting phenotypes in horses. *PLoS Genet* 8:e1002653
8. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11:499–511
9. Kijas JW (2013) Detecting regions of homozygosity to map the cause of recessively inherited disease. *Methods Mol Biol* 1019:331–345

10. Rentería ME, Cortes A, Medland SE (2013) Using PLINK for genome-wide association studies (GWAS) and data analysis. *Methods Mol Biol* 1019:193–213
11. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 11:415–425
12. van Driel MA, Cuelenaere K, Kemmeren PP et al (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res* 33:W758–W761
13. Tranchevent LC, Barriot R, Yu S et al (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res* 36:W377–W384
14. Okou DT, Steinberg KM, Middle C et al (2007) Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* 4:907–909
15. Gnirke A, Melnikov A, Maguire J et al (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* 27:182–189
16. Stitzel NO, Kiezun A, Sunyaev S (2011) Computational and statistical approaches to analyzing variants identified by exome sequencing. *Genome Biol* 12:227
17. Li H, Ruan J, Durbin RM (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18:1851–1858
18. Li H, Durbin RM (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760
19. Langmead B, Trapnell C, Pop M et al (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25
20. Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967
21. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res* 21:936–939
22. Fonseca NA, Rung J, Brazma A et al (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics* 28:3169–3177
23. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25:2078–2079
24. Nielsen R, Paul JS, Albrechtsen A et al (2011) Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 12:443–451
25. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814
26. Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249
27. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164
28. Cingolani P, Platts A, Wang LL et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6:80–92
29. McLaren W, Pritchard B, Rios D et al (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26:2069–2070
30. Earl D, Bradnam K, St John J et al (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 21:2224–2241
31. Davey JW, Hohenlohe PA, Etter PD et al (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* 12:499–510
32. Nordström KJ, Albani MC, James GV et al (2013) Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. *Nat Biotechnol* 31:325–330

Chapter 2

Production and Analytic Bioinformatics for Next-Generation DNA Sequencing

Richard James Nigel Allcock

Abstract

The bioinformatics requirements within the clinical environment are very specific, and analytic techniques need to be fit for purpose, robust, and predictable. At the same time, the bewildering amount of information produced during these analyses needs to be carefully managed, used and interpreted correctly. The challenge for clinical laboratories now is to implement production analytical processes that are capable of handling different experimental approaches on current equipment, as well as to incorporate ways for these systems to evolve to take account of developments likely to make impacts in the near future. This is complicated by the many options available at each of the critical processing steps and a clear method needs to be developed to assemble appropriate pipelines. Here, I discuss the issues relevant to the development of an informatics pipeline that meets these criteria that should allow individual laboratories to assess their proposed strategies.

Key words Annotation, Bioinformatics, DNA, Filtering, Indel, Mapping, NGS, Sequencing, SNP, SNV, Variants

Abbreviations

BAM	Binary version of SAM file
CNV	Copy number variant
NGS	Next-generation sequencing
Q	Quality score
QC	Quality control
SNV	Single nucleotide variant
SSV	Variant instance
VCF	Variant call format
WES	Whole-exome sequencing
WGS	Whole-genome sequencing

1 Introduction

The past 10 years have seen a revolution in the techniques for DNA sequencing. Where once high-throughput sequencing was the domain of a few highly specialized core facilities, next-generation sequencing (NGS) and benchtop sequencing have evolved to such a point that individual laboratories are capable of sequencing panels of genes, exomes, and even whole genomes with relatively minimal cost and effort (*see* refs. 1, 2 for reviews of sequencing technologies). This has completely transformed the research world, especially in the area of genetic diseases (inherited disease, as well as cancer). The obvious potential of NGS and its promise of the *\$1,000 genome* is so great that many clinical DNA testing laboratories have also sought to evaluate and implement NGS techniques. This represents a significant change in the approach of clinical laboratories where new techniques have only entered routine use after significant periods of stability within the research world. Aside from the obvious differences in laboratory approaches and sequencing technologies, the use of genomic techniques has been paralleled by a significant increase in the complexity of data analysis, now called *bioinformatics* (also known as *informatics* or *computational biology*).

A wide array of new methods have been developed in a very short space of time and bioinformatics is emerging as a field in its own right. This presents the other significant challenge for implementation in the clinical laboratory—mastery and assembly of a robust set of analytical techniques that will meet the criteria and standards of modern medical and diagnostic practice [3].

One of the critical features enabled by NGS is the move away from the analysis of individual genes (genetics) to analysis of much larger regions or even whole genomes (genomics). Within the clinical laboratory this is proving disruptive, with clinicians able to order very powerful tests that have significant diagnostic capability, albeit with a manyfold increase in analytical complexity, requiring much more significant input from laboratory scientists and bioinformaticians [4]. Until relatively recently, the required analytical techniques were not stable and the subject of much research in the literature. There is recent evidence that this is changing, with broad acceptance of critical techniques beginning to emerge [5–7]. However, there are still often multiple ways to implement individual methods and the particular combination of software packages needs to be resolved on an individual laboratory basis, depending on the type of testing being performed, the equipment utilized, and the computing resources available, as well as the specific requirements of medical professionals and regulatory agencies.

For the clinical laboratory, the following features are critically important:

- A stable bioinformatic/computational environment, with defined inputs and outputs. Thought should be given as to how the system can be updated/improved and the effect of differences measured and quantified. Ideally data should be analyzed along predefined, automated pathways or pipelines.
- A defined series of analytical metrics allowing tracking and troubleshooting of the process.
- The process must be tuned to support the particular instrument platforms used in the laboratory.
- The process should be designed so that alternative sequencing platforms could be used in the laboratory. This allows for a changing laboratory environment, including new equipment that is in the pipeline but does not yet exist.

There are a wide variety of laboratory approaches within the genomics discipline. A major focus of clinical laboratories will be the detection of variations in the genome, be they variations in chromosome numbers, large indels/rearrangements, and small-scale variants such as single nucleotide variants (SNV) and indels. The specific characteristics of the variants sought after by individual laboratories will determine a range of laboratory techniques that may be appropriate, including (1) Whole-genome sequencing (WGS), (2) Whole-exome sequencing (WES), (3) Targeted panel/gene sequencing, and (4) Individual gene sequencing.

With the exception of WGS, there are many different laboratory approaches for each of the other techniques, each with advantages and disadvantages, i.e., hybridization, small-scale PCR, massively parallel PCR, as well as specific ways in interpreting and troubleshooting them. Appropriate QC metrics will need to be developed by individual laboratories relevant to the approaches used.

Overall, six distinct processes can be defined leading from DNA to interpretable data:

- Sequence generation, whose primary output is sequence “reads” in a standard format, e.g., FASTQ, BAM.
- Quality control (QC).
- Alignment of reads to a reference sequence.
- Multiple analyses of the aligned data generating single nucleotide variants (SNV), CNVs, and other structural variants.
- Annotation of variants. Annotations may include gene, location, effect on amino acids, presence and frequency in relevant databases, others.
- Positive identification of known variants in genes of interest and/or negative selection (filtering) to remove unimportant variants (according to precisely defined parameters).

Some of these processes will occur without human intervention, according to previously set parameters, whilst others may require a degree of interpretation from laboratory scientists or clinicians. Some of the steps (especially **step 1**) will occur on the chosen sequencing instrument, while others may occur on a variety of computing platforms. The combination of hardware and software requirements to perform NGS is the source of much confusion. However, while there is a vast array of programs for performing individual tasks, it is important to distinguish between the underlying algorithms performing a particular task and the program in which it is implemented. There is also a mixture of commercial and open-source programs that can be used. Care should be taken to ensure that a particular program is not used merely because it is easily available, fits with a particular IT policy or other criteria—ultimately, the process must first be capable of performing the intended task.

2 Materials

2.1 *Sequence Reads*

Reads are the primary input into a bioinformatics pipeline and can be either single-end or paired-end. In addition, individual base quality scores (so called “Q” scores) are required. Reads and Q scores can be large files, are often compressed and can come in a number of different formats. The most common formats include FASTQ (a text file containing reads, together with ASCII-33 offset Q scores; [8]) and BAM (a binary compressed format often used for alignment; <http://samtools.sourceforge.net/SAMv1.pdf>; [9]). Occasionally, reads and Q scores are contained in separate files. Rapid, accurate convertors are available to transpose data between formats.

2.2 *A Read Mapper*

This is the program that will perform the mapping of reads to a reference sequence. The chosen mapper must be capable of interpreting the read format used. Some sequencing systems perform read mapping on instrument, whilst others do not, in which case an appropriate mapper needs to be run on the read file. Mappers can be “tuned” to change the stringency. Some mappers only use the base and quality information provided in the read file, whilst others take account of other data produced during primary sequence generation to enhance mapping. The specific parameters used for mapping reads will likely be very different depending on the sequencing system used, and will need to be independently determined. A variety of mappers, both proprietary and open-source, are available (*see Note 1*).

2.3 Mapping Assessment Tools

After mapping, specific evaluation of the mapping must be performed. This will be performed in association with files specifying the regions being analyzed.

2.4 Variant Callers

These are the programs that will evaluate specific sites of interest for the presence of variants, i.e., differences between the sample and the chosen reference sequence. A variety of variant callers each with advantages and disadvantages is available (*see Note 2*). Different variant callers may be required for different classes of variants and the results of individual analyses may need to be aggregated for further analysis and filtering. The emerging data format for variants is called VCF—variant call format [10]. VCF allows a variety of data fields regarding variant quality to be conveniently stored and manipulated.

2.5 Annotation and Filtering

These are the programs that will be used to add additional information to variants, as well as to identify specifically or filter out particular variants of note. A variety of programs can perform annotation and be used for filtering. Depending on the scale of the sequencing being performed, the filtering step may be performed on csv files in commonly used programs such as Microsoft Excel, whilst more sophisticated approaches will be required for larger data sets, e.g., WES, WGS (*see Note 3*).

2.6 Reference/ Database Files

Regardless of the methods and programs chosen, a variety of static files will also be required. For simplicity, wherever possible, well-curated, standardized files and databases should be used. However, as these change on a reasonably regular basis, it may be convenient to download the appropriate files from a central repository and use these static files until updated files can be tested and validated. It is also important to note that most databases and reference sequences are linked via specific versions and it is essential to obtain the correct set of files and databases. The minimum requirements are as follows:

- *A reference sequence*—the Human Genome hg19 build (GRCh37; <https://genome.ucsc.edu>) is the most commonly used reference sequence. It is likely to be superseded by hg20 (GRCh38) in future, although hg19 will continue to be usable for a substantial period. If a whole-genome reference is not to be used, laboratories should clearly identify the origin and version number of the chosen reference sequence and maintain local copies.
- *An annotation database*. A number are available and they change occasionally. The particular database chosen should be shown to contain accurate annotation for genes of interest. It is worth noting that there may be sequence differences between individual gene-specific references used in clinical laboratories

and the hg19 genome reference. There may also be differences in what is regarded as the primary transcript for a particular gene.

- *A file of regions/genes being targeted*, listed by chromosome and position. This file will be in standard BED format [11].
- *Databases of known and/or important variants*. These will be local databases, but may also contain bigger data sets acquired elsewhere, e.g., the Human Gene Mutation Database or HGMD (*see* also Chapter 15).
- *Databases of known and/or unimportant variants*. These are useful as they identify very common variants in the population. These are generally assumed to be unimportant in certain diseases (mostly rare diseases). Common databases include the 1000 genomes database (www.1000genomes.org; [12]), the Exome Sequencing Project (<http://evs.gs.washington.edu/EVS>), and the SNP database (dbSNP; www.ncbi.nlm.nih.gov/SNP).
- *Other databases containing information that might be considered useful*. Commonly used databases include precomputed matrices of scores derived from programs such as SIFT [13], Polyphen2 [14], and MutationTaster [15] which can be used to assess the possibility that if a particular variant is pathogenic or not (*see* Chapters 13 and 14 for further discussion).

3 Methods

3.1 Perform Quality Control on Reads

Having performed the required laboratory preparations and sequencing, the raw data will usually be stored in FASTQ or BAM format. Other formats are possible but are increasingly rare and it is unlikely that most clinical laboratories will make use of them for very much longer. An example is the Life Technologies SOLID-specific XSQ format, which is the equivalent of a binary compressed colorspace FASTQ file. Depending on the particular laboratory setup and choice of sequencing platform, various QC steps may have already been performed automatically. If not, they will need to be performed before proceeding. QC will usually take the form of filtering and trimming.

1. First, a number of reads may be filtered, i.e., removed entirely because of quality issues. The quality of reads can be quite variable and laboratory/platform-specific criteria for inclusion/exclusion will need to be determined.
2. Second, poor quality bases, often at the 3' end of a read pass may fall below a defined quality threshold and are automatically trimmed off as part of signal processing/base calling on the sequencer. Where this does not happen automatically,

the Q scores will need to be assessed and the reads trimmed accordingly. In some sequencing platforms, e.g., Illumina GAI, Hiseq, MiSeq, the base-quality scores decline predictably across a read and hence reads can simply be trimmed from the 3' end until the first base with a Q score above a defined threshold is reached. This trimming strategy will result in reads of uneven length and so some trimming strategies simply trim all reads back to a set length such that all bases are above a particular threshold. In other sequencing platforms, e.g., Roche GS-FLX, Life Technologies Ion Torrent, it is possible for Q scores to vary substantially across a read and trimming using such a strategy deletes most of the data at the expense of the high quality bases throughout the remainder of individual reads. Hence, in this case, read trimming is performed using a sliding window, trimming reads back only from the 3' end until a threshold average is reached. This strategy explicitly results in reads of varying lengths and with variable base qualities internal to individual reads. It is important to note that these features directly affect the choice of mapping and variant calling algorithms, as they need to be able to take account of these varying qualities. Hypothetically, if mappers and variant callers were able to completely incorporate Q scores in the statistical assessment of reads and positions, read trimming would be unnecessary.

3. Finally, it is worth noting that in the older SOLID sequencing system (being phased out, but still in use in some clinical laboratories given its extremely high individual read accuracy), reads are not trimmed at all until alignment with a reference sequence, i.e., there is no Q score-based trimming, and that specialized mappers and variant callers capable of handling colospace data are required.
4. The program FASTqc (<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) has been used to assess Q score distributions and other features of larger read sets including GC/AT distribution across reads (should be consistent across most reads) and this can also be used to assist read trimming. Regardless of the specific trimming/filtering strategy used within a particular laboratory, a number of metrics should be recorded here. These include the following features:
 - Total number of reads.
 - Total number of reads after filtering, i.e., complete removal of reads.
 - The number of bases removed by trimming.
 - The final number of reads and their average length (or some other indication of the read distribution).

For a given sequencing application, e.g., WES, these metrics should remain consistent from run to run. If not, it suggests variations in the process. The specific variations may be indicative of the part of the process in which a problem has occurred.

3.2 Map the Reads to a Reference Sequence

After QC filtering and read trimming, the reads are mapped to a reference sequence. Specific parameters for the read mapper should be set based on experiments with particular laboratory applications. For bulk mapping of reads to the hg19 human genome reference sequence, there is significant experience and guidance to be obtained from the literature, specific to individual sequencing platforms and applications. For DNA derived from blood or tissues, >95 % of reads should map to the human genome reference sequence. The percentage of reads mapping to the reference sequence should be recorded. Large reductions in this metric suggest contamination with nonhuman DNA. This is unlikely from blood and most other tissues, but a substantial problem in saliva or buccal-derived DNA [16].

For most NGS-applications, the best choice of reference sequence will usually be the entire human genome. Of course there will be some applications where an individual gene sequence is a superior choice, but this will be based on the specific characteristics of the gene(s) being analyzed. Most NGS preparative techniques (WES, multiplex PCR and so on) are inherently “noisier” than previous techniques which relied on the ability to specifically amplify a single sequence. When working with human exomes, e.g., the hybridization process is not absolute and highly similar regions can be enriched. Mapping reads to the entire genome should not be mistaken for performing whole-genome analysis. Once reads have been mapped, further analyses should be restricted to genes of interest using bed files specifying chromosomal positions of interest. The failure to allow reads to map to the appropriate genomic loci may result in reads being mis-mapped, which may cause problems in further analyses, usually because reads from an alternative locus have been forced to align to a single-gene reference sequence. This is a fundamental difference between traditional editing/alignment of Sanger chromatograms performed semi-automatically or manually, and bulk read-mapping performed computationally in NGS. Given the wide range of different local sequence-specific contexts, it is not possible to ensure that every read is completely properly mapped—only to ensure that a read is mapped to its most likely location in the genome. Greater specificity can be achieved at the cost of a substantial reduction in the number of reads mapped to a particular gene or region, with some regions essentially becoming “unmappable” because of the specific sequence context.

3.3 Determine the Distribution of Reads in Targets of Interest

After mapping, the number and proportion of reads assigned to the targeted regions/genes must be determined. This task is usually performed by the same software performing the read mapping. A number of metrics should be recorded at this step and are strongly indicative of the success or failure of the laboratory enrichment protocol. The specific ranges of value are application dependent rather than sequencing-platform dependent. Many laboratories record *average coverage* within a given set of targets. However, this metric alone is often not sufficient to guarantee consistent enrichment across multiple gene/region targets. It is often also necessary to quantify a number of other factors such as the proportion (and identity) of regions and bases with zero coverage (either failures of enrichment or failures in sequencing) or substantially lower than average coverage. Other metrics can also be recorded, such as the proportion of bases covered at greater than a predetermined threshold. The consistency of poorly performing regions should be identified during assay workup and monitored during production. One metric that has emerged as proxy for the consistency of coverage in exome sequencing is the proportion of the exome covered at $>20\times$ [17]. Most regions with $20\times$ coverage should result in reliable variant-calling, whilst regions with fewer reads are much more unreliable. The relationship between average coverage and a metric such as % base $>20\times$ is highly application dependent and should be determined and used to identify the necessary balance between appropriate coverage and volume of sequence required.

3.4 Call Variants and Merge Outputs

Generating and mapping large volumes of sequence data, even as far as sequencing an entire human genome is a relatively simple process in the laboratory. However, in clinical settings it is important to distinguish between the most efficient laboratory techniques (efficiency can be defined in terms of reducing the different number of assays, costs, staff requirements, and equipment requirements) and the genes that have been requested to be analyzed. Hence, once reads have been mapped and the analysis restricted to regions/genes of interest, the specific targets of interest can be analyzed for variants such as SNVs, indels (small and large) and CNVs. As an extreme example, it may soon be likely that the most efficient and cheapest way to analyze a single large gene might actually be to sequence the entire exome (or even entire genome) from a particular sample and then electronically extract only the information from the gene of interest. This electronic extraction is best performed at the variant-calling step, with regions or genes of interest defined using bed files.

There are a number of approaches to SNV calling and the area is still developing. Fundamentally, variant-callers examine all the reads covering a particular base and counts the number of reads containing the reference base and the number of reads containing a variant or non-reference base. The differences in approach to variant

calling result from the way different information regarding reads and bases is used. In the first approach, reads are heavily QC'ed to remove those with poor quality bases and those that map poorly. A variety of other factors can be used to exclude reads. The number of reads containing different bases at each position is then counted and assessed statistically to determine the likelihood. A second approach excludes some reads that pass below various parameter thresholds, but then also weights the remaining reads according to the values of the Q scores of individual bases covering the region. Other factors such as the presence of other variants within a read and proximity to the 3' or 5' end of a read can be used to reduce the weighting of a particular base. The emerging standard is the development of a genotype quality (similar to base Q scores) as a statistical assessment of a variant call at a particular position.

Analyses of small and large indels and/or CNVs should also be performed where required. The inputs to all of these algorithms are the BAM files produced during read-mapping. Certain analyses such as CNV-calling may require control datasets, i.e., a large number of normal samples with which a test sample can be compared.

After different kinds of variant-calling have been performed, it may be necessary to combine and aggregate the different analyses into a single output for further consideration. The file format that has emerged as a standard is VCF—<http://vcftools.sourceforge.net/specs.html> [10]. VCF files can be viewed in many software packages and the VCFTools suite of program can be used to manipulate VCF files. Metrics to be recorded in variant calling include the total number of variants, homozygous/heterozygous ratio, number of SNVs/indels, and the breakdown of variants per chromosome. Variations in these metrics can be indicative of a number of issues and problems including strand biases, poor sequencing quality, enrichment issues, as well as samples that are highly disparate from the mostly Caucasian-derived hg 19 genomes reference sequence.

3.5 Annotate and Filter Variants According to Predefined Criteria

Various analyses can be performed on BAM files to generate variant candidates. It is implicit that the generated variant candidates will contain a mixture of real variants and variants that arise as a result of specific platform, analytical, or sequence-context issues. All NGS platforms suffer from these issues [17]. Whilst this is of concern to some, the next stage in analyzing genome-wide variant is the annotation and filtering of variants to derive a prioritized list for further consideration. There are a variety of ways to perform annotation and a variety of annotations that can be added to variants, including (but not limited to) gene, location within gene, effect on amino acid, i.e., synonymous versus non-synonymous, identity and frequency within various databases (commonly HGMD, dbSNP, 1000 genomes, Exome Sequencing Project), identity and frequency within various in-house databases of known

pathogenic variants as well as known systematic sequencing errors. Once annotated, variants can then be filtered by various criteria. These criteria will vary between laboratories and even within laboratories, dependent on the characteristics of genes and/or diseases being investigated. It is critical to record the various filtering steps and the numbers of variants retained or excluded at any point. Most programs with inbuilt filters allow a series of filters to be predefined, and these can then be run automatically with the user observing the outputs from each filtering step.

Two approaches to filtering are possible.

1. Variants may be filtered against a known list of disease-causing/pathogenic variants and hence important variants in the sample are positively identified. This requires a database of such variants and there are numerous efforts to produce comprehensively annotated pathogenic variants for specific diseases and genes. This approach is relatively straightforward. Following positive identification of specific variants in genes, they may need to be confirmed by an orthogonal method such as Sanger sequencing. It is important to define the logic by which a particular variant or pair of variants is declared as pathogenic. In the past, following a clinical diagnosis, a single gene might have been sequenced and any variants within declared pathogenic simply because they were found in a gene known to cause the disease. However, the introduction of NGS also includes the much wider screening of many genes at once. In addition, databases are available to provide a wealth of information. These are particularly relevant for the second method of filtering variants.
2. The exclusion approach allows variants unlikely to be important to be removed, enriching for likely important variants. Many approaches are possible for this task and the individual workflows are beyond the scope of this article. However, they should be clearly defined, rigorously justified and implemented. Using this negative filtering approach will likely result in a list of potential gene candidates, with priorities for further investigation.

Other investigations may include showing that the inheritance of particular variants is consistent with disease status in a family as well as functional assays for particular variants.

3.6 Keeping Pace with Developments

Genomic techniques have developed rapidly over the past decade. There have been advances in almost every aspect of the process, from laboratory techniques, sequencing platforms, databases and approaches to analysis. Whilst many of these techniques are entering clinical use, they have not yet finished developing and every month brings with it a host of new changes, updates, and improvements to some or all aspects of the process. The continuing challenge for the clinical laboratory running a

production informatics pipeline is to build a process which can change and grow, accommodating changes and improvements with minimal interruption, allowing individual laboratories to take advantage of improvements in accuracy, workflow, time, and cost as the various parts of the system evolve. Individual changes to the pipeline will need to be evaluated and validated to show that they do not degrade the performance in some aspects whilst improving others or introducing new components. A systematic program to maintain and develop these pipelines should be implemented by laboratories to ensure that this happens, taking account of individual circumstances.

4 Notes

1. *Read Mapping/alignment software.* A significant number of software packages are available to align reads to a reference sequence. It is beyond the scope of this article to provide and review an exhaustive list. There is a substantial literature on read alignment software [18], online references are often more up to date. Useful Web pages for read aligners include http://en.wikibooks.org/wiki/Next_Generation_Sequencing_%28NGS%29/Alignment and http://wwwdev.ebi.ac.uk/fg/hts_mappers/.
2. *Variant callers.* Variant callers take as their base input aligned sequence reads in BAM format. The assumptions underlying the particular variant-calling strategy are critically affected by the specific methodology used for variant calling. Unlike traditional forms of variant detection, i.e., manual inspection of Sanger electropherograms, different algorithms must be performed independently to call different variant types, e.g., SNVs and indels. Each of these algorithms will differ in their sensitivity and specificity, and this will also be influenced by the laboratory workflow, e.g., WES versus WGS, and mapping parameters. There is currently no universal standard variant caller applicable to all data types. A useful discussion of the differences between some variant callers can be found in ref. 19.
3. *Annotation and filtering.* The critical step in annotation and filtering is the definition of the process. The program used to implement the process is often chosen based on convenience or cost. This is an active and growing area, with a significant number of commercial software companies producing packages which may perform these tasks. The final choice of software will often depend on the interactions of consideration of cost and convenience. A large number of groups have made use of the Annovar package [20], a freely available program, well supported by its author and developer which is powerful

enough to annotate and filter very large datasets, including whole-genome data on even modest computing hardware. Its principal drawback is the requirement for operation via the command line, although there are now also freely available Web-based implementations of the algorithm—<http://wannovar.usc.edu/> [21].

Acknowledgments

I would like to acknowledge the long-standing support of Mr Neill Hodgen and the Department of Clinical Immunology, Royal Perth Hospital for their past and ongoing support.

References

- Metzker ML (2010) Sequencing technologies: the next generation. *Nat Rev Genet* 11:31–46
- Liu L, Li Y, Li S et al (2012) Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012:251364
- Kamalakaran S, Varadan V, Janevski A et al (2013) Translating next generation sequencing to practice: opportunities and necessary steps. *Mol Oncol* 7:743–755
- Hong H, Zhang W, Shen J et al (2013) Critical role of bioinformatics in translating huge amounts of next-generation sequencing data in personalized medicine. *Sci China Life Sci* 56:110–118
- Yang Y, Muzny DM, Reid JG et al (2013) Clinical whole-exome sequencing for the diagnosis of Mendelian disorders. *N Engl J Med* 369:1502–1511
- Bromberg Y (2013) Building a genome analysis pipeline to predict disease risk and prevent disease. *J Mol Biol* 425:3993–4005
- Guo Y, Ye F, Sheng Q et al (2013) Three-stage quality control strategies for DNA re-sequencing data. *Brief Bioinform*. doi: [10.1093/bib/bbt069](https://doi.org/10.1093/bib/bbt069)
- Cock PJA, Fields CJ, Goto N et al (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38:1767–1771
- Li H, Handsaker B, Wysoker A et al (2009) The sequencer alignment/map format and SAMtools. *Bioinformatics* 16:2078–2079
- Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. *Bioinformatics* 27:2156–2158
- Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842
- 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073
- Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814
- Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249
- Schwarz JM, Rodelsperger C, Schuelke M et al (2010) MutationTaster evaluates disease-causing potential of sequence alternations. *Nat Methods* 7:575–576
- Quinque D, Kittler R, Kayser M et al (2006) Evaluation of saliva and a source of human DNA for population and association studies. *Anal Biochem* 353:272–277
- Boland JF, Chung CC, Roberson D et al (2013) The new sequencer on the block: comparison of Life Technology’s Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Hum Genet* 132:1153–1163
- Pavlopoulos GA, Oulas A, Iacucci E et al (2013) Unravelling genomic variation from next generation sequencing data. *BioData Min* 6:13–38
- Liu X, Han S, Wang Z et al (2013) Variant callers for next generation sequencing data: a comparison study. *PLoS One* 8:e75619
- Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164
- Chang X, Wang K (2012) wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* 49:433–436

Chapter 3

Analyzing the Metabolome

Francis G. Bowling and Mervyn Thomas

Abstract

Metabolites, the chemical entities that are transformed during metabolism, provide a functional readout of cellular biochemistry that offers the best prediction of the phenotype and the nature of a disease. Mass spectrometry now allows thousands of metabolites to be quantitated. The targeted or untargeted data from metabolic profiling can be combined with either supervised or unsupervised approaches to improve interpretation. These sophisticated statistical techniques are computationally intensive. This chapter reviews techniques applicable to metabolomics approaches to disease.

Key words Mass spectrometry, Metabolites, Metabolomics, Pathway analysis, Targeted metabolomics, Untargeted metabolomics, Supervised analysis, Unsupervised analysis

Abbreviations

AUC	Area under curve
LC	Liquid chromatography
MS	Mass spectrometry
NMR	Nuclear magnetic resonance
ROC	Receiver operating characteristic

1 Introduction

Metabolites are small molecules that are chemically transformed during metabolism and, as such, provide a functional readout of the cellular state. Unlike genes and proteins, whose function is subject to epigenetic regulation and posttranslational modifications respectively, metabolites serve as direct signatures of biochemical activity. Thus, they are easier to correlate with phenotype [1]. In this context, metabolite profiling, or metabolomics, has

become a powerful approach that has been widely adopted for clinical diagnostics because of the correlation of biochemical changes with phenotype. Metabolomic techniques are designed to measure rapidly thousands of metabolites simultaneously from only minimal amounts of sample.

Metabolomics can be used for diagnosis, monitoring therapy, and predicting the natural history of a disease state. A *targeted metabolomics approach* which focuses on a set of predetermined molecular species is particularly useful for inherited metabolic disorders, which are a group of diseases caused by defects in biochemical pathways [2]. An *untargeted approach* can be used for more general diseases, e.g., acute renal failure, where the nature of the molecular disruption may be more diverse than that in a single pathway disorder, but where the predictive value of combining many independent molecular markers is greater than the value of measuring a single marker [3]. With an untargeted approach, it is not necessary to know the identity of the disease related metabolites.

The application of these technologies has revealed system-wide alterations of unexpected metabolic pathways in response to phenotypic perturbations. Moreover, many of the molecules detected are currently not included in databases and metabolite repositories, indicating the extent to which our picture of cellular metabolism is incomplete [4].

2 Materials

2.1 Case Ascertainment

As the general principle with metabolomics is to compare specific samples against a matched control group, careful consideration should be given to the composition of the study groups. Affected may be compared against unaffected. Treated cohorts may be compared against an untreated set, or longitudinally in a case control series. Cross-sectional data may be useful to dilute the effect of confounding factors such as other disease conditions or disturbances to metabolism. Generally, cohorts should be matched for age, particularly in pediatric studies, and possibly for gender. It may also be necessary to match samples for fasting state, exercise, medications, and diet. Population based reference data can be used for comparison also (*see Note 1*).

2.2 Samples

Serum samples have been used most commonly and are generally considered to be the most informative [2]. Other fluids such as urine and cerebrospinal fluid can be studied. Investigation may now be also undertaken on tissue samples, either from biopsy or by in vivo techniques such as magnetic resonance spectroscopy. Sampling of the affected tissue is the most informative. Again, sample collection, preparation, and storage should be standardized to prevent artifactual changes in metabolites (*see Note 2*).

2.3 Equipment

Developments in mass spectrometry (MS) and nuclear magnetic resonance (NMR) offer distinct advantages for performing targeted metabolomic studies because of their specificity and quantitative reproducibility. However, there are many analytical tools available for measuring metabolites that could in principle be considered such as high performance liquid chromatography and ultraviolet-visible spectroscopy [5]. Triple quadrupole MS used to perform selected reaction monitoring experiments are now available to analyze most of the metabolites of amino acids, organic acids, lipids, steroids, fatty acids, glycans, and purines at their naturally occurring physiological concentrations. These techniques are highly sensitive and robust methods able to measure a significant number of biologically important metabolites with relatively high throughput. However, their applications are often limited by lack of understanding of informatics and statistics (*see Note 3*).

Untargeted metabolomics methods are global in scope and have the aim to measure simultaneously as many metabolites as possible from biological samples without bias. Although untargeted metabolomics can be performed using NMR, liquid chromatography (LC) followed by MS (LC/MS) enables the detection of the most metabolites and has therefore become the technique of choice for global metabolite profiling efforts [6]. Most frequently data are collected on a quadrupole time-of-flight (QTOF) mass spectrometer or an Orbitrap mass spectrometer, but other time-of-flight and ion trap instruments can also be used.

3 Methods

3.1 Overview

The first step in performing metabolomics is to determine the number of metabolites to be measured and the context in which they are to be measured. In some instances, it may be of interest to examine a defined set of metabolites by using a *targeted* approach (*see Note 4*). In other cases, an *untargeted* or global approach may be taken in which as many metabolites as possible are measured and compared between samples without bias. Ultimately, the number and chemical composition of metabolites to be studied is a defining attribute of any metabolomic experiment that then shapes experimental design with respect to sample preparation and choice of instrumentation.

3.1.1 Targeted Metabolomics

The principle behind a targeted approach is to combine the information from multiple (known) metabolites to improve data interpretation. The combination of multiple (independent) metabolites is more discriminatory than a single metabolite. Classical biochemical metabolism shows metabolites arising as substrates, intermediates, and products in pathways. This suggests that sequential pathway metabolites may not be independent variables.

However, regulatory mechanisms, flux analysis, alternative pathways, organelle distribution of pathway steps, and intermediate pools show that the relationships between pathway metabolite concentrations are not simply linear. Hence, combining their values may be more informative than simply relying on limited traditional disease markers [7].

Targeted analyses are usually conducted to answer well defined clinical questions, examining the differences between clinical groups which are defined a priori, e.g., healthy control subjects versus individuals with mitochondrial disease. In this situation, the primary statistical analysis should be based on a supervised technique. Supervised techniques are designed to exploit the known a priori clinical structure of the data. A comprehensive analysis will usually include some unsupervised techniques, e.g., cluster analysis or principal components, principally for quality checking. In short, *supervised* techniques look for the differences we expect to find in the data; *unsupervised* techniques look for unexpected differences.

3.1.2 *Untargeted Metabolomics*

Untargeted metabolite profiles require more analytical effort but offer a great discriminatory yield if the experiments are tightly controlled. This approach will require a greater number of samples or combination of data through meta-analysis techniques. Non-supervised statistical analyses may be employed in untargeted metabolomics experiments for research or exploratory reasons. They can also be used for quality checking of the assumptions and identification of features in a supervised analysis of the same set, e.g., Identification of clusters of metabolites not known to be related to the phenotype or disease presentation in question. Artifacts in sample collection and handling may be identified through this untargeted analysis.

3.2 **Sample Preparation and Data Acquisition**

3.2.1 *Targeted Metabolomics*

The aim in design of a targeted metabolomic analysis is to ensure correct quantitation of a known set of molecular species. Commercial kits are available for this purpose. Assay preparation may be performed in multi-well plates with isotope labeled internal standards. Plasma samples are derivatized by agents such as phenylisothiocyanate and extracted with an organic solvent. Alternatively, to reduce cost, laboratories may establish a calibration curve from varying concentrations of each metabolite. These curves are stored against the metabolite in the MS library for that laboratory and are referenced against an internal standard such as trophic acid added to each run. This approach reduces the effect of inter-assay variation and does not require isotopic standards to be used for every compound in every run.

For MS, a standard flow injection method may be used without chromatographic separation. Two injections, one for the positive and one for the negative detection mode analysis can be employed [8].

For laboratories investigating inherited disorders of metabolism, a QTRAP 5500 tandem mass spectrometer (AB SCIEX, USA) with electrospray ionization is often employed. Multiple reaction monitoring detection allows for identification and quantification of approximately 200 endogenous metabolites from different metabolite classes.

3.2.2 *Untargeted Metabolomics*

The first step in the untargeted metabolomic workflow is to isolate metabolites from biological samples. Various approaches, depending on the experimental design, involving sample homogenization and protein precipitation may be utilized. Prior to MS analysis, isolated metabolites are separated chromatographically by using relatively short solvent gradients (in the order of minutes) that allow for high-throughput analysis of large numbers of samples [9]. Because of physicochemical differences in the variety of molecular species constituting the metabolome, multiplexing extraction and separation methods maximizes the number of metabolites detected. Sample extraction using both organic and aqueous solvents increases the yield of hydrophobic and hydrophilic compounds [10]. Reverse-phase chromatography is better suited for the separation of hydrophobic metabolites, whereas hydrophilic-interaction chromatography separates hydrophilic compounds. Given the challenge of predicting a triple quadrupole fragmentation pattern for most metabolites, untargeted metabolomic profiling typically acquires only the mass-to-charge ratio (m/z) of the intact metabolite. By using chromatographic separation in combination with mass spectrometry, thousands of peaks with a unique mass-to-charge ratio and unique retention time are detected from biological samples [11].

3.3 *Metabolite Identification*

3.3.1 *Targeted Metabolomics*

For targeted metabolomics approaches, it is necessary to identify each metabolite. The metabolites are defined by their retention on the chromatographic method and their ion spectra. Usually, each laboratory will validate the retention time for their conditions and enter the data into a metabolite library. In commercial kits with standardized conditions, the complete analytical process may be performed using integral software, e.g., MetIQ software. For in-house assays, bioinformatic tools such as MetaboAnalyst allow for identification of metabolites [12]. Metabolic software programs are available to identify metabolite features that are differentially altered between sample groups. These can include methods for peak picking, nonlinear retention time alignment, visualization, relative quantitation, and statistics. XCMS is an example of publically available metabolomic software [13]. Users can upload data, perform data processing, and browse results within a Web-based interface.

3.3.2 *Untargeted Metabolomics*

Untargeted metabolomic software does not output metabolite identifications. Typically, it provides a list of features with relative changes derived from the difference in relative intensity between samples. A separate analysis must be undertaken to determine the identity of a feature of interest. The mass of the compound is searched in metabolite databases such as the Human Metabolome Database [14] and METLIN [15, 16]. A database match represents only a tentative identification that must be confirmed in a separate experiment or by comparing the retention time and mass spectrometry data of a model (*see Note 5*). The Human Metabolome Database includes detailed data for each of its included metabolites (~8,550). In addition to having molecular weights and experimental NMR spectra, the biochemical pathway, biological concentration, tissue/cellular location, involved enzymes, and related disorders are included. METLIN contains experimental data for approximately 45,000 compounds.

3.4 *Statistical Analysis*

3.4.1 *Targeted Metabolomics*

In a targeted approach, there will be a well-defined set of candidate analytes used to differentiate (clinical) sample groups. These candidate analytes will be known a priori from understanding of the pathway involved or from biomarkers already observed in the disease state. Metabolites related to the altered pathway may be increased proximal to a block, decreased distal to a block, generated from alternate branched pathways, or may even be formed by mechanisms not known to be linked to the blocked pathway.

Classical approaches, before the advent of metabolomics technologies, typically considered one analyte at a time. Often the distribution of values in the clinical cohort was compared against the distribution in a reference cohort. The concentration distributions for an analyte may not have a normal distribution, and the sample and control distributions will often overlap. For single analytes, *Receiver Operating Characteristic* (ROC) curves may be constructed and the *Area Under the Curve* (AUC) calculated [17]. This allows for comparison of the performance of single (or combined) biomarkers (*see Note 6*).

The powerful advantage of metabolomics is the ability to consider many analytes simultaneously. This may be intended either for screening purposes (to narrow down the set of candidate analytes to a small number that show strong diagnostic promise), to improve diagnostic performance by combining information from multiple analytes, or some combination of the two. Whatever the objective, the multiplicity which provides the advantage of metabolomics also poses special challenges.

For screening purposes it is essential to remember that when a very large number of analytes are examined, some will show very large between-group-separations by chance alone. For example, consider a hypothetical study with ten samples, five from a control group and five from a disease group. Assume there are 1,000 independently distributed analytes, and there are no true differences

between the groups for any analyte. Then the probability that there will be at least one analyte which is perfectly separated between the groups (i.e., a ROC AUC of 1.0) is 0.98.

One useful strategy for dealing with this is to first screen analytes by comparing groups using the Mann–Whitney U statistic [18]. Only consider ROC AUCs for those analytes which are significant following p -value adjustment using Holm’s method [19]. Holm’s method ensures that the probability of falsely rejecting the null hypothesis of no group difference for at least one analyte is maintained at the desired significance level (0.05 by convention). This is referred to as the family-wise type I error rate. Holm’s method is an extremely conservative procedure, and for many analytes it may substantially reduce the chance of finding any metabolites which are differentially represented in the clinical groups. A somewhat less conservative approach would be to use the Benjamini Hochberg adjustment [20]. This approach controls the false discovery rate rather than the family-wise Type I error. The false discovery rate is the proportion of analytes flagged as different that are not truly different between clinical groups.

If the objective is to improve diagnostic power by integrating information across a set of biomarkers, then the process is to use a machine learning algorithm to induce a classification rule. There are many different machine learning algorithms, based on very different statistical models (and some based only on simple heuristics with no well-articulated model at all) [21]. No single machine learning algorithm is uniformly better than all the others; their relative performance is problem dependent (*see* also Chapter 16).

Nevertheless there are a few approaches which usually are amongst the best for any given problem. The rigor with which results are validated, and spurious diagnostic power is avoided are much more important than the choice of machine learning algorithm. These general issues are discussed before making recommendations about specific machine learning algorithms.

The starting point for the machine learning exercise is a data set with multiple analytes for each subject, and a clinical group label, e.g., control or disease for each subject. The naïve approach is to develop a diagnostic rule using the entire data set, and then to assess its performance by re-substituting the analyte values from the data set into the rule. This generates a predicted label for each subject, or a probability of being in the disease group. Sensitivity and specificity are then calculated, or perhaps a ROC AUC using the probability of being in the diseased group as a multivariate index biomarker [22]. Unfortunately, this naïve strategy produces results which are so optimistic as to be meaningless. In the extreme case, where there are at least as many analytes as subjects, many algorithms will produce a classification rule which works without error in re-substitution. Such rules generally have very poor performance for future cases, independent of those used to generate the diagnostic rule.

The ideal solution to this problem is to have to have two independent data sets. The diagnostic rule is developed (or “trained”) using one data set, and it is evaluated using the independent test data set. The two data sets may be generated by a random partition of the original data set. This approach leads to unbiased estimates of diagnostic performance (*see Note 7*).

Unfortunately, many observations may not be available in early stage investigations. An alternative strategy is to use cross-validation [23]. In k -fold cross-validation, the data set is randomly partitioned into k subgroups. Typically, this is stratified by the disease state such that each of the k random groups has approximately the same proportion of disease and control cases. Each of the k subsets is dropped in turn, and the diagnostic rule is developed using the remaining $k-1$ groups. The classifier developed using those $k-1$ groups is then used to predict the dropped group (either the disease state or the probability of disease). The dropped subgroup is reinstated, and the process repeated with the next subgroup. Predictions are recorded across the k subgroups. At the end of this process, each observation has been used to develop the classifier (actually $k-1$ times), and each observation has been used to test the classifier. But now observation has been used to train and test the classifier at the same time. The resulting estimates of diagnostic performance (sensitivity, specificity, ROC AUC or any other success metric of choice) are almost as unbiased as those obtained from an independent test data set. In the extreme case, k may be set to the number of observations, and each observation is let out in turn. This is known as *leave one out* cross-validation. k -fold cross-validation may be repeated many times with different random partitions into the k groups. This is the usual practice (*see Note 8*), and should always be adopted because although cross-validation error estimates may be nearly unbiased in small samples they have very high variance [24]. Repeating cross-validation many times with different random folds goes some way towards ameliorating the problems of highly variable error or performance estimates; but this variability is an inherent limitation of small study sizes.

Machine learning algorithms differ in motivation and approach. Some, like the Elastic Net [25] or Generave [26], use a generalized linear model structure, with a variable selection engine based on a penalty function designed to ensure a sparse solution. That is, they produce classification models with relatively few analytes included. This is particularly useful when the final clinical platform will not be MR, and will quantify relatively few analytes (such as in in vivo magnetic resonance spectroscopy).

Other methods such as partial least squares discriminant analysis [27], or penalized discriminant analysis [28] retain all the analytes, but project into a low dimensional space. Kernel methods, such as support vector machines [29, 30], adopt yet another strategy. Some of the most powerful methods are Meta learners; which

combine information from an ensemble of relatively weak classifiers. Examples include random forests [31] and LogitBoost methods [32, 33].

In general, it is advisable to explore at least one method from each class of algorithm. A good default selection would be the Elastic Net, penalized discriminant analysis, support vector machines, and LogitBoost. The number of iterations in LogitBoost, the kernel width and penalty factors of support vector machines and the penalty weights of penalized discriminant analysis will all be chosen by cross-validation. It is important that this tuning is not allowed to bias performance estimates. If sample numbers allow, it is by far the best strategy to use a separate test set for validation and to use cross-validation within the training set for model tuning (*see Note 9*).

3.4.2 *Untargeted Metabolomics*

In contrast to targeted metabolomic results, untargeted metabolomic data sets are exceedingly complex with file sizes on the order of gigabytes per sample for some new high-resolution MS instruments. Manual inspection of the thousands of peaks detected is impractical and complicated by experimental drifts in instrumentation. In LC/MS experiments, for example, there are deviations in retention time from sample to sample as a consequence of column degradation, sample carryover, small fluctuations in room temperature and mobile phase pH. These variations present difficulty for interpreting untargeted profiling data. Metabolomics software such as MetaboAnalyst [12], MathDAMP [34], MetAlign [35], MZMine [36], and XCMS [37] allow for an approach to these data.

In untargeted metabolomics, supervised training analysis may be employed, as with the targeted approach. The problem will be more challenging computationally (simply because of the larger number of analytes), but not beyond the bounds of feasibility. The problems of very high dimensionality will be more pressing with untargeted metabolomics than with targeted metabolomics, but the same approaches and considerations apply. For research or exploratory investigations, and for quality checking purposes, unsupervised techniques may be employed to find hidden and unexpected structure in the unlabeled data. There are many approaches for unsupervised data exploration, and a range should be employed in any analysis.

Principal components analysis [38], projection pursuit [39], and independent components analysis [40] project the data into a low dimensional space that seeks to preserve “interesting” features of the data [11]. The techniques differ in terms of the criteria that define the interest under consideration. In principal components analysis, the criterion is maximum variance (subject to orthonormality constraints), in independent components analysis, the criterion is a likelihood function based on a mixture of non-normal

distributions, and in projection pursuit the criterion is negative entropy (which can be interpreted as a measure of difference from normality). Subjects are plotted in the space of the summary variables produced by these techniques; and groups of subjects are noted. These summary variables may also be plotted against aspects of the experimental protocol which are unrelated to clinical features, such as sequence of samples through the mass spectrometer, length of time the sample has been in storage, assay kits, site from which the sample was sourced. Occasionally, such checking procedures will reveal substantial sources of bias quite unrelated to the clinical questions.

In addition to dimension reduction techniques, attempts may be made to represent the multivariate similarity between samples and analytes by means of *cluster analysis* [41]. There are two broad categories of cluster analysis: Hierarchical and non-hierarchical. Hierarchical cluster analysis usually works by combining sets of observations in a tree structure. Non-hierarchical cluster analysis usually works by separating observations into a predetermined number of groups. The most popular non-hierarchical method is *k*-means clustering [42]. The question often arises as to how many clusters are suggested by the data. For *k*-means the most widely used method is Tibshirani's Gap statistic [43], but this method can work badly for very high numbers of clusters [44].

Hierarchical clustering algorithms differ in terms of the difference metric they use, and the algorithm used to combine sets of observations. The three algorithms most commonly used are single linkage, average linkage and complete linkage. Single linkage often produces cluster solutions represented by uninterpretable long chains of observations joined together [45]. For this reason, many researchers avoid single linkage. Rather than being a weakness of single linkage algorithms, however, this can be a strength. If a clear cluster structure is found with single linkage it is a strong indication that the structure is real. Complete linkage can usually be relied upon to provide aesthetically pleasing beautifully balanced cluster trees for which the researcher finds it all too easy to provide spurious post hoc rationalizations.

More modern, model based techniques are available. They are computationally more demanding but may give a better insight into the appropriate number of clusters [46].

In general, it is suggested that cluster solutions should be produced with multiple algorithms and multiple metrics. They should only be interpreted if the broad features of the solution are common across the analyses.

3.5 *Meta-analysis*

Untargeted metabolomics profiling of a particular disease entity can reveal alterations that are unlikely to have mechanistic implications. In rare diseases, which are not fully understood, the range of sample data may be limited. Because of the effort of

additional experiments needed to identify both known and unknown compounds, strategies to reduce lists of potentially interesting features increase the efficiency. Meta-analysis, by which untargeted profiling data from multiple studies are compared, increases the data available and the case and control sample sizes. By comparing multiple models of a disease, for example, features that are not similarly altered in each of the comparisons may be de-prioritized as being less likely to be related to the shared phenotypic pathology. To automate the comparison of untargeted metabolomic data, software such as metaXCMS [47] may be used.

3.6 Statistical Software

Undoubtedly the most comprehensive software for metabolomics data processing and analysis is found in the extensions to the R package [48, 49], especially in the Bioconductor project [50, 51]. In addition to a comprehensive suite of tools for supervised and unsupervised high dimensional data analysis in R, Bioconductor provides packages designed to facilitate the preliminary spectrum recalibration, peak identification, normalization, and preprocessing requirements of data from a range of mass spectroscopy instrument technologies. It also supports comprehensive pathway analysis tools.

R and Bioconductor are open source projects and are free to download. They are supported by an enthusiastic and active community which includes some of the leading researchers in applied statistics and bioinformatics. R is the tool of choice for most professional statisticians and statistical bioinformaticians. However, R does have a somewhat steep learning curve. Nevertheless, for anyone who will be involved in regular analysis of metabolomic data it is well worth while investing the time and effort to become skilled with R.

4 Notes

1. It is important that assay techniques are standardized to ensure the same molecular species can be identified with similar yield and quantitated (against a reference standard).
2. These changes may be introduced through cellular leakage, ongoing metabolism, contamination, or oxidation.
3. Additionally, triple quadrupole methods are quantitatively reliable and allow for absolute quantitation of low-concentration metabolites that are difficult to detect with less sensitive methods such as NMR.
4. With targeted metabolomics, a specified list of metabolites is measured, typically focusing on pathways of interest. Targeted metabolomics approaches are commonly driven by a specific biochemical question or hypothesis that motivates the

investigation of a particular pathway. This approach can be effective for pharmacokinetic studies of drug metabolism as well as for measuring the influence of therapeutics or genetic modifications on a specific enzyme. Although the term *metabolomics* has only recently been coined, examples of targeted studies of metabolites date back to the earliest of scientific inquiries.

5. If necessary, MS/MS data for features selected from the profiling results are obtained from additional experiments and matching of MS/MS fragmentation patterns is performed manually by inspection. These additional analyses are time intensive and represent the rate limiting step of the untargeted metabolomic workflow. Additionally, although metabolite databases have grown considerably over the last decade, a substantial number of metabolite features detected from biological samples do not return any matches. Identification of these unknown features requires de novo characterization with traditional methods. Taken together, it should be recognized that comprehensive identification of all metabolite features detected by LC/MS is currently impractical for most samples analyzed.
6. When comparing the ROC curves calculated using different analytes measured on the same samples it is important to use a statistical method which models the dependence appropriately, e.g., Venkatraman and Begg [52].
7. The issue of bias in estimates of diagnostic performance is complex [53]. Independent test data sets will ensure that the estimate of performance is unbiased *for the sample from which the test data set is drawn*. In early stage research this may very well not be the target clinical population. For example, it may contain especially severe disease cases, and healthy controls who are active young adults. Estimates based on this population will almost certainly exhibit spectrum bias, even if an independent test set is used.
8. Although cross-validation provides a high degree of protection against over optimistic results it is possible to subvert that protection. There are several common errors which lead to this subversion.
 - Cross-validation may be applied after the biomarkers have been filtered, and only those biomarkers showing large between group differences have been retained. This approach destroys the benefits of cross-validation and leads to major biases. If the analysis strategy requires filtering, then the filtering must take place *inside* the cross-validation [54].
 - Cross-validation works well when observations are independent. It will break down when there are strong

dependencies between observations—as would be the case with longitudinal data where there are multiple observations on each subject. It is possible to adopt more sophisticated cross-validation strategies, but professional statistical advice is necessary.

- Many machine learning algorithms involve one or more tuning parameters (a penalty factor, kernel bandwidth, number of dimensions, etc.), which are often selected using cross-validation. If cross-validation is used to select tuning parameters, then this should be achieved in a two stage cross-validation, with parameters set in an inner cross-validation loop and performance estimated using an outer loop. Again, professional statistical input is desirable.
9. Any researcher intending to become seriously involved in metabolomics would be well advised to invest some time in acquiring the basics of machine learning. One of the more accessible texts is Witten, Hastie, and Tibshirani's "An Introduction to Statistical Learning: with Applications in R" (Springer Texts in Statistics) [55]. This book by leading researchers in the field manages to avoid unnecessary mathematics, and yet is an authoritative resource.

References

1. Patti G, Yanes O, Siuzdak G (2012) Innovation: metabolomics: the apogee of the omic trilogy. *Nat Rev Mol Cell Biol* 13: 263–269
2. Janecková H, Hron K, Wojtowicz P et al (2012) Targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders. *J Chromatogr A* 1226: 11–17
3. Robinson AB, Robinson NE (2011) Origins of metabolic profiling. *Methods Mol Biol* 708:1–23
4. Kind T, Scholz M, Fiehn O (2009) How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS One* 4:e5440
5. Dudley E, Yousef M, Wang Y et al (2010) Targeted metabolomics and mass spectrometry. *Adv Protein Chem Struct Biol* 80:45–83
6. Yanes O, Tautenhahn R, Patti GJ et al (2011) Expanding coverage of the metabolome for global metabolite profiling. *Anal Chem* 83: 2152–2161
7. Suhre K, Shin SY, Petersen AK et al (2011) Human metabolic individuality in biomedical and pharmaceutical research. *Nature* 477: 54–60
8. Nordstrom A, Want E, Northen T et al (2008) Multiple ionization mass spectrometry strategy used to reveal the complexity of metabolomics. *Anal Chem* 80:421–429
9. Buescher JM, Moco S, Sauer U et al (2010) Ultrahigh performance liquid chromatography-tandem mass spectrometry method for fast and robust quantification of anionic and aromatic metabolites. *Anal Chem* 82:4403–4412
10. Want EJ, O'Maille G, Smith CA et al (2006) Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry. *Anal Chem* 78:743–752
11. Patti GJ (2011) Separation strategies for untargeted metabolomics. *J Sep Sci* 34:3406–3469
12. Xia J, Psychogios N, Young N et al (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res* 37:W652–W660
13. Tautenhahn R, Patti GJ, Tinehart D et al (2012) XCMS Online: a web based platform to process untargeted metabolomic data. *Anal Chem* 84:5035–5039
14. Wishart D, Tzur D, Knox C et al (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Res* 35:D521–D526

15. Smith CA, O'Maille G, Want EJ et al (2005) METLIN: a metabolite mass spectral database. *Ther Drug Monit* 27:747–751
16. Zhu ZJ, Schultz AW, Wang J et al (2013) *Nat Protoc* 8: 451–460. Scripps Centre for Metabolomics and Mass Spectrometry: METLIN. <http://metlin.scripps.edu/>
17. Hanley JA, McNeil BJ (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143: 29–36
18. Lehmann EL (1975) *Non parametric statistical methods based on ranks*. Holden-Day, San Francisco, CA, Section 1.2
19. Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6:65–70
20. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc B* 57:289–300
21. Hastie T, Tibshirani R, Friedman J (2001) *The elements of statistical learning: data mining, inference and prediction, 1st edn*, Springer series in statistics. Springer, New York
22. Zhang Z, Chan DW (2010) The road from discovery to clinical diagnostics: lessons learned from the first FDA-cleared in vitro diagnostic multivariate index assay of proteomic biomarkers. *Cancer Epidemiol Biomarkers Prev* 19:2995–2999
23. Stone M (1974) Cross-validated choice and assessment of statistical predictions. *J Roy Stat Soc B* 36:111–147
24. Braga-Neto UM, Dougherty ER (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics* 20:374–380
25. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. *J Roy Stat Soc B* 67:301–320
26. Kiiveri HT (2008) A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. *BMC Bioinformatics* 9:195
27. Ding B, Gentleman R (2005) Classification using generalized partial least squares. *J Comput Graph Stat* 14:280–298
28. Hastie T, Buja A, Tibshirani R (1995) Penalized discriminant analysis. *Ann Stat* 23:73–102
29. Cortes C, Vapnik V (1995) Support-vector networks. *Mach Learn* 20:273–297
30. Cristiani N, Taylor JS (2000) *An introduction to support vector machines*. Cambridge University Press, Cambridge
31. Breiman L (2001) Random forests. *Mach Learn* 45:5–32
32. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. *Ann Stat* 28:337–374
33. Blanchard G, Lugosi G, Vayatis N (2003) On the rate of convergence of regularized boosting classifiers. *J Mach Learn Res* 4:861–894
34. Baran R, Kochi H, Saito N et al (2006) MathDAMP: a package for differential analysis of metabolite profiles. *BMC Bioinformatics* 7:530
35. Lommen A (2009) MetAlign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. *Anal Chem* 81:3079–3086
36. Katajamaa M, Miettinen J, Oresic M (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22:634–636
37. Smith C, Want E, O'Maille G et al (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 78:779–787
38. Jolliffe I (1986) *Principal components analysis*. Springer, New York
39. Friedman JH, Tukey JW (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans Comput* 23:881–889
40. Hyvärinen A, Oja E (2000) Independent component analysis: algorithms and applications. *Neural Netw* 13:411–430
41. Everitt B, Landau S, Leese M (2001) *Cluster analysis, 4th edn*. Edward Arnold, London
42. Hartigan J, Wong M (1979) A K-means clustering algorithm. *J Roy Stat Soc C-App* 28:100–108
43. Tibshirani R, Walther G, Hastie T (2001) Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc B* 63:411–423
44. Feng Y, Hamerly G (2006) PG-means: learning the number of clusters in data. In: Scholkope B, Platt J, Hofmann T (eds) *Advances in neural information processing systems* 19. MIT, Cambridge, MA, pp 393–400
45. Sibson R (1973) SLINK: an optimally efficient algorithm for the single-link cluster method. *Comput J* 16:30–34
46. The Comprehensive R Archive Network: R Sources (2014) <http://cran.r-project.org/>. Accessed 14 Apr 2014
47. Tautenhahn R, Patti G, Kalisiak E et al (2011) metaXCMS: second-order analysis of untargeted metabolomics data. *Anal Chem* 83:696–700
48. The Comprehensive R Archive Network. <http://cran.r-project.org/>

49. R Development Core Team (2003) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, www.R-project.org
50. Bioconductor: High Throughput Assays (2014) <http://www.bioconductor.org/>. Accessed 14 Apr 2014
51. Gentleman RC, Carey VJ, Bates DM et al (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5:R80
52. Venkatraman E, Begg CB (1996) A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* 83:835–848
53. Begg CB (1987) Biases in the assessment of diagnostic tests. *Stat Med* 6:411–423
54. Ambroise C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 99:6562–6566
55. Witten JG, Hastie T, Tibshirani R (2013) An introduction to statistical learning with applications in R. Springer, New York

Chapter 4

Statistical Perspectives for Genome-Wide Association Studies (GWAS)

Jennifer H. Barrett, John C. Taylor, and Mark M. Iles

Abstract

In this chapter we consider some key elements in conducting a successful genome-wide association study or GWAS. The first step is to design the study well (Subheading 3.1), paying particular attention to case and control selection and achieving adequate sample size to deal with the large burden of multiple testing. Second, we focus on the crucial step of applying stringent quality control (Subheading 3.2) to genotyping methods. The most crucial potential confounding factor in GWAS is population stratification, and we describe methods for accounting for this in study design and analysis (Subheading 3.3). The primary association analysis is relatively straightforward, and we describe the main approaches to this, including evaluation of results (Subheading 3.4). More comprehensive coverage of the genome can be achieved by using an external reference panel to estimate genotypes at untyped variants using imputation (Subheading 3.5), which we consider in some detail. We finish with some observations on following up a GWAS (Subheading 3.6).

Key words Genome-wide association study, GWAS, Imputation, Multiple testing, Population stratification, Quality control

Abbreviations

GWAS	Genome-wide association study
HWE	Hardy–Weinberg equilibrium
LD	Linkage disequilibrium
PCA	Principal component analysis
QC	Quality control
SNP	Single-nucleotide polymorphism

1 Introduction

Genome-wide association studies (GWAS) have been successfully used to investigate many major diseases and traits. The basic idea is a simple one: genotype a set of individuals with the disease of interest (cases) and a set of individuals from the same population

without the disease (controls) for many (typically ~1 million) common variants across the genome and compare genotype distributions between the two groups, thus identifying genetic variants that are associated with disease risk. These studies have only been possible since about 2006 due to advances in genotyping technology. Since then over 1,800 publications have reported associations with over 13,000 single-nucleotide polymorphisms (SNPs) [1].

Although most common diseases have now been investigated using large sample sizes, many populations have not been well studied, and many outcomes that are potentially genetically controlled (such as adverse response to treatment) remain to be investigated, so GWAS are likely to continue to be of value for some time to come. In this chapter we outline the main steps on how to carry out successfully and interpret a GWAS (*see* also Chapter 5).

2 Materials

A great deal of specialist software is freely available to assist in the statistical analysis of a GWAS. The following programs are all referred to later in the text. The most widely used program for association analysis is PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink/>), which also has many other features including tools for data management and QC.

Population stratification can be handled using an approach based on principal components using Eigenstrat, implemented in EIGENSOFT software (www.hsph.harvard.edu/faculty/alkes-price/software/), or mixed effects models using EMMAX (<http://genetics.cs.ucla.edu/emmax/>) or TASSEL (www.maizegenetics.net/).

Several programs are available for imputation: IMPUTE (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html), MACH (www.sph.umich.edu/csg/abecasis/MACH/), and BEAGLE (<http://faculty.washington.edu/browning/beagle/beagle.html>).

Finally, results at a particular locus can be illustrated using LOCUSZOOM (<http://csg.sph.umich.edu/locuszoom/>).

3 Methods

3.1 Study Design

Most GWAS investigate a specific disease (a binary phenotype) and use a case–control study design (*see* **Note 1**). There has been some relaxation of the usual principles of case–control study design in GWAS, so that controls are sometimes ascertained quite differently to the cases; this can be justified in view of the need for large sample sizes and the fact that an individual’s genotype does not differ over time and is less subject to confounding than many environmental factors.

3.1.1 Selection of Cases

The cases are a set of individuals from the population(s) with the disease in question; it is not generally important to collect incident cases, as it may be in a classical case-control study, since issues of recall bias do not arise when investigating genes. Although the ascertainment criteria for case recruitment may be broad, it is important that this is clearly specified in order to facilitate comparisons between studies. Many “diseases” are complex and variable, and phenotype heterogeneity may be a contributing factor to differences in results between studies and the failure to replicate.

The power of a GWAS to detect risk factors may be increased by attempting to select “genetically enriched” cases, for example those with a family history of disease or early disease onset [2]. Risk estimates from such studies may not be generalizable to disease in the population, but the primary aim of a GWAS is to identify associated genetic variants or regions rather than estimating risk. In any case estimates from a GWAS are likely to be upwardly biased because of the winner’s curse phenomenon [3]. Follow-up studies of particular variants in an independent sample can be used to provide unbiased risk estimates.

3.1.2 Selection of Controls

Similarly, the principles behind selection of controls are often more relaxed in a GWAS context, partly because of the expense of genotyping a separate large set of controls for each disease studied. This is also acceptable for a strictly genetic study, but care must be taken, e.g., over possible socioeconomic differences between cases and controls, if environmental factors are also considered at a later date (*see Note 2*). In addition a control group consisting of unselected samples from the population is likely to include cases, which will reduce power when studying very common conditions such as high blood pressure. The most important criterion for control selection for a purely genetic study is that the controls be selected from the same population as cases with respect to ethnic origin (Subheading 3.3).

3.1.3 Sample Size

Since GWAS consider a huge number of variables and aim to detect variants with modest or even low effect sizes, the sample size needs to be large to obtain adequate power to compensate for multiple testing. With one million SNPs genotyped, 50,000 would be expected to be significantly associated with outcome at the usual 5 % significance level just by chance. Additionally imputation is often used (Subheading 3.5), increasing the number of tests perhaps tenfold.

A simple Bonferroni correction for the number of tests conducted is likely to be conservative because of correlation between SNPs, which increases as marker density increases, e.g., through imputation. In practice, a significance threshold of 5×10^{-8} has been accepted by many as a threshold for “genome-wide significance” (*see Note 3*). In Table 1 power calculations are presented

Table 1
Power calculations for a Cochran–Armitage trend test assuming a
significance level of 5×10^{-8} , an additive genetic model, and a baseline
disease risk of 5 %

Minor allele frequency	Number of cases (= number of controls)	Per allele genetic relative risk	Power (%)
0.01	2,000	≤ 1.5	0
		2.0	4
0.05		≤ 1.2	0
		1.3	1
		1.5	19
		2.0	100
0.1		≤ 1.2	0
		1.3	6
		1.5	79
		2.0	100
0.3		1.1	0
		1.2	7
		1.3	66
		≥ 1.5	100
0.5		1.1	0
		1.2	13
		1.3	79
		≥ 1.5	100
0.01	5,000	≤ 1.3	0
		1.5	2
		2.0	75
0.05		1.1	0
		1.2	1
		1.3	17
		1.5	96
		2.0	100
0.1		1.1	0
		1.2	11

(continued)

Table 1
(continued)

Minor allele frequency	Number of cases (= number of controls)	Per allele genetic relative risk	Power (%)
		1.3	76
		≥1.5	100
0.3		1.1	2
		1.2	82
		≥1.3	100
0.5		1.1	3
		1.2	92
		≥1.3	100

Estimates are based on 10,000 simulations for each set of parameters. Values are in bold where at least 80 % power is achieved

using this significance level for a range of different minor allele frequencies and effect sizes assuming an additive genetic model (Subheading 3.4). It can be seen that with 2,000 cases and 2,000 controls, only common variants with quite a strong effect can be reliably detected at this level; with 5,000 cases and controls, common variants of modest effect can be detected, although the power to detect rare variants (minor allele frequency <0.05) is still quite low.

3.2 Quality Control

With the large number of SNPs analyzed, any unchecked systematic errors occurring during the genotyping process are likely to result in a number of false-positive signals of association. A number of quality control (QC) steps concerning both samples and SNPs are essential to prevent this.

3.2.1 Identifying Failed SNPs

Genotype calls from genome-wide chips are based on allele intensities; a high intensity for only one allele indicates a homozygote for that allele, while intermediate intensities for both alleles indicate a heterozygote. Plotting the intensities of one allele versus the other for all samples for a particular SNP should result in three distinct clusters of points (Fig. 1) (see Note 4). The boundaries of the clusters may be provided by the chip manufacturer but may also be defined applying a clustering algorithm to the users' own data. Genotypes are called depending on the cluster the sample falls within, and samples falling outside of the cluster boundaries are

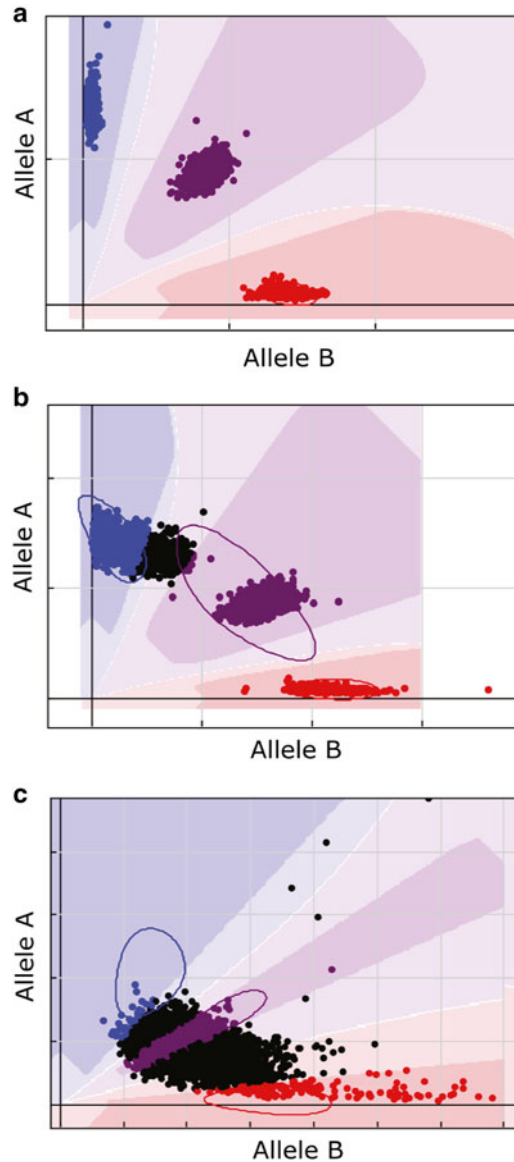


Fig. 1 Cluster plots generated from GenomeStudio (www.illumina.com/software/genomestudio_software.ilmn), showing (a) a well-called SNP with three defined clusters, (b) a poorly called SNP with a large proportion of samples (*black points*) falling outside the cluster boundary for the common homozygote (*blue points*), and (c) a poorly called SNP where no cluster separation is observed

not called for that particular SNP. Poorly clustered SNPs may not be identified by the calling algorithm, but there are a number of methods to identify these. SNPs should be excluded if they have a low call rate (<95 % or possibly stricter) or show deviation from Hardy–Weinberg equilibrium (HWE) in control samples (*see Note 5*).

3.2.2 Identifying Problem Samples

Studies involving a large number of samples or coming from several laboratories have an increased potential for sample mix-ups or labelling errors. The gender of samples can be estimated from the genotype calls on the X and Y chromosome and checked against the recorded gender. Males should have a near-zero heterozygote rate on the X chromosome. Females should have heterozygote rate $>20\%$ on the X chromosome and missing Y chromosome data. The recorded gender of any discrepant samples should be checked. An additional check on potential sample mix-ups is to look at concordance rates between any existing genotype data available for the samples.

With large numbers of SNPs on a relatively homogeneous sample, pairwise estimates of the proportion of alleles shared identical by descent can be obtained and used to identify closely related and duplicate samples. A closely related pair of individuals will share a far greater number of alleles identical by descent than an unrelated pair. Duplicate samples may arise from sample mix-ups or unknowingly recruiting the same sample more than once. Generally the sample within the duplicate or the closely related pair (or group) that has the highest call rate will be retained and the other sample(s) dropped, although related samples may be retained and adjusted for in the analysis [4].

3.2.3 Data from Different Sources

When combining genotype data from several different sources, QC should be carried out separately for each dataset. This is because a SNP may be of poor quality in one of the datasets but not in the other(s). Performing QC after combining could have two consequences if genotyping is poor on a subset of samples: (1) The SNP may be unnecessarily dropped from all the datasets as it fails QC on the combined set, or (2) the marker passes QC on the combined dataset, leading to unreliable results.

3.3 Population Stratification

Ideally the samples used in a study should be collected with the aims of that study in mind. However, given that GWAS are so large, existing sample collections are often used. The collection of these samples may thus not have taken into account issues of specific importance to GWAS, most notably the problem of population stratification. Population stratification occurs when a sample consists of distinct subpopulations between which there is little mating, so that allele frequencies may differ between subpopulations. If the proportions of each subpopulation differ between cases and controls, any variants that differ in frequency between these subpopulations will appear to be associated with disease risk (*see Note 6*).

Differential sampling across subpopulations may occur as a result of bad design, by chance, or because one subpopulation has a higher incidence of disease (for cultural, environmental, or genetic reasons). The differences in allele frequencies seen in samples

from different parts of Europe are too small to have affected candidate gene association studies but may give rise to false positives in GWAS, that are designed to find variants of smaller effect. Thus, case–control matching in GWAS should be tighter than merely being from the same continent, e.g., matching by country may be appropriate within Europe (*see* **Note 7**).

The simplest method for identifying population stratification is to check for deviation from HWE, although this will only pick up strong stratification.

Another simple approach is that of the Q–Q (quantile–quantile) plot. Here the test of association is conducted, and the ordered test statistics for all SNPs (having excluded those of low quality) are plotted against the corresponding quantiles of the distribution expected under the null hypothesis of no association. While some SNPs may genuinely be associated with disease, the vast majority should not. Thus, a deviation from the $x = y$ line other than at the highest quantiles indicates that there is some overall inflation of the test statistics. Such inflation is usually measured by dividing the median of the test statistics by its expectation under the null hypothesis, denoting this ratio λ . This gives rise to the genomic control method for correcting for such inflation, whereby each test statistic is divided by λ [5, 6]. However, this is a rather crude approach that is only really suitable for quite distinct subpopulations [7].

The most commonly used approach for detecting and adjusting for population stratification is the application of principal component analysis (PCA) [7] (using EIGENSTRAT, *see* Subheading 2). The idea is that individuals who are geographically close are likely to be more correlated in terms of genotypes (i.e., closely related) than those who are far apart. Even if there is only a slight correlation on a SNP-by-SNP basis, when this is considered genome wide, it may be enough to distinguish between subpopulations (when these are distinct) or reveal gradients in SNP frequencies when there are no distinct subpopulations. The first principal component gives the linear combination of genotypes that best captures the variation in the data. The second principal component is the orthogonal combination that best captures the remaining variation in the data, and so forth.

The sample may be combined with data from across the world, for example the HapMap data (www.hapmap.org), which include samples from Europe, Asia, and Africa, and PCA applied to the combined sample. This will identify individuals who are ethnically distinct from the rest of the sample. For example, in a study of a European population, combining with the HapMap data and applying PCA will produce principal components 1 and 2 that separate out the continents into three distinct clusters. Anyone who is not of European origin will appear in one of the non-European clusters [8]. Ethnic outliers from the sample can then be excluded from further analyses.

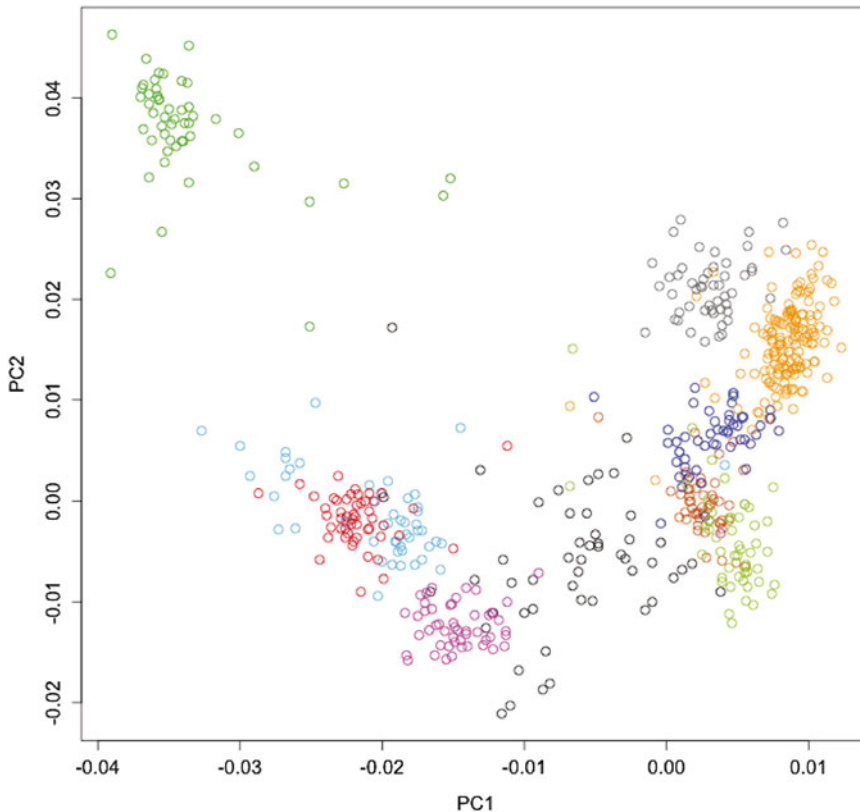


Fig. 2 Plot of principal components 1 and 2 for a sample of European and Israeli origin. Each circle indicates a single sample, and color indicates the country/city from which the sample was collected: England (*dark brown*), Genoa (*light blue*), Cesena (*red*), the Netherlands (*dark blue*), France (*black*), Sweden and Norway (*orange*), Spain (*magenta*), Scotland (*light green*), Israel (*dark green*), and Poland (*grey*). Note that not only does PCA group individuals from the same country/city together, but it also arranges countries according to their geographic location, allowing rough north–south and east–west gradients to be observed

Once outliers have been removed to give a more homogeneous sample, PCA can be reapplied to detect more subtle stratification. It has been shown [9, 10] that the first and second principal components are likely to correspond to orthogonal two-dimensional geographical axes, such as latitude and longitude. Further principal components may correspond to further orthogonal axes, depending on the structure of the population. It is important when applying PCA that the SNPs are first thinned so that linkage disequilibrium (LD) is minimized; otherwise, the principal components may just pick out regions of strong LD. In Europe, PCA distinguishes individuals from different countries extremely well [10, 11] (Fig. 2). The principal components may then be used to adjust for population stratification, for example by including as covariates in a logistic regression on disease status.

Another, more recent, approach is to use a mixed effects model. Here the SNP genotypes and other phenotypic variables of interest

(sex, age, etc.) are modelled as fixed effects, while the population structure is modelled as a random effect based on estimated relatedness between individuals (from genotypic information). It is only recently that such approaches have become computationally feasible (EMMAX [12] and TASSEL [13], *see* Subheading 2).

While such approaches to correcting for stratification work well for common genetic variants, there is evidence that they may be less successful for rarer variants. Rare variants are likely to have arisen more recently and so correlate more weakly (if at all) with the broad demographic history recreated by PCA or mixed effects modelling. This means that testing of such variants (say those with minor allele frequency <0.05) may result in an excess of false-positive associations with disease status [14]. To date there has been little investigation into this as commercial chips have typically focussed on capturing common variation. However, with the advent of much denser chips and large-scale sequencing, rarer variants are becoming more widely studied, and careful study design becomes ever more important.

3.4 Association Analysis

Post-QC genotype data are commonly analyzed (*see* Note 8) using the Cochran–Armitage trend test, which has 1 degree of freedom and assumes a log-additive mode of inheritance. This has consistently good power over a wide range of genetic models (although of course specific models would have greater power for particular SNPs) [15].

3.4.1 Alternative Analyses

Logistic regression of case–control status on genotype (measured as a continuous trait taking the value 0 for the common homozygote, 1 for the heterozygote, or 2 for the rare homozygote) is asymptotically equivalent to the trend test. This is useful in order to adjust for other factors, such as age, gender, geographical origin, or the first few principal components (Subheading 3.3). Similarly, if the study outcome is a continuous trait, such as body mass index, linear regression can be performed.

3.4.2 Examination of Results

The p -values from the test used are often plotted on a log scale against the SNP chromosomal position (Fig. 3). The peaks of the Manhattan plot enable initial identification of significant regions. Once analysis has been performed, any significant SNPs should be checked thoroughly against the above QC criteria. Particular attention should be paid to those SNPs that are highly significant but whose neighboring SNPs show no evidence of association, which will appear as isolated points on the Manhattan plot. This could only be a possible true association in the unlikely event that the highly associated SNP was not in LD with any of the neighboring SNPs; a more likely explanation is genotype error. Finally, the cluster plots (Fig. 1) of the remaining significant SNPs should be examined before any attempt at replication analysis.

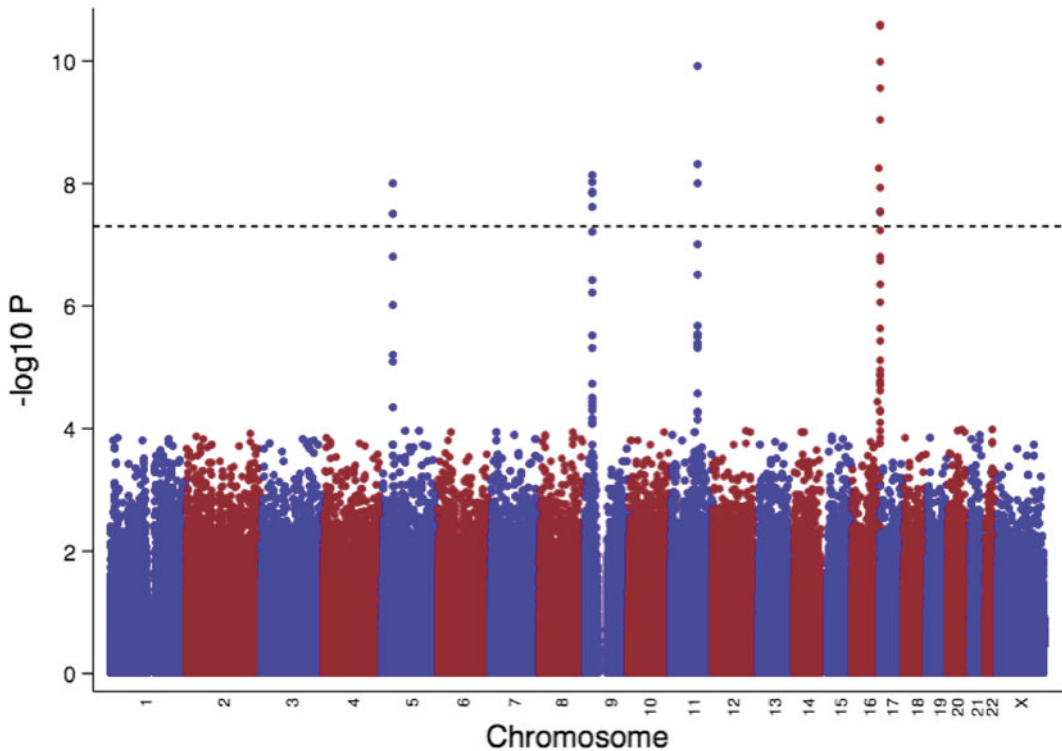


Fig. 3 Manhattan plot of p -values from a Cochran–Armitage trend test showing four regions on chromosomes 5, 9, 11, and 16 that are significantly associated with outcome at a level of 5×10^{-8} , indicated by the *dashed line*

3.5 Imputation

Even though commercial genotyping arrays now include a huge number of polymorphisms, these are still a subset of all known genetic variation. Many of the untyped genetic variants will be of less interest to researchers, either because they are so rare that even if they are associated with disease risk a typical GWAS will be underpowered to detect this or because they are in such strong LD with genotyped markers that they add little information. However, more exhaustive coverage of genetic variation may be of interest for two reasons: (1) a region may be associated with disease risk, but only the ungenotyped markers have a strong enough signal to reach the significance threshold and so the association is missed using data from the array alone, and (2) a region is established as being associated with disease, but there is interest in fine-mapping the region to understand better the source of the association, ideally to identify the functional variants, but at least to delimit the associated region (Fig. 4).

Even for a small region, either genotyping all variants or sequencing all individuals in the GWAS may be prohibitively expensive, particularly if the region in question is large. One alternative is to impute the genotypes at those SNPs that have not been typed. By using a dataset such as the 1000 Genomes Project

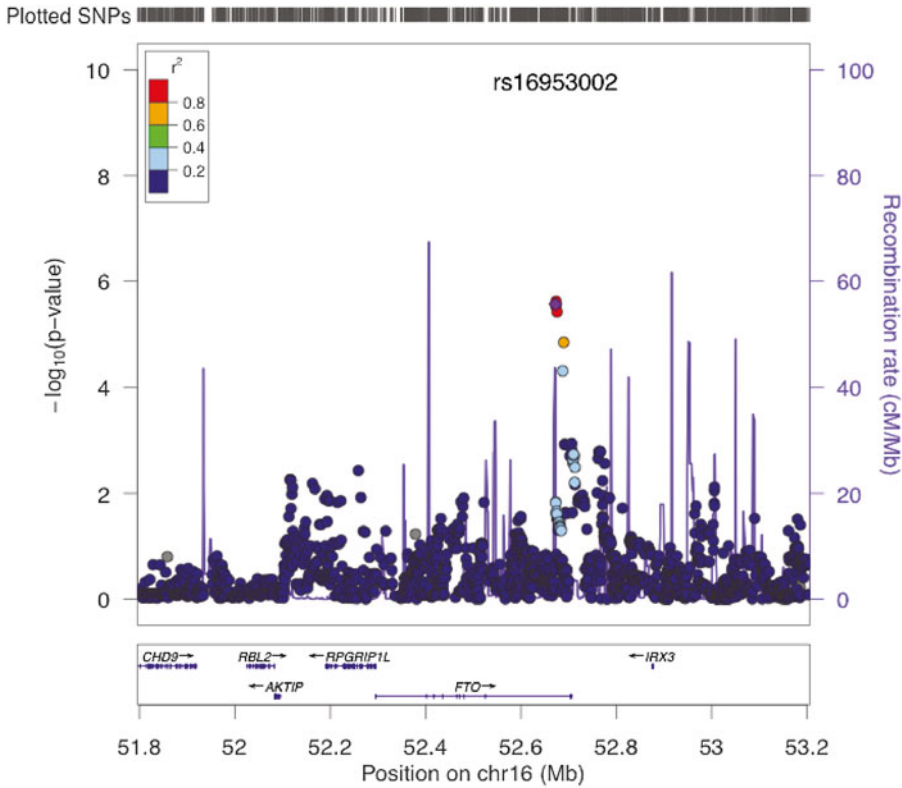


Fig. 4 A Manhattan-style plot of the region around the *FTO* gene, which has been associated with melanoma risk after imputation of a GWAS case-control study [21]. The *x*-axis shows position along the chromosome, the *left-hand y*-axis (and the *points*) $-\log_{10} p$ -value, and the *right-hand y*-axis (and the *line*) the estimated recombination rate. The most significant SNP (rs16953002) is colored *purple*, and the remaining SNPs are colored according to the degree of LD with this SNP. Plotting performed using LocusZoom [22]

(www.1000genomes.org), which contains far denser genotype data than is currently available on commercial SNP chips, the pattern of LD between nearby SNPs can be established and then applied to the sample data so that genotypes of untyped SNPs can be estimated. The most popular programs for this are IMPUTE [16], MACH [17], and BEAGLE [18] (*see* Subheading 2). Those SNPs that are estimated with suitable confidence in sufficient samples can be treated as though genotyped and the data analyzed as usual, although it is more correct to account for this uncertainty, using either the full posterior distribution or, as an approximation, the expected genotype count (or dosage). Researchers should of course be more cautious about the results at SNPs that are imputed rather than genotyped. Before imputation, the usual QC should be applied and low-quality markers excluded, although it is advisable to apply more stringent QC to genotyped markers before imputing, since a poorly genotyped marker may adversely impact a number of imputed markers. Post-imputation QC should also be applied to identify poorly imputed SNPs (*see* Note 9).

Reporting either a novel associated region or fine-mapping of a region of interest based solely on imputed genotype data is likely to be treated with some scepticism, particularly for rare markers. Thus, it is good practice to genotype the key variant(s), at least in a subset of samples, to establish that the imputation is working reliably.

3.6 Next Steps

In practice loci will not be fully established as associated with a disease or a trait until a high level of statistical significance has been reached and the finding has been replicated in independent samples. Once this has been achieved there is much work still to do in understanding what underlies the finding. Detailed consideration of this is beyond the scope of this chapter. An important element is fine-mapping, where imputation and denser genotyping are used to narrow down the association signal and identify the most parsimonious model(s) that explains the observed associations. This may turn out to consist of a single SNP, but it is also common to find that more than one variant and even more than one gene show independent association within the region. Additionally, bioinformatics can be used to mine the publically available databases to evaluate the potential function of any SNPs identified by fine-mapping.

4 Notes

1. For a continuous outcome, such as blood pressure, a population sample could be analyzed, although it may be more powerful to select subjects with extreme phenotypes [19].
2. The relaxed attitude to case ascertainment has, however, limited the wider utility of many large GWAS, for example in looking at survival or incorporating information on environmental exposure.
3. Although a significance level of 5×10^{-8} is now generally regarded as the benchmark for “genome-wide significance,” like all p -value thresholds, this is only useful if interpreted intelligently. In particular, there will still be some false-positive results (especially with serial testing); conversely, a result just failing to meet this threshold but with prior or external supporting evidence is unlikely to be a false positive.
4. This assumes that allele frequencies are common enough to generate heterozygotes and both sets of homozygotes. Since three clusters are expected, this can lead to errors in calling the genotypes when the minor allele is very rare.
5. In addition, the log R ratio (ratio of observed to expected intensity based on other samples) and the B allele frequency can be used to detect copy number variation and other

chromosomal anomalies. It is also wise to check for unexpected differences (in allele frequencies or effect sizes) between batches or plates.

6. Subpopulations, even within Europe, might be quite distinct, e.g., population isolates such as Sardinia. More common, and more of a problem, is subtle stratification, as it is harder to detect and more difficult to correct for analytically.
7. Existing datasets may not record information on ethnicity, or the definition of ethnicity may not be precise enough to identify homogeneous groups: for example samples from the USA of “European origin” may be quite diverse. Fortunately, methods exist for detecting unobserved (cryptic) population stratification and correcting for this, as discussed here, but it should be remembered that such approaches are never as satisfactory as a well-matched case–control set.
8. Specialized software such as PLINK [20] is required to reduce analysis times for a large number of SNPs.
9. The most commonly used QC metric for imputed markers is to estimate how much of the variation in a marker has been captured by imputation as compared to how much would be expected if the marker were genotyped (also interpretable as the correlation between the imputed genotype and the actual genotype, were it available). Both the INFO score in IMPUTE and R^2 in MACH and BEAGLE are versions of this metric. There are no strict thresholds for this metric, and values from 0.3 to 0.9 have been utilized. Given that this is, in effect, a ratio of two estimated variances, the score will be less reliable for rarer variants.

References

1. Hindorff LA, MacArthur J, Morales J et al (2013) A catalog of published genome-wide association studies. www.genome.gov/GWASstudies. Accessed 11 Apr 2014
2. Antoniou AC, Easton DF (2003) Polygenic inheritance of breast cancer: implications for design of association studies. *Genet Epidemiol* 25:190–202
3. Xiao R, Boehnke M (2009) Quantifying and correcting for the winner’s curse in genetic association studies. *Genet Epidemiol* 33:453–462
4. Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. *Stat Sci* 24:451–471
5. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55:997–1004
6. Devlin B, Roeder K (2001) Genomic control: a new approach to genetic-based association studies. *Theor Popul Biol* 60:155–166
7. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
8. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678
9. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. *Nat Genet* 40:646–649
10. Novembre J, Johnson T, Bryc K et al (2008) Genes mirror geography within Europe. *Nature* 456:98–101

11. Bishop DT, Demenais F, Iles MM et al (2009) Genome-wide association study identifies three loci associated with melanoma risk. *Nat Genet* 41:920–925
12. Kang HM, Sul JH, Service SK et al (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348–354
13. Zhang Z, Ersoz E, Lai CQ et al (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42:355–360
14. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44:243–246
15. Iles MM (2010) The impact of incomplete linkage disequilibrium and genetic model choice on the analysis and interpretation of genome-wide association studies. *Ann Hum Genet* 74:375–379
16. Marchini J, Howie B, Myers S et al (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. *Nat Genet* 39:906–913
17. Li Y, Willer CJ, Ding J et al (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34:816–834
18. Browning SR (2008) Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet* 124:439–450
19. Wallace C, Chapman JM, Clayton DG (2006) Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *Am J Hum Genet* 78:498–504
20. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
21. Iles MM, Law MH, Stacey SN et al (2013) A variant in *FTO* shows association with melanoma risk not due to BMI. *Nat Genet* 45:428–432
22. Pruim RJ, Welch RP, Sanna S et al (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26:2336–2337

Bioinformatics Challenges in Genome-Wide Association Studies (GWAS)

Rishika De, William S. Bush, and Jason H. Moore

Abstract

Genome-wide association studies (GWAS) are a powerful tool for investigators to examine the human genome to detect genetic risk factors, reveal the genetic architecture of diseases and open up new opportunities for treatment and prevention. However, despite its successes, GWAS have not been able to identify genetic loci that are effective classifiers of disease, limiting their value for genetic testing. This chapter highlights the challenges that lie ahead for GWAS in better identifying disease risk predictors, and how we may address them. In this regard, we review basic concepts regarding GWAS, the technologies used for capturing genetic variation, the *missing heritability* problem, the need for efficient study design especially for replication efforts, reducing the bias introduced into a dataset, and how to utilize new resources available, such as electronic medical records. We also look to what lies ahead for the field, and the approaches that can be taken to realize the full potential of GWAS.

Key words Data imputation, Epistasis, Electronic medical records, Filtering, Gene–gene interactions, GWAS, Meta-analysis, Missing heritability, Replication

Abbreviations

EMR	Electronic medical record
GWAS	Genome-wide association study/studies
LD	Linkage disequilibrium
MAF	Minor allele frequency
SNP	Single nucleotide polymorphism

1 Introduction

In the field of genetics and epidemiology, genome-wide association studies (GWAS) have become a standard approach for querying the genetic basis of disease susceptibility. This study design measures and analyzes a million or more DNA sequence variations such as single nucleotide polymorphisms (SNPs) that capture

much of the common variation in the genome, in an effort to identify genetic risk factors for diseases [1]. Moreover, technological advances that have lowered the cost of genotyping have also fueled an increase in the number of GWAS over the years. In 2012 alone, the National Human Genome Research Institute (NHGRI) GWAS catalog recorded 1,350 published studies [2]. GWAS provide us with a unique opportunity to make disease risk predictions for the general population on the basis of the disease susceptibility loci that are identified. Knowledge of these loci may also provide clues to the biological basis for various diseases, and open up new avenues for prevention and treatment strategies. The key steps involved in conducting a GWAS are summarized in Fig. 1.

The GWAS approach is not *hypothesis-free*; it is based upon the Common Disease—Common Variant (CD–CV) hypothesis. This hypothesis ties together the basic principle of GWAS and the design of genotyping chips. It states that common diseases are caused in part by genetic variations that are also common in the population [3]. Testing the CD–CV hypothesis provides an insight into the underlying genetic architecture of common diseases, e.g., type 2 diabetes, rheumatoid arthritis, or essential hypertension, and some evidence that they are driven by multiple susceptibility alleles. If common variants have a small effect size but common diseases show a strong inheritance in families (high heritability), then almost by definition the disease must be influenced by multiple genetic factors. For example, if a disease shows a heritability of 30 %, this indicates that 30 % of the total variance in the disease risk comes from genetic factors. Hence, if a SNP has a modest effect on disease risk, it can only account for a small portion of the total variance due to genetic factors. Consequently, the total risk of disease due to common genetic variation then must be distributed over multiple susceptibility alleles.

Published concurrently with family-based linkage studies, one of the earliest GWAS success stories was the identification of Complement Factor H as a major risk factor for age-related macular degeneration [4–7]. This study not only showed that DNA sequence variations in the gene were associated with the disease but also provided a new insight into the biological basis for the disease. However, despite the moderate success of the risk variants identified for age-related macular degeneration, most loci identified by GWAS are known to be associated with small increases in disease risk, thereby limiting their value for genetic testing [2, 8, 9].

The example of breast cancer best highlights the failures and successes of GWAS during its tumultuous history. Familial breast cancer, a rare disease with high heritability, is believed to have a simple underlying genetic architecture. In 2007, Easton et al. identified five significant associations by GWAS that were also replicated in multiple independent samples [10]. In a follow-up study two additional susceptibility loci were identified.

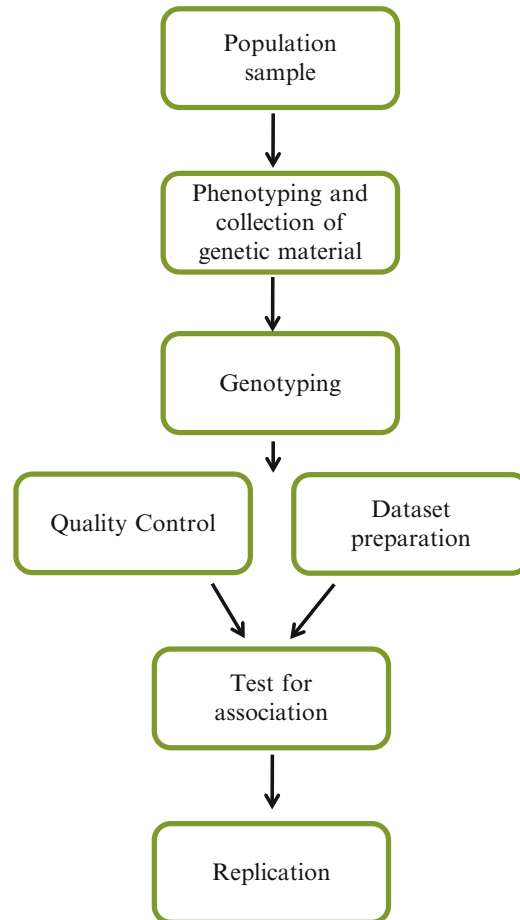


Fig. 1 Overview of the GWAS process. A sample of individuals, e.g., a group of families or cases/controls, is selected from the population to study a disease or phenotype of interest. After strict criteria have been established, phenotypic information and genetic material are collected from the study participants. This is followed by genotyping of the collected material using popular genotyping platforms such as those available from Illumina or Affymetrix. Genotypic data obtained for genetic variants such as SNPs (single nucleotide polymorphisms) are cleaned using quality control procedures such as MAF (minor allele frequency) or LD (linkage disequilibrium) filtering. Data are also adjusted for various covariates and population stratification if required. Next, single locus or multi-locus association tests can be performed to identify genetic variants associated with the phenotype of interest. Ultimately, identified genotype–phenotype associations must be replicated in an independent dataset to assert their credibility

These two loci accounted for <1 % of the familial risk of breast cancer [11]. When these loci were combined with previously known genetic risk factors, together they were able to explain only 5.9 % of the familial risk of breast cancer. On the other hand, the *BRCA1* and *BRCA2* mutations together account for 20–40 % of familial breast cancer, and have been very successful as markers for

genetic testing. Hence, although the susceptibility loci identified via GWAS analyses have been useful in providing a new insight regarding the biology of the disease, they have not resulted in new genetic tests.

Similar to the discouraging results with familial breast cancer, GWAS has had limited success in detecting genetic variants that account for a large portion of the heritability of any common disease trait. This chapter will highlight the challenges that lie ahead for GWAS in identifying genetic risk factors that are better classifiers of disease, and how these may be addressed.

First, we must address this *missing heritability* problem by specifically designing our studies to search for nonlinear interactions amongst SNPs. However, this can be a very computationally intensive problem due to the enormous number of pair-wise combinations possible from a GWAS dataset. Linkage disequilibrium (LD) patterns within a dataset may be used to devise strategies for prioritizing SNPs for inclusion in an analysis, reducing this computational burden. Second, we must improve the study design of our GWAS to ensure we increase our statistical power to detect true genetic effects and replicate them by using methods such as meta-analysis and data imputation. Additionally, we must make use of new resources such as electronic medical records (EMRs) to unravel a wealth of phenotypic detail that was previously unavailable. Third, to reduce the biases in GWAS design, we must establish strict criteria for defining phenotypes, adjust for confounding variables that may affect the phenotype of interest, and correct for multiple hypothesis testing (*see* also Chapter 4).

2 Materials

2.1 Genotyping Platforms

Much of the growth and success of GWAS reflects the technology behind the thumb-sized DNA microarray chips designed to probe one million or more SNPs dispersed throughout the genome.

The genotyping platforms used by most GWAS belong to one of two commercial companies: Illumina (San Diego, CA) or Affymetrix (Santa Clara, CA). The two companies' products differ slightly in their approaches to measure SNP variation, and provide researchers with options in terms of cost, coverage, amount of target DNA required, and protocol complexity.

Affymetrix chips use a printed array format, where each spot on the array, representing a locus or allele, contains a cluster of 25-mer oligonucleotides. This platform also offers a cost-effective approach for high-volume GWAS, as most costs are mainly up front. Illumina, however, produces chips that consist of an ordered array of beads, each representing 50-mer oligonucleotides. Even though this platform offers higher sensitivity, it comes at a cost—the arrays are more expensive and the protocols for decoding bead

positions are time intensive. Ragoussis et al. provide an excellent review of these genotyping platforms and their unique strengths and weaknesses [12]. Ultimately, GWAS using either of these platforms have been equally successful in the search for genetic risk factors for common, complex diseases.

2.2 Electronic Medical Records

EMRs, which were primarily designed for hospital administrative processes, have recently given way to a new model of genetic discovery. These records are being used to extract relevant phenotypic information for a subject population. Medical centers can leverage this information for genetic studies by linking these data to biological samples to create large-scale biobanks.

EMRs are a rich resource for different types of information, such as—billing data, diagnosis codes, laboratory results, vital signs, provider documentation from reports and tests, and medication records. The billing data and certain laboratory results are made available as structured “name-value” pair data. The clinical documentation such as test results and medication records are provided as narrative or semi-narrative texts. Provider documentation and medical records form an important resource for correct phenotype characterization. Many hospitals are also installing barcodes to keep records of each drug administration for all patients, which may improve accuracy of pharmacogenomic traits [13].

3 Methods

3.1 Basic Concepts for GWAS

3.1.1 Single Nucleotide Polymorphisms and Minor Allele Frequency

A major goal for both the International HapMap Project as well as the 1000 Genomes Project was to capture and catalog sequence variation in the human genome [14, 15]. SNPs, which are single base pair changes in the DNA sequence, have now become the modern unit of genetic variation. Currently, the public catalog of variant sites (dbSNP Build 138) contains approximately 44 million SNPs [16].

SNPs have been shown to have important functional consequences such as affecting mRNA transcript stability or transcription factor binding affinity [17]. However, it is the ability of SNPs to explain much of the genetic diversity observed amongst humans that makes them ideal candidates for use as markers of a genomic region in GWAS.

For each SNP location, there are two or more allele possibilities. The frequency of the less common allele is referred to as minor allele frequency (MAF). The MAF, along with the minor allele, can be specific to a population. Variants can be classified as common or rare: SNPs with a MAF $\geq 5\%$ are usually referred to as common variants, and those with MAF $< 5\%$ are rare. For example, a SNP with a minor allele (A) frequency of 0.30 indicates that 30% of the population carries the A allele at the SNP location, instead of the

more common allele. Traditionally, most GWAS focus on common variants as witnessed by the long list of validated examples—*FTO* (type 2 diabetes and body mass index) [18–20], *GCKR* (triglycerides) [21], and *APOE4* (Alzheimer disease) [22]. Nevertheless, it has been suggested that rare variants play a role in disease and hereditary risk as well [23, 24].

3.1.2 Linkage Disequilibrium

Linkage disequilibrium (LD) is a measure of correlation between SNP alleles at one site and the specific alleles carried at variant sites nearby. Likewise, a particular combination of alleles along a chromosome is termed a *haplotype*. The concept of LD is closely related to chromosomal linkage, where two markers on a chromosome are physically linked through multiple generations of a family. Both these properties can be eroded by recombination and mutation events across multiple generations, thereby breaking up any contiguous stretches of chromosome. The LD observed in a population is also dependent upon its ancestry. Consequently, populations of African descent show smaller regions of LD, as they are more ancestral compared to Asian and European populations and have undergone greater extents of recombination.

The most common measures for LD are— D' and r^2 . Both measures try to capture the difference in the observed frequency of two alleles that occur together, and how often they would be expected to occur together if they were independent of each other [14, 25].

Measures of LD are extremely useful in GWAS design. r^2 values are used to select *tag SNPs*, which are variants selected specifically because they are in strong LD with other variants surrounding them (*see* **Notes 1** and **2**). This advantageous property of tag SNPs allows them to be used for capturing the variation in that specific stretch of LD. Tag SNPs have been especially useful in reducing genotyping costs for GWAS. According to HapMap, more than 80 % of commonly occurring SNPs in populations of European descent can be captured by a set of 500,000 to a million SNPs spread across the genome [26, 27]. This is what forms the basis of selecting a panel of markers for genotyping chips.

An understanding of LD is also essential for correct comprehension of results from a GWAS analysis. Positive results from a GWAS may represent two types of associations—direct or indirect. A direct association involves a SNP that was directly genotyped in the study. Such a SNP is also referred to as a *functional SNP* or the causal variant. An indirect association is a positive association where the SNP of interest was not directly genotyped in the GWAS. This association represents a tag SNP that was genotyped in the study and is in strong LD with the variant altering the biology of the organism. Usually, follow-up tests, such as resequencing the specific genomic region or performing functional studies to examine the role of the variant in the disease, are required to distinguish between these two possibilities [1, 28].

3.2 GWAS Study Design

In this section we begin with an overview of GWAS design and the steps that we can take to reduce the bias introduced into a dataset—such as setting a rigorous criteria for defining phenotypes. Moreover, we talk about the exciting opportunity of extracting phenotypic information from EMRs and the unique challenges that presents.

3.2.1 Case–Control Design Versus Quantitative Design

Case–control GWAS utilize categorical phenotypes, which are often binary outcomes such as case/control or affected/unaffected. The case group includes individuals who have been diagnosed with the disease phenotype of interest. However, the control group can be chosen in one of two valid ways—individuals who are unaffected by the disease or randomly selected from the population. To avoid false positive results, cases and controls must be matched carefully (*see Note 3*). Overall, the case–control study design compares the frequency of SNPs or alleles between the two study groups. A higher frequency of a SNP within the cases instead of controls is indicative of that SNP being associated with increased disease risk [28, 29].

As the name suggests, *quantitative* study designs assess quantitative or *continuous* traits that can be measured, to obtain a quantitative value such as HDL (high-density lipoprotein) and LDL (low-density lipoprotein) cholesterol levels [30]. This study design is statistically more powerful for detecting genetic effects. Quantitative traits also make it easier for researchers to obtain a precise measurement, and provide an outcome that is clinically easier to interpret. Ultimately, such a study design measures if the frequency of a SNP or allele is associated with a certain amount of change in the quantitative trait being studied [31].

3.2.2 Standardizing Phenotype Criteria

For any GWAS, it is important to establish measures to standardize the criteria for defining the phenotype of interest. This is especially true for diseases that do not have well-established quantitative measures to describe the disease phenotype, such as multiple sclerosis. In such a situation, patients are usually classified as either being *affected* or *unaffected* by the disease in question. However, in these cases a simple misclassification error of categorizing someone as a *case* instead of a *control* can have more serious consequences than an error in recording a precise quantitative measure.

Despite a complex clinical phenotype that is difficult to diagnose, multiple sclerosis studies have been successful [32]. This is mainly because these studies use a rule list based on various clinical variables such as the McDonald criteria to establish case–control status [33]. This is especially important when studies are based on collaborations between multiple institutions and centers. In such cases, strict criteria ensure that phenotype definitions are applied uniformly across various clinicians, thereby avoiding any site-based effects. This brings to notice that the success of a GWAS does not

always depend on the nature of the phenotypic outcome being analyzed, but rather on the awareness of the specific challenges each phenotypic category presents.

3.2.3 *Extracting Phenotypes from EMRs*

In recent years, the growth of EMR-linked DNA databanks has presented an exciting new avenue for genetic research. EMRs provide an alternative source for researchers to derive phenotype information about a large population of individuals. They are especially appealing as they contain a longitudinal record of robust clinical data due to routine clinical care of patients. Furthermore, EMR-linked DNA databanks provide researchers with the unique opportunity of reusing genetic information to investigate additional phenotypes. Nevertheless, identifying phenotypes from EMRs presents its own set of challenges, because these records were designed with the logistical problems and billing practices of hospitals in mind [13, 34].

The first step in phenotype extraction involves the use of an initial selection algorithm that chooses a subset of records from the bio-repository through text mining of unstructured text or by making use of structured data fields such as billing codes, in the EMR. The choice of billing codes available for use in the EMR is also important in ensuring the accuracy of the phenotype information extracted or the diagnosis established from the record. The CPT (Current Procedural Terminology) coding system is known to have a higher specificity and lower sensitivity, in comparison to the ICD (International Classification of Diseases) coding system. Though the availability of a single type of code is usually sufficient for identifying a phenotype, often a combination of the codes works better as ICD codes also provide the reason for a clinical encounter or procedure [13].

Similarly, to complement the information from billing and procedure codes, they can be combined with free text in the EMR. Such free text can be parsed using Natural Language Processing (NLP) procedures, which apply syntactic and semantic rules to extract structured information. They do so by connecting the text with medical concepts from a controlled vocabulary such as the Unified Medical Language System (UMLS) or with medication information from vocabularies such as RxNorm [13, 35–37] (see also Chapter 16).

Ultimately, as a gold standard measure, clinicians and phenotype experts examine the accuracy of the results obtained from the subset of EMR records selected for the study. A measure of precision, the positive predictive value (PPV) of the initial selection algorithm is assessed. The algorithm is then continually refined based on the feedback from these experts. This process continues until the desired PPV is achieved [13, 34]. This approach has not only been applied to various pharmacogenomic and clinical conditions [38–41], but has also successfully replicated established genotype–phenotype relationships [42].

3.3 Testing for an Association

In addition to ensuring a strong study design, there are a few challenges that must be addressed at the level of association analysis in a GWAS. In this section we describe the steps involved in preparing a dataset prior to an association analysis and in adjusting for confounding variables that may also affect the phenotypic outcome of interest. Moreover, the *missing heritability* problem is addressed, as are the steps that can be taken during an association analysis to prioritize SNPs and search for nonlinear interactions.

3.3.1 Dataset Preparation Prior to an Association Analysis

Prior to testing for a genetic association with a disease outcome of interest, researchers must go through a few steps to prepare their dataset for this analysis. First, a method must be chosen for encoding the genotype information in the dataset, as this may have important implications on the statistical power of an association test. As such, association tests can test for either allelic or genotypic associations. Allelic associations look for an association between an allele and the phenotype of interest, whereas genotypic associations search for associations between genotypes or genotypic classes and the phenotype. There are several ways to form these genotypic classes—using a dominant, recessive, multiplicative, or additive model [29, 31].

Datasets must also be adjusted for a range of factors or covariates—age, sex, clinical covariates like Body Mass Index (BMI) or the study site used for data collection—that are known to affect the phenotype outcome, to prevent spurious associations from being detected. Regression methods are a popular choice for covariate adjustment; logistic regression is used for binary traits and linear regression for examining quantitative traits. These methods calculate the “residuals” for the trait of interest, after covariate adjustment. This is the portion of the trait that is not accounted for by the covariates [43].

Population substructure is one of the more important covariates to address in a dataset, especially when the population comprises various ethnicities. The prevalence of a disease phenotype, as well as allele frequencies, can vary between different human subpopulations. Due to this, within a dataset of multiple ethnicities, ethnic-specific SNPs may show up to be associated with a trait due to population stratification [44]. To prevent any false associations, the ancestry of each subsample needs to be measured using one of various methods such as STRUCTURE [45] or EIGENSTRAT [46]. These methods compare genome-wide allele frequencies with ethnic-specific frequencies on HapMap. This allows for samples to be excluded if they are found to be similar to a nontarget population. As an alternative, EIGENSTRAT can also use a statistical method such as principle component analysis (PCA) to generate principle component values or *ethnicity scores*, which can then be used as covariates for adjustments.

3.3.2 Testing for an Association: Single Locus Versus Multi-locus

The popular approach for analyzing GWAS data includes a series of single-locus statistical tests, which compare the genotype distributions for cases and controls, one SNP at a time. On the whole, these methods aim to identify an association between a SNP and the disease/phenotype of interest. However, the type of association test chosen is dependent upon the phenotypic class (case–control or quantitative) being studied.

Binary traits and case–control study designs are analyzed using a contingency table method or logistic regression. For a set of cases and controls, a contingency table summarizes the number of individuals within each genotypic group for a single biallelic SNP [28]. It searches for a deviation from the *null hypothesis* that there is no association between the phenotype and genotype. Popular statistical tests using this method are the chi-square test or the Fisher’s exact test. In addition, contingency tables can be analyzed using standard statistical software packages such as SAS, SPSS, Stata, or Microsoft Excel [29]. As for logistic regression, it is an extension of linear regression where the phenotypic outcome studied is transformed using a logistic function. This method predicts the probability of an individual having a *case* status, given their genotype class. Moreover, logistic regression is often the method of choice as it allows for covariate adjustment.

A popular method for analyzing quantitative traits is the Analysis of Variance (ANOVA), which is similar to linear regression with a categorical predictor variable. For single-SNP analysis, ANOVA functions under the null hypothesis, which states that there is no difference between the trait means for any genotype group. However, ANOVA does function on the basis of certain assumptions: it assumes that the trait is normally distributed, the variance of the trait is the same within each group, and that the groups are independent.

For such GWAS analysis, PLINK is a popular and useful software. It has robust features to handle large amounts of data. It can perform association tests per SNP using either the allelic or inheritance model, or by using the Cochran-Armitage test (a contingency table method). Most importantly, PLINK provides a very detailed user manual that is easy to follow [47].

As mentioned earlier in this chapter, the field of GWAS has had limited success in detecting genetic variants that explain a large portion of the heritability for any given trait. This has led researchers to propose potential sources of *missing heritability*. One such possibility is that *missing heritability* may be found within epistatic interactions between various genes [48]. *Epistasis* is usually defined in one of two ways—biological or statistical. Biological epistasis refers to the physical interactions between biomolecules that are influenced by multiple genetic variants. Statistical epistasis is the term for the nonadditive interactions between multiple genes, each of which affects disease susceptibility, and the environment [49, 50].

The *missing heritability* problem may be exacerbated by GWAS approaches that use a linear modeling framework to analyze SNPs one at a time, thereby failing to recognize the genetic and environmental context of each SNP [51, 52]. Hence, this has led to the adoption of more holistic approaches that recognize the complex landscape of the genotype–phenotype relationship and examine nonlinear interactions between genetic variants throughout the genome. This is referred to as a *multi-locus analysis*, which brings with it a new set of challenges [53, 54]. Amongst these, the biggest challenge is that the exhaustive examination of all pair-wise interactions involving 500,000 SNPs can be very computationally intensive. This often makes it necessary to use specific criteria to filter the 500,000 markers to make the problem computationally tractable.

Traditionally, most GWAS approaches using a chip of this size perform an initial filtering based on MAF, LD, and other initial quality control checks [47]. Even though these steps reduce the number of markers greatly, a researcher may still be left with about 300,000 SNPs in the dataset. In such cases, a single SNP analysis can be performed to select markers with main effects (these are single SNPs that show a strong association with the disease outcome), based on an arbitrary threshold set as the *significance criteria*. This creates a manageable data subset for an unbiased search of all pair-wise interactions.

Conversely, the dataset can also be filtered so that only those multi-marker interactions will be examined that fit within a certain biological context such as a biological pathway, protein family, and group of genes or proteins involved in a certain molecular function. For example, the Biofilter algorithm combines biomedical knowledge from multiple public repositories with statistical methods such as logistic regression or multifactor dimensionality reduction (MDR) method to analyze SNP–SNP combinations [55]. MDR is a novel method that detects and characterizes higher order combinations of genetic and environmental factors that may be predictive of a phenotype or clinical outcome of interest [56]. Another similar method is INTERSNP, which uses logistic regression, log-linear, and contingency table methods to assess SNP–SNP models [57]. However, it is important to keep in mind that any dataset filtering based on particular criteria will introduce its own biological bias into the dataset (*see Notes 4 and 5*).

3.3.3 Post Analysis: Correcting for Multiple Hypothesis Testing

A *p*-value is defined as the probability of observing a test statistic that is equal to or greater than the observed test statistic, if the null hypothesis is true. It is generated for each statistical test that is carried out. A common *p*-value cut off (α) that is used in scientific literature is 0.05. When a *p*-value is equal to or falls below this α cut off, the null hypothesis is rejected. This means that 5 % of the time, when the null hypothesis is rejected, it will actually be true,

representing a false positive. This probability value is with regard to a single hypothesis or statistical test. However, for a GWAS study that tests numerous hypotheses and applies many statistical tests, each of these tests has their own false positive probability. Hence, the combined likelihood of a GWAS result being a false positive is a lot higher than from one test. This brings to light the importance of correcting for multiple hypothesis testing and adjusting the p -value threshold accordingly.

There are a few popular ways to approach correction for multiple testing:

- *The Bonferroni correction.* This is the most stringent of the three; it assumes that each association test in a GWAS is independent of all the others. It corrects an $\alpha = 0.05$ to $\alpha = (0.05/k)$, where k is the number of statistical tests performed. However, this assumption of independence between all the association tests is not necessarily true, due to the presence of LD between markers. For a GWAS with 500,000 markers, the statistical significance threshold for an association would be corrected to $1e-7$.
- *Adjusting the False Discovery Rate (FDR).* Developed by Benjamini and Hochberg this provides an estimate of the proportion of the statistically significant results that are false positives, at an $\alpha = 0.05$ [58]. The approach essentially corrects for this expected number of *false discoveries*, giving the user an idea of the proportion of true associations within their results. The FDR approach is less stringent than the Bonferroni correction as it allows for a proportion of false positive results rather than calculating the probability of observing one or more false positive results over the entire analysis. These procedures have been used extensively in GWAS and also extended in a variety of ways [59].
- *Using permutation testing to adjust the significance threshold.* Although it is computationally intensive, it is the best approach for generating an empirical distribution of test statistics for a given dataset when the null hypothesis is true. The dataset is permuted by rearranging the phenotype labels for all the individuals, but leaving the genotypic information intact. This breaks up any genotype–phenotype relationship within the dataset. However, this technique ensures that the inherent genotype architecture of the dataset is kept intact. This rearrangement of the phenotype labels is done N times (a prespecified number). Each time the labels are rearranged, it represents a new permuted dataset, i.e., a possible sampling of individuals under the null hypothesis. There are a number of software packages that can perform permutation testing for GWAS such as—PLINK [60], PRESTO [61], and PERMORY [62].

3.4 Replication of Results

The biggest concern regarding GWAS results has been the lack of replication of genotype–phenotype associations in an independent study. But an equally formidable challenge is to ensure that a replication study has sufficient statistical power to detect the initial finding. Accordingly, meta-analysis and data imputation procedures can help to tackle this type of challenge.

3.4.1 Statistical Replication

The sole purpose of a replication study is to evaluate an initial positive finding from a GWAS and replicate it to assert its validity and give the association higher credibility. But, despite the general consensus regarding its importance, what actually constitutes a replication is still up for debate. This was the topic of a National Human Genome Research Institute (NHGRI) working group—to outline various criteria involved in defining a replication of a GWAS result [63].

One of the first criteria for establishing a positive replication is that the sample size of the replication study be large enough to detect the effect of the susceptibility allele. This is especially crucial because the effects detected in the original GWAS are often overestimated in the study population it was identified in, as compared to the general population, due to a phenomenon called *winner's curse* [64]. Hence, in reality the sample size required to detect this effect, would have to be much larger than the original study population. This is especially true when trying to distinguish the proposed effect from no effect.

The replication study must be carried out in an independent dataset derived from the same population to avoid any introduction of bias due to differences in ethnicity. Additionally, identical criteria should be used in the replication set to define the phenotype in question. Since the ultimate goal is to replicate a statistical model—a given SNP with a given phenotypic effect—using even slightly different phenotypic definitions can adversely affect the interpretation of the replication results.

Since GWAS markers are chosen based on LD patterns, researchers should aim to replicate a *genomic region*, and not necessarily the original SNP from the initial study. All SNPs in high LD with the original SNP would be considered as candidates for replication. However, a strong rationale should be provided regarding the SNPs being selected for replication, based on linkage disequilibrium, published literature, or putative functional significance. To be considered a successful replication, the magnitude and direction of the genetic effect should be similar across both discovery and replication studies.

3.4.2 Meta-analysis

Meta-analysis is a statistical method for combining several different studies to provide one summary result. It is a widely applied technique in the GWAS field; it allows researchers to increase the power to detect association signals by increasing sample size and

examining a larger number of variants across the genome. Ultimately this helps reduce the chances of false positive findings. An essential component to combining multiple GWAS for a meta-analysis is that all the studies should be *examining the same hypothesis*. A key advantage to the meta-analysis method is the inherent protection of patient and clinical data. It only requires the transfer of statistical results and not the original data that other parties may not have permission for.

In the initial stages of a meta-analysis, researchers should set up strong collaborative agreements ahead of time. Accordingly, a detailed analysis plan should be formulated to avoid any heterogeneity being introduced into the study (*see Note 6*). There are various statistical measures to quantify heterogeneity and to measure how much the various combined studies differ from each other. Some typical measures of heterogeneity are Cochran's Q or the I^2 statistic [65, 66]. The Cochran's Q statistic aims at revealing whether there is statistically significant heterogeneity or not. It is the weighted sum of squared differences between individual study effects and the summary effect across studies. However, the statistic is often underpowered when too few studies are involved in the meta-analysis.

The I^2 statistic, which is favored more in recent studies, measures the proportion of heterogeneity between studies that is true and not due to chance. A major advantage is that the power of the statistic is not dependent upon the number of studies combined in the meta-analysis. I^2 values may fall within low (<25), medium (>25 and <75) and high (>75) heterogeneity values. These ranges are helpful in identifying which studies may need to be removed from the meta-analysis (*see Note 7*).

3.4.3 Data Imputation

A meta-analysis aims to examine the effect of the same allele across all studies. However, this proves difficult when the combined studies have been carried out using different genotyping platforms, each using a different set of markers. To ease this challenge, GWAS can use data or genotype imputation to generate results for a common set of SNP across all the combined studies. The imputation procedure makes use of the known LD and haplotype patterns in reference panels such as HapMap and the 1000 Genomes project, to estimate genotypes for SNPs that were not directly genotyped within a study (*see Note 8*) [67, 68].

Some popular algorithms for genotype imputation are BimBam [69], IMPUTE [70], MaCH [71], and Beagle [72]. The underlying principle for these algorithms is similar to that of haplotype phasing algorithms, which estimate the contiguous set of alleles that lie on a specific chromosome. Genotype imputation algorithms identify the shared underlying haplotypes between the study population and the reference panel. This set of shared haplotypes is then used to calculate haplotype frequencies within the

genotyped SNPs. The phased haplotypes are next compared with a reference set of haplotypes such as those from the HapMap or 1000 Genomes projects. The matched reference haplotypes are also able to provide genotypic information for surrounding markers that were not directly genotyped. Additionally, haplotypes from the study sample may match more than one reference haplotype. In such cases, the surrounding genotypes are given a score or probability of a match, based on the amount of overlap. These scores are also useful for getting an idea about the amount of uncertainty in the genotype imputation process.

4 Future Directions

Irrespective of its victories and failures, GWAS have ushered in an exciting era in the field of genetics and has added new knowledge to our understanding of various diseases and their underlying mechanisms. Although, as the content of genotyping chips, cohort sizes, and biobanks grow even larger, the challenges of data manipulation, quality control, strong study design, and strict phenotypic definitions grow more complex. Hence, moving forward human geneticists will have to develop bioinformatics infrastructure and expertise to overcome such challenges. Most importantly, scientists will have to combine their bioinformatics efforts with genetics, biochemistry and cell biology to confirm the functional consequence and biological relevance of the genotype–phenotype associations that are identified. Ultimately, the translation of GWAS findings into clinical practice will rely upon correct assumptions regarding the genetic architecture of complex traits especially in the context of gene–gene and gene–environment interactions.

5 Notes

1. An r^2 value of 1 is a sign of complete LD and that the alleles at these two associated markers have identical frequencies. To select a tag SNP, an r^2 value of 0.8 or greater is considered to be high and appropriate for using one SNP to tag another in a GWAS [73, 74].
2. LD structures vary between populations, hence, tag SNPs picked for one population may not work for another. Accordingly, populations with high LD will require fewer tag SNPs to capture their variation.
3. Appropriate matching of cases and controls in a GWAS is crucial for preventing any genetic difference between the two groups from being detected due to biased sampling. Researchers must ensure that cases and controls share the same ethnicity, and, if possible, come from the same geographical area.

4. A created dataset based on SNPs that show main effects, enriches for markers that first show a strong association on their own, before searching for pair-wise interactions. This will prevent the detection of certain *purely epistatic* multi-marker interactions—i.e., interactions between markers which by themselves may not have a detectable main effect, and a large part of the heritability is concentrated in their interaction, not individual effects [53].
5. An obvious drawback of filtering datasets based on biological criteria is the reliance upon existing biomedical knowledge, and the quality of this knowledge in public databases. However, SNP combinations identified from the examination of such a data subset are easier to interpret within a biological context.
6. There are several measures that can be taken to avoid introducing heterogeneity in a meta-analysis. The general design of each included study, the quality control procedures, covariate adjustment, and phenotypic definition applied should be the same across all studies. Similarly, the SNP analysis strategies at the level of each individual study should also follow near-identical procedures. Most importantly, the samples added from each study should be independent of each other. Lastly, all results from the individual studies should be reported relative to a common genomic build and reference allele [66, 75].
7. As is true with using any statistical values, these measures should only be used as guides to identify studies introducing an obvious bias. For example, a study may examine a different hypothesis or it may be unduly influential as an outlier. Furthermore, removing a study based solely on a statistical score increases the chances for false discoveries, as it does not make correct use of an agnostic statistical procedure designed to reduce such bias.
8. The reference panel chosen for genotype imputation should be derived from a population with the same ethnicity as the study population to avoid poor quality of the haplotype matches. Additionally, the reference allele for each SNP must be identical between the study population and the reference panel used.

References

1. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* 6: 95–108
2. Hindorff L, MacArthur J, Morales J et al. A catalog of published genome-wide association studies. www.genome.gov/gwastudies/
3. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet* 17:502–510
4. Edwards AO, Ritter R, Abel KJ et al (2005) Complement factor H polymorphism and age-related macular degeneration. *Science* 308: 421–424

5. Haines JL, Hauser MA, Schmidt S et al (2005) Complement factor H variant increases the risk of age-related macular degeneration. *Science* 308:419–421
6. Klein RJ, Zeiss C, Chew EY et al (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–389
7. Maller J, George S, Purcell S et al (2006) Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. *Nat Genet* 38:1055–1059
8. Williams SM, Canter JA, Crawford DC et al (2007) Problems with genome-wide association studies. *Science* 316:1841–1842
9. Jakobsdottir J, Gorin MB, Conley YP et al (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS Genet* 5:e1000337
10. Easton DF, Pooley KA, Dunning AM et al (2009) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* 447:1087–1093
11. Ahmed S, Thomas G, Ghossaini M et al (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* 41:585–590
12. Ragoussis J (2009) Genotyping technologies for genetic research. *Annu Rev Genomics Hum Genet* 10:117–133
13. Denny JC (2012) Mining electronic health records in the genomics era. *PLoS Comput Biol* 8:e1002823
14. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–1320
15. The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A et al (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073
16. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29:308–311
17. Griffith OL, Montgomery SB, Bernier B et al (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* 36:D107–D113
18. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678
19. Scuteri A, Sanna S, Chen W-M et al (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* 3:e115
20. Frayling TM, Timpson NJ, Weedon MN et al (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316:889–894
21. Saxena R, Voight BF, Lyssenko V et al (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316:1331–1336
22. Corder EH, Saunders AM, Strittmatter WJ et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261:921–923
23. Bansal V, Libiger O, Torkamani A et al (2010) Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11:773–785
24. Gibson G (2012) Rare and common variants: twenty arguments. *Nat Rev Genet* 13:135–145
25. Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29:311–322
26. Li M, Li C, Guan W (2008) Evaluation of coverage variation of SNP chips for genome-wide association studies. *Eur J Hum Genet* 16:635–643
27. Distefano JK, Taverna DM (2011) Technological issues and experimental design of gene association studies. *Methods Mol Biol* 700:3–16
28. Lewis CM, Knight J (2012) Introduction to genetic association studies. *Cold Spring Harb Protoc* 3:297–306
29. Lewis CM (2002) Genetic association studies: design, analysis and interpretation. *Brief Bioinform* 3:146–153
30. Teslovich TM, Musunuru K, Smith AV et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466:707–713
31. Bush WS, Moore JH (2012) Genome-wide association studies. *PLoS Comput Biol* 8:e1002822
32. Habek M, Brinar VV, Borovečki F (2010) Genes associated with multiple sclerosis: 15 and counting. *Expert Rev Mol Diagn* 10:857–861
33. Polman CH, Reingold SC, Edan G et al (2005) Diagnostic criteria for multiple sclerosis: 2005 revisions to the “McDonald Criteria”. *Ann Neurol* 58:840–846
34. Kohane IS (2011) Using electronic health records to drive discovery in disease genomics. *Nat Rev Genet* 12:417–428

35. Sager N, Lyman M, Bucknall C et al (1994) Natural language processing and the representation of clinical data. *J Am Med Inform Assoc* 1:142–160
36. Friedman C, Hripcsak G, Shablinsky I (1998) An evaluation of natural language processing methodologies. *Proc AMIA Symp* 855–859
37. Haug PJ, Ranum DL, Frederick PR (1990) Computerized extraction of coded findings from free-text radiologic reports. *Work in progress. Radiology* 174:543–548
38. Kullo IJ, Fan J, Pathak J et al (2010) Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. *J Am Med Inform Assoc* 17:568–574
39. Ding K, de Andrade M, Manolio TA et al (2013) Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. *G3 (Bethesda)* 3:1061–1068
40. Wilke RA, Berg RL, Linneman JG et al (2010) Quantification of the clinical modifiers impacting high-density lipoprotein cholesterol in the community: Personalized Medicine Research Project. *Prev Cardiol* 13:63–68
41. McCarty CA, Wilke RA (2010) Biobanking and pharmacogenomics. *Pharmacogenomics* 11:637–641
42. Ritchie MD, Denny JC, Crawford DC et al (2010) Robust replication of genotype–phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 86:560–572
43. Dubé JB, Hegele RA (2013) *Genetics 100 for cardiologists: basics of genome-wide association studies.* *Can J Cardiol* 29:10–17
44. Price AL, Zaitlen NA, Reich D et al (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11:459–463
45. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multi-locus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587
46. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38:904–909
47. Sale M, Mychaleckyj JC, Chen W (2009) Planning and executing a genome wide association study (GWAS). *Methods Mol Biol* 590:403–418
48. Eichler EE, Flint J, Gibson G et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11:446–450
49. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet* 11:2463–2468
50. Moore JH, Williams SM (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 27:637–646
51. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
52. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26:445–455
53. Moore J, Ritchie M (2004) The challenges of whole-genome approaches to common disease. *J Am Med Assoc* 291:1642–1643
54. Moore JH (2004) Computational analysis of gene–gene interactions using multifactor dimensionality reduction. *Expert Rev Mol Diagn* 4:795–803
55. Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac Symp Biocomput* 368–379
56. Ritchie MD, Hahn LW, Roodi N et al (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* 69:138–147
57. Herold C, Steffens M, Brockschmidt FF et al (2009) INTERSNP: genome-wide interaction analysis guided by a priori information. *Bioinformatics* 25:3275–3281
58. Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. *Stat Med* 9:811–818
59. van den Oord EJ (2008) Controlling false discoveries in genetic studies. *Am J Med Genet Part B Neuropsychiatr Genet* 147B:637–644
60. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81:559–575
61. Browning BL (2008) PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. *BMC Bioinformatics* 9:309
62. Pahl R, Schäfer H (2010) PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. *Bioinformatics* 26:2093–2100

63. Chanock SJ, Manolio T, Boehnke M et al (2007) Replicating genotype–phenotype associations. *Nature* 447:655–660
64. Zollner S, Pritchard JK (2007) Overcoming the winner’s curse: estimating penetrance parameters from case-control data. *Am J Hum Genet* 80:605–615
65. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F et al (2006) Assessing heterogeneity in meta-analysis: Q statistic or I² index? *Psychol Methods* 11:193–206
66. Evangelou E, Ioannidis JP (2013) Meta-analysis methods for genome-wide association studies and beyond. *Nat Rev Genet* 14:379–389
67. Li Y, Willer C, Sanna S et al (2009) Genotype imputation. *Annu Rev Genomics Hum Genet* 10:387–406
68. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11:499–511
69. Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. *PLoS Genet* 4:e1000279
70. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 5:e1000529
71. Biernacka J, Tang R, Li J et al (2009) Assessment of genotype imputation methods. *BMC Proc* 3(Suppl 7):S5
72. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84:210–223
73. Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. *Nat Genet* 38:659–662
74. Pe’er I, de Bakker PI, Maller J et al (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* 38:663–667
75. Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. *Pharmacogenomics* 10:191–201

Studying Cancer Genomics Through Next-Generation DNA Sequencing and Bioinformatics

Maria A. Doyle, Jason Li, Ken Doig, Andrew Fellowes, and Stephen Q. Wong

Abstract

Cancer is a complex disease driven by multiple mutations acquired over the lifetime of the cancer cells. These alterations, termed somatic mutations to distinguish them from inherited germline mutations, can include single-nucleotide substitutions, insertions, deletions, copy number alterations, and structural rearrangements. A patient's cancer can contain a combination of these aberrations, and the ability to generate a comprehensive genetic profile should greatly improve patient diagnosis and treatment. Next-generation sequencing has become the tool of choice to uncover multiple cancer mutations from a single tumor source, and the falling costs of this rapid high-throughput technology are encouraging its transition from basic research into a clinical setting. However, the detection of mutations in sequencing data is still an evolving area and cancer genomic data requires some special considerations. This chapter discusses these aspects and gives an overview of current bioinformatics methods for the detection of somatic mutations in cancer sequencing data.

Key words Bioinformatics, Cancer, Copy number alterations, Next-generation sequencing, Somatic mutations, Structural rearrangements

Abbreviations

CNA Copy number alterations
CNV Copy number variants
SNV Single-nucleotide variants
SR Structural rearrangements

1 Introduction

Cancer, the uncontrolled growth of cells commonly arising from a series of mutations, is caused by a number of factors including inherited genetic defects, lifestyle factors, certain infections, exposure to radiation, and environmental pollutants. The mutations the

cancer cells acquire range from single-nucleotide mutations to large structural rearrangements. In the last few years, one technology has significantly increased our ability to detect mutations: next-generation sequencing (NGS). NGS allows many millions of short DNA sequence reads to be generated from a single sample and has greatly expanded our knowledge of the mutations in cancer. The first cancer genome sequenced was in 2008 [1]; and this has been followed by a plethora of cancer sequencing studies from around the world, especially from multicenter collaborations such as the Cancer Genome Atlas (TCGA) [2] and the International Cancer Genome Consortium (ICGC) [3].

Successes from cancer sequencing studies include the surprising identification of a metabolic enzyme, *IDH1*, as a cancer-driving oncogene in glioblastoma [4], which has defined the important roles of metabolic genes in cancer, and the identification of the *BRAF* V600E mutation in hairy cell leukemia patients [5], opening up the possibility of using targeted therapeutics already approved for treating patients with *BRAF* mutant melanoma. Notably, these studies have revealed that, depending on the type of cancer, there can be a large assortment of different genes mutated that can provide a selective advantage for tumor growth.

The DNA variants that confer the growth advantage are referred to as *driver* mutations as opposed to mutational events which are present in a tumor but do not provide a selective advantage and are referred to as *passenger* mutations. While the number of mutations can vary depending on the tumor type, over a hundred cancer-driving genes have been identified from these sequencing studies with the average tumor containing two to eight driver mutations [6]. For any given tumor, multiple genes may need to be analyzed to determine what mutations are present. Current molecular diagnostic tests screen for mutations in selected loci only and often evaluate only one locus at a time which is costly and inefficient. Sanger sequencing is, at present, the main technique employed for clinical sequencing, but it lacks sensitivity. Thus, the greater sensitivity and high-throughput nature of NGS make it a very attractive technology for clinical diagnostics.

However, while researchers are increasingly using sequencing, its uptake into the clinical environment has been slower, in a large part due to difficulties in bioinformatics analysis. Inherent issues associated with tumor samples including heterogeneity, ploidy, and cellularity (purity) present major challenges for the analysis. Because cancer can contain a broad and diverse variety of somatic alterations, a large range of bioinformatics tools are also required. The aim of this chapter is to give an overview of somatic mutation detection in cancer NGS data and the bioinformatics methods being used to decipher cancer complexity.

2 Materials

2.1 Tumor Samples

Tumor material, when used for NGS, has a number of associated challenges that are often not present in other sample types such as blood. These factors—tumor heterogeneity, purity, ploidy, and sample quality, which often vary vastly between samples—affect the ability of the bioinformatics programs to detect all types of somatic mutations and are critical determinants of whether the sequencing depth will be enough to identify a mutation. These factors are discussed further below.

2.1.1 Heterogeneity

Tumors are not a homogenous mass of cells, all containing the same genetic profile. This is because tumor cells constantly undergo mutational changes and evolve to compete for growth advantages, to metastasize, to avoid immune responses, and to survive therapy. A tumor can, and usually does, contain multiple subpopulations of cells that differ in the mutations they harbor. These subpopulations arise when a cancer cell acquires a new mutation(s) and goes on to produce a population of cells different from the parent clone, a subclonal population. As well as subclonal heterogeneity within a primary tumor, there can also be heterogeneity between a primary tumor and its metastases and between the metastases themselves, and, as tumors are evolving populations, there can also be heterogeneity between samples taken at different time points.

Heterogeneity has important implications for patient care, as a subclone resistant to a drug could expand in number and lead to relapse in the patient. Ding et al. showed using sequencing that a subclone representing just 5 % of the initial tumor cell population in a leukemia patient expanded during treatment and led to relapse of the disease [7]. In another study in chronic lymphocytic leukemia, the presence of a subclone with a cancer-driving mutation was a risk factor for faster disease progression [8]. Thus, being able to obtain a comprehensive genetic profile of a tumor and its subpopulations may be able to help predict if relapse or poor prognosis is likely or even identify additional targetable mutations.

Heterogeneity, however, may also be a cause of failure to detect a mutation, as a single biopsy may not capture all mutations present in a patient's cancer. An important study by Gerlinger et al. analyzed tumor samples taken from spatially separated sites in renal cancer patients and discovered that only ~30 % of mutations were found in all regions sampled [9]. Therefore, the level of heterogeneity in a patient's tumor may lead to some mutations being missed. Programs and methods for assessing the level of heterogeneity in NGS data are now emerging such as tumor heterogeneity analysis (THetA) [10] and mutant-allele tumor heterogeneity (MATH) [11].

2.1.2 Purity

A further challenge, particularly for solid cancers, is that tumors are usually contaminated with normal cells that surround the tumor cells, such as stromal tissue. Normal contamination makes mutation detection more difficult as it reduces the number of tumor cells in the sample. The level of tumor cellularity should be assessed by a pathologist before a sample is sequenced as low cellularity could mean that a mutation may be below the sensitivity of detection and therefore missed. The large omics studies have tended to aim for >60 % tumor purity [12]. For low-purity samples, dissection methods can be used to isolate tumor cells from surrounding normal tissue to increase the purity of the tumor DNA prior to sequencing. Programs such as PurityEst [13] and PurBayes [14] can be used to estimate tumor purity using NGS data.

2.1.3 Ploidy

With the exception of germ cells, normal human cells are diploid containing two copies of each chromosome, one from each parent. Tumor cells, however, can contain abnormal numbers of chromosomes, a state termed aneuploidy. Aneuploidy is very common in cancer and ubiquitous in solid tumors; one quarter of the genome of the average tumor sample contains gains or losses of whole chromosomes or chromosome arms [15], and genome doubling (tetraploidization) is a common occurrence in epithelial tumors such as breast, lung, ovarian, esophageal, and colorectal [16]. Alterations in ploidy will affect the amount of sequence obtained for the regions of interest. For example, in a fully tetraploid tumor sample every chromosome present would be represented by half the reads as in a normal diploid sample if equal amounts of DNA were used.

2.1.4 Formalin-Fixed and Paraffin-Embedded Tumor Biopsies: Sample Quality and Quantity

The quality and quantity of DNA extracted from tumor samples can vary quite considerably. This is because solid tumor biopsies are typically formalin fixed and paraffin embedded. Formalin fixation is essential to preserve tissue and necessary to diagnose if a patient has cancer based on the morphology of a specimen. However, it can lead to degradation (fragmentation) of genomic material, reducing the amount of DNA template that can be sequenced. Low amounts of input DNA can often result in amplification of a reduced number of regions during the PCR step prior to sequencing leading to low overall sequencing coverage. Formalin fixation can also result in non-reproducible base changes in the sample DNA that can mimic true variants [17]. This can be a particular issue for amplicon sequencing (where specific loci of interest are amplified, generally using primers) as duplicate reads which may contain artifacts are not removed (*see Note 1*).

2.2 Matched Normal

The use of a matched normal sample is a common practice in cancer research sequencing studies (Fig. 1). The matched normal sample is taken from the same patient as the tumor and is usually a blood sample, unless it is a leukemia or a lymphoma being studied; if so, another tissue must be utilized. Matched normal samples are

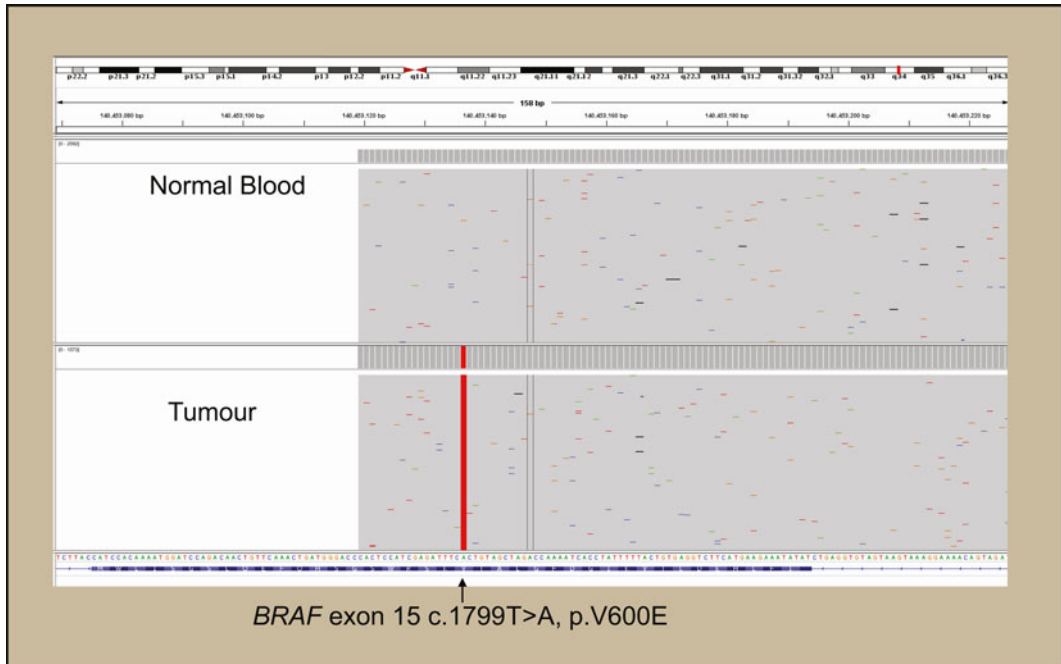


Fig. 1 Somatic variant in NGS data. Integrative Genomics Viewer (IGV) [54] screenshot showing reads from a melanoma tumor and its matched normal sample aligned to the human reference genome. A single base change *BRAF*V600E mutation is highlighted in the tumor reads. It is not present in the normal sample indicating that it is a somatic variant

used to define the germline status of the individual. Every individual has polymorphisms in their germline genome, i.e., sequence changes present at conception that differ between individuals. These are not the mutations acquired by the cancer and are generally not relevant to cancer development and progression. Exome (where only exons are sequenced) and whole-genome sequencing typically identify tens of thousands of germline variants per sample. Sequencing a tumor and matched normal enables the germline variants to be subtracted, narrowing down the number of variants for follow-up analysis. However, this doubles the sequencing costs, as two samples (tumor and normal) must be analyzed for each individual. Most of the current cancer bioinformatics programs have been developed for tumor samples with matched normal, although some will work without a matched normal sample. When no matched normal is available databases of common variants (such as dbSNP, 1000 Genomes, and Exome Variant Server) can be used to help eliminate likely germline variants and identify the tumor-specific variants for downstream analysis and interpretation.

2.3 Whole Genome, Exome, or Targeted Gene Panels

The decision as to the type of sequencing to be performed is largely driven by cost. While whole-genome and exome sequencing allow for a broader analysis of the patient's tumor genome since they

Table 1
Currently available cancer panels

Panel name	Targets
Illumina TruSeq Amplicon Cancer Panel	Mutational hotspots across 48 oncogenes
Illumina TruSight Tumor Panel	175 exonic regions from 26 genes
NimbleGen SeqCap EZ Comprehensive Cancer Panel	578 genes from Cancer Gene Census and NCBI Gene Tests
Agilent HaloPlex Cancer Research Panel	COSMIC mutations within 47 genes
Ion AmpliSeq Cancer Hotspot Panel v2	2,800 COSMIC mutations across 50 oncogenes and tumor-suppressor genes
Ion AmpliSeq Comprehensive Cancer Panel	Exons in >400 oncogenes and tumor-suppressor genes
RainDance ONCOSeq Cancer Panel	142 cancer genes, including >90 % of the genes from the Cancer Gene Census
Ambry Genetics Somatic Mutation Analysis (SOMA) Panel	Clinically actionable mutations in 26 genes

cover more genomic regions, they remain too costly for everyday clinical use, both economically and in terms of effort required for analysis. Targeted gene panels on the other hand allow focused analysis of selected cancer genes. A benefit of this approach is that the reduction in the size of the genomic region being analyzed allows sequencing to be carried out to a much greater depth (several hundred- or thousandfold depth), increasing the ability to identify low-frequency mutations.

Therefore, targeted gene panels are the current method of choice for clinical cancer research as reflected in the growing number of commercially available assays (Table 1) and tests such as the CLIA-certified test from Foundation Medicine, which is a targeted gene panel of >200 genes. Current targeted cancer gene panels can be either hybridization capture based or amplicon based. Hybridization capture panels have a similar bioinformatics workflow compared to exome and whole-genome data with the exception that the higher sequencing depth may mean that the parameters of programs such as variant callers need to be modified to take the higher depth into account. In contrast, amplicon sequencing requires some modification to the bioinformatics analysis (*see* **Notes 1** and **2**). For detection of copy number variations, current methods have largely been developed for whole-genome and exome data, and for structural rearrangement detection, whole-genome sequencing has usually been used. In principle, the general theories these methods are based on should be applicable to all types of sequencing data. A list of the bioinformatics programs mentioned below can be found in Table 2 (*see* **Note 3**).

Table 2
Examples of programs used in cancer genomics (see Note 3)

Application	Programs
SNV detection	MuTect, SomaticSniper, JointSNVMix, Strelka, Varscan2
Indel detection	GATK Somatic Indel Detector, Strelka, Varscan2
Structural rearrangement detection	CREST, BreakDancer, PRISM
Copy number detection	ExomeCNV, ADTE _x , Control-FREEC
Annotation and interpretation	Polyphen, SIFT, COSMIC, My Cancer Genome, Genomics of Drug Sensitivity in Cancer

3 Methods

3.1 Alignment and Processing of Reads

The following pipeline is commonly used for somatic variant detection in whole genome, exome, and many targeted hybridization methodologies. A typical analysis starts with quality checking of the raw sequence reads, i.e., trimming low-quality base calls at the ends of reads including removal of contaminating sequencing adapters and primers. Next the reads are aligned (mapped) to the human reference genome to identify the genomic location of the reads. Reads for the tumor and the normal sample are aligned separately to the reference genome using an aligner such as BWA [18] (see Note 4). After alignment, post-processing steps may be carried out. This may include duplicate read removal (see Note 1), local realignment of reads (see Note 5), and base quality score recalibration, as these steps generally improve the accuracy of the final results.

3.2 Detection of Single-Nucleotide Variants and Indels

Single-base changes are the easiest type of variant to detect since they require only the consistent detection of a mismatched base compared to the aligned reference sequence. However, there are still challenges in distinguishing true somatic variants from background errors, such as sequencing errors which typically occur at about 1 % frequency. Unlike germline mutations, which would be expected to be present at 50 % frequency for heterozygous mutations or 100 % for homozygous mutations, somatic variants can be present at a range of frequencies, the frequency of the variant being dependent on aspects previously discussed—the purity, ploidy, and heterogeneity of the tumor.

As the two most commonly employed variant callers in NGS, GATK's Unified Genotyper [19] and SAMtools [20], were not specifically developed for somatic variants they assume normal diploid samples. With these programs, tumor and normal samples should be analyzed separately followed by a subtraction step to

remove variants common to both samples. This method has been used in early cancer sequencing studies [21]. However, analyzing the tumor and normal separately is not ideal as tumor variants also present in the normal sample but just below the detection threshold will be falsely called somatic.

Several callers have been released in recent years that aim to detect single-nucleotide variants (SNVs) in cancer samples: Varscan 2 [22], MuTect [23], SomaticSniper [24], JointSNVMix [25], and Strelka [26]. They are designed to analyze the tumor and normal sample together and to detect variants present at lower frequencies than would be expected in germline samples. These callers assess the total number of reads covering a site in the tumor and normal samples and the number of those reads that contain the SNV, with different callers applying different thresholds for read numbers.

Variant callers will also typically apply filters, for example, excluding bases with low base-call quality scores, as these reflect low confidence of the sequencer in calling the base, and excluding reads with low mapping quality scores, as these reflect low confidence of the aligner in assigning the genomic location for the read. Other features that may be used in deciding to call a variant include DNA strand information, which refers to whether the variant has been seen in reads mapped to both the forward and reverse strands. Some sequencing artifacts show strand bias, being seen in reads from one strand only [27, 28].

Varscan takes a relatively simple approach to detect somatic SNVs using read depth and base quality thresholds followed by Fisher's exact test to compare the tumor and normal samples and identify the variants that are tumor specific. Other somatic callers use Bayesian approaches which incorporate prior knowledge of what would be expected, for example, the probability of seeing particular genotypes for the tumor and the normal.

Sequencing studies typically use one variant caller, but a recent comparison of four variant callers (Varscan, Strelka, SomaticSniper, and JointSNVMix) found that they differed considerably in their output, with only a small fraction of variants identified in common by all four callers [29]. SNVs identified with high confidence by some callers were identified with low confidence by others. Indeed, large discrepancies in variant calls among programs have been noted by others [30], including the TCGA benchmarking studies [27]. Low concordance between programs is not just an issue for somatic variant calling; it has also been reported for germline variant detection [31]. To address this issue, some groups have developed pipelines that incorporate multiple somatic variant callers in an attempt to integrate the different caller outputs [30] and because mutations detected by more than one caller are more likely to be true variants [27, 30]. The considerable differences in results obtained from different callers emphasize the need for rigorous testing and validation of variant calling pipelines.

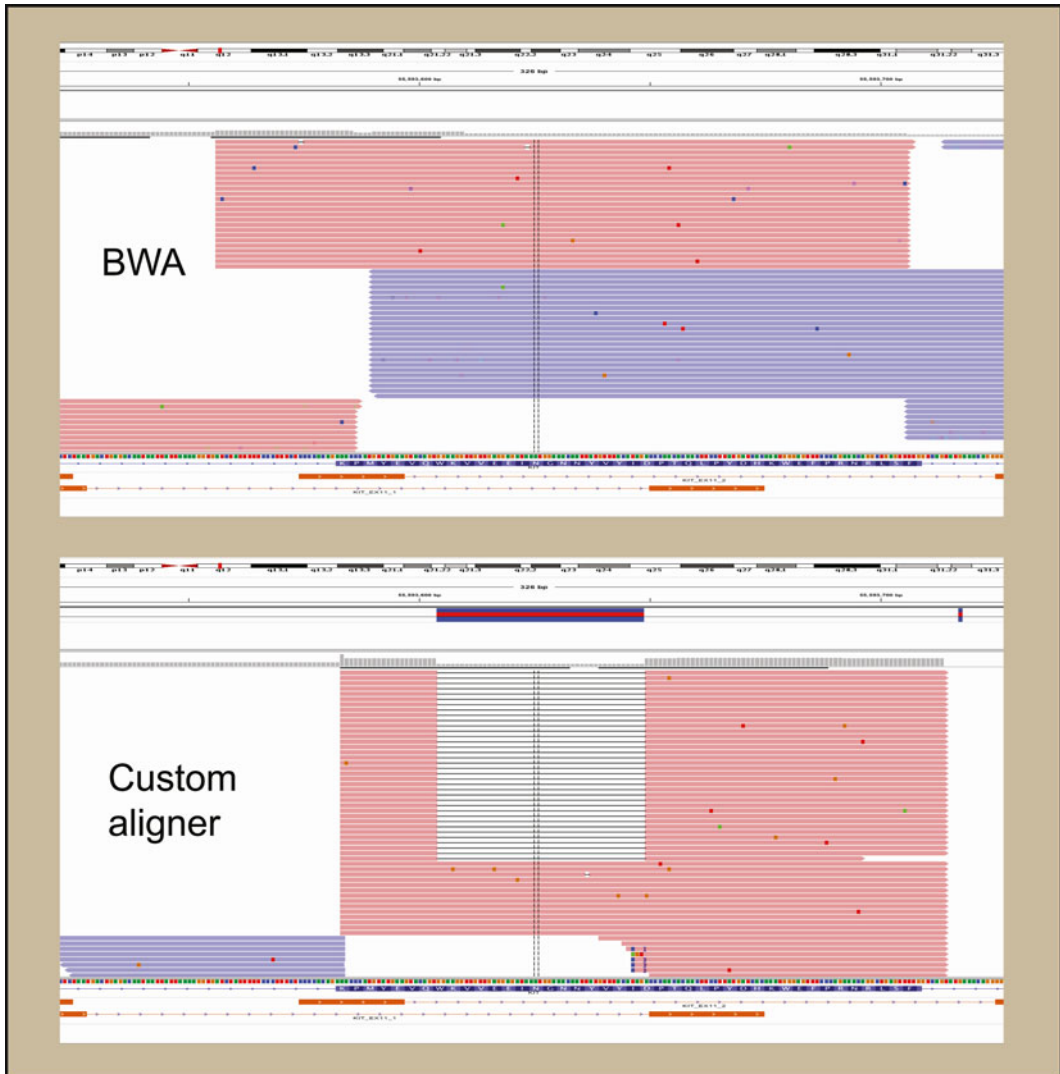


Fig. 2 Indel in NGS data. IGV screenshot showing reads from a tumor sample aligned with BWA (*top panel*) and with a custom in-house aligner (*bottom panel*). BWA failed to map the reads containing the large indel in the *KIT* gene

While several programs have been released in the last few years for the detection of somatic SNVs, the number of programs available for calling indels (small insertions and deletions typically <50 bp) is more limited. GATK's Somatic Indel detector, Varscan, and Strelka are among the few currently available options. The reason for the paucity of programs is that detection of indels even in normal samples is challenging [31, 32]. Reads containing large indels may not align to the reference (Fig. 2), or the exact location of the indel may be difficult to identify if it falls within a region

containing a repeated sequence of bases. For example, if a read contains a deletion of an AT dinucleotide and the reference genome contains a stretch of ATATAT at that location, there are three possible locations where the aligner can place the deletion and an arbitrary decision will be made. Realignment of reads around indels with a program like GATK, prior to indel calling, is useful for improving calls. Some programs such as Strelka perform their own local alignment of reads prior to calling indels.

3.3 Detection of Copy Number Alterations

Herceptin (trastuzumab) is used to treat breast cancers that have overexpression of the ERBB2 (HER2) receptor detected by fluorescent in situ hybridization (FISH). This can be caused by amplification of the *ERBB2* gene through its duplication in the genome of the cancer cell. Such large duplications or deletions are called copy number variants (CNVs) which are found within the class of variants called structural variants—these also include structural rearrangements (*see* Subheading 3.4). CNVs are common in the normal population with about 12 % of the human genome affected by CNVs [33]. Somatic copy number changes are sometimes termed copy number alterations (CNAs) to distinguish them from germline CNVs. CNAs are widespread in tumors with one-third of the genome of the average cancer sample affected by CNAs [15]. In the past, CNAs have typically been identified using comparative genomic hybridization (CGH) arrays (*see* Chapter 8) where the tumor and normal samples are hybridized to probe sequences and differences in fluorescence signal intensity measured.

NGS provides the potential to obtain copy number information alongside information on the presence of other variants like SNVs and rearrangements, which cannot be obtained from copy number arrays alone. While methods for CNA detection for clinical use are still in their infancy, programs are being developed to identify these alterations in NGS data. Copy number changes are identified in NGS data through detection of genomic regions that have increased or decreased numbers of reads in tumor samples relative to normal samples, suggesting amplifications or deletions, respectively (Fig. 3). Programs that have been written to detect CNAs in cancer using sequencing data include Control-FREEC [34], ExomeCNV [35], and ADTEx (originally known as CoNVEX) [36]. These programs detect copy changes using a windowing approach, counting reads within genomic regions of defined size after normalization for variable coverage across the genome.

Variable coverage is the biggest issue for copy number detection, and it arises from differences in GC content, read mapping, and differences in bait capture efficiencies in targeted sequencing, leading to biases in the numbers of reads mapping to different regions [37]. Read mapping issues result from repetitive regions of the genome where reads may not map at all or map ambiguously causing difficulties with detecting copy number changes in these regions.

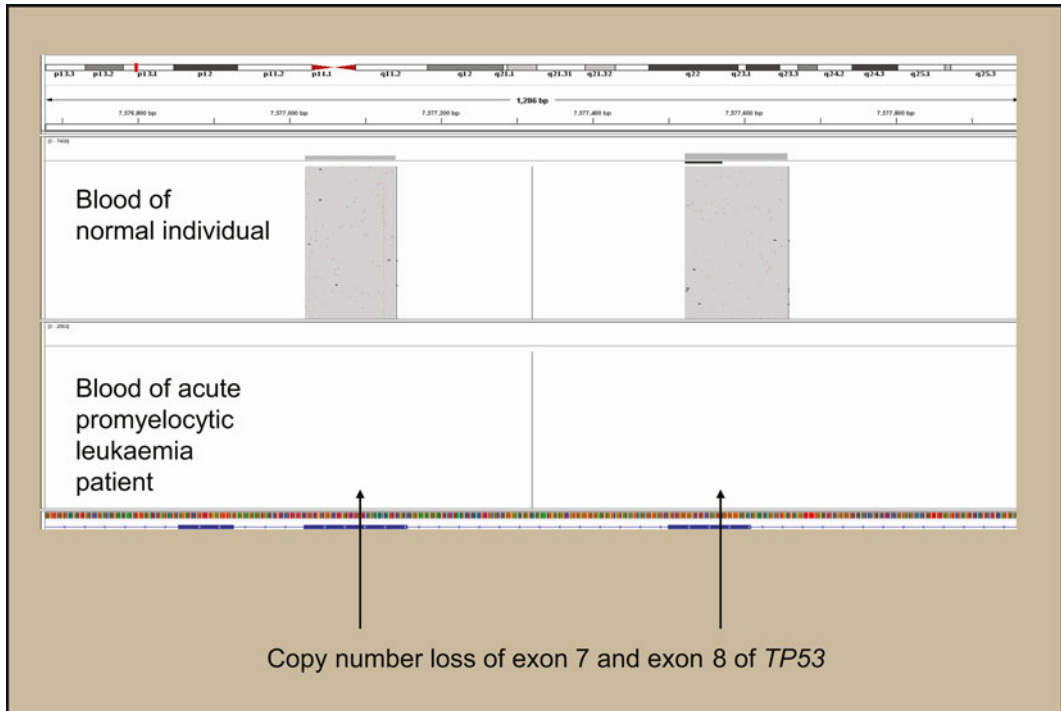


Fig. 3 CNA in NGS data. IGV screenshot showing reads from the *TP53* gene from an acute promyelocytic leukemia patient compared to a normal individual. The arrows indicate where no reads have aligned for the leukemia patient suggesting copy number loss of this region

Longer reads such as from newer sequencing chemistries may help with this issue. How different somatic copy number callers compare in performance has not been comprehensively assessed.

3.4 Detection of Structural Rearrangements

Structural rearrangements (SRs) occur where large pieces of the genome have moved or been rearranged as in the case of translocations and inversions. The Philadelphia chromosome, an abnormality seen in some types of leukemia, results from a translocation between chromosomes 9 and 22 and gives rise to the BCR-ABL fusion gene, which is a target for the drug imatinib (Gleevec). In comparison to other methodologies such as FISH, sequencing can provide base pair resolution of SRs. In 2008 Campbell and colleagues showed how whole-genome sequencing could be used to detect SRs in cancer [38].

SRs can be identified in paired-end sequencing from reads that are not *properly paired*. *Paired-end sequencing* results from the sequencing of opposite ends of DNA fragments. *Properly paired reads* are reads that map to the genome within an expected distance of each other (related to the size of the fragments sequenced) and in the expected orientation (the forward and reverse reads of the pair oriented towards each other on opposite strands).

Most aligned reads would be expected to be of the properly paired type. *Discordant paired reads*, on the other hand, are reads that do not map in the correct orientation or where the reads from the pair map with an unexpected distance, for example, far apart on the same chromosome or to different chromosomes, suggesting a translocation. *Split reads* are also used to identify SRs. *Split reads* are reads that map across the junction (breakpoint) of the structural rearrangement, where one part of the read maps to one region of the genome while the second part of the read belongs to a different region.

Programs like CREST [39] use a split read strategy to detect SRs, while others like BreakDancer use the discordant pair approach [40]. PRISM [41] is a newer program that combines paired-end and split read strategies to increase detection accuracy and can be used to identify SRs in cancer genomes. As with copy number detection, current issues with detection of rearrangements include the repetitive regions of the genome and the difficulties they cause for read mapping [42]. Moreover, no comprehensive assessment of the various programs has been performed.

3.5 False Negatives and False Positives in Somatic Mutation Detection

False negatives in cancer mutation detection most likely result from sample-specific issues, such as the quality, heterogeneity, ploidy, or purity of the tumor sample, resulting in sequencing lacking sufficient depth to detect the mutation. False negatives may also arise in regions of low or zero coverage, such as those that do not amplify or capture well, or regions where it is difficult for reads to map. The thresholds and criteria used by the various variant detection programs are important to keep in mind as variants may be missed if they are just outside cutoffs or fail some filter used by the program.

False positives can be a concern because cancer mutations may be present at low frequencies making it difficult to separate true mutations from background error. This would be a particular challenge for early detection of cancer through screening of circulating tumor DNA [43] where mutations may be present at extremely low frequencies. False positives generally result from sequencing error [44, 45], artifacts introduced during the sample preparation from steps such as PCR amplification [17, 46] and DNA shearing [47], and misalignment of reads.

Ensuring that false negatives and false positives are as low as possible is critical to the adoption of sequencing for clinical use, and it will be essential to test and thoroughly validate any analysis pipeline. The New York State Department of Health have released guidelines for somatic variant detection from sequencing (available at http://www.wadsworth.org/labcert/TestApproval/forms/NextGenSeq_ONCO_Guidelines.pdf). They have recommendations for QC of the analysis pipeline and should be a useful and evolving point of reference.

Adding to these complications, sequencing from formalin-fixed samples often results in a low number of reads due to the limited amount of DNA available for sequencing. Moreover, sequencing

from formalin-fixed samples also results in the appearance of non-reproducible sequencing errors caused by formalin-induced cross-linking of cytosine bases. Beyond applying strict filtering thresholds for variant calling, enzymatic [17] and sequence tagging [48] approaches have been developed over the last few years that have made variant calling in formalin-fixed samples more reliable.

3.6 Interpretation of Cancer Variants

The clinical interpretation of variants identified by NGS is moving towards a tier-based classification scheme that is dependent on the level of clinical and biological relevance. While many variants are clearly actionable and indicate clinical intervention or drug administration, others are difficult to classify due to lack of scientific and clinical understanding, in some cases because it is the first time the mutation has been identified; i.e., it is novel. Bioinformatics programs which predict the effect of a mutation on protein function and structure such as PolyPhen-2 [49] and SIFT [50] aid in the interpretation of these unclassified variants by predicting the likelihood of the mutation causing a functional deleterious effect. However, results from these types of predictive *in silico* programs need to be interpreted with considerable caution as described elsewhere (*see* Chapters 13 and 14).

Two useful resources providing information on cancer variants are My Cancer Genome [51] and the Catalogue of Somatic Mutations in Cancer (COSMIC) [52]. My Cancer Genome is a new online website aimed at providing physicians and patients with information on cancer mutations and available therapies. It can inform whether there is a drug available to target a mutation, and it also lists information on whether there are clinical trials for which the patient may qualify. While the information currently available is limited, it will likely grow to be a valuable resource. COSMIC is a database that aims to catalogue all human somatic mutations reported in the literature and is a useful resource for finding information on whether a variant has been previously linked to cancer. The current COSMIC version v66 (August 2013) contains information on >1.2 million unique variants gathered from >900,000 samples. Other databases such as the Genomics of Drug Sensitivity in Cancer [53], which is linked to COSMIC, are useful in determining the sensitivity and resistance to specific therapies based on mutational data.

Ultimately, clinical translation of sequencing results to guide diagnosis and treatment decisions will require multidisciplinary collaborations with expertise across many disciplines.

4 Notes

1. In whole-genome sequencing, or exome, or targeted sequencing using hybridization capture, the sequence reads will align to the genome in an overlapping pattern. Reads that align to exactly the same site, termed duplicate reads, are considered

PCR amplification artifacts and are typically removed. However, in amplicon sequencing the reads for any one amplicon will all align to exactly the same site (as in Fig. 1), so the duplicate removal step cannot be employed.

2. With amplicon sequencing, if primers are used to amplify the targets of interest, either the primer sequences should be trimmed from the reads or variants should not be called from the primer regions, since that sequence reflects the sequence of the primer and not the sample. In addition, if an amplicon is shorter than the combined length of the read pairs, e.g., if 150 bp paired-end sequencing is performed on a <300 bp amplicon, the paired reads will overlap. The overlapping regions will result in double counting of those bases, so a consensus of the region must be used. We developed a custom in-house aligner for amplicon data as no currently available solution met our needs.
3. Somatic variant detection is a rapidly evolving field, new programs are being released at a frenetic rate, and current programs are regularly being updated. Therefore, to obtain up-to-date guidance on the best programs to use it is advisable to check a recent review in the specific area and popular online forums like <http://www.seqanswers.com> or <http://www.biostars.org>.
4. The Novoalign aligner (<http://www.novocraft.com>) could be used for possibly greater alignment accuracy, but it is substantially slower than aligners like BWA, so it is best used if there are only a few samples or if computational cost and time are not an issue.
5. Indel realignment is a local alignment of reads around indel sites. This step is usually performed because the initial alignment aligns each read separately to the reference genome, potentially resulting in slight differences in the indel position between aligned reads. This step then realigns an indel seen in multiple reads to the same genomic position. For somatic variant detection, it is preferred to realign both tumor and normal samples together so that an indel present in both samples will be realigned to the same location.

References

1. Ley TJ, Mardis ER, Ding L et al (2008) DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. *Nature* 456:66–72
2. Cancer Genome Atlas Research Network, Kandoth C, Schultz N et al (2013) Integrated genomic characterization of endometrial carcinoma. *Nature* 497:67–73
3. International Cancer Genome Consortium, Hudson TJ, Anderson W et al (2010) International network of cancer genome projects. *Nature* 464:993–998
4. Parsons DW, Jones S, Zhang X et al (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science* 321:1807–1812

5. Tiacci E, Trifonov V, Schiavoni G et al (2011) BRAF mutations in hairy-cell leukemia. *N Engl J Med* 364:2305–2315
6. Vogelstein B, Papadopoulos N, Velculescu VE et al (2013) Cancer genome landscapes. *Science* 339:1546–1558
7. Ding L, Ley TJ, Larson DE et al (2012) Clonal evolution in relapsed acute myeloid leukemia revealed by whole-genome sequencing. *Nature* 481:506–510
8. Landau DA, Carter SL, Stojanov P et al (2013) Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152:714–726
9. Gerlinger M, Rowan AJ, Horswell S et al (2012) Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med* 366:883–892
10. Oesper L, Mahmoody A, Raphael BJ (2013) THetA: inferring intra-tumor heterogeneity from high-throughput DNA sequencing data. *Genome Biol* 14:R80
11. Mroz EA, Tward AD, Pickering CR et al (2013) High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer* 119:3034–3042
12. Mardis ER (2012) Genome sequencing and cancer. *Curr Opin Genet Dev* 22:245–250
13. Su X, Zhang L, Zhang J et al (2012) PurityEst: estimating purity of human tumor samples using next-generation sequencing data. *Bioinformatics* 28:2265–2266
14. Larson NB, Fridley BL (2013) PurBayes: estimating tumor cellularity and subclonality in next-generation sequencing data. *Bioinformatics* 29:1888–1889
15. Beroukhi R, Mermel CH, Porter D et al (2010) The landscape of somatic copy-number alteration across human cancers. *Nature* 463:899–905
16. Carter SL, Cibulskis K, Helman E et al (2012) Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 30:413–421
17. Do H, Wong SQ, Li J et al (2013) Reducing sequence artifacts in amplicon-based massively parallel sequencing of formalin-fixed paraffin-embedded DNA by enzymatic depletion of uracil-containing templates. *Clin Chem* 59:1376–1383
18. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25:1754–1760
19. McKenna A, Hanna M, Banks E et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303
20. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079
21. Pleasance ED, Cheetham RK, Stephens PJ et al (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463:191–196
22. Koboldt DC, Zhang Q, Larson DE et al (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22:568–576
23. Cibulskis K, Lawrence MS, Carter SL et al (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31:213–219
24. Larson DE, Harris CC, Chen K et al (2012) SomaticSniper: identification of somatic point mutations in whole genome sequencing data. *Bioinformatics* 28:311–317
25. Roth A, Ding J, Morin R et al (2012) JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumor paired next-generation sequencing data. *Bioinformatics* 28:907–913
26. Saunders CT, Wong WS, Swamy S et al (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28:1811–1817
27. Kim SY, Speed TP (2013) Comparing somatic mutation-callers: beyond Venn diagrams. *BMC Bioinformatics* 14:189
28. Minoche AE, Dohm JC, Himmelbauer H (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome Biol* 12:R112
29. Roberts ND, Kortschak RD, Parker WT et al (2013) A comparative analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics* 29:2223–2230
30. Rashid M, Robles-Espinoza CD, Rust AG et al (2013) Cake: a bioinformatics pipeline for the integrated analysis of somatic variants in cancer genomes. *Bioinformatics* 29:2208–2210
31. O’Rawe J, Jiang T, Sun G et al (2013) Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med* 5:28
32. Lam HY, Clark MJ, Chen R et al (2011) Performance comparison of whole-genome sequencing platforms. *Nat Biotechnol* 30:78–82
33. Redon R, Ishikawa S, Fitch KR et al (2006) Global variation in copy number in the human genome. *Nature* 444:444–454

34. Boeva V, Popova T, Bleakley K et al (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28:423–425
35. Sathirapongsasuti JF, Lee H, Horst BA et al (2011) Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 27:2648–2654
36. Amarasinghe KC, Li J, Halgamuge SK (2013) CoNVEX: copy number variation estimation in exome sequencing data using HMM. *BMC Bioinformatics* 14(Suppl 2):S26
37. Teo SM, Pawitan Y, Ku CS et al (2012) Statistical challenges associated with detecting copy number variations with next-generation sequencing. *Bioinformatics* 28:2711–2718
38. Campbell PJ, Stephens PJ, Pleasance ED et al (2008) Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat Genet* 40:722–729
39. Wang J, Mullighan CG, Easton J et al (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. *Nat Methods* 8:652–654
40. Chen K, Wallis JW, McLellan MD et al (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6:677–681
41. Jiang Y, Wang Y, Brudno M (2012) PRISM: pair-read informed split-read mapping for base-pair level detection of insertion, deletion and structural variants. *Bioinformatics* 28:2576–2583
42. Raphael BJ (2012) Chapter 6: structural variation and medical genomics. *PLoS Comput Biol* 8:e1002821
43. Forsshew T, Murtaza M, Parkinson C et al (2012) Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma DNA. *Sci Transl Med* 4:136ra168
44. Meacham F, Boffelli D, Dhahbi J et al (2011) Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* 12:451
45. Nakamura K, Oshima T, Morimoto T et al (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res* 39:e90
46. Kanagawa T (2003) Bias and artifacts in multi-template polymerase chain reactions (PCR). *J Biosci Bioeng* 96:317–323
47. Costello M, Pugh TJ, Fennell TJ et al (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res* 41:e67
48. Schmitt MW, Kennedy SR, Salk JJ et al (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* 109:14508–14513
49. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* Chapter 7, Unit 7.20
50. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073–1081
51. Swanton C (2012) My Cancer Genome: a unified genomics and clinical trial portal. *Lancet Oncol* 13:668–669
52. Forbes SA, Bhamra G, Bamford S et al (2008) The Catalogue of Somatic Mutations in Cancer (COSMIC). *Curr Protoc Human Genet* Chapter 10, Unit 10.11
53. Yang W, Soares J, Greninger P et al (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* 41:D955–D961
54. Robinson JT, Thorvaldsdottir H, Winckler W et al (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26

Chapter 7

Using Bioinformatics Tools to Study the Role of microRNA in Cancer

Fabio Passetti, Natasha Andressa Nogueira Jorge, and Alan Durham

Abstract

High-throughput sequencing (HTS) has emerged as a promising method to study gene expression in neoplastic and normal tissues. Using HTS, many research groups have described transcript variants as well as discovering new transcribed loci and noncoding RNAs, including microRNAs. In oncology, expression profiling of microRNAs in matched tumor and normal tissues has been used to detect differential expression of microRNAs in cancer. We present one approach for laboratories with few bioinformatics support to assist in the analysis of microRNA HTS data focused in oncology. This approach can also be adapted to study other systems.

Key words Bioinformatics, High-throughput sequencing, microRNA, miRNA databases, miRNA target

Abbreviations

HTS High-throughput sequencing
miRNA microRNA
ncRNA Noncoding RNA

1 Introduction

A fraction of the transcriptome in eukaryotic cells is translated into proteins via coding RNA. The majority of the transcriptome is not translated although biological function for this so-called *dark matter* is starting to be described. Some noncoding RNAs (ncRNAs) are essential to the translation process (rRNA and tRNA), and in the past few years many other types of ncRNAs have been studied. One of the most important classes of small ncRNA is microRNA (miRNA). This comprises ncRNAs 22 nucleotides in length that may prevent mRNA translation and lead to its degradation [1]. Because miRNAs can be preserved in either formalin-fixed,

paraffin-embedded (FFPE) or fresh tissues samples, many studies have focused on the identification of potential molecular markers of diseases using the expression profile for this class of ncRNA. Differentially expressed miRNAs have been found in tumor samples using microarray, qPCR, and, more recently, high-throughput sequencing (HTS) technologies.

Many of these technologies can be accessed through core facilities in universities and research institutes, and there are software packages available that can be used by individual researchers or small groups to do first round analyses. If the datasets are not very large, a standard workstation suffices to perform these analyses.

In this chapter, we present a step-by-step protocol to analyze miRNA HTS data and the usage of bioinformatics databases and tools to start the process of unveiling the biological meaning of differentially expressed miRNAs.

2 Materials

This chapter consists of three parts: HTS data quality control and mapping, identification of differentially expressed genes, and functional annotation of a selected miRNA. To perform each step, we will guide you on how to use the Galaxy Web site [2–4], the Bioconductor’s package EdgeR [5], the statistical environment R, and the public Web sites Rfam [6], miRBase [7], TargetScan [8], and Tarbase [9].

We will adopt the data published by Witten et al. [10] as an example of miRNA HTS data. The authors sequenced the miRNAs present in the normal and cervical tumors of 29 patients from the Gynecologic Oncology Group Tissue Bank (PA, USA); 19 paired sequencing runs are available at the NCBI SRA database [11].

Galaxy is an open Web-based platform for genomic research with easily accessible tools and storage for RNA-Seq experiments [2]. In order to have access to all Galaxy’s site functionality, such as saving and sharing objects and increasing the data quota, one must be a registered user. To register, access the Galaxy Web site (<http://main.g2.bx.psu.edu>), go to “User” and “Register”, enter an e-mail, user name, password and press the submit button.

Both the Bioconductor’s package and the R environment may be used in either Unix-like (Linux, MacOS X) or Windows operating systems. In this chapter, the user will need to use some basic Unix command-line programs such as `mkdir` (create a directory), `cd` (change directory), `cp` (copy), and `mv` (move). For an introduction to the use of Linux, we recommend the following tutorial (<http://www.ncbi.nlm.nih.gov/books/NBK6827/>).

All files and scripts which are expected to be created during this chapter are available for download at http://lbbc.inca.gov.br/ClinicalBioinformatics2ed/supporting_files.zip.

3 Methods

3.1 Quality Checking and Genome Mapping

The first step is quality checking and genome mapping. For this, you will use the Galaxy Web site to upload the data, to eliminate from the dataset information that can be erroneous and mislead you into creating artifacts, and finally to map the ncRNA sequences into the human genome to find from where the miRNA was originally transcribed. This mapping phase is important to identify the miRNAs present in the samples and to create measures of the expression level of each miRNA. These measures are the counts of reads mapped by each genomic region. This step assists making expression data more precise.

3.1.1 Getting the Data

To import the Witten et al. data [10] into Galaxy, go to the left panel, click on “Get Data”, then “EBI SRA”. In the field “Text search”, type SRP002326, and then press “Search”. This will open a screen with details of the study. In the column “Fastq files (galaxy)”, click on “File 1” for all available runs. The Galaxy Web site also allows users to upload their own data (*see Note 1*).

3.1.2 Changing Quality Format

Tools available in the Galaxy Web site are preprogrammed to use the Sanger quality format (*see Note 2*), but some sequencing platforms, such as Illumina prior to version 1.8, produce data in a different format. In these cases, the user must convert the quality data.

The data from the Witten dataset were generated by the Illumina platform. In order to convert the quality format from Illumina to Sanger, you should use the Galaxy Web site. In Galaxy’s home page, the user must click on “NGS: QC and manipulation”, on the left panel, then on “FASTQ Groomer”. Next, choose the uploaded files to groom, keep the “Input FASTQ quality scores type” as “Sanger & Illumina 1.8+”, and execute it. Repeat this step for all uploaded files. This step will generate several groomed FASTQ files, which are the original FASTQ files in Sanger quality format.

3.1.3 Quality Analysis

HTS generates miscalled or unidentified bases, bases with poor quality, adapter contamination, and artifacts. All these features must be removed prior to alignment to save computational time and, more importantly, to avoid incorrect mapping [12]. To this end, you need first to perform the analysis of the sequencing data, which may be performed in the Galaxy Web site. First, click on the “FastQC: Read QC”, on the left panel. Select the groomed data and execute it.

This step will generate a quality report on the selected data. The user can download this report by clicking on the job, then on the floppy disk icon; extract the downloaded file and open the

“FastQC_FASTQ_Groomer_on_data_X_html” file, where X is the number of the groomed data provided. This file contains basic statistics, per base quality score, GC content, N content, duplication level, overrepresented sequences, among others.

3.1.4 Quality Processing

Next, you will remove from your dataset information that is not useful and can be detrimental to subsequent analysis, such as sequence artifacts, adapter and barcode sequences (from the sequencing process), and low-quality sequences as well.

- *Removing sequencing artifacts.* To remove sequencing artifacts, go to the “Remove Sequencing artifacts” link, on the left panel. Select the groomed data and execute it for each groomed run. This tool will remove reads with an excessive number of identical bases; in this case, the read will be kept if it has more than three bases that are different from the rest of the bases of the read.
- *Removing the adapter sequence.* Usually small RNAs are smaller than the sequenced read size; therefore, it is usual to find a subsequence of the 3' adaptor at the end of the read [13]. Because this sequence does not belong to the studied organism, it must be removed prior alignment to avoid incorrect mapping. On the left panel, click on “Clip”, choose the library without sequencing artifacts and keep the minimum length of 15 nucleotides. In the field “Source”, choose “Enter custom sequence” and type the adapter sequence (in the case of our data, the sequence should be CTGTAGGCACCATCAATAGATCGGAAGAGCTCG) (*see Note 3*). Be sure not to keep any bases after the adapter by choosing “yes” on “discard sequences with unknown bases” and “Output only clipped sequences”, i.e., sequences which contained the adapter. This step will also discard reads with unidentified bases.
- *Barcode trimming.* Barcoding a sequence emerged as an option to allow the sequencing of more than one sample in a HTS, since each run is likely to produce more data than necessary for analysis. Witten et al. [10] added one of four barcode sequences (AAA, TTT, CCC, and GGG) after the 5' adapter in all runs. To identify the barcodes in the library, the user must first create a text file with the barcode's identification and its sequence, separated by a tab (*see Note 4*). One such example can be found in Fig. 1, where *B.A* stands for *Barcode AAA*.

B.A	AAA
B.T	TTT
B.C	CCC
B.G	GGG

Fig. 1 Example of file for barcode identification

The text file with the barcodes must be uploaded to Galaxy (*see Note 1*), then, on the left panel, click on “Barcode splitter”. On “Barcodes to use”, choose the uploaded file; on “Library to split” choose the clipped data; and change the number of allowed mismatches to zero. Once the process is completed, go to the right panel and click on the eye icon of the barcode splitter data; this reveals a table showing how many reads there are for each of the barcodes listed; the last column is a link for the sequences. In the case of Witten et al. [10], there is only one barcode per run, so it can be trimmed directly by choosing “Trim sequences”, on the left panel; on “Library to keep” choose the clipped data; the first base to keep is 4 and the last is 36 (*see Note 5*).

- *Filtering by quality.* Low quality bases may arise from poor quality libraries or sequencing faults [14]. Due to their uncertainty they must be removed prior to alignment. On the left panel, choose “Filter by quality”, in “Library to filter”, choose the trimmed data and set the “Quality cut-off value” to 20 and “Percent of bases in sequence that must have quality equal to (higher than a cut-off value)” to 90 (*see Note 6*).

3.1.5 Mapping

As stated earlier, mapping sequences into the genome is important to identify the RNAs expressed in the samples and to obtain expression levels. The alignment of thousands of small sequences onto a reference genome or transcriptome can be a computationally demanding task. This motivated the creation of many aligners to deal specifically with the task of genome mapping of large datasets of small sequences. Linder and Friedel [14] have performed a comparison of the most common alignment programs.

On the left panel, click on “NGS: Mapping” and “Map with BWA for Illumina”. Use a built-in index and select the reference genome as “Human (*Homo sapiens*): hg18 Canonical”, and choose the filtered data as “FASTQ file” (*see Note 7*).

3.2 Finding Differentially Expressed miRNAs

The second part of this chapter is the analysis of differential expression. In the previous part, you have identified a set of miRNAs that are being expressed for each clinical condition. However, these numbers have to undergo a statistical significance analysis before we can conclude which of these miRNAs can be considered to be differentially expressed with some degree of certainty. This analysis will depend, among other factors, on the relative abundance of miRNAs in each clinical condition.

3.2.1 SAM to BAM

Most mapping software produce a result file in SAM format [15] (for detailed explanation about the SAM file format, we recommend reference [16]). This file must be converted to its binary form (BAM) in the next step to reduce space usage. The conversion can be done using the “NGS: SAM Tools” and “SAM-to-BAM” link. Just choose the mapped data and execute it.

3.2.2 Counting and Identifying Differentially Expressed miRNA Genes

In HTS, the measure of a gene's expression level in a sample is determined by counting how many reads were mapped in the same region of the gene. This can be performed by comparing the mapped regions of the BAM file with a gene annotation file. However, raw read counts are subject to sample and experimental variation, and therefore, they must be normalized before being compared with other samples [17]. After normalization, the proper statistical method can be applied to identify the differentially expressed genes.

1. *Getting the miRNA gene annotation file.* The annotation file for miRNAs can be downloaded from the Web site ncrna.org [18]. Access the NcRNA.org Web site (<http://www.ncrna.org>) and click in the “UCSC Genome Browser for Functional RNA” icon in the center of the page. On the new screen, click on “Tables”; then choose the “Mar 2006 NCBI36/hg18” assembly (*see Note 8*); in “group” choose “miRNA-related Tracks”; in “output format” choose “BED—browser extensible data”; check the “Send output to Galaxy” box; and click on the “get output” button. On the next screen, check the “coding exons” box and click on “Send query to Galaxy” button. This will send to Galaxy a BED format file (*see Note 9*) with the coordinates of the mature miRNAs.
2. *Counting the reads that overlap annotated regions.* In the mapping step, you have identified the genomic regions expressed in each sample and in this step you will identify the microRNA genes annotated in those regions. Because a BED annotation file is uploaded, click on “BEDTools”, on the left panel, and “Count intervals in one file overlapping intervals in another file”. In the box “Count how many intervals in this BED or BAM file (source)” choose the converted BAM data, and on “Overlap the intervals in this BED file (target)” box choose the uploaded BED file. In the “Count” field, choose “Only overlaps occurring on the ****same**** strand” and execute it.

This command will generate another BED file with one additional column at the end of the line with the read counts overlapping that annotation. Download all BED files by clicking on the floppy disk icon.

3.2.3 Creating a Count Table

In this step, you will create a text file with the count table to use as input for the statistical test. In this table, columns correspond to samples and rows to microRNAs. As Witten et al. [10] made available 38 files corresponding to normal tissue and tumor sample for 19 patients, the table will have 39 columns (the 38 patients plus a column for miRNA identification). The user must know the order the samples will be added to the table because its group (type of sample and patient's identification) must be provided subsequently to the statistical software.

To create a count table, we must use the Unix (Linux or Mac) command line. The following commands will create the count table by obtaining the read counts and miRNA identification list, and adding a header line (*see* **Notes 10** and **11**). Open a terminal in your Linux-like operating system and go to the directory where the saved BED files are (to do so type “cd” followed by the folder name and press the enter key in your keyboard), then type the commands:

- `awk -F '\t' '{print $4}' GalaxyX_file>miRNA_ids.txt`

Where “GalaxyX_file” is one of the BED files downloaded from Galaxy and “miRNA_ids.txt” is the name of the new file which will contain the miRNAs identification;

- `awk '{a[FNR]=(a[FNR] ? a[FNR] FS : "") $7} END {for (i=1;i<=FNR;i++) print a[i]}' *.bed>counts.txt`

This command will create a text file with all read counts.

- `paste miRNA_ids.txt counts.txt>count_table.txt`
- `sed -i '1imiRNA_id\tN1\tT1\tN2\tT2\tN3\tT3\tN4\tT4\tN5\tT5\tN6\tT6\tN7\tT7\tN8\tT8\tN9\tT9\tN10\tT10\tN11\tT11\tN12\tT12\tN13\tT13\tN14\tT14\tN15\tT15\tN16\tT16\tN17\tT17\tN18\tT18\tN19\tT19' count_table.txt`

This command line creates a header for the count table; the first column is “miRNA_id”; the second is “1n”, indicating that those values belong to the normal sample of the first patient; the third is “1t”, indicating it belongs to the tumor sample of the first patient, and so on. The column labels are separated by “\t” which is a symbol for tab.

Note that the files were downloaded and joined in the order of patient and type of sample. The user must be aware of the joining order to be able to properly identify the samples and perform the differential expression test.

3.2.4 Identifying Differentially Expressed Genes

The normalization and statistical test can be undertaken using the R environment (*see* **Note 12**) and the Bioconductor’s package EdgeR (*see* **Note 13**). The following commands will identify the different conditions for each sample, normalize the count values, and apply the negative binomial test to identify differentially expressed genes. These and other commands can be found in the EdgeR User’s manual at <http://www.bioconductor.org/packages/2.12/bioc/html/edgeR.html> (*see* **Notes 14** and **15**).

So to enter the R environment, simply type R in the command line window. Your terminal will enter an interactive mode with the R system and all commands typed will be forwarded to R. You should then type the following series of R commands:

- `count.table=read.table (“count_table.txt”, header=T)`
- `count.table[is.na(count.table)]<-0`

- `rownames(count.table)=count.table$miRNA_id`
- `count.table$miRNA_id=NULL`
- `library(edgeR)`
- `y=DGEList(counts=count.table)`
- `keep=rowSums(cpm(y)>1)>=1`
- `y=y[keep,]`
- `y$samples$lib.size=colSums(y$counts)`
- `y=calcNormFactors(y)`
- `sample=as.factor(c("N","T","N","T","N","T","N","T","N",
"T","N","T","N","T","N","T","N","T","N","T",
"N","T","N","T","N","T","N","T","N","T","N",
"T","N","T"))`
- `patient=factor(c(1,1,2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9,10,10,
11,11,12,12,13,13,14,14,15,15,16,16,17,17,18,18,19,19))`
- `design=model.matrix(~sample+patient)`
- `rownames(design)=colnames(y)`
- `y=estimateGLMCommonDisp(y,design)`
- `y=estimateGLMTrendedDisp(y,design)`
- `y=estimateGLMTagwiseDisp(y,design)`
- `fit=glmFit(y,design)`
- `lrt=glmLRT(y,fit)`
- `de=topTags(lrt,n=length(lrt))`
- `write.table(de,file="differentially_expressed_genes.txt",sep="\t")`

This script will only keep genes with a count per million >1 in ≥ 2 samples. Please notice that the objects `sample` and `patient` must be created in the same order as the `count.table` object with N and T symbolizing the normal and tumor samples, respectively, and the identification for which patient the sample belongs to; that is why the `patient` object has two 1, 2, 3, until 19, indicating that the first two samples belong to patient 1.

The last command will create a text file with the differentially expressed genes ordered by their p -value. This file can be opened by any text editor or spreadsheet software.

3.3 Finding the Biological Function

The last part of this chapter is directed to finding the biological function of selected miRNAs. In particular, we are going to guide you in trying to discover more about the miRNAs that are differentially expressed in cancer patients. You will perform analyses of miRNAs using Web-based databases.

There is a plethora of Web sites designed to help the analysis of ncRNAs. Web sites can be searched in many different ways. In particular, you can search based on the miRNA's name or on its

sequence. Searching by sequence is the only option when you have a sequence that is a potential miRNA, but has not been characterized before. In the example for this chapter, we have restricted our analysis to miRNAs that have been previously annotated as miRNAs and which have assigned names, so we will not describe search by similarity. However, all sites that will be described in this chapter can be searched by sequence similarity and the reader is encouraged to use them if he or she has an unfiltered dataset with uncharacterized sequences.

To find Web sites that can assist the analysis of ncRNAs in general, and miRNAs in particular, a good starting point is the Web site NRDR (<http://www.ncrnadatabases.org>). This Web site provides a comprehensive list of published databases for ncRNAs [19], including miRNA, piRNA (Piwi-interacting), and lncRNA (long noncoding), and was designed to assist researchers to find ncRNA databases that satisfy their research needs. The user can search Web resources based on miRNA types, the source of information, and other options.

You will be guided to configure a search to find miRNA repositories with data obtained both from experimental data and in silico prediction. If you are interested in miRNAs, you can either configure to NRDR to search databases that are specific for miRNAs, or for databases that contain multiple classes of ncRNAs. To do this, first click on the SEARCH button that appears just below the Web site's title; you will be directed to the search page (Fig. 2).

In this case, first select "miRNA" in the RNA Families list (A), check the boxes "Experimental" (B) and "Prediction" (C) on the information source list, and check the boxes "TAG" (D), "keyword" (E), and "similarity" (F) on the search method list. Once we click on the "Search" button (G), you will have as a result a page with links to the description of 34 different databases. Of these, you will concentrate your attention on three: miRBase, dbDEMOC and TarBase. Clicking on the Web resource's name in the result page will lead you to a page with a detailed description that includes a summary of its content, information source, search methods, and a link to the Web site, among others.

As mentioned previously, if you are interested in miRNAs, you should also search for Web resources that include multiple types of ncRNAs. This is done by changing the RNA family choice to "Multiple classes" in the original search page. This particular search will provide nine different results, from which we will choose RFAM.

Next, we will proceed to describe the analysis of one of our miRNAs using the three Web sites mentioned and also a literature search using PubMed. However, the reader is encouraged to read the description of the Web sites listed in the two searches described above, and also to try different searches.

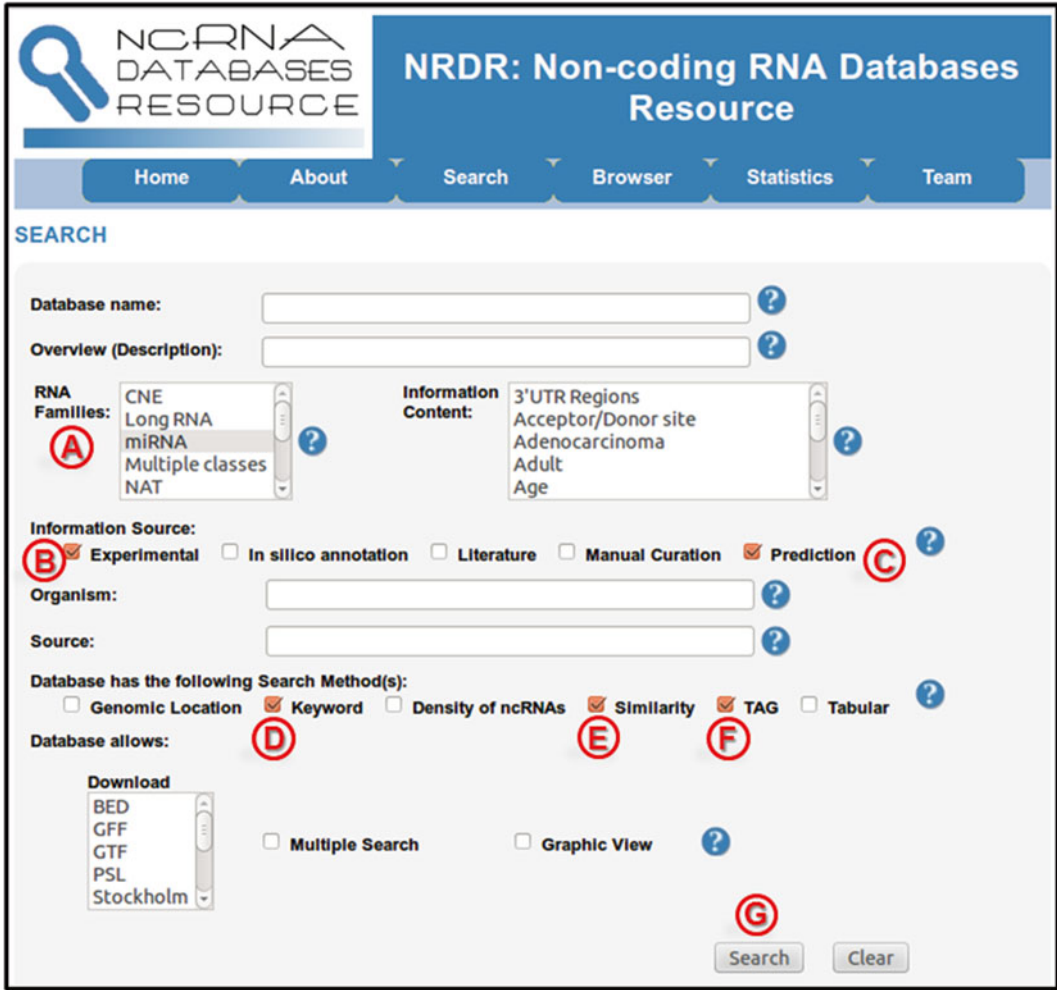


Fig. 2 Using NDNR to search for miRNA databases

The analysis performed in Subheading 3.2 showed that compared to others, hsa-miR-21 is more expressed in tumor than in normal tissue samples (*see Note 16*). You will use this miRNA to model your analysis. Next, we describe the use of each site.

3.3.1 *miRBase*

The Web-based miRBase [7] is one of the central repositories for miRNA data housing >24,000 different entries. The Web site contains information collected from the scientific literature, the Ensemble site and the UCSC Genome browser. The site is divided in three parts: (1) Predicted target genes, (2) miRNA sequences and annotations, and (3) A registry with the option to provide a unique name for novel miRNA genes discovered. Searches can be performed by miRNA name or nucleotide sequence.

To use miRBase, go to the Web site <http://www.mirbase.org> and click on the search tab on top of the page. A new page with all

the search options will appear. The database can be searched by miRNA identifier, genomic location, by clusters, by tissue expression, and by sequence. Type hsa-miR-21 in the first box to search for miR-21 human data. A new page will open presenting information both on the precursor stem-loop sequences and the mature sequences. Additional resources provided include a community annotation box that links to a Wikipedia entry, the secondary structure of the precursor stem-loop, deep sequencing data, genomic coordinate, validated and predicted targets, and literature references.

As mentioned, the same database can be searched by the nucleotide sequence of the miRNA of interest. This is particularly useful if the reader has an unidentified miRNA. To exemplify this search highlight the nucleotide sequence of the mature hsa-miR-21-5p (UAGCUUAUCAGACUGAUGUUGA), go back to the search tab (top of the miRBase home page), and paste the sequence in the last search box (“By sequence”). Next, select “SSEARCH” in the “Search method” selection, which is recommended for short sequences, and click on the “search miRNAs” button on the bottom left part of the box. The result will be a page with a table summarizing all database hits, including the accession and ID, and data on the alignment. Clicking on the first description of the table will lead to the original hsa-miR-21-5p description page. Other entries contain miRNAs from different organisms for this highly conserved family.

3.3.2 *dbDEMC*

This Web resource [20] is particularly relevant for researchers interested in cancer, as it is a repository of differentially expressed miRNAs in human cancers. It includes information on tumor, cell line, tissue and annotation of the differentially expressed miRNAs.

To use dbDEMC go to the Web page <http://159.226.118.44/dbDEMC/index.html> and click on the “Search db” link on the left box and type hsa-miR-21 in the search box. The results page will show the expression profiles in different cancers. In our case the Web site registers 29 different experiments that contain hsa-miR-21, including 11 different cancer types. A table at the bottom summarizes the results of each experiment, including the cancer type, organism, cell line, and expression status (upregulated or downregulated). All lines in the table include the miRNA ID, which is a link to another description page with more details such as expression data from various experiments, summaries of the expression profiles, levels of differential expression across various cancers, predicted targets, and validation. The validation summary shows that miR-21’s expression has been validated by three different publications using quantitative RT-PCR and northern blot. The original articles can be accessed by links to PubMed.

As with miRBase, information in dbDEMC can also be searched by sequence similarity. To do this, click on the “Blast”

link on the “Web Pages” box in the left part of the page. This will open the Blast search page; paste the original hsa-miR-21 sequence (UAGCUUAUCAGACUGAUGUUGA) in the big box and click the “submit” button. The page will show the text output of the blast program. It will list any sequences showing significant alignments, if there are any. In our case the software reports what follows: “hsa-miR-21 j MI0000077 *Homo Sapiens* miR-21”. Now you can copy the miRNA’s name hsa-miR-21 and go to the “Search DB” link as we have performed at the beginning of this section.

The result page for the “Search DB” link also includes target prediction links. We can focus on the results using TargetScan. Following the link will lead us into the TargetScan Web site result page for hsa-miR-21. In the result page, you will find a table with a list of 186 target genes. One important column to look carefully at is Total context+score (*see* **Notes 17** and **18**). By default, the TargetScan result table is sorted by this column. The target gene with the lower value in the Total context+score column is *ZNF367*, a gene which produces a protein that has affinity based on DNA sequence but no association with tumor development. The TargetScan result table can also be sorted by the “Aggregate Pct” column. Higher values of Aggregate Pct permit the selection of stronger sites. In the case of hsa-miR-21, the gene having the higher value for Pct is *TIMP3*, which is a metalloproteinase inhibitor. According to NCBI Gene summary for the *TIMP3* gene (<http://www.ncbi.nlm.nih.gov/gene/7078>), “proteins encoded by this gene family are inhibitors of the matrix metalloproteinases, a group of peptidases involved in degradation of the cellular matrix”. Hence, if hsa-miR-21 targets to *TIMP3* mRNA, it may reduce TIMP3 protein, which will not inhibit degradation of the extracellular matrix, an important step during cancer metastasis.

3.3.3 TarBase

The identification of a miRNA target sequence in an mRNA is still a demanding task and only a few predictions have been experimentally confirmed to date. Hence, TarBase [9] provides a list of experimentally defined targets for a given miRNA. Accessing TarBase Web site (<http://diana.cslab.ece.ntua.gr/tarbase/>), select “Human” in the first dropdown, “miR-21” in the second dropdown, and “Any gene” in the last dropdown. A new page will open and you will find four human genes with experimental support for being targets of miR-21. They are *TPM1* (variant 1 and 5), *SERPINB5*, *PTEN*, and *PDCD4*. If you compare this list with that provided by TargetScan, we will find that TargetScan also detected the *PDCD4* gene as a target for miRNA. If sorted by Pct column, *PDCD4* will be the 10th best ranked, with a 67 % probability to be correct (check the “Aggregate Pct” column). According to NCBI Gene report for *PDCD4* (<http://www.ncbi.nlm.nih.gov/gene/27250>), this gene is a *programmed cell death 4 (neoplastic transformation*

inhibitor). Hence, if hsa-miR-21 targets *PDCD4* mRNA, it may reduce *PDCD4* protein level, leading to less inhibition of neoplastic transformation. These findings are supported by recent articles, such as by Qiu et al. (2013) [21] in which they associate the hepatitis B virus X protein upregulating miR-21 and downregulating *PDCD4* in hepatocellular carcinoma.

3.3.4 RFAM

RFAM is a large repository of ncRNAs from many classes, grouped in >2,200 families characterized by probabilistic models called Covariation Models, since Rfam has not been designed to provide data for any specific species. Each ncRNA family has a *seed* dataset of manually curated sequences used to build the initial structural alignment required to build the probabilistic models and the *complete* dataset that includes sequences that were automatically aligned.

The Web site has links to the ncRNA Wikipedia entry, views on the abundance of the ncRNA across different species, phylogenetic tree of the ncRNA family, visualization of the structural alignment of the ncRNA family, and consensus secondary structure.

To use RFAM [6] go to the Web page <http://rfam.sanger.ac.uk> and use the keyword miR-21 in the search field in the main page of Rfam. A new page will open describing this ncRNA family. The page will have a left box with nine tabs:

1. *Summary*—the tab selected initially, shows the Wikipedia page describing the miR-21 family;
2. *Sequences*—shows a table listing the seed sequences and a second table with all the sequences deposited in the database that are included in this family. The table lists accession number, alignment information, description, and species;
3. *Alignments*—shows the structural alignments of the family;
4. *Secondary structure*—shows the consensus secondary structure for the family, with a color scheme indicating sequence conservation for the various nucleotides in the structure;
5. *Species*—shows the sequences in the database across different species in two views: *sunburst* and *tree*.
6. *Trees*—shows a phylogenetic tree for all sequences from that family.
7. *Structures*—shows 3D structures when available
8. *Database references*—shows links to Literature references in PubMed, and annotation information in the form of three links: GO terms in the Gene Ontology database, SO terms in the Sequence Ontology database, and miRBase [7] entries.
9. *Curation*—detailed information on the RFAM [6] family characterization.

As we have mentioned, RFAM's families are based on a probabilistic technique called *Covariation Models*. These models are a development on the well-known Hidden Markov Models (HMMs) that have the advantage of including secondary structure information on the probabilistic data. As a consequence, a more sensitive similarity search for miRNA precursor sequences becomes possible. The searches in RFAM take into considerations sequence similarity, but allow for more variation than a general BLAST search, provided the sequence maintains the family's base-pairing structure. Unfortunately, RFAM cannot be searched using only the mature miRNA sequence. To perform a similarity search we need to obtain a sequence that should include at least part of the precursor miRNA secondary structure. We obtain the precursor sequence by mapping the miRNA onto the human genome and extracting a genomic region including the mapped sequence and the neighboring nucleotides (100 nucleotides in each direction should be sufficient). To illustrate the RFAM similarity search, we will use the precursor sequence for hsa-miR-21 available at miRBase (UGUCGGGUAGCUUAUCAGACUGAUGUUGACUGUUGAAUCUCAUGGCAACACCAGUCGAUGGGCUGUCUGACA).

To perform the similarity search, you will go back to the search page, select the tab "Sequence", type the precursor sequence in the "Sequence" box and click on the "Submit" button. The result page will show all matches found among RFAM families. In your case, there is just one, not surprisingly miR-21. The summary table displays the family ID, accession number, alignment information and a button to show the alignment. Clicking on the button will show a multiple alignment with the secondary structure in the first line, the family consensus sequence in the second, the matching information on the third and the submitted sequence in the last line. Clicking on the Accession number of the table will lead to the family description page described above.

3.3.5 Literature Search

PubMed is a large repository indexing literature in the fields of Biology and Medicine. Using PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>) to search for articles having the keywords miR-21 and cancer, you will find the work by Defteros et al. [22] reporting that miR-21 increased expression is associated with poorer clinical outcome. Moreover, they identified lower concentrations of PDCD4 protein in invasive cell carcinoma than in the early stages of disease. Therefore, combining the experimental small ncRNA-seq by Witten et al. [10] with other independent findings, we can conclude that miR-21 is overexpressed in cervical cancer and there is evidence this miRNA inhibits the production of the PDCD4 protein. It is also important to note that although we described how to do this type of analysis in cervical cancer data,

other experimental data show that hsa-miR-21 is frequently more expressed in tumor than in normal tissues (reviewed by Buscaglia and Li, 2011) [23].

4 Notes

1. If your data are <2 GB, you may upload them via “Get Data”, then “Upload File” link. Larger datasets should be uploaded via FTP. The instructions are found at <http://wiki.galaxyproject.org/FTPUpload?action=show&redirect=Learn%2FUplod+via+FTP>.
2. To upload data already in the Sanger format, the user must choose “fastqsanger” in the “File Format” field, otherwise Galaxy will not detect the quality format automatically and conversion will be necessary.
3. Adapter sequence must be known for every HTS dataset.
4. There is no name standard for this file format, so you can use any that makes sense to you.
5. Some like to trim the bases with poor quality from the 3’ end. This can be done by changing the last base to keep in “Trim sequences”.
6. Some select other values such as 99 % [24]. Others only trim the 3’ end of the read.
7. According to the Galaxy Web site, the Full version of the Genome contains all primary chromosomes, plasmid, and other sequences, while the Canonical version contains only the primary chromosomes.
8. The UCSC is currently under migration to the latest version of the Human Genome sequence (hg19/NCBI37) [15], so we used the previous version. The hg18 version of the human genome was chosen because there is no annotation of miRNA genes on the hg19 version in the ncRNA.org database.
9. A BED file is a gene annotation file format with its chromosomes, genomic coordinates, strand and other information. A detailed description of the BED format can be found in <http://genome.ucsc.edu/FAQ/FAQformat.html#format1> (Accessed in September 2013).
10. The count table file provided is an example of is accepted by EdgeR. Errors during the execution of Galaxy’s procedures may alter the initial file order. Therefore, users must be aware of the order of their files and to which sample they correspond. The second awk command will look for any file with Galaxy and put its counting in the counts.txt file. To see the order, simply type “ls | grep Galaxy” in the command line of your

Linux or Mac terminal. This command returns a list of files in that directory starting with Galaxy. The corresponding header for this experiment is the order the Galaxy file was created in the sed command. If the user's order is any different, he or she must alter this line after "lmiRNA_id\t" to his or hers corresponding order. We also recommend against creating any other file with "Galaxy" on the name in the working directory, because this may cause the script to add the content of the unrelated file in counts.txt. If the user wants to use the provided shell script, he or she must type the following command line: `./creating_a_count_table.sh`

11. The EdgeR software was designed to identify the differentially expressed genes in an experiment from the raw read count in each sample. Do not insert modified or normalized values because this may lead to wrong assumptions about differentially expressed genes.
12. The R environment can be downloaded from: <http://www.r-project.org/>, on the "CRAN" link on the left panel; instructions for installation are also available on this site on the "Manuals" link.
13. To install the EdgeR bioconductor's package, while running R: `source(http://bioconductor.org/biocLite.R)`
`biocLite("edgeR")`.
14. This R script is specific for the type of experiment described by Witten et al. [10] and for the order in which the files were added to the count table. If the user's count table is in a different order than ours, he or she must alter the sample and patient lines to the corresponding order in the user's count table. If the user has a different type of experiment to analyze, we suggest reading EdgeR's User's Guide found on the EdgeR's Bioconductor's page (<http://bioconductor.org/packages/2.12/bioc/html/rdger.html>). In this file the user will find detailed explanations about EdgeR's methodology, and examples of its applications.
15. If the user decides to use the provided R script, he or she must type the following command on the command line:
`R--o-save--lave<de_script.r`
This command will execute EdgeR and write the differentially_expressed_genes.txt file.
16. Usually a miRNA identified in the human genome has its name prefixed by hsa- for *Homo sapiens* followed by the miR identifier. For example, miR-21 in humans is hsa-miR-21. miR-21 in mouse (*Mus musculus*) would be mmu-miR-21.
17. TargetScan uses a computational approach based on assumptions in evolution to provide more reliable target genes.

Target sequence conservation among distinct species is one feature to enable a prediction to be made in more than one organism. Based on this, there are two subdivisions in the prediction sites in the result table of Target scan: Conserved sites and Poorly conserved sites. TargetScan gives more weight to predictions in Conserved sites than in Poorly conserved sites. This is based on the evolutionary assumption that conserved sequences among species are more likely to have a biological significance than the rest. A miRNA target prediction in a Poorly conserved site is not necessarily a wrong finding.

18. TargetScan works on the knowledge that the base-pairing efficacy of miRNA and its target mRNA depends on how the first eight miRNA nucleotides align to the target sequence. There are different miRNA target site types and TargetScan presents three of them in the prediction results page: 8mer, 7mer-m8, and 7mer-1A. For the purpose of this chapter, consider that the target site recognition efficacy is based on the following hierarchical rule: miRNA aligning to a target sequence based on the 8mer rule has a stronger specificity than 7mer-m8, which in turn has stronger specificity than 7mer-1A. A complete and detailed explanation regarding miRNA target site rules is found in [25].

Acknowledgements

F.P. acknowledges the support of Fundação Carlos Chagas de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) and Fundação do Câncer. F.P. and A.D. acknowledge the support of Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). N.A.N.J. is supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

References

1. Cuperus JT, Fahlgren N, Carrington JC (2011) Evolution and functional diversification of MIRNA genes. *Plant Cell* 23:431–442
2. Giardine B, Riemer C, Hardison RC et al (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res* 15:1451–1455
3. Blankenberg D, Von Kuster G, Coraor N et al (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr Protoc Mol Biol* 19(10):1–21
4. Goecks J, Nekrutenko A, Taylor J et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol* 11:R86
5. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140
6. Burge SW, Daub J, Eberhardt R et al (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res* 41:D226–D232
7. Kozomara A, Griffiths-Jones S (2011) miR-Base: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39:D152–D157

8. Grimson A, Farh KK, Johnston WK et al (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* 27:91–105
9. Papadopoulos GL, Reczko M, Simossis VA et al (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res* 37:D155–D158
10. Witten D, Tibshirani R, Gu SG et al (2010) Ultra-high throughput sequencing-based small RNA discovery and discrete statistical biomarker analysis in a collection of cervical tumours and matched controls. *BMC Biol* 8:58
11. Leinonen R, Sugawara H, Shumway M (2011) The sequence read archive. *Nucleic Acids Res* 39:D19–D21
12. Creighton CJ, Reid JG, Gunaratne PH (2009) Expression profiling of microRNAs by deep sequencing. *Brief Bioinform* 10:490–497
13. Givan SA, Bottoms CA, Spollen WG (2012) Computational analysis of RNA-seq. *Methods Mol Biol* 883:201–219
14. Lindner R, Friedel CC (2012) A comprehensive evaluation of alignment algorithms in the context of RNA-seq. *PLoS One* 7:e52403
15. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079
16. Hoffmann S (2011) Computational analysis of high throughput sequencing data. *Methods Mol Biol* 719:199–217
17. Majer A, Caligiuri KA, Booth SA (2013) A user-friendly computational workflow for the analysis of microRNA deep sequencing data. *Methods Mol Biol* 936:35–45
18. Mituyama T, Yamada K, Hattori E et al (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res* 37:D89–D92
19. Paschoal AR, Maracaja-Coutinho V, Setubal JC et al (2012) Non-coding transcription characterization and annotation: a guide and web resource for non-coding RNA databases. *RNA Biol* 9:274–282
20. Yang Z, Ren F, Liu C et al (2010) dbDEMC: a database of differentially expressed miRNAs in human cancers. *BMC Genomics* 11(Suppl 4):S5
21. Qiu X, Dong S, Qiao F et al (2013) HBx-mediated miR-21 upregulation represses tumor-suppressor function of PDCD4 in hepatocellular carcinoma. *Oncogene* 32:3296–3305
22. Deftereos G, Corrie SR, Feng Q et al (2011) Expression of mir-21 and mir-143 in cervical specimens ranging from histologically normal through to invasive cervical cancer. *PLoS One* 6:e28423
23. Buscaglia LE, Li Y (2011) Apoptosis and the target genes of microRNA-21. *Chin J Cancer* 30:371–380
24. Dreszer TR, Karolchik D, Zweig AS et al (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res* 40:1–6
25. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *Cell* 136:215–233

Chromosome Microarrays in Diagnostic Testing: Interpreting the Genomic Data

Greg B. Peters and Mark D. Pertile

Abstract

DNA-based Chromosome MicroArrays (CMAs) are now well established as diagnostic tools in clinical genetics laboratories. Over the last decade, the primary application of CMAs has been the genome-wide detection of a particular class of mutation known as copy number variants (CNVs). Since 2010, CMA testing has been recommended as a first-tier test for detection of CNVs associated with intellectual disability, autism spectrum disorders, and/or multiple congenital anomalies...in the post-natal setting. CNVs are now regarded as pathogenic in 14–18 % of patients referred for these (and related) disorders.

Through consideration of clinical examples, and several microarray platforms, we attempt to provide an appreciation of microarray diagnostics, from the initial inspection of the microarray data, to the composing of the patient report. In CMA data interpretation, a major challenge comes from the high frequency of clinically irrelevant CNVs observed within “patient” and “normal” populations. As might be predicted, the more common and clinically insignificant CNVs tend to be the smaller ones <100 kb in length, involving few or no known genes. However, this relationship is not at all straightforward: CNV length and gene content are only very imperfect indicators of CNV pathogenicity. Presently, there are no reliable means of separating, a priori, the benign from the pathological CNV classes.

This chapter also considers sources of technical “noise” within CMA data sets. Some level of noise is inevitable in diagnostic genomics, given the very large number of data points generated in any one test. Noise further limits CMA resolution, and some miscalling of CNVs is unavoidable. In this, there is no ideal solution, but various strategies for handling noise are available. Even without solutions, consideration of these diagnostic problems per se is informative, as they afford critical insights into the biological and technical underpinnings of CNV discovery. These are indispensable to any clinician or scientist practising within the field of genome diagnostics.

Key words CGH, CMA, CNV, Deletion, Duplication, LOH, Microarray, SNP, VOUS

Abbreviations

BAF	B allele frequency
CGH	Comparative (competitive) genomic hybridization
CMA	Chromosome microarrays
CNV	Copy number variant

DLR	Derivative log ratio
LCSH	Long continuous stretches of homozygosity
LOH	Loss of heterozygosity
SNP	Single nucleotide polymorphism
UPD	Uniparental disomy
VOUS	Variant of unknown significance

1 Introduction

CMA testing [1–3] is concerned primarily with mutations that change gene copy number but not the primary DNA sequence. These are referred to collectively as copy number variants (CNVs).

In genetics and population biology, the concept of CNV has had a long history, perhaps underappreciated since the advent of molecular biology. Germane to this is the concept of *chromosome balance*, established from the 1920s [4] and based on studies of various plant and insect species. This evolved to become the *gene balance hypothesis* [5]. Clinical cytogenetics studies of the 1970s and 1980s [6, 7] extended the balance concept to the human karyotype. Critical to it, and the related *gene dosage hypothesis* [8, 9], is the notion that for any species, deviations from the “normal” (= euploid) chromosome or gene copy number can often lead to some form of developmental abnormality.

When viewed under the light microscope, human chromosome imbalances are visible only at the resolution of the G-banded mitotic karyotype. However, following such observation of many trisomies and monosomies, often detected as unbalanced autosomal translocations [6, 10], cytogeneticists confirmed that heterozygous gain or loss of almost any discernible genomic region would result in a clinically significant phenotype. These phenotypes ranged in severity through mild intellectual handicap, to multiple congenital abnormalities, to first trimester death in utero [10]. It also became apparent that deletions generally have more severe effects on phenotype, when compared with duplications of comparable length. But, for both deletions and duplications, the more severe phenotypes were associated with the larger imbalances [10]. Accordingly, whole autosome duplication (trisomy) results in early miscarriage although trisomies 13, 18, and 21 can be exceptions [7]. Significantly, common to all live-born cases of chromosome segment imbalance is some degree of intellectual impairment. Thus, it appears likely that many loci scattered through the genome contribute, in some degree, to this critical and complex trait.

The scale of resolution available via the G-banded karyotype was no better than 5–10 Mb, as expressed in DNA bases. Following the Human Genome Project, much higher resolution became possible, at the molecular level. When the *chromosome balance/gene dosage* concept is extended to those submicroscopic monosomies

and trisomies that can now be discerned as the larger CNVs (of size ~100 kb to ~5 Mb), intellectual impairment is still the most consistent clinical association [11]. Indeed, these associated phenotypes may be categorized under the general term “developmental brain dysfunction” [12, 13]. This category includes developmental delay, autism spectrum disorders, neurodevelopmental defects, personality disorders, and others. As mentioned earlier, the severity of the phenotype is (very roughly) associated with the size of the CNV detected, and so larger CNV imbalances can additionally produce severe dysmorphic phenotypes and multiple defects, some lethal.

For any genes implicated under the *gene balance hypothesis*, it is not sufficient that the DNA sequence be normal. It is also critical that the normal sequence be present in the standard number, which, for the autosomes of all diploid species, is two copies (= disomy). Such genes are said to be *dosage-sensitive*, and may themselves be of two classes: haplo-insufficient (HI) or triplo-sensitive (TS) [11]. These two classes need not be mutually exclusive. As the names imply, these two states involve (respectively) gene loss/deletion, i.e., partial haploidy, partial monosomy, or hemizygoty, and gene gain/duplication (partial trisomy), when having a phenotypic effect.

Well over 30,000 human CNVs are now recorded (for example the ISCA database <https://www.iscaconsortium.org>), dispersed through the genome. Their effects range from pathological to nil, and their ubiquity, even among “normal” individuals, is indeed staggering [14, 15]. Using a high resolution array, with molecular confirmation, ~1,100 validated CNVs were found per individual, among 41 normal subjects. Of mean length 2.9 kb, these CNVs ranged in size from ~500 bases to 1.3 Mb. 8,600 different CNV loci were confirmed overall, and 40 % included at least one known gene [14].

On this basis, one might predict that a sample of 41 *patients* referred for CMA testing would carry an equally large number of presumably benign CNVs, and perhaps 14–18 % of these 41 people would carry those very few, but critical CNVs responsible for their abnormal phenotypes. Overall, the clinician and/or scientist will be confronted with a vast number of clinically irrelevant CNVs. Some form of compromise using data filters is necessary, and a current view is that only CNVs ≥ 200 kb are practicable targets for routine CMA testing [16]. This indeed is consistent with the genome balance concept, whereby most smaller CNVs might be expected to be benign.

In the light of these arguments, some labs may report only those CNVs of >100–200 kb. This renders unreportable the shorter 95 % of the research-observed size range mentioned above [14], but still leaves multiple CNVs per normal individual, as potential false positives on routine testing. While not straightforward, there are other ways of dealing with these.

A second filter is based on the assumption that the most common or polymorphic CNVs (of population frequency >1 %) are likely to be benign, regardless of their length or gene content [17, 18]. While this is an arbitrary distinction, it has broad acceptance in the field.

A third type of filter involves selective microarray design. One might, for example, “target” those loci known to be relevant to dosage-dependent disease, using a higher array probe density than for the remainder of the genome [19, 20]. On this basis, microarrays are often designed with a genome-wide “backbone” probe set, at lower density, within which are regions of higher density, at the targeted loci. Despite their widespread acceptance, targeted designs offer no panacea (as yet), because many pathogenic and dosage-sensitive loci still await discovery. And perhaps their discovery will be postponed, if all laboratories choose to use only targeted arrays!

As argued here, it is in the nature of CNV testing that compromises (by filtering) need to be applied—even in the knowledge that they will inevitably result in occasional diagnostic errors. As more of the genome’s pathological and dosage-sensitive loci are discovered, errors will be reduced, but the ultimate solution remains a long way off. Until then how are dosage-sensitive genes recognized, when interpreting CNV data in routine CMA testing?

An important preliminary question concerns the magnitude of the problem: what proportion of all genes might be dosage-sensitive? Given that very many of the common and benign CNVs do contain genes, one might predict that many genes are *not* HI (or TS). Using complex bioinformatics study [21], it has been estimated that the chance of any one human gene being haplo-insufficient is ~20 %. Consistent with this, the authors noted that their defined haplo-insufficient genes were more likely to be associated with dominant inheritance, than chance would predict. The latter is important, since the vast majority of detected CNVs are heterozygous. A more recent study makes a very similar estimate—21 % genome-wide haplo-insufficient frequency [22]. In the diagnostic laboratory, these are important findings. If ~80 % of genes are not haplo-insufficient, then any CNV deletion involving, say, one gene, will most likely be benign. Such CNVs might therefore be very common in normal subjects, and as seen earlier, there is much evidence to support this.

From the study quoted above [14], comes another result, equally important in CMA diagnostics: Genomic mapping indicates a paucity of CNVs (especially deletions) overlapping recognized structural genes, when compared with the density expected under models of random distribution. This relative absence, it is argued, reflects a history of (purifying) selection, against those putative and *no-longer-extant* CNVs that included some of these genes. The corollary is that the vast numbers of CNVs remaining are those to which such selection did not apply, i.e., these are, in

effect, selectively neutral. On this argument, one is most likely to find those CNVs sensitive to selection by detecting them before the selection process is complete. Accordingly, one should expect the selection-sensitive CNVs to be most frequent among de novo mutations. And that, indeed, informs much of our work in CMA diagnostics, for it is in the *patient* population that these CNVs under current selection are most likely to be encountered.

The arguments above are consistent with the *structural-variant disease hypothesis* [23], i.e., many larger CNVs (>100 kb) may arise recurrently during gametogenesis, but will come under strong selective constraint in any progeny bearing them. Therefore, those CNVs detected as de novo are, by that fact, those more likely to cause disease. (This hypothesis does not imply that all familial CNVs are innocuous, as we shall see).

Consistent also with this hypothesis is the de novo paradigm for mental retardation [24], which relies on parallel evidence in respect of point mutations. On considering all the above, and despite some contrary views in the literature [25], we are drawn to conclude that the majority of the *extant* CNVs are likely to be selectively neutral, serving no selectable function and having no effect on fitness, either deleterious or otherwise.

As argued so far, the routine practice of CMA diagnostics has come to rely heavily on considering: *What is the likely clinical significance (if any) of the diverse copy number imbalances detected?* At present, no bioinformatics aids can provide reliable a priori evidence whether novel CNVs can be described as disease-causing, or dismissed as benign, within the patient report. But these studies do warn us that many CNVs are likely to be clinically insignificant, even when including known genes.

In this chapter, we present examples of pathological and other CNV findings from our own laboratories, illustrating some of the common problems encountered in preparing the final patient report. We emphasize methods to sort likely disease-causing CNVs from the clinically insignificant or laboratory artifacts. We do not aim to deal with CNV testing in the prenatal or malignancy settings, nor do we deal with any microarrays designed to measure levels of gene expression.

2 Materials and Methods

This chapter emphasizes the *in silico* aspects of CMA copy number analysis, rather than the wet lab processes that generate the data. However, since the quality of any array data set is largely determined by either the patient sample and/or the laboratory process, any clinical interpretation of the CMA output must first ask: Is the quality of the output data sufficient to meet the level of resolution expected by the clinician referral? In this context, it is

critical to appreciate that no CMA data set will ever be perfect [20, 26, 27]. Bearing this in mind, we must quantify the imperfection—and hence consideration of quality control (QC) parameters is always obligatory, as the first step in the analysis. Further description of CMA hardware, including variations on microarray design, is widely accessible from the literature [19, 20], and the Internet, e.g., <https://earray.chem.agilent.com>.

The *Materials* here comprise the CMA data output file, which is presented to the pathologist or scientist in the form of a graphical user interface (GUI), examples of which are provided in the *Methods* section as cases 1–5. As included within the array manufacturers' analytical software suites, these interfaces also offer alternative displays for the same CMA data, in the form of a long list, where probe name, map location, \log_2 ratio, quality control data, and various other parameters may be viewed as text.

In the field of DNA copy number analysis, microarray platforms fall into two major categories. Cases 1 and 2 of this chapter provide examples of the first type, often described as *CGH arrays* (CGH—comparative genomic hybridization). Cases 3–5 present more complex data sets from *SNP arrays* (SNP—single nucleotide polymorphism). The latter can generate genome-wide genotype data [28] in addition to the same copy number data produced by CGH arrays. We will consider the diagnostic applications of two commercial packages, with emphasis on their roles in: (1) Quality control parameters critical to CMA data acceptance; (2) Presentation of array data to the user; (3) Their flagging of statistically significant CNVs.

2.1 CGH Arrays

We will here use *Agilent* arrays as examples. These currently utilize the manufacturer's *Cytogenomics* software suite (<http://www.genomics.agilent.com>). We also present data displays from a hardware-independent software suite: *CGH Fusion* software (*InfoQuant*, UK: <https://www.infoquant.com/index/cghfusion>). Though we are using these two software packages here for CGH data only, they also have SNP array capabilities.

Despite the high abundance of CNVs overall, it is not difficult, in theory, to distinguish these from the copy-number-normal majority of the genome. This is because, over any one genome, the vast majority of array probes tested should be in the same *normal* state, i.e., diploidy (= 2 copies), and the theoretically possible deviations from this (generally) occur as discrete (integer) entities, i.e., 0, 1, 3, etc., copies. Furthermore, the latter states are few in number since germ-line copy number increase is limited, in practical terms, to about six copies only. However, this is not true for the CNVs of cancer cells.

For these reasons, almost all plotted array data will cluster along the *normal* line, where the \log_2 ratio = $2/2$, or 0. Hence, the chance that, for example, ten consecutive probes will cluster around some distinctively different, and theoretically predetermined value (say,

$\log_2 = 3/2$, or $+0.58$, for 3 copies) is extremely small. Obviously, the power to discriminate, say, 0 from $+0.58$, will diminish if: (1) Dealing with smaller CNVs, detected only by stretches of say, <5 consecutive probes, or (2) Technical issues including noise in the data arising from suboptimal DNA sample (Cases 2 and 4); problems with the hybridization experiment itself; or properties inherent to some probe sequences chosen for the array design [16, 27].

In order to detect significant deviations from the normal diploid state, the minimum number of adjacent probes required can range from two (for the earlier-used, large BAC probes, of >50 kb in length) up to perhaps five or more (for oligomeric CGH probes of say 40–60 bases). These minima are usually suggested in the manufacturer's specifications. Diagnostic laboratories may apply their own more stringent rules in some cases, but less stringent rules may be problematic. The user is able to set the software to any preferred in-house minimum probe number, n (for Cases 1 and 2, $n=5$ was used). Other parameters may also be varied at the user's discretion. These include the desired level of statistical significance, minimal acceptable probe signal intensity, maximum background signal, minimum signal-to-noise ratio, and so on. It is critical that each laboratory decides on these values in a systematic manner, based upon local needs, and determined by the laboratory's own series of validation cases. Once these parameters are set, they should not be altered, without revalidation. These last points also apply to SNP array use.

2.2 QC with CGH Arrays

As to the measurement of data quality per se, various CMA quality control parameters have been devised, and included in software packages both commercial and in the public domain. These QC parameters calculate microarray data noise, appreciation of which is critical to matters considered here. For a recent general account of some quality and performance parameters *see* refs. 20, 26.

For those CGH arrays manufactured by *Agilent Technologies*, the major QC parameter is known as the derivative log ratio, or DLR, which may be described notionally, as follows: Consider the difference in \log_2 ratio for any two adjacent probes, e.g., the data in Fig. 1a. The mean log ratio difference for all such adjacent data pairs can be calculated as the average for the entire autosomal data set. When this mean value rises, it should reflect the increase in map-position-*independent* sources of noise in the CMA data, and is thus used as a quality control measure. An example of such noise is that due to a low signal-to-background ratio, across the test chip.

DLR values of <0.20 usually indicate acceptable data, but <0.15 is desirable. With suboptimal DNA sources, e.g., from formalin-fixed or postmortem specimens, it may be necessary to accept higher DLRs in certain circumstances, e.g., Case 2. But for optimal specimens (freshly collected whole blood in EDTA), the laboratory may reject the data (and/or the specimen) if the $\text{DLR} > 0.20$. One should always strive for the lowest achievable DLR levels. However, because some noise is inevitable, there is a

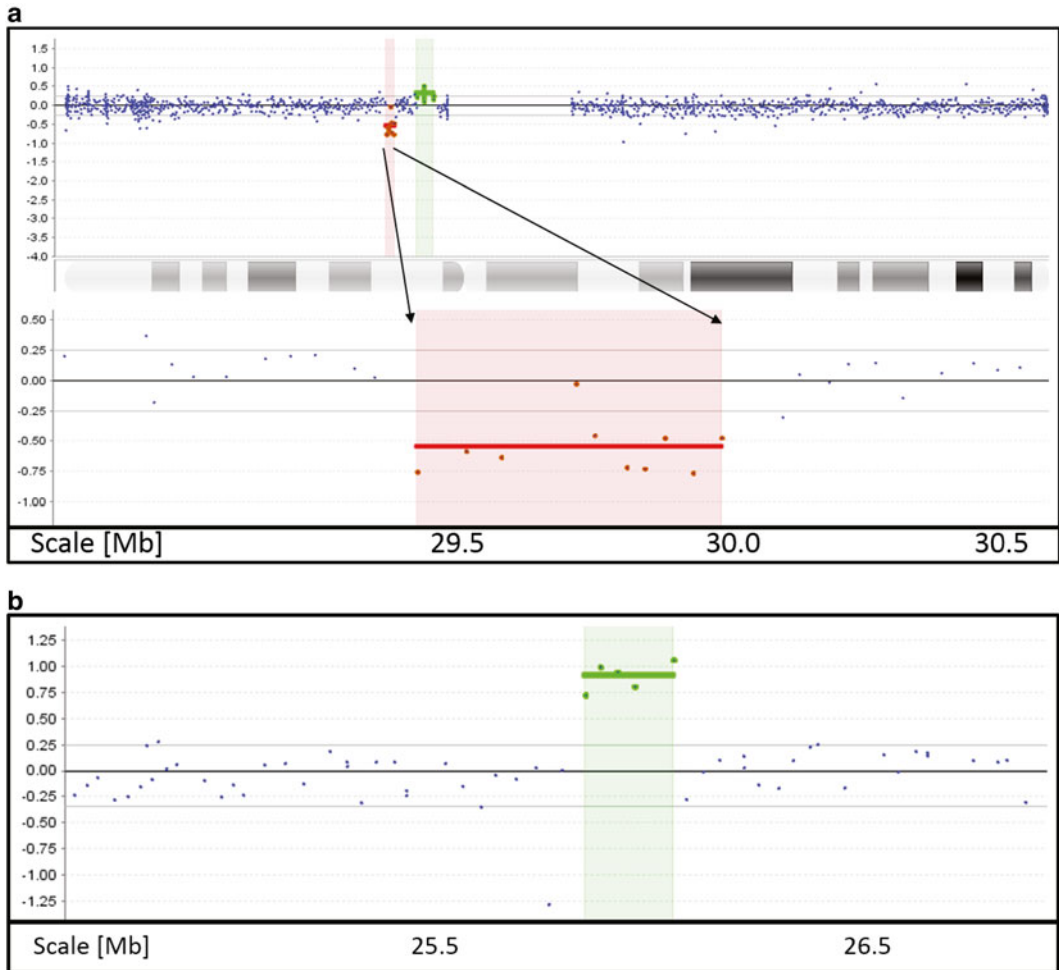


Fig. 1 (a) Case 1: CMA data for chromosome 16. *Upper panel:* Array data for the whole of chromosome 16. *Y-axis* presents the \log_2 ratio (patient signal/control signal). *X-axis:* all chromosome 16 probes on the “Agilent SurePrint G3 Human CGH Microarray 8x60K,” as analyzed initially by Agilent CytoGenomics software. However, the data are displayed here via a different software suite: “CGH Fusion” software (InfoQuant, UK). Each data point represents the \log_2 ratio for one chromosome 16 probe (other chromosomes not shown). These data points are plotted in genomic order, from the p terminus of chromosome 16 (*extreme left*) to the q terminus (*extreme right*). Scale: total length of chromosome 16 is 90.3 Mb. Below this is a horizontal schematic of chromosome 16, showing the G-band ideogram. Both a deletion (*red*) and a duplication (*green*) have been auto-detected by the algorithm (as delimited by the *red bar*). Note the array includes no probes in the centromeric region (= central gap of 10–15 Mb), where highly repeated sequences and extreme segmental polymorphism would render CMA data unreadable. This is true for all chromosomes, but the gap is larger here than for most others. *Lower panel:* expansion of the deleted region (*red*) shows the 0.71 Mb deletion of 16p11.2, and its flanks. (b) Detail of CMA data for Case 1, showing the 0.27 Mb “?triplication” CNV (*green*) in chromosome 10, flagged here by only five consecutive probes. Note that for both a heterozygous triplication, or a homozygous duplication, the \log_2 ratio expected ($\log_2(4/2) = +1.0$) is observed here. Other features are as for (a). The algorithm-detected CNV is here shown as a *green bar*. Note that each end of this bar corresponds with one data point. For any CMA data, this pair of data points defines only the *minimum* length of the CNV, because

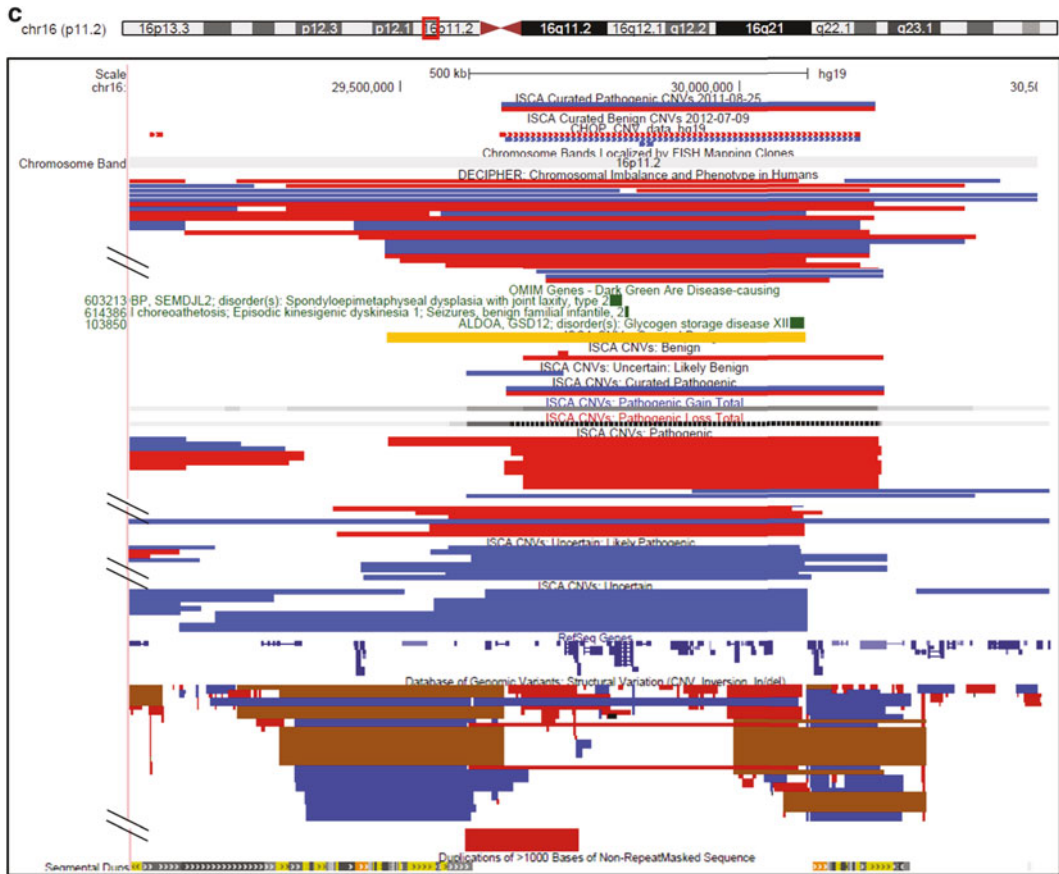


Fig. 1 (continued) we cannot know exactly how far it extends beyond them, in either direction. What we do know is that it does not extend as far beyond them as the next (or flanking) pair of data points, and this flanking pair thus defines the *maximum* extent of the CNV. *Caveat*: We are relying on a single data point to define a CNV boundary. But due to the inherent noise in the data, any single array probe alone does not comprise a reliable estimator of copy number, and hence such measurements of CNV length are approximations. (c) Image from <http://genome.ucsc.edu>, showing the genes, CNVs, etc. for the genomic segment 29.0–30.5 Mb (hg19), within chromosome band 16p11.2 (also indicated as the *small red box* within the ideogram at the *top* of the picture). The *horizontal orange bar* in the *upper center* has been added, to indicate the extent of the deletion detected for Case 1. *NB*: in order to fit this image on the page, it has been greatly truncated in the *vertical axis* (at four places, as shown), removing the great majority of both deletions (*red*) and duplications (*blue*) known for this 16p11.2 recurrent CNV. (d) Image from <http://genome.ucsc.edu>, showing the genes, CNVs etc. for the genomic segment from 25.3 to 26.7 Mb (hg19), within chromosome band 10p12.1. The *horizontal orange bar* in the *lower center* indicates the 10p triplication detected for Case 1. Other features are as for (c), but this screenshot has not been truncated, reflecting the relative paucity of CNV reports for this region. Note that no CNV here (*red or blue*) matches the 10p CNV of Case 1 (*orange bar*). Also, no OMIM disease-causing gene lies within this CNV. The only gene that does (*in part*) is *KIAA1136* also called *GPR158* (*the latter name is truncated, at left edge*), which as yet has no disease association. The horizontal blue bars under “ISCA CNVs Pathogenic” etc., extend beyond either side of this figure, indicating longer length and additional genes. That these CNVs are pathogenic does not, therefore, imply any necessary pathogenicity for the much smaller CNV detected

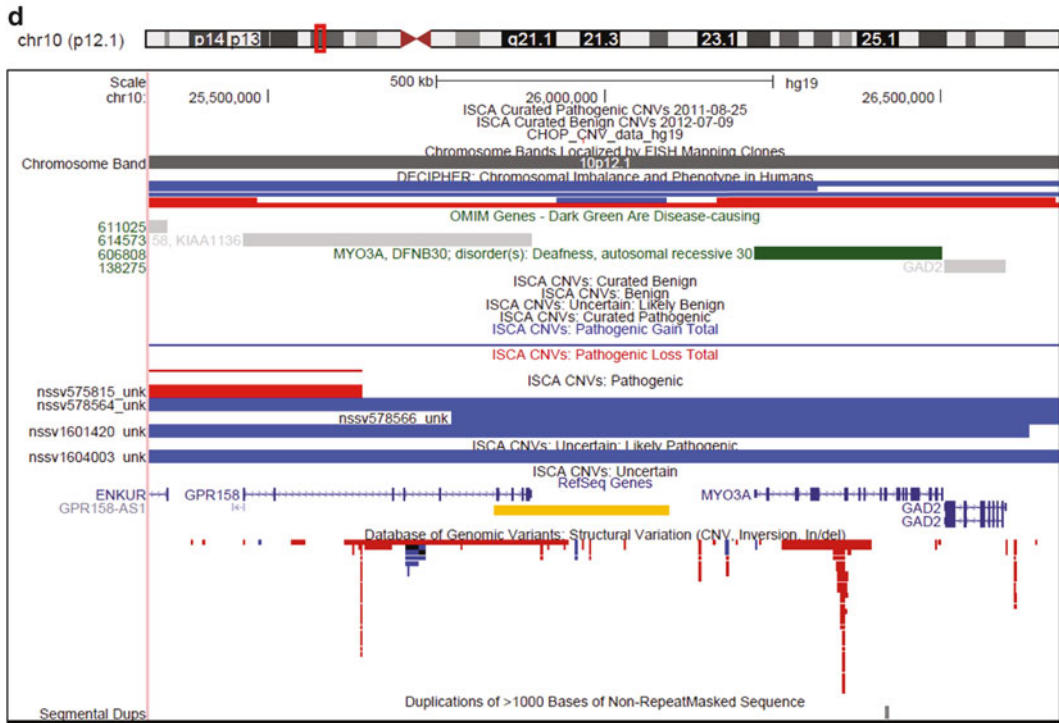


Fig. 1 (continued)

limit to what is routinely achievable, at a minimum DLR of approximately 0.10. Noise in the data may also be map-position-dependent, for which the DLR is a poor measure. Probe-specific (and GC-content related) correction factors have been built into some second generation algorithms such as ADM2, and can assist in controlling for this noise source [29].

2.3 SNP Arrays

In the cytogenomics market, two companies, *Illumina* (<http://www.illumina.com/>) and *Affymetrix* (<http://www.affymetrix.com/>), are well recognized for producing several generations of SNP-based microarrays for use by clinical laboratories. Both manufacturers employ their own assay chemistry and proprietary software to generate copy number and genotyping calls, but the basic principles are similar. The *Illumina* assay uses a single base extension method for generating differentially labelled SNP probes, while *Affymetrix* utilizes a base mismatch hybridization assay [30–32].

For this review, the *Illumina Infinium BeadChip* platform (*HumanCytoSNP-12* microarray) and *Illumina KaryoStudio* software are used to provide examples of SNP-based microarray data analysis. The *Illumina Infinium* assay employs 50-mer oligo probes bound to the *BeadChip* array. Once fragmented, the unlabelled patient DNA is hybridized to the array, and an enzymatic single base extension step is used to incorporate a differentially labelled fluorescent nucleotide at the site of the SNP (one color for each of the two SNP alleles).

For all SNP arrays, copy number is estimated from a normalized ratio of probe total fluorescence intensity (R) relative to a reference set of probes from many control samples, while genotyping calls (allelic ratios) are determined from the relative fluorescence intensities of the two separate SNP allele probes included for each SNP locus. This pair of probes has two 50-mers differing at a single base at the polymorphic site. Therefore, targeted SNP probe sites for microarray analysis are bi-allelic. All SNP array allele pairs are referred to arbitrarily (in the software displays) as allele A or B.

The addition of the genotyping data provides a powerful tool that expands the clinical utility beyond the detection of copy number imbalances alone. Deviations from the expected allele frequencies among adjacent bi-allelic SNPs allows for the detection of some copy-neutral changes that may have clinical relevance in the identification of uniparental disomy (UPD) (Case 4), chimerism [33], low grade mosaicism (Case 5), consanguinity and, indirectly, DNA sequence mutations for recessive diseases [28, 34–36]. Furthermore, some other whole of genome copy number abnormalities such as triploidy can also be detected, which is not possible with CGH. Somatically acquired copy-neutral loss of heterozygosity (LOH) events are the hallmark of many cancers, and these are also readily identified using SNP-based whole genome microarrays [37–40]. They are not dealt with here.

The analytical process differs from that of the CGH array, in that a reference control genome is not hybridized competitively with each test sample. Accordingly, Log R ratio (normalized intensity values, where $\text{Log R ratio} = \text{Log}_2 (R_{\text{observed}}/R_{\text{expected}})$) and allelic intensity ratio (B-allele frequency: BAF) are determined not by comparison with the control genome, but by canonical cluster position information, derived from a reference set of many normal data samples, which are included in the software itself (*see* **Notes 1** and **8**). Hence, each bi-allelic SNP probe, when used to interrogate the normal (= diploid) two copy state, will produce three positional clusters representing the genotypes AA, AB, and BB.

In an analytical setting, copy number loss and gain are seen to alter both the expected BAF and Log R intensity ratio values (Case 3), whereas copy neutral changes, such as uniparental isodisomy, will alter the BAF but not copy number (Case 4). Mosaicism for a copy number change may alter BAF, and may perhaps alter the Log R intensity ratio, depending on the relative proportion of normal and abnormal cells (Case 5). With respect to trisomy mosaicism, it is usually possible to determine whether the trisomic cells have a meiotic or mitotic origin based on the genotyping profile generated by the SNP array data (also Case 5).

2.4 QC with SNP Arrays

The *Illumina Infinium HD* microarray software presents the user with the parameter LogRDev, which is the key quality metric used for measuring the level of noise in these SNP arrays. This metric is a measure of the standard deviation of the Log R ratios for all autosomal SNPs. The manufacturer recommends a value

<0.3, however, our experience suggests LogRDev should not be >0.22. As a guide, excellent data are obtained at values between 0.07 and 0.12 when requiring a 0.2 Mb effective resolution, for the *Illumina 300K HumanCytoSNP-12* array. In our laboratory, reports are qualified at 0.5 Mb resolution if LogRDev is in the range ≥ 0.16 –0.22, while samples are generally failed if LogRDev exceeds 0.22. Other monitored quality metrics are BAFDev (standard deviation of all B allele frequency calls), which should be ≤ 0.03 , and SNP calling rate (=the proportion of successfully genotyped SNPs), which should be >99 %.

2.5 Databases

The cases presented are discussed with reference to several CNV and genomic databases or browsers, available via the following links: The databases of “International Standards for Cytogenomic Arrays” (<https://www.iscaconsortium.org>); the Decipher Consortium (<http://decipher.sanger.ac.uk>); OMIM (<http://www.omim.org>); the UCSC genome browser (<http://genome.ucsc.edu>); the Toronto “Database of Genomic Variants” (<http://dgv.tcag.ca/dgv/app/home>); the “Copy Number Variation 3 Project of the Children’s Hospital of Philadelphia” (<http://cnv.chop.edu>), and ECARUCA (<http://www.ecaruca.net/>).

Among the above, the CHOP and DGV databases deal exclusively with normal (or non-patient) cases. A thorough familiarity with these databases is important. Most are publicly available, to some extent, although registered users may have access to additional data which is preferred. The CMA results for each case are expressed in the standardized terms of the “*International System for Human Cytogenomic Nomenclature*” [41], and all CMA laboratories should be familiar with this nomenclature, which can be very complex.

3 Case Studies

Cases 1 and 2 were tested by CGH array, on *Agilent* platforms, (Case 1: *SurePrint G3 Human CGH Microarray 8x60K*, and Case 2: *SurePrint G3 Human CGH Microarray 2x400K*). Cases 3–5 are SNP array tests (with details given below). The CGH arrays were analyzed by *Agilent Cytogenomics* software, with user interface settings as follows: “Aberration Algorithm: ADM-2, Threshold: 6.0, Centralization: ON, Bin Size: 10, Centralization Threshold 6.0, Fuzzy Zero: ON, Combine Replicates (Intra Array): OFF, Genome: hg19, Aberration Filters: minProbes = 5 AND minAvgAbsLogRatio = 0.25 AND maxAberrations = 30 AND percentPenetration = 0, Expand Non Unique Probes: OFF”.

3.1 CGH

Example: Case 1

Reason for referral: Speech problem, ?autistic features, joint hypermobility.

DNA source: whole peripheral blood in EDTA.

CGH array: SurePrint G3 Human CGH Microarray 8x60K, resolution 0.2 Mb

Quality score (QC): DLR=0.136 (=acceptable)

ISCN result: arr 10p12.1(25,843,396-26,106,162)×4, 16p11.2(29,478,050-30,190,508)×1

Case 1: Sample CMA Report (CGH Array)

Test type and reason for referral: Chromosome Microarray (CMA); female child referred for “speech problem, autistic features, joint hypermobility.”

Specimen: DNA from peripheral blood, in EDTA.

Test platform: “Agilent SurePrint G3” Targeted Microarray 8x60K (60-mer oligo probes, mean effective resolution: 0.2Mb). Array serial number: ##0042R4 1_4

QC data: DLR=0.136

Results: A likely triplication, and a deletion were detected on chromosome microarray.

The ISCN (2013) description is:

arr 10p12.1(25,843,396-26,106,162)×4,
16p11.2(29,478,050-30,190,508)×1

Interpretation:

This CMA test found a probable heterozygous triplication (= 4 copies) within chromosome 10, band p12.1, and a heterozygous deletion (1 copy) within chromosome 16, band p11.2.

The 10p12.1 triplication has minimum length 0.27 Mb, and extends from position 25.84 to 26.11 Mb (Max. length is 0.37 Mb). This CNV contains part of one gene only (GPR158), and is regarded as a variant of uncertain significance (VOUS). No equivalent CNV is represented in the relevant databases. Even duplication CNVs are not well-established, at this locus. Given however, that the patient’s deletion in 16p is likely to account for her clinical picture, this 10p finding is regarded as likely benign and incidental. The absence of relevant genes supports this interpretation, but a clinical significance cannot be excluded entirely.

The heterozygous deletion of 16p11.2 has minimum size 0.71 Mb and extends from map position 29.48–30.19 Mb (Max. size is 0.95 Mb). This well-known deletion includes the region known as the “16p11.2 autism susceptibility locus” (<http://www.omim.org/entry/611913>). This deletion includes here approx. 33 genes, from *LOC388242* to *MAPK3*. Three of the 33 genes are recorded in OMIM as disease causing, namely, *KIF22*, *PRRT2*, and *ALDOA*.

But, the included gene *KCTD13* may also be relevant, as a likely candidate for the ASD phenotype (Zufferey et al., *J Med Genet*, 49:660–668, 2012) [42].

In the present clinical context, we note that *KIF22* has been associated with joint laxity, and mutations of this gene are known to be dominant in effect (see <http://www.omim.org/entry/603213>). *PRRT2* mutation is also dominant, associated with phenotypes of paroxysmal kinesigenic dyskinesia (PKD) and infantile convulsions, which may be seen in company with the other features of this “autism susceptibility locus.” Patients can also present with obesity, or a range of dysmorphisms (Zuffery et al: cited above).

Indications for follow-up testing:

CMA testing of the proband’s parents is recommended.

The 16p11.2 deletion represents a susceptibility locus only, and inheritance via a normal (or mildly affected) carrier parent is known. If any parent carrier is confirmed, then a significant risk of transmission to other offspring is established, and appropriate counselling can be considered.

This same parental array testing can also exclude de novo mutation of the 10p12.1 ?triplication. If a normal parent is found to carry this rare CNV, then its interpretation as a benign and incidental finding will be strongly supported.

Technical note: Map data presented above are based on the February 2009 human genome assembly (GRCh37, or hg19). Gene numbers and names are based on the “NCBI ReferenceSequence” (or RefSeq).

Standard Caveat for post-natal CGH arrays: Our current reporting policy is to omit copy number changes that do not contain genes, are well-established polymorphisms, or are smaller than 0.20 Mb (unless associated with a gene of known or suspected clinical significance). This test does not exclude balanced rearrangements, DNA sequence mutations, or Fragile X syndrome.

3.1.1 Results

Some CGH data collected in the CMA test for Case 1 are shown in Fig. 1a. Three CNVs were detected on this array platform. Of these, only (1) and (3) were reported (*Box*).

1. *A well-known and likely pathogenic 16p11.2 deletion* [42].
2. *A common and innocuous duplication, also in 16p11.2* (see above link to the *Database of Genomic Variants*).
3. *A very rare 10p12.1 triplication (or homozygous duplication), of unknown significance.*

The upper and lower panels of Fig. 1a show respectively the CGH array data, for all of chromosome 16, and the 16p11.2 deletion

region only. As the former shows, there is both a deletion (red) and a duplication (green), located just to the left of the chromosome 16 centromere, in the proximal p arm. The 16p11.2 duplication can be excluded immediately since it is a very well-known, benign, and thus incidental finding. As the various databases show, such CNVs are especially common, and polymorphic, in the pericentromeric (as here) or sub-telomeric regions of many chromosomes. Curiously though, the apparent “duplication” here can sometimes reflect cross-hybridization with paralog sequences in band Xq28 [43]. Originally, the paralogous region was mapped to 16p11.1, later corrected to 16p11.2. Irrespective of the cause, this finding is benign.

The 16p deletion (lower panel: red dots) includes ten consecutive data points, with a mean ratio of -0.55 , auto-detected by the software algorithm as significantly below the expected “normal” log ratio of 0.0 (as flagged by the red horizontal bar). Despite the acceptable QC value for these data (DLR= 0.136), note that there is considerable noise within these ten data points, and that the mean ratio within the deletion (-0.55) is higher than the -1.0 ratio theoretically expected for a heterozygous deletion (since $-1.0 = \log_2(1/2)$). Such deviations from the expected can arise through various sources of noise (discussed below) and can affect CMA data, even if QC parameters are satisfactory as seen here. For the heterozygous triplication/homozygous duplication CNV in 10p (Fig. 1b), note that all five abnormal probes (green) fall close to $\log_2(4/2)$, or $+1.0$, as should be expected for four copies of the region in question.

The report for Case 1 (*Box*) has been written as an educational aid and so is longer than one routinely generated in a high-throughput laboratory with an automated reporting pipeline. But the subcategories listed (in bold) should all be included. As is required for any patient report that includes reference to *variants of unknown significance* (VOUS), this one was composed via reference to several databases, as shown here in Fig. 1c, d, for the deletion in 16p, and the query triplication in 10p respectively.

The ISCN result [41] is an obligatory feature of all CMA reports (*Box*). All details of this complex nomenclature cannot be included here, but the reader should be able to infer the following: (1) This is an array (“arr”) result, rather than a FISH or karyotype result; (2) The chromosome bands involved 10p12.1, meaning chromosome 10, band p12.1, and (3) The CNV’s genomic map coordinates for the relevant chromosome, which are placed immediately after the band designation. Thus, coordinates here are for chromosome 10, from base 25,843,396–26,106,162, indicating a *minimum* length of 0.27 Mb. Equivalent results are given for the second CNV (as ordered by chromosome number): the deletion in 16p11.2. In the legend to Fig. 1b, an explanation is given why these are *minimum* lengths for both the CNVs described.

Last, the number of copies for each CNV is indicated by the two numerical characters at the end of each CNV descriptor, thus $\times 4$ indicates four copies for the “query heterozygous triplication”

(or homozygous duplication) of 10p and $\times 1$ indicates one copy for the heterozygous deletion in 16.

In respect of the map coordinates, a patient report must state to which version of the human genome build it refers. This may be stated as a suffix to the ISCN result itself, but in the present report (*Box*), it is given near the bottom, in the Technical Note.

The degree to which any individual genes are discussed in the patient's CMA report will depend upon several factors. First, the size of the CNV: as very large CNVs will contain many genes making it impractical to list them. In these cases the size of the CNV will likely attest to pathogenicity per se and this can be stated, but the relevant critical gene(s) may be impossible to identify. If an included gene's phenotype appears relevant to the referral context, then it should be mentioned as for example, "possibly worthy of consideration" especially if it is known to be dominantly expressed.

If the CNV corresponds to a well-established contiguous gene syndrome, e.g., *Case 1*, it may be unnecessary to mention individual genes. A link to a site such as OMIM, in which the overall syndrome is discussed, may suffice. For some of the contiguous gene syndromes, the critical gene(s) may not be known. Indeed, this is true in part for the present deletion in 16p11.2, i.e., for its autism-related aspects. In all cases an extra comment is needed confirming that the size and position of the reported CNV is consistent with (or not) what would be expected for that syndrome.

In almost all cases, it is important to include the first and last genes which map (completely or in part) within the *minimum* detected CNV. This gives the reader a simple means of inferring the gene content of the CNV including which genes are *not* present, even if unfamiliar with reading genomic map coordinates. It is *very* important here to consider the *minimum*-length CNV only. If a gene closely flanks a minimum length CNV, there are rare circumstances where this gene might also be mentioned, e.g., if relevant to the clinical picture. But this is discouraged without a specific justification, and any such mention must carry the specific caveat: *NB: the present CNV may not include these genes at all.*

For a CNV as well-known as is this particular 16p deletion (Fig. 1c), it is not often necessary to review any databases. But it is recommended periodically, as the databases are frequently updated. Little description of the clinical and population aspects of this 16p deletion are included here, but further details can be sought from the literature [12, 42].

For any very unusual CNV, like the 10p here (which is regarded as a VOUS), extensive database reference is obligatory. Laboratories that have accumulated their own in-house CMA databases will also refer to these, at such times. No OMIM disease-causing gene lies within this CNV, although there is one gene nearby, namely, *MYO3A* (Fig. 1d). By clicking on the dark green icon representing this gene in the UCSC interface, one is taken to the OMIM Web site, where we find it is a deafness gene involving recessive inheritance.

Deafness was not mentioned in the referral, and in any case, this gene is not in the CNV (even at its maximum length), and is unlikely to be affected by it. The “?triplication” does lie partly within one other gene (*GPR158*: see legend of Fig. 1d), but this has no known disease association.

In the present setting, we conclude that this novel 10p CNV is very unlikely to be of any clinical significance. But there is one aspect that bears slight consideration. Strictly speaking, the CMA result here does not distinguish between a heterozygous triplication, and a homozygous duplication (as both imply four copies). But the latter is extremely unlikely, given the overall rarity of CNVs at this locus (see **Note 2**).

Database details for both CNVs are summarized in Fig. 1c, d, which comprise screenshots taken from the UCSC gene browser at <http://genome.ucsc.edu>, as viewed after linkage from <https://www.iscaconsortium.org>. Display options switched “on” include nine categories of CNVs from the ISCA database and others from the CHOP database, the Decipher Consortium, the OMIM disease genes, the DGV database, and segmental duplications of “>1,000 bases” (bottom Fig. 1c). Note that two groups of segmental duplications flank the common CNV locus, and it is within these that non-allelic homologous recombination can occur [14, 44]. Under the categories “CHOP CNV...” and “ISCA CNVs benign” are included single examples of 16p11.2 deletions equivalent to *Case 1*, but reported from presumed normal individuals. These two likely reflect the occasional incomplete penetrant normal carrier of this “curated pathogenic” deletion. They may also represent possible entry errors in the database files.

When *deletion* CNVs are found as VOUS, another matter that may need to be considered is the possibility of recessive disease [11] as these genes will be hemizygous within the deleted region. For more discussion see **Note 3**, and for reporting of incidental CMA findings involving genes of later onset disease (see **Note 4**).

3.1.2 Results of Follow-up Testing

In the patient report (*Box*), follow-up testing was indicated, leading to results (1) and (2) below. The significance of these findings, and the reasons for their request are discussed, with a view to establishing a more general *modus operandi* for CMA follow-up.

1. *Follow-up array testing of the parents showed that neither was a carrier of the critical deletion in 16p11.2—i.e., this is a de novo mutation.* The absence of a deletion 16p11.2 carrier parent here is not at all surprising in view of what is known about this CNV [12, 18, 42]. But of course, the result has a major impact on the estimated risk of recurrence in future progeny, and so is an important justification for follow-up. If a heterozygous carrier parent had been found, the progeny’s risk of inheriting the same heterozygous CNV would comprise the Mendelian expectation of 50 %. But absence of a carrier parent implies a much smaller

recurrence risk of ~1 %. In clinical genetics, the risk of recurrent dominant inheritance from a non-carrier parent is not regarded as zero. It is customarily quoted as 1 %, merely to account for the indisputable possibility of germ-line mosaicism, for the dominant mutation (CNV, or rearrangement), in either of the apparently “non-carrier” parents. Of course, this assumes paternity is not an issue.

It is known that this 16p deletion is associated with incomplete penetrance [12, 17, 42], i.e., “normal” parent carriers are described. Penetrance for this 16p11.2 deletion has been reported as high as 93 % [17] to a lower estimate of 47 % [45]. The matter is complex and the topic of current research [12, 42, 46] (*see Note 5*). But regardless of these details, follow-up testing of the parents was appropriate since “normal” parent carriers have been reported and so testing was necessary to establish *risk* for other offspring.

In terms of a clinical *diagnosis*, this follow-up result would have no bearing on our interpretation that the 16p11.2 deletion was responsible for the proband’s phenotype. With or without a carrier parent, the evidence here is equally compelling.

We have used the recurrent 16p11.2 deletion to exemplify the very common class of pediatric CMA referrals involving features of autism spectrum disorder/developmental delay. However, despite being the second most common individual CNV associated with autism, this particular deletion accounts for only 0.5 % of cases [46]. This is not unexpected as many genomic regions can affect these phenotypes [2].

We are also interested in the more general case that involves novel, or poorly characterized CNVs (or VOUS). For these, different rules of interpretation apply. We would in this context regard any *de novo* finding as evidence supporting *likely pathogenicity*. In this regard, we are primarily informed by the structural-variant disease hypothesis [23] described in Subheading 1.

2. *On parental follow-up, the proband’s 10p triplication was shown to be familial as one “normal” parent carried it.* Because of the deletion 16p finding and this rare CNV’s minimal gene content (Fig. 1d), it is presumed to be a benign incidental finding despite its very rare and unusual nature. Even in the absence of the 16p deletion, we would have reached the same conclusion, because of its familial nature.

One other point arises from follow-up here. We now know that the 10p CNV is a heterozygous triplication rather than a homozygous duplication, since both the proband’s extra copies of this CNV have come from the one parent. As suggested earlier this is also as expected, given the relatively remote possibility of both parents being carriers. In our experience, appreciation of such nuances may be of more than academic interest, especially in some regions of the genome, where it can be difficult to

distinguish duplications from normal, and triplications from duplications. A relevant example is at a clinically well-known locus in chromosome 7q11.23 [47, 48]. Such problematic regions may be characterized by atypical DNA parameters including GC content (*see* **Note 6**).

3.1.3 Regarding the Possible Limitation of Scope in Data Analysis, for CMA Follow-ups

Because VOUS are so often encountered, some labs may severely limit the scope of their routine follow-up CMA analysis, to prevent indefinite proliferation of follow-ups, and in deference to the possibilities raised in **Note 4**. In *Case 1*, for example, such limitation of scope would permit follow-up reporting of array data for bands 16p11 and 10p12 only, since CNVs within both regions were included in the proband report. Under such a practice: Whether or not the proband's CNV is found, the parent or relative tested will receive a "limited scope" CMA report, which must include the critical caveat: *This CMA analysis was limited to chromosome band... only*. Alternative to this though, many labs do not limit the analytical scope of their CMA follow-ups, and some others may conduct follow-ups via a locus-specific test, such as FISH.

3.2 CGH Example: Case 2

Reason for referral: Neonatal death, ?alveolar capillary dysplasia.

DNA source: formalin-fixed tissue

CGH array: SurePrint G3 Human CGH Microarray 2x400K: typical resolution 0.06 Mb.

Quality Score (QC): DLR=0.457 (unacceptable in almost all circumstances)

ISCN Result: arr 16q24(86,201,128-86,326,963)×1

3.2.1 Results

Figure 2 presents the 16q24 deletion data for Case 2 using a higher resolution *Agilent* array. These data demonstrate a 16q24.1 deletion of minimum size 0.13 Mb, mapping to the approximate interval 86.20–86.33 Mb. However, the QC value is so poor that the data would normally be rejected. They were not, due to the positive and disease-specific result obtained, consistent with the very unusual referral context. The patient report was issued with the caveat that independent confirmation was obligatory because of the poor data quality. On the other hand, if no positive CNV finding had been made, a report stating *No abnormality detected* would not have been issued. Instead, the report would state *No result obtainable, due to unacceptable QC values*.

The gene *FOXF1* is implicated in alveolar capillary dysplasia (<http://www.omim.org/entry/601089>). Because of the nature of the referral, the specimen could not be recollected, and poor quality DNA was likely from the fixed tissue sample available. For this type of referral a higher resolution array might be chosen from the start because: (1) A single gene was likely to be involved, and

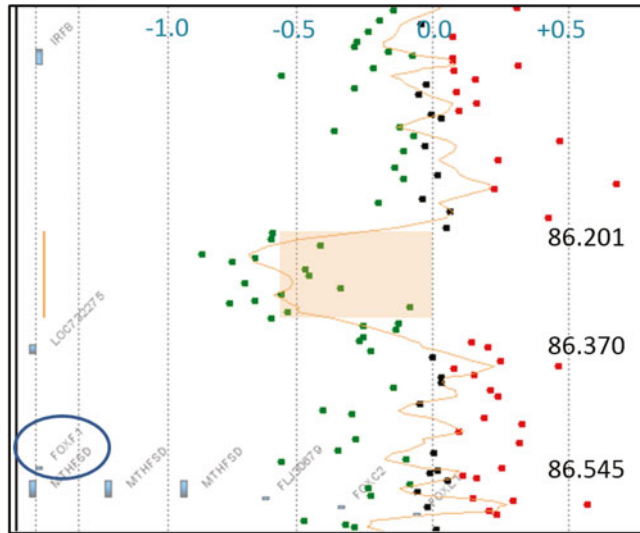


Fig. 2 Detail of the 16q24 deletion for Case 2, as presented via *Agilent Cytogenomics* software. The *X-axis* here plots the \log_2 ratios, ranging from -1 to $+0.5$ (*top*), while the *Y-axis* indicates the hg19 map position within the genome (in megabases: *at far right*). Map coordinates of the two genes *LOC732275* and *FOXF1* are indicated, and *FOXF1* is circled. The ADM2 algorithm (supplied with the *Agilent* software) flags an apparent 0.14 Mb deletion in chromosome 16, band q24.1, indicated here by the *vertical orange bar at left*, and the *rectangular color box at center*. The *dots* are data points for all probes in this region, with *green* and *red dots* indicating \log_2 values significantly lower or higher than zero, respectively. The *darker orange trace line* indicates the smoothed mean \log_2 ratio, with window size here set at $n=5$ consecutive data points. Note the scatter of data points, as reflected in the unacceptably high DLR of 0.457 (*see text*). Indeed, \log_2 ratios for the five right-most data points are consistent with duplication: an obviously false interpretation, given the broad scatter overall

hence, a small CNV is possible; (2) To compensate, for the loss of resolution inherent in the poorer quality data.

In Fig. 2, note that the scatter of all data points is considerable, and that the mean log ratio within the flagged deletion segment ($\log_2 = -0.57$) does not reach the expected value of -1.0 . The high DLR value here ($=0.457$) is attributable to the DNA quality, and map-position-independent noise is expected. Despite this, the probe concentration of this 400K array is sufficient for the algorithm's auto-detection of this deletion. The deleted region includes no known genes, but does not correspond to any known benign CNV. However, relevant to the present context is the proximity of the gene *FOXF1* located 0.22 Mb from its distal end (Fig. 2).

Earlier in this chapter, we stressed that the ubiquity of small CNVs in the normal population necessitates the general application of size filters to CMA data, so that vast numbers of small CNVs will be excluded from patient reports. However, we present

here a case in which the *minimum size rule* for the lab reporting Case 1 (*Box*) has been ignored, very reasonably, on grounds of a highly specific clinical context. Also ignored is another Case 1 caveat concerning gene content.

The implications here are that for very general phenotypes (like developmental delay and so on), a filter on CNV size, and gene content, may be applicable. But if, a priori, the clinical context indicates a specific disease locus, then these filters may be relaxed. Important here also is the distinction made between a disease (or phenotype) locus, and the mapped limits of a known structural gene (although either may be referred to as a genetic locus). In some cases, mutations at sites outside of some structural gene, but usually not far from it, may affect gene regulation to produce the disease phenotype [49]. Thus, the small pathogenic 16q24.1 deletion detected here contains no known genes and lies 0.22 Mb from the structural gene *FOXF1* (Fig. 2).

The same specimen was retested and the deletion confirmed as part of a research project which studied further this aspect of *FOXF1* regulation [49]. In a routine setting though, this finding would have been confirmed by other means, e.g., FISH, if an appropriate sample had been available. While a carrier parent is not likely here, the possibility that a parent carried some balanced rearrangement involving this locus could probably be excluded, also by FISH (*see Note 7*).

In CMA diagnostics, it is useful to know that small non-gene containing CNVs can still be associated with severe clinical effects. However, proving that such a segment includes a cryptic functional element does not mean they all do. In fact, it seems likely that the great majority do not [50].

3.3 SNP

Example: Case 3

Reason for referral: History of miscarriage in a 27 year old female.

DNA source: Fetal tissue (miscarriage sample)

Quality score: LogRDev = 0.08 (excellent)

ISCN result: arr 9p24.3p24.1(36,587-8,187,045)×1, 11p15.5p15.4(193,788-6,197,926)×3 dn

3.3.1 Results

Figure 3 shows the effect of a deletion and duplication (both heterozygous) on Log R Ratio and BAF (B allele frequency) values in an *Illumina HumanCytoSNP-12* microarray analysis. These data are generated from a late first trimester miscarriage sample in which two pathogenic copy number abnormalities are detected; a terminal deletion of approximately 8.2 Mb from chromosome region 9p24.3p24.1 and a terminal duplication of ~6.2 Mb from chromosome region 11p15.5p15.4. While both are large, their comparable sizes, with reciprocal gain and loss, render them cryptic under the light microscope.

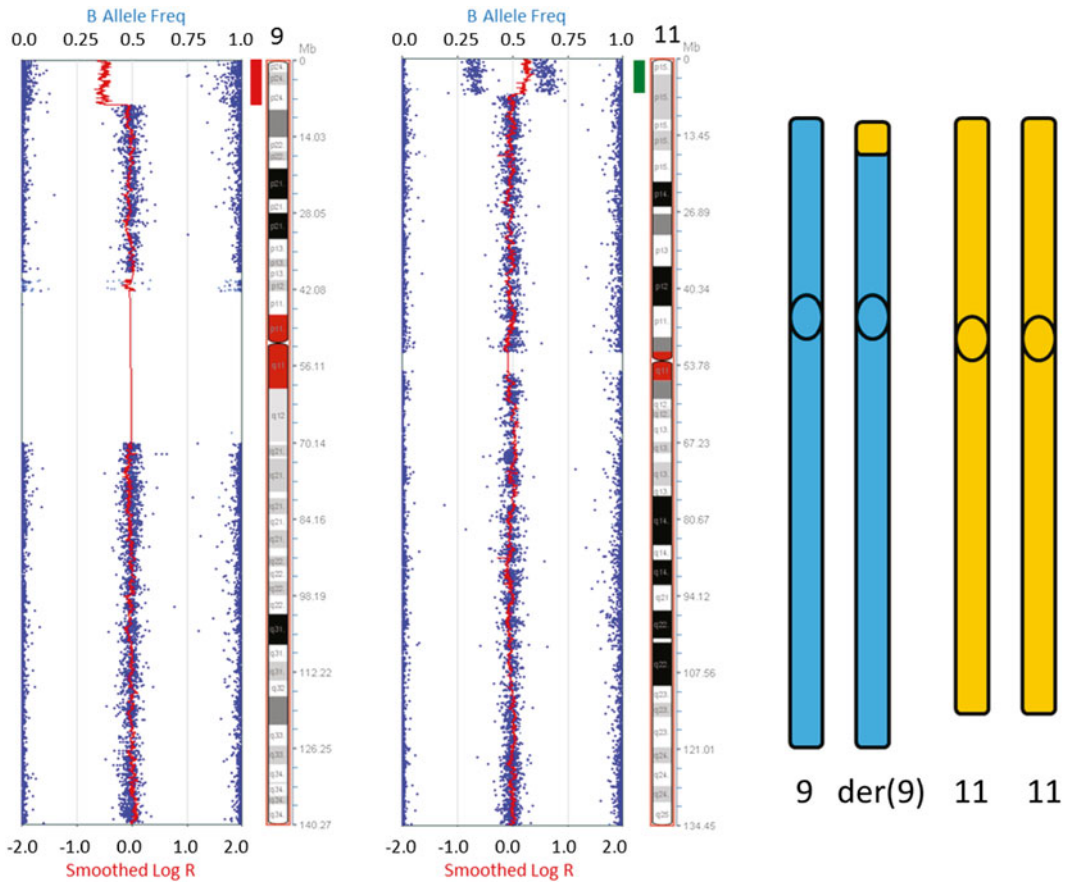


Fig. 3 *KaryoStudio* SNP microarray data plots for chromosomes 9 and 11 (Case 3). The display incorporates Smoothed Log R (*lower X-axis*) and B Allele Frequency (BAF) (*upper X-axis*). The smoothed Log R is a sliding window of average Log R intensity ratio and is represented by a continuous *red line* that runs the length of the chromosomes. The Smoothed Log R decreases in value at the point of the terminal 9p deletion. The deletion call is supported by the genotyping data (SNP probes represented by *blue dots*), where heterozygous AB genotyping calls (BAF = 0.5) are absent. On chromosome 11p, the Smoothed Log R increases in value and is associated with BAF genotyping calls at 0.0 (AAA), 0.33 (AAB), 0.66 (ABB), and 1.0 (BBB), consistent with duplication of this region. At *right* is a schematic of chromosome 9 and 11 homologues showing the duplicated 11p terminal region translocated onto the derivative (*der*) chromosome 9

In excess of 350 genes are involved overall. The deleted 9p region is associated with a decrease in Log R ratio (Smoothed Log R) and an absence of heterozygous SNP calls at BAF=0.5, consistent with one allele being deleted at each SNP site assayed. Thus, the BAF for genotype A/- is 0.0 and for B/- is 1.0. Note these values are the same as those generated for the two homozygous states (AA or BB). Any deleted region will necessarily exhibit loss of heterozygosity (LOH) since only one allele can be present at any locus within this heterozygous deletion. In this context then, the LOH is explained by hemizygosity of the deleted region.

The duplicated region from chromosome 11p is associated with an increase in Log R ratio, while BAF values of 0.0 (AAA), 0.33 (AAB), 0.66 (ABB), or 1.0 (BBB) are consistent with a duplication (*see Note 8*).

Readers familiar with cytogenetics will be aware that these copy number abnormalities suggest the presence of an unbalanced translocation, where the terminally duplicated segment of chromosome 11p is translocated onto the deleted chromosome 9p, resulting in a derivative chromosome (Fig. 3). An unbalanced translocation can be inherited from a balanced carrier parent, or less frequently it might arise *de novo*. In this case a subtelomere FISH analysis on parental samples indicated neither parent carried the translocation, demonstrating the rearrangement arose as a *de novo* event with low (1 %) risk for recurrence.

Although Case 3 was tested here by SNP array, the same result could have been achieved by CGH array testing.

3.4 SNP

Example: Case 4

Reason for referral: Unexplained fetal death in utero

DNA source: Fetal tissue

Quality score: LogRDev=0.39 (very poor, indicating failure to meet standard QC guidelines)

ISCN result: upd(15)mat.arr 15q11.2q13.3(21,361,700-29,256,215)×2 hmz, 15q21.3q26.1(54,114,19-88,153,073)×2 hmz

3.4.1 Results

As with Case 2, this example is characterized by poor quality DNA. Evidence of this poor quality, and its effects on the CMA analysis, are considered.

SNP-based microarray analysis has the advantage of being able to identify copy neutral changes, where allele frequency is altered but copy number remains unchanged [28, 34, 35, 51]. The “abnormality” here appears as one or more *long continuous stretches of homozygosity* (LCSH). If the latter is observed on many chromosomes it suggests consanguinity, where identical segments of the genome are inherited from a common parental ancestor (*see Note 9*). For example, the offspring from a first cousin union will have, on average, 6.25 % (1/16) of their genome *identical by descent*. LCSH in this context is most often benign, and not in itself diagnostic for any condition. Nevertheless, it is associated with an increased risk for recessive disease [28, 34, 52].

The present case deals with a completely different scenario. Here, only one chromosome is involved, and there is no history of consanguinity. Therefore, this finding is more likely to reflect segments of isodisomy associated with uniparental disomy (UPD). UPD occurs when both members of one homologous pair of chromosomes are inherited from the one parent, with no contribution from the other parent [53, 54]. UPD findings may be clinically

significant per se, but only if the chromosome involved is among those few that are known to include imprinted segments, as is the case for chromosome 15 [54, 55]. However, as for the LCSH that arises via consanguinity, the LCSH of UPD can also be associated with recessive disease, but only within the isodisomic region itself.

Figure 4 shows the microarray profile of chromosome 15 from a mid-trimester fetal death in utero. The fetal DNA is degraded, which is reflected in the very poor array QC score ($\text{LogRDev} > 0.30$) and the erratic Smoothed Log R along the length of the chromosome. Despite this, the allele frequency data are of sufficient quality to identify two large (7.8 Mb and 33.9 Mb) regions of LCSH on chromosome 15, without evidence for LCSH on any other chromosome (*see Note 10*). This finding is suggestive of UPD15, where regions of LCSH likely reflect segments of isodisomy that arose from recombination between chromosome 15 homologues at meiosis I, prior to conception. Confirmation of UPD15 can be achieved by several molecular based approaches [56], including a comparative analysis of parental and proband DNA using microsatellite or SNP data, or by using a methylation-specific assay that exploits the differential methylation of maternal and paternal DNA within the chromosome 15 imprinted region. In Case 4 a methylation-specific PCR (MS-PCR) assay was used to confirm maternal UPD15 (Fig. 4) which is associated with Prader-Willi syndrome [56].

3.5 SNP

Example: Case 5

Sample: Placental chorionic villi (miscarriage sample)

Quality score: $\text{LogRDev} = 0.10$ (very good);

ISCN Result: $\text{arr}(16) \times 2\text{--}3$

3.5.1 Results

Figure 5 shows the complex SNP microarray profile (left panel) of a first trimester miscarriage sample, associated with whole chromosome (=trisomy 16) mosaicism.

This case illustrates meiotic, mitotic and copy number complexities that can arise in SNP CMA data analysis. Understanding this case allows some insight into what might be found in the even more complex SNP analysis of cancer cells where many simultaneous CNVs, both heterozygous and homozygous, may evolve, even within a single chromosome.

Genomic mosaicism in the noncancer setting is typically characterized by the presence of two cell lines in the one individual. The mosaicism itself arises from post-zygotic, and hence mitotic, abnormalities of cell division [57]. In the case of trisomy mosaicism, presence of sufficient trisomic cells can (as for non-mosaic trisomies) result in miscarriage. However, with a sufficient proportion of normal cells, the pregnancy can persist, possibly resulting in morphological and/or developmental defects in the live-born child [34, 35, 58].

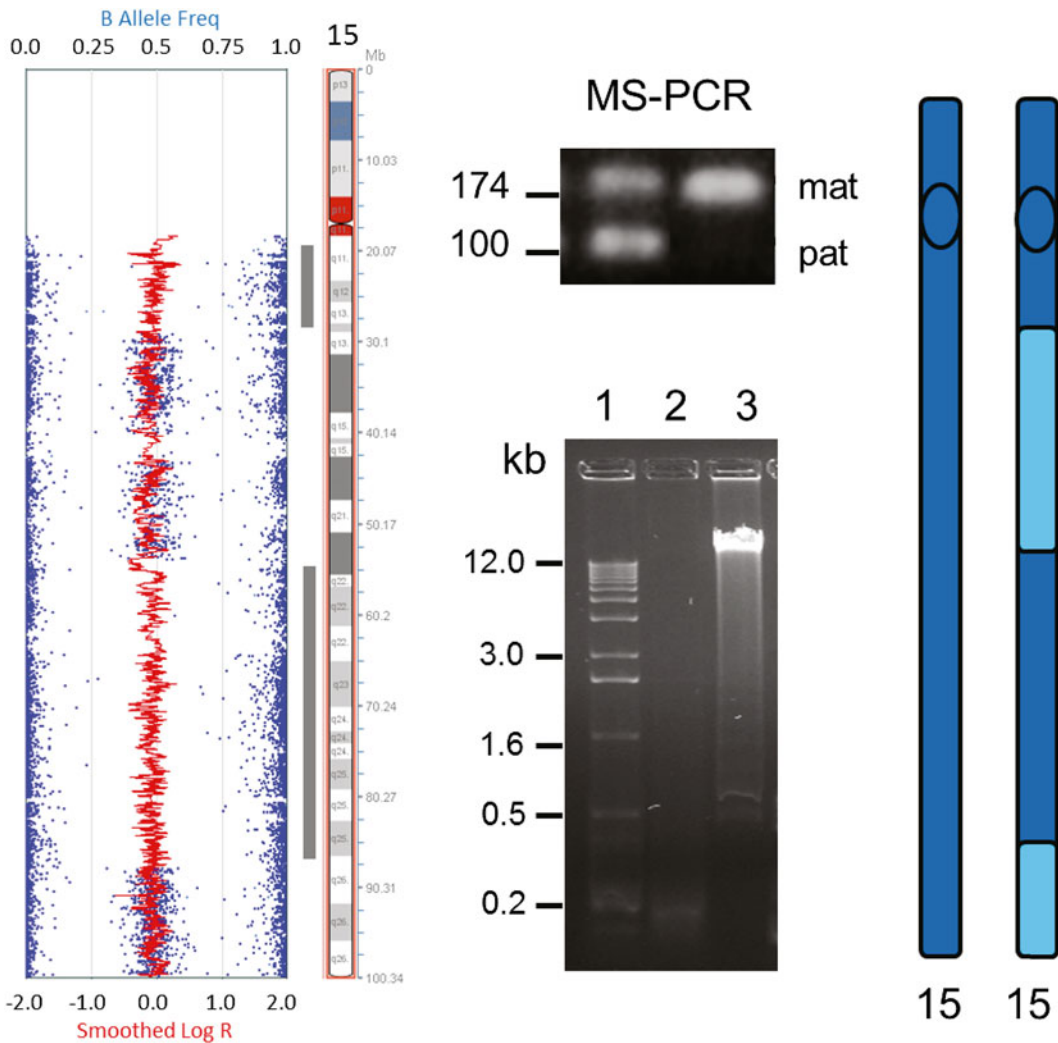


Fig. 4 *KaryoStudio* SNP microarray data plot for chromosome 15 from a fetal death in utero sample. Two large segments of long continuous stretches of homozygosity associated with uniparental isodisomy (Case 4) can be seen. The isodisomic segments show a loss of heterozygous SNP calls (BAF = 0.5%, *upper X-axis*) in association with normal copy number (Smoothed Log R = 0.0, *lower X-axis*). Maternal UPD15 was ascertained using MS-PCR (*see upper gel* photo), in which the normal control sample produces paternal (100 kb) and maternal (174 kb) bands, but the test case produces the maternal band only. In the *lower gel*, the degraded DNA used for the array shows a very light smear around 0.2 kb (*see lane 2*), compared with the non-degraded, high molecular weight (>12 kb) control sample in lane 3. (Lane 1 shows DNA size markers). *At right*, a schematic view of the chromosome 15 homologous pair, highlighting regions of uniparental isodisomy. This isodisomy is limited to the regions where both copies of 15 share *dark blue segments*

Commonly, trisomy mosaicism involves one abnormal plus one normal (= diploid) cell line. If the original (= zygotic) cell line is the abnormal one, then the one extra chromosome itself has a pre-zygotic (and usually meiotic) origin, in the germ cells of one parent or other. Detailed analysis of the SNP data here reveals that

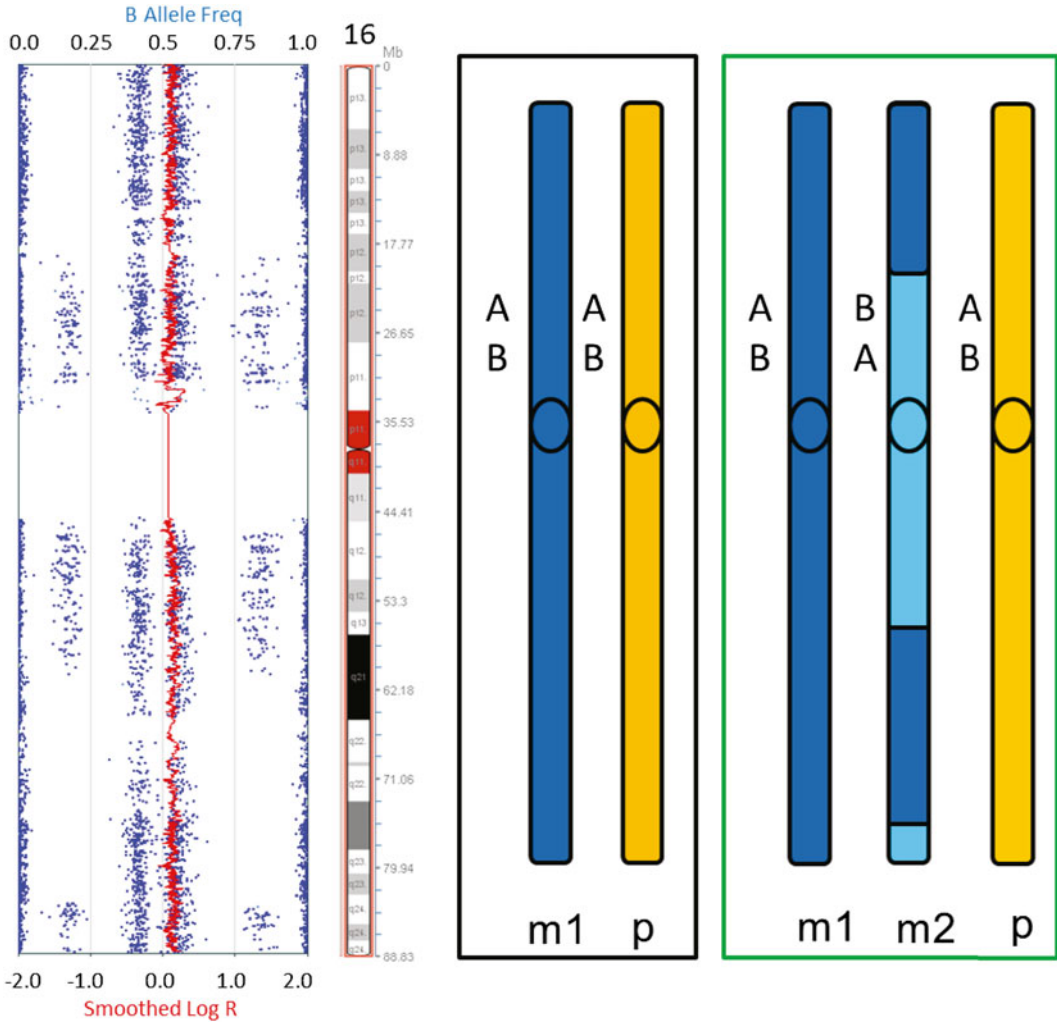


Fig. 5 *Left panel:* KaryoStudio SNP microarray data plot for mosaic trisomy 16 in a first trimester miscarriage sample (Case 5). As per the *lower X-axis* scale, the smoothed Log R (*vertical red line*) plots at >0.0 and <0.5 throughout its length, consistent with presence of mosaicism for a disomic and a trisomic cell line, with the former predominant. *Central panel (black frame):* the two schematic chromosomes indicate the inferred SNP genotypes of the two copies of 16 (m1 and p) remaining in the disomic cell line, after the third copy was lost. In the *right panel (green frame)* is shown the original trisomic (and bi-maternal) cell line, with genotypes, and positions of meiotic chiasmata in m2, as inferred from the SNP data at *left*. In m2, the changes from *dark to light blue* indicate each chiasma position, with the *light blue* around the centromere indicating heterozygosity in that region, and thus a meiosis I nondisjunction, as the cause of the initial trisomy (*see text*). The regions of recombination are inferred from the changes in BAF profiles, in the *left panel*. Note: the data here do not indicate in which parent the trisomy arose. The maternal origin depicted was established by separate parental testing, not shown

an extra chromosome 16 arose here during gametogenesis (in meiotic first division), and one copy (from the same parent) was later lost, by post-zygotic mitotic nondisjunction, to form a normal disomic and *biparental* second cell line (and hence there is no LCSH to suggest UPD in this SNP data).

In Fig. 5, the smoothed Log R ratio (red vertical line, left panel) is raised marginally above 0.0 along the length of the chromosome, indicating a mosaic copy number, of between two and three, while the BAF genotyping data are clearly abnormal.

We explain the abnormal BAF data as follows: Any trisomic cell line must have two chromosomes contributed by one parent (here as m1, m2), and one by the other (p, *see* green frame in Fig. 5). Given the multi-segment nature of the SNP data in this chromosome (left panel, Fig. 5), let us assume that the trisomic cell line here was the original one, and the extra chromosome had a meiotic origin. From these assumptions, this argument follows: For the two uniparental copies m1 and m2, some segments may be iso-allelic, or identical by descent (shown dark blue), while others are heterozygous (light blue), the distinction being due to inferred meiotic chiasmata, at the dark blue/light blue boundaries. Such chiasmata reflect exchange between non-sister (=light blue here) and sister (dark blue) chromatid segments, during meiotic prophase.

In the dark blue regions, therefore, m1 and m2 always carry the same allele (=identical by descent), and so there must be at least two copies of the same allele, whether A or B, for each trisomic SNP locus here (green box). Accordingly, when m2 is lost, to form the normal (disomic) cell line m1/p (black frame), *all initially heterozygous loci* (AAB or ABB) in the dark blue region can only become AB. Hence, as the new disomic cell line increases in frequency, genotype ratio (expressed as mean BAF) here shifts up from $0.33 \rightarrow 0.5$ (at some loci), or down, from $0.66 \rightarrow 0.5$ (in others). This means that the number of “bands” of BAF data visible does not change from the initial number (= 4), although two of them will shift their mean plotted position towards 0.5 (Fig. 5, left panel, as aligned with dark blue segments at right).

In the light blue region, however, loci on m1 and m2 do not necessarily carry the same alleles, and so initial heterozygotes AAB and ABB may, with loss of m2, become either AA or AB, and AB or BB, respectively. So, with increase in disomic cell frequency, the BAF may shift up (from $0.66 \rightarrow 1.0$, and $0.33 \rightarrow 0.5$), or down (from $0.66 \rightarrow 0.5$, and $0.33 \rightarrow 0$). In the process, it is the presence of a third allele on m2 (not identical to that of m1 or p), that accounts for the two additional bands of genotype data. Thus the SNP data originally seen as four bands of BAF ratios will come to occupy six bands, as we see for the light blue regions of Fig. 5 (right panel). Hence, our initial assumptions above are consistent with this interpretation of the SNP data. Generally, whenever six genotypic (BAF) bands appear anywhere within the SNP data for the extra chromosome, a meiotic origin is indicated. On the other hand, if the *entire chromosome* is comprised only of 4-banded regions, we regard this as tantamount to proof of a mitotic (and post-zygotic) origin of the extra chromosome.

Both the degree to which the BAF bands are shifted, and the mean position of the smoothed Log R ratio, allow us to estimate the percentage of the abnormal cell line. In this case, it was about 40 %. SNP arrays can, it is claimed, detect mosaicisms of ~10 %. Interested readers are directed to literature which provides detailed summaries [28, 35].

Finally, the importance of inspecting the SNP array data *manually* must be emphasized [34]. Some copy-neutral (or near copy neutral) aberrations, particularly low frequency mosaicism, and chimerism, may not be recognized automatically by the *Illumina* (or other) software algorithms. The trisomy 16 mosaicism reported here was detected by manual data inspection only, and this is true for most cases of low to moderate level mosaicism involving whole chromosomes, or chromosomal segments. Although the data are less complex, the same caveats can apply to mosaicism detected by CGH arrays.

4 Notes

1. Quality of SNP array data can be improved by users creating their own cluster or reference files from in-house clinical samples, rather than relying on reference sets supplied by manufacturers. In our experience this is true for both *Illumina* and *Affymetrix* microarrays.
2. In routine practice, homozygous duplications (and deletions) are not unusual for some very common CNVs. But for the rarer examples, dup homozygosity is so unlikely that one favors heterozygous triplication (or failing that, consanguinity or UPD) as an alternative explanation. Regardless of origins though, a homozygous duplication in the present CNV of 10p might perhaps be significant. It could imply sequence disruption for both copies of the one gene involved (*GPR158*: Fig. 1d), and therefore, effective nullisomy. Any CNV that might involve nullisomy, regardless of the gene involved, is at least worthy of a second thought. For the various reasons stated, none of this is likely to be significant for Case 1, but in the context of this chapter the implications are worth exploring.

Even when the CNV is heterozygous, determining significance of any duplication (or triplication) that includes *only part* of a gene is fraught with many unknowns. For example: is it a tandem duplication/triplication, or an inverted one? Is it a duplication/triplication in situ, or is an extra copy transposed elsewhere? None of these are known but will affect the gene sequence within and around the start or end point of the duplication (or triplication). Finally, is the start/end point intronic or exonic? This is something we might know from the CMA test itself, if using a very high resolution array: but this is not

usually the case for routine analysis. Without such information, we really have no idea what sequence will be expressed (if any), at or around the site of the duplication itself.

The smaller novel duplications are thus a challenge in CMA diagnostics. They rarely turn out to be of clinical significance, but a disproportionate effort can be required to consider their possible effects on gene expression. Often, a clinical significance cannot be excluded entirely, and they are reported, and even followed up, as VOUS.

3. For any deletion CNV that contains at least one gene, it is always possible that because of hemizygoty, a point mutation in the one remaining allele might result in a recessive disease. Awareness of this is important, as laboratories do occasionally find such examples [11]. On the other hand, no deletion CNV should be reported (or even commented upon) on grounds of speculation alone. The only time this is justified is when: (1) The deleted gene is known to result in a well-characterized recessive disorder, *and* (2) the patient referral indicates a phenotype suggestive of that disorder. Thus, these CNVs should not be reported if it is the laboratory's policy not to report CNVs that are classified as *likely benign* in the International Standards for Cytogenomics Arrays (ISCA), and are recorded in say, more than two studies. Policies are not simply changed on the grounds of speculation that one gene within a CNV might be mutated in the other allele (unless points (1) and (2) apply).

Consider also the case where parents are tested by CMA, as follow-up to a VOUS result for the proband, and one parent is found to be a carrier. In this circumstance, the follow-up report should not speculate: *This VOUS may thus be an incidental finding for the proband, but the VOUS does include a particular gene...that might be mutated on the proband's other allele.* Instead, the lab's general rule is followed as above, and the report should say only: *Assuming the carrier parent is of normal phenotype, this VOUS can be regarded as an incidental finding for the proband.*

4. There is one other class of CNV (usually a deletion) that requires special mention in the context of report writing. Most pathogenic CNVs contain multiple genes, some even a hundred or more, and any such CNV is likely to have caused the patient phenotype described in the referral. However, most of the genes included in the CNV will not be contributing to that phenotype, i.e., these gene deletions on their own are incidental findings. So it is not usually necessary to comment (or even list) them in the patient report. But if any gene deletion here predisposes to a later-onset disease, that single-gene finding might eventually have clinical significance to the patient.

Questions then arise concerning how the laboratory (and the referring clinician) should treat this sensitive and unsought information. One way is to first discuss the result with the referring clinician who may call in other specialists such as cancer geneticists, to reassess the data. Then, an agreed-upon report may be issued.

With special reference to genome sequencing rather than CMA, a recent US recommendation [59] states that the laboratory has an obligation to report these types of incidental findings, and suggests a list of 57 genes (plus others “as deemed appropriate by the laboratory”), which should be dealt with in this manner. This reporting should occur even in the absence of the patients’ prior agreement. However, the latter aspect is controversial and remains under debate, in terms of the patient’s “right not to know” any information that was not sought [60].

These problems around incidental findings require the CMA testing laboratory to be ever alert for the presence of such genes among all CNVs detected, even those smaller CNVs that might otherwise be regarded as benign. Fortunately, the latter are very rarely implicated, since the 57 genes are, by their nature, very unlikely to map within a recurrent CNV.

5. As seen in Case 1, the strategies adopted in considering CMA results can turn on whether incomplete penetrance has been established for the CNV in question. Consider in this context an alternative scenario to Case 1, where such a CNV is found in a proband, but the published phenotype is inconsistent with the patient’s clinical picture. In this case, either the proband may be: (1) Expressing a rare or novel variant form, of that CNV’s possible range of phenotypes, or (2) A carrier of this CNV, in its non-penetrant form. For the latter, the detected CNV may be an incidental finding, despite its established pathogenicity. This is certainly possible, and might be suggested in the report. Remember, at least 80 % of CMA tests do *not* solve the diagnostic problem.

The important consideration here is to be aware of the clinical notes, and so avoid reporting simply that: *A well-known CNV was found, which is consistent with the following features: ...*A machine-generated report might look like this, and such reports are seen. The CMA analysis software usually includes options for the auto-generation of patient reports, but they are not recommended for any complex or unusual findings.

When faced with rare cases of apparent genotype/phenotype mismatch, it is necessary to consider also the possibility that two (or even three) pathological and unlinked mutations of uncertain penetrance are present, some possibly segregating among the family members. Surprisingly, such families are not that rare. The “other” mutation(s) may be other CNVs or they

may be sequence mutations, not detectable on CMA testing. The challenge then is to consider how these multiple factors might interact.

For the 16p11.2 deletion there is debate as to whether incomplete penetrance or variable expressivity (or both) are the relevant terms [12], and it is suggested that “the ultimate phenotype of the child is probably affected by his/her genetic background and other environmental factors, the vast majority of which are unknown and cannot be tested” [11, 45]. When microarray testing identifies an additional CNV, it is likely to be even more difficult to predict the resultant phenotype. Such comments exemplify well the caveats often accompanying studies of recurrent CNVs (and other mutations), where the degree of penetrance (or expressivity) is inherently uncertain. For both the 16p11.2 and 22q11.2 deletions (the two most common CNVs of clinical significance), evidence now suggests that variance in one aspect of their phenotypes (mean reduction in IQ) is partly explicable in terms of parental variance for that same trait [12].

Consider also a related scenario: the *presumed normal parent* (on follow-up) is found to carry a known pathogenic but incompletely penetrant CNV that the proband does *not* carry. This circumstance is not extremely rare, and can present difficulties for reporting. Of course, if your lab has adopted a policy of “limited scope” CMA analysis (or FISH) on follow-up (as per Subheading 3.1.3), then this scenario is unlikely to arise.

A further related situation is this: the parent is found to carry an unexpected CNV *of late onset disease*. Clearly, such a finding would require special consideration (*see Note 4*).

6. In various regions of the genome, GC-related genomic waves represent a substantial source of false positive identifications [27, 61]. The GC percentage varies across the genome’s DNA, and array platform designers go to considerable lengths to overcome probe related noise related to this source via hardware [61] and software modifications [29]. GC-rich (also gene rich) regions correspond roughly with the pale (G-negative) bands of the G-banded karyotype [62], and many are of size order 3–20 Mb. GC-dependent noise in the log₂ ratios of CMA data can thus appear to rise and fall in “GC waves,” and hence the oft-described “waviness” problem. These are map-*position-dependent* sources of CMA noise, and an acceptable QC value does not necessarily guarantee their absence. The problem may be exaggerated when conditions are suboptimal [7, 26].

One example of GC-associated noise occurs at the Williams–Beuren Syndrome (deletion) locus in chromosome 7q11.23 [47, 48]. Here deletions, duplications, and triplications are all possible, and are variously significant, in a clinical sense. CMA differentiation of the 2, 3, and 4 copy states can be unreliable,

and so likely errors to be aware of here are reporting a normal as a duplication, a duplication as a triplication or vice versa

A second error-prone region lies within the q terminal band of the X chromosome—Xq28. This region exhibits high GC percentage in places, but the values vary over a shorter scale [62]. Affected are regions to both sides of the gene *MECP2*, and particularly the proximal gene *SLC6A8* itself (Fig. 6a). Both genes are associated with mental retardation, and so are often targeted on arrays. Unfortunately, this results in much poor data from the *SLC6A8* probes, which can have a disproportionate effect on the local \log_2 ratio. The lower panel of Fig. 6a depicts one attempt (by algorithm modification) to overcome the problem [29].

In regard to Xq28 itself, the most likely error is reporting a true normal as an Xq28 duplication (Fig. 6a, upper panel), or less likely, the reverse. The likelihood of error, though, is partly dependent on the patient's sex. Regrettably, both authors have experienced misreporting of the Xq28 region in the early years of CMA testing. But with the many improvements, and the accumulation of databases since that time, this type of error should not now occur!

An additional confounding factor in Xq28 is the segmental duplication intrinsic to the region, including *SLC6A8* itself, which shares paralogous sequences with proximal 16p11.2 [43]. Recall that we noted the same 16p11.2 segment appearing as a ?*duplicated* CNV in Case 1 (Fig. 1a). Of course, reciprocal errors of analysis might be expected in 16p11.2, but unlike Xq28, this particular part of 16p11.2 is not of clinical significance.

Figure 6b shows, in an abnormal case, some confusing \log_2 ratio scores, embedded within a large “true” duplication in distal Xq. On first analysis, these data were interpreted erroneously

Fig. 6 (continued) has been rescaled here, to best overlie the *blue trace*. Note that the gene *SLC6A8* has high percentage GC unlike *MECP2*. Note also the close fit of the two smoothed means, indicating that most of the noise within this CMA data is percentage GC-dependent. *Lower panel*: X-axis: ~300 of the same Agilent array probes, from approximately the same region of Xq28, with positions of two genes indicated (not to map scale). Y-axis: Scale indicates Agilent *Probescores*, as designed into the algorithm ADM2, to correct for GC bias across the genome. Probes with high %GC have compensatory low *Probescores*, designed bioinformatically as weighting coefficients, to discount the undue influence of high GC regions and so reduce false CNV calls [29]. The very extreme weightings shown for *SLC6A8* reflect its highly problematic nature. **(b)** X-axis: ~3,000 consecutive *Agilent* array probes are listed in genome order only, as for **(a)**. Probes for the distal ~50 Mb of the Xq arm are shown, with Xqter to the right. The Y-axis (\log_2 ratios) scores the CMA data for an unbalanced translocation carrier, with der(3)t(X;3)(q27.1;q29), and duplication for the distal 16 Mb of Xq (*yellow arrow* indicates inferred breakpoint). For this CMA test, QC was acceptable (DLR = 0.20). As expected for a male, the duplicated X region should show a \log_2 ratio of approx. +1.0, for data to the *right of the yellow arrow*. Unexpectedly, however, at the point indicated by the *black arrows*, this value drops to around +0.35, then rising gradually back toward +1

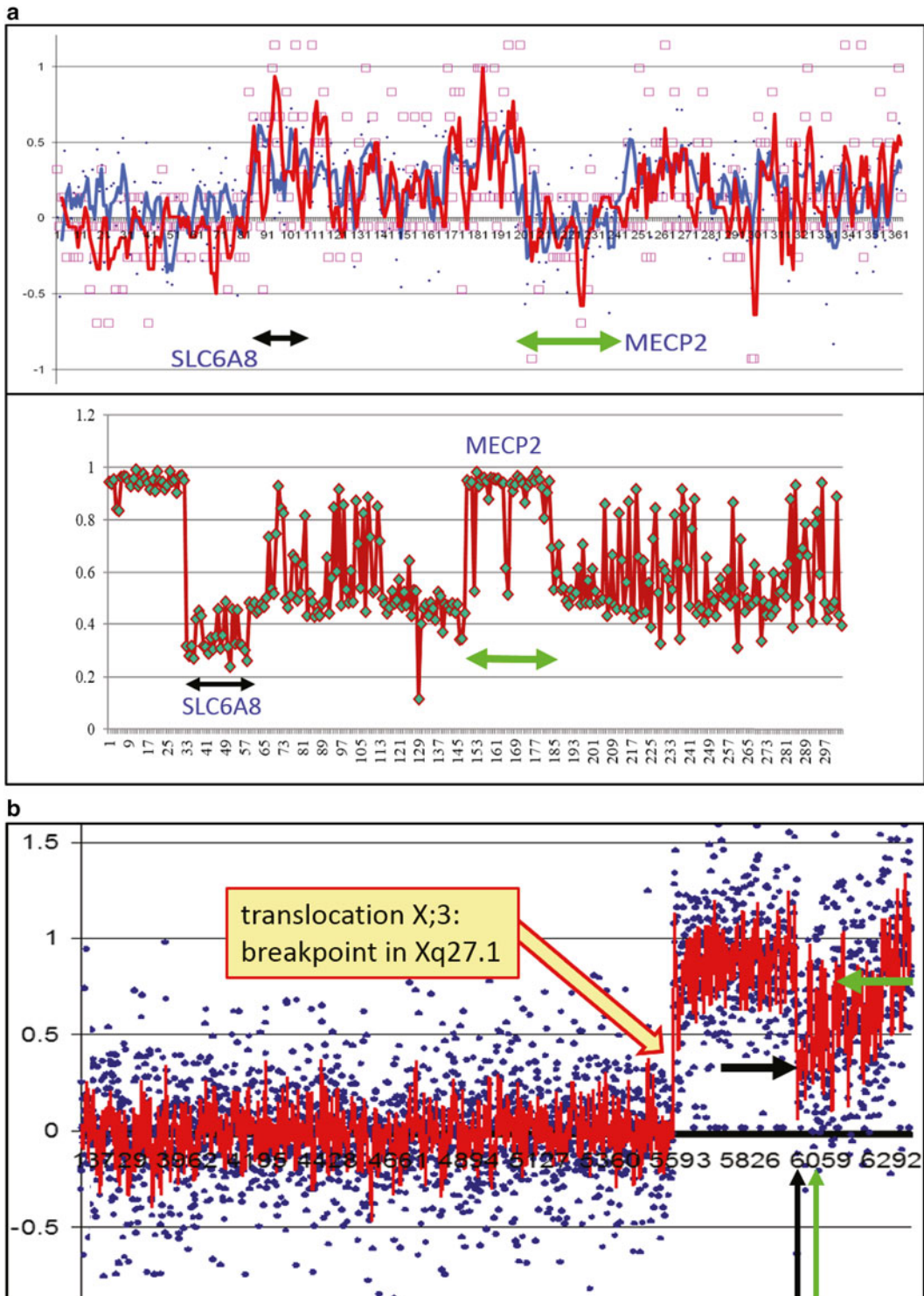


Fig. 6 (a) Effects of percentage GC variation in band Xq28: *Upper panel*: X-axis: 370 consecutive Agilent array probes from a 4.2 Mb region within band Xq28, listed in genomic order only (data plotted here with MS Excel, and not scaled to genomic map position). Probes within the two genes *SLC6A8* and *MECP2* are indicated by the black and green arrows. Y-axis: scale shows the log₂ ratio data for the blue dots and (and the blue smoothed mean) indicating CMA data from an Agilent 60K test, with marginal DLR of 0.24. The pink squares (and red line=smoothed mean) show, in relative terms, the percentage GC contents for the 370 probes. The red trace

as indicating a complex rearrangement, or an interstitial, rather than terminal Xq duplication. However, subsequent karyotyping revealed an unbalanced X;3 translocation. The vertical and horizontal arrows of Fig. 6b show respectively the probe location and mean \log_2 ratios of the *SLC6A8* (black) and *MECP2* (green) array probes. Consistent with Fig. 6a, the *SLC6A8* probes perform poorly. But while they often overestimate copy number in the normal case (*see* Fig. 6a), they are underestimating it here. Such error reversals are not unknown for other scenarios involving GC-dependent noise [26]. Underlying this may be a probe performance overly sensitive to DNA sample concentration or the critical matters of dynamic range and target saturation, which may be more problematic for some high GC ranges [16].

7. If a proband's duplication or deletion CNV appears de novo on parental CMA, further follow-up testing may be ordered for the parents and/or proband, in the form of a FISH test. This may be necessary to exclude the very rare case where the parents are balanced for the proband's CNV, but one parent has one of their two copies (or alleles) transposed or translocated to elsewhere in the genome (usually to a different linkage group). In such cases, meiotic segregation or recombination can result in the proband (or siblings) inheriting 2 or 0 copies from this one parent, plus 1 copy from the other. FISH testing can detect the ex situ allelic copy in the balanced-carrier parent, whereas CMA will not. The same FISH follow-up may be applied to the proband, but this is useful for duplications only. Note also that due to technical limitations, standard FISH testing may not be feasible for CNVs of size <100 kb. Finally, this type of follow-up may be unnecessary for those well-known (curated) CNV loci where de novo reports are common, and the recurrent mutation can be expected to arise in situ. This is true for many examples of non-allelic homologous recombination [14, 44], where a pair of commonly recombining non-allelic, homologous segments map to either flank of the CNV's usual locus. This is true for Case 1's pathogenic deletion in 16p11.2 (Fig. 1c: note flanking segmental duplications at bottom of figure).
8. The B allele frequency (BAF) is calculated according to the proportion of B alleles for a given SNP, using the formula $B/(A+B)$. Thus, the genotypes AA, AB, and BB will produce expected BAF values of 0.0 ($0/2+0$), 0.5 ($1/1+1$), and 1.0 ($2/0+2$), respectively, while the Log R intensity ratio (Log2 ratio) will be 0.0, representing two copies (just as it does for CGH arrays). In comparison to copy number analysis, the statistical power inherent in tests for the three possible genotype frequencies is relatively limited. Unlike the case with copy

number per se, there is no definable “abnormal” state for the SNP genotype at any one locus. Although one allele will always be more frequent than the other, any of the three genotypic outcomes is “normal,” and none is particularly rare. Additionally, as is consistent with Hardy—Weinberg equilibrium (and $BAF < 0.5$) the majority of SNP loci will be homozygous, even in the normal (=control) state, and so some stretches of homozygosity are always present due to chance alone. Resolution for LOH is thus much poorer than resolution for CNVs, and LOH regions of < 5 Mb may not be reliably detectable, and even if detected, they may be of no clinical or biological significance. A resolution of 5 Mb is adequate for clinical analysis, and the array software is usually set to detect LOH > 3 – 10 Mb [28]. The *Illumina HumanCytoSNP-12* array, with higher density probe coverage (10 kb mean spacing), high SNP heterozygosity (30–40%), and accurate, reproducible call rates ($> 99\%$) can detect LOH reliably at 2–3 Mb.

The caveat above applies only to LOH testing in a (presumed) balanced genome. But, when testing for copy number *imbalance*, high density SNP-based arrays are advantageous because the Log R and allelic ratios can be used as controls for each other. In particular, the genotyping data produced by heterozygous (AAB or ABB) duplication calls provides powerful evidence that an increase in Log R intensity ratio represents a true copy number gain. Likewise, deletions resulting in monosomy should always be associated with a loss of heterozygous SNP calls. In effect, these mutual corroborations increase resolution, above what it would be for Log R intensity alone.

9. The use of SNP-based arrays also identifies excessive homozygosity as caused by parental relatedness [28, 63]. The percentage of the scoreable genome affected by LCSH can be estimated by summing the lengths (Mb) of all autosomal LCSH segments of size > 3 – 5 Mb, and dividing this by the total length of the euchromatic autosomal genome (approximately 2,881 Mb, per GRCh37/hg19). The American College of Medical Genetics has recently released standards and guidelines for documenting suspected consanguinity as an incidental finding of genomic testing [64].
10. Clinical laboratories generally set their microarray software filters to detect LCSH of lengths > 5 Mb, which is above the size of autozygous segments seen in demonstrably outbred populations [65]. Segments of LCSH associated with UPD are often large (> 10 Mb) and will be confined to a single chromosome [51, 66]. Even so, in a routine setting, SNP-based arrays will not identify every case of UPD, but only those with sufficiently large segments of isodisomy to raise suspicion for further investigation.

We have seen several instances of PWS associated with maternal uniparental heterodisomy with no evidence for isodisomy (unpublished data). Similar findings have been reported by others [66]. Such cases demonstrate a normal chromosome 15 profile, and hence do not raise suspicion for UPD. For Case 4, poor DNA quality was a confounding factor (Fig. 4). Thus, it was fortunate the isodisomic regions were large, as smaller CNVs would not have been detected. Had we found no abnormality this case would not be reported as a straightforward normal result. Depending on the quality of the array data itself, one might issue a qualified report, stating that resolution was limited, and hence certain (listed) abnormalities could not be excluded. Otherwise the report would state: *No result available due to poor QC, likely related to poor DNA quality*. Although a clinically significant CMA was reported, it is not possible to conclude that this finding explains the fetal death in this case. Prader–Willi syndrome is usually detected during infancy or childhood, and hence the present finding may be incidental to the fetal death in utero.

Acknowledgments

We thank the laboratory staff at VCGS, Melbourne; the laboratory staff of *Sydney Genome Diagnostics* at Children’s Hospital Westmead (CHW) including A. Darmanian, D. Hung, L. St. Heaps, D. Wright, and Assoc Prof. B. Bennetts; other staff at CHW: Department of Clin. Genetics, including Dr F. Collins, Dr. M. Wilson, Dr. R. Jamieson, and Prof. D. Sillence; Prof. R. Dale (CHW Department of Neurology); Prof. S. Arbuckle (CHW Anat. Path); and Prof. J. Christodoulou, Head of the Western Sydney Genetics Program. For Case 2: thanks also to Drs M. Chopra and J. Pinner. For Fig. 1c, d we acknowledge Kent W.J., Sugnet C.W., Furey T.S. et al. (2002) The human genome browser at UCSC. *Genome Res.* **12**, 996–1006.

Note added in proof: The International Standards for Cytogenomics Arrays (ISCA) has recently become the International Collaboration for Clinical Genomics (ICCG). At the time of writing, their database is in the process of migration to a new web address: www.iccg.org.

References

1. Miller DT, Adam MP, Aradhya S et al (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital abnormalities. *Am J Hum Genet* **86**:749–764
2. Kearney H, Thorland E, Brown K et al (2011) American College of Medical Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet Med* **13**:680–685
3. Hochstenbach R, Buizer-Voskamp J, Vorstman J et al (2011) Genome arrays for the detection of copy number variations in idiopathic mental retardation, idiopathic generalized epilepsy and neuropsychiatric disorders: lessons for diagnosis

- tic workflow and research. *Cytogenet Genome Res* 135:174–202
4. Blakeslee AF, Belling J, Farnham ME (1920) Chromosomal duplication and Mendelian phenomena in *Datura* mutants. *Science* 52:388–390
 5. Birchler JA, Veitia RA (2012) Gene balance hypothesis: connecting issues of dosage sensitivity, across biological disciplines. *Proc Natl Acad Sci U S A* 109:14746–14753
 6. Warburton D (1991) De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. *Am J Hum Genet* 49:995–1013
 7. Hassold T, Abruzzo M, Adkins K et al (1996) Human aneuploidy: incidence, origin and etiology. *Environ Mol Mutagen* 28:167–175
 8. Stingle S, Stoehr G, Peplowska K et al (2012) Global analysis of genome, transcriptome and proteome reveals the response to aneuploidy in human cells. *Mol Syst Biol* 8:608
 9. Sheltzer J, Amon A (2011) The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends Genet* 27:446–453
 10. Daniel A (1979) Structural differences in reciprocal translocations: potential for a model of risk in *Rcp. Hum Genet* 51:171–182
 11. Poot M, van der Smagt J, Brilstra E et al (2011) Disentangling the myriad genomics of complex disorders, specifically focusing on autism, epilepsy and schizophrenia. *Cytogenet Genome Res* 135:228–240
 12. Moreno-De-Luca A, Myers S, Challman T et al (2013) Developmental brain dysfunction: revival and expansion of old concepts based on new genetic evidence. *Lancet Neurol* 12:406–414
 13. Moreno-De-Luca D, Sanders SJ, Willsey AJ et al (2012) Using large clinical data sets to infer pathogenicity for rare copy number variants in autism cohorts. *Mol Psychiatry* 18:1090–1095
 14. Conrad D, Pinto D, Redon R et al (2010) Origins and functional impact of copy number variation in the human genome. *Nature* 464:704–712
 15. Kidd JM, Cooper GM, Donahue WF et al (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64
 16. Vermesch J, Brady P, Sanlaville D (2012) Genome-wide arrays: quality criteria and platforms to be used in routine diagnostics. *Hum Mutat* 33:906–915
 17. Cooper GM, Coe BP, Girirajan S et al (2011) A copy number variation morbidity map of developmental delay. *Nat Genet* 43:838–846
 18. Kaminsky EB, Kaul V, Paschall J et al (2011) An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. *Genet Med* 13:777–784
 19. Ledbetter DH, Riggs ER, Martin CL (2012) Clinical applications of whole-genome chromosomal microarray analysis. In: Ginsburg GS, Willard HF (eds) *Genomic and personalized medicine*, vol. 1, ch 11, 2nd edn. Elsevier, New York, pp 133–144
 20. Gambin T, Stankiewicz P, Sykulski M et al (2013) Functional performance of aCGH designs for clinical cytogenetics. *Comput Biol Med* 43:775–785
 21. Huang N, Lee I, Marcotte E et al (2010) Characterising and predicting haploinsufficiency in the human genome. *PLoS Genet* 6:e1001154
 22. Meader S, Ponting F, Webber C (2012) Prediction of 3551 human haploinsufficient genes. *ASHG 2012 San Francisco*, program # 445F
 23. Mefford HC, Clauin S, Sharp AJ et al (2007) Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *Am J Hum Genet* 81:1057–1069
 24. Vissers LE, de Ligt J, Gilissen C et al (2010) A de novo paradigm for mental retardation. *Nat Genet* 42:1109–1112
 25. Tang Y-C, Amon A (2013) Gene copy-number alterations: a cost-benefit analysis. *Cell* 152:394–405
 26. Scharpf RB, Beaty TH, Schwender H et al (2012) Fast detection of de novo copy number variants from SNP arrays for case-parent trios. *BMC Bioinformatics* 13:330
 27. Mulle J, Patel V, Warren S et al (2010) Empirical evaluation of oligonucleotide probe selection for DNA microarrays. *PLoS One* 5:1–7
 28. Kearney HM, Kearney JB, Conlin LK (2011) Diagnostic implications of excessive homozygosity detected by SNP-based microarrays: consanguinity, uniparental disomy, and recessive single-gene mutations. *Clin Lab Med* 31:595–613, ix
 29. Lipson D, Tsalenko A, Yakhini Z et al (2005) Interval scores for quality annotated CGH data. In: *Proceedings of the genomic signal processing and statistics workshop (GENSIPS 2005)*, [Online] May 2005 (2005-05), XP007906019 Newport, Rhode Island. www.cs.technion.ac.il/~dlipson/abs.html#intervals
 30. Cooper GM, Zerr T, Kidd JM et al (2008) Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet* 40:1199–1203

31. McCarroll SA, Kuruwilla FG, Korn JM et al (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet* 40:1166–1174
32. Peiffer DA, Le JM, Steemers FJ et al (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res* 16:1136–1148
33. Wilson M, Peters G, Bennetts B et al (2008) The clinical phenotype of mosaicism for genome-wide paternal uniparental disomy: two new reports. *Am J Med Genet A* 146:137–148
34. Bruno DL, White SM, Ganesamoorthy D et al (2011) Pathogenic aberrations revealed exclusively by single nucleotide polymorphism (SNP) genotyping data in 5000 samples tested by molecular karyotyping. *J Med Genet* 48:831–839
35. Conlin LK, Thiel BD, Bonnemann CG et al (2010) Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum Mol Genet* 19:1263–1275
36. Wierenga KJ, Jiang Z, Yang AC et al (2013) A clinical evaluation tool for SNP arrays, especially for autosomal recessive conditions in offspring of consanguineous parents. *Genet Med* 15:354–360
37. Mei R (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. *Genome Res* 10:1126–1137
38. O’Keefe C, McDevitt MA, Maciejewski JP (2010) Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood* 115:2731–2739
39. Tuna M, Knuutila S, Mills GB (2009) Uniparental disomy in cancer. *Trends Mol Med* 15:120–128
40. Walter MJ, Payton JE, Ries RE et al (2009) Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci U S A* 106:12950–12955
41. Shaffer LG, McGowen-Jordan J, Schmid M (eds) (2013) *ISCN (2013): an International System for Human Cytogenetic Nomenclature*. S. Karger, Basel
42. Zufferey F, Sherr EH, Beckmann ND et al (2013) A 600 kb deletion syndrome leads to energy imbalance and neuro-psychiatric disorders. *J Med Genet* 49:660–668
43. Eichler EE, Lu F, Shen Y et al (1996) Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. *Hum Mol Genet* 5:899–912
44. Itsara A, Wu H, Smith J et al (2010) De novo rates and selection of large copy number variation. *Genome Res* 20:1469–1481
45. Rosenfeld JA, Coe BP, Eichler EE et al (2012) Estimates of penetrance for recurrent pathogenic copy-number variations. *Genet Med* 15:478–481
46. Walsh KM, Bracken MB (2011) Copy number variation in the dosage-sensitive 16p11.2 interval accounts for only a small proportion of autism incidence: a systematic review and meta-analysis. *Genet Med* 13:377–384
47. Sanders SJ, Ercan-Sencicek AG, Hus V et al (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron* 70:863–885
48. Hayes JL, Tzika A, Thygesen H et al (2013) Diagnosis of copy number variation by Illumina next generation sequencing is comparable in performance to oligonucleotide array comparative genomic hybridisation. *Genomics* 102:174–181
49. Szafranski P, Dharmadhikari A, Brosens E et al (2013) Small noncoding differentially methylated copy-number variants, including lncRNA genes, cause a lethal lung developmental disorder. *Genome Res* 23:23–33
50. Doolittle WF (2013) Is junk DNA bunk? A critique of encode. *Proc Natl Acad Sci U S A* 110:5294–5300
51. Papenhausen P, Schwartz S, Risheg H et al (2011) UPD detection using homozygosity profiling with a SNP genotyping microarray. *Am J Med Genet A* 155A:757–768
52. Hamamy H, Antonarakis SE, Cavalli-Sforza LL et al (2011) Consanguineous marriages, pearls and perils: Geneva international consanguinity workshop report. *Genet Med* 13:841–847
53. Engel E (1980) A new genetic concept: uniparental disomy and its potential effect, isodisomy. *Am J Med Genet* 6:137–143
54. Robinson WP (2000) Mechanisms leading to uniparental disomy and their clinical consequences. *Bioessays* 22:452–459
55. Yamazawa K, Ogata T, Ferguson-Smith AC (2010) Uniparental disomy and human disease: an overview. *Am J Med Genet C Semin Med Genet* 154C:329–334
56. Cassidy SB, Schwartz S, Miller JL et al (2012) Prader-Willi syndrome. *Genet Med* 14:10–26
57. Biesecker LG, Spinner NB (2013) A genomic view of mosaicism and human disease. *Nat Rev Genet* 14:307–320
58. Jinawath N, Zambrano R, Wohler E et al (2011) Mosaic trisomy 13: understanding origin using SNP array. *J Med Genet* 48:323–326

59. Green RC, Berg JS, Grody WW et al (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 15:565–574
60. Wolf SM, Annis GJ, Elias S (2013) Point-counterpoint. Patient autonomy and incidental findings in clinical genomics. *Science* 340:1049–1050
61. Curtis C, Lynch AG, Dunning MJ et al (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. *BMC Genomics* 10:588. doi:[10.1186/1471-2164-10-588](https://doi.org/10.1186/1471-2164-10-588)
62. De Sario A, Geigl EM, Palmieri G et al (1996) A compositional map of human chromosome band Xq28. *Proc Natl Acad Sci U S A* 93:1298–1302
63. Grote L, Myers M, Lovell A et al (2012) Variability in laboratory reporting practices for regions of homozygosity indicating parental relatedness as identified by SNP microarray testing. *Genet Med* 14:971–976
64. Rehder CW, David KL, Hirsch B et al (2013) American College of Medical Genetics and Genomics: standards and guidelines for documenting suspected consanguinity as an incidental finding of genomic testing. *Genet Med* 15:150–152
65. Kirin M, McQuillan R, Franklin CS et al (2010) Genomic runs of homozygosity record population history and consanguinity. *PLoS One* 5:e13996. doi:[10.1371/journal.pone.0013996](https://doi.org/10.1371/journal.pone.0013996)
66. Tucker T, Schlade-Bartusiak K, Eydoux P et al (2012) Uniparental disomy: can SNP array data be used for diagnosis? *Genet Med* 14: 753–756

Chapter 9

Bioinformatics Approach to Understanding Interacting Pathways in Neuropsychiatric Disorders

Ali Alawieh, Zahraa Sabra, Amaly Nokkari, Atlal El-Assaad, Stefania Mondello, Fadi Zaraket, Bilal Fadlallah, and Firas H. Kobeissy

Abstract

Bioinformatics-based applications have been incorporated into several medical disciplines, including cancer, neuroscience, and recently psychiatry. Both the increasing interest in the molecular aspect of neuropsychiatry and the availability of high-throughput discovery and analysis tools have encouraged the incorporation of bioinformatics and neurosystems biology techniques into psychiatry and neuroscience research. As applied to neuropsychiatry, systems biology involves the acquisition and processing of high-throughput datasets to infer new information. A major component in bioinformatics output is pathway analysis that provides an insight into and prediction of possible underlying pathogenic processes which may help understand disease pathogenesis. In addition, this analysis serves as a tool to identify potential biomarkers implicated in these disorders. In this chapter, we summarize the different tools and algorithms used in pathway analysis along with their applications to the different layers of molecular investigations, from genomics to proteomics.

Key words Algorithms, Bioinformatics, Computational, Data mining, Genomics, Omics, Pathways, Phenotype, Phenomics, Polymorphisms, Proteomics, Psychiatry, Tools, Transcriptomics

Abbreviations

CNVs Copy number variants
GWAS Genome-wide association study
SNPs Single-nucleotide polymorphisms

1 Introduction

After the great success in the sequencing of the first human genome in 2003 [1], the use of bioinformatics tools in biological and medical research has escalated enormously. The main reason behind this increase was the availability of large clinical and molecular datasets that could no longer be handled manually and required high

throughput processing tools [2–4]. The importance of computational applications in biological and clinical research has brought up several collaborative efforts between mathematicians, engineers, statisticians, biologists, and physicians and has boosted the field of bioinformatics. This field aims to provide the fastest and most reliable way to handle, analyze, store, and visualize massive datasets as well as to evaluate their clinical implications.

Bioinformatics has been incorporated in several medical disciplines including recently psychiatry. Psychiatry is a discipline with complex and heterogeneous diseases where computational bioinformatics tools are urgently needed. Until now, the field of *molecular psychiatry* is still in its infancy and insight into the etiology and pathophysiology of psychiatric diseases is poor due to their multifactorial etiology and the influence of multiple environmental factors [5–7]. Similar to molecular psychiatry, psychiatric practices face major challenges involving the lack of objective techniques for disease diagnosis and classification along with the need for personalized medicine due to the unpredictable outcome of disease pharmacology. In this context, bioinformatics has enabled a new direction for understanding pathogenesis in molecular and clinical psychiatry through the ability to comprehend and draw inferences from data on genomics, transcriptomics, proteomics, and phenomics as well as other high-throughput data acquisition techniques [5]. These data inputs have been used in two major interrelated applications: biomarker discovery and pathway analysis. In this chapter, we review the methods and applications of bioinformatics tools in molecular psychiatry with major emphasis on pathway discovery and network analysis.

2 Methods in Pathway Analysis

In addition to the challenges listed above, simple inspection of large datasets with basic visual computational aids and low-order statistical metrics will not extract the full informational potential of the available data, hence the importance of using computational methods for data mining. Data mining concerns information extraction technique from large and huge datasets using mathematical methods [8]. In bioinformatics, data mining is used in applications such as finding keynotes in sequences to provision patterns, finding genome pathways of diseases, and finding cluster rules for DNA and protein sequences [9]. Data mining methodologies and algorithms differ in techniques and goals from one application to another. The following is a list of the most important categories of data mining methodologies with associated algorithms, techniques, and sample applications.

2.1 Mining Methodologies

2.1.1 Data Cleaning, Preprocessing, and Integration

Since we are dealing with large datasets, it is necessary to separate the useful information from the distributed and heterogeneous data [9]. Many algorithms can be used to preprocess the data prior to analysis. Time-series filtering, outlier detection, data cleaning algorithms, and data normalization algorithms are some of the applied mathematical algorithms depending on the bioinformatics applications [8].

The cleaning stage aims at ensuring the quality of raw data by manipulations based on answering some important questions related to the delivery, gathering, and analysis of the data. A microarray processing procedure represents a data quality continuum example [10].

The preprocessing stage involves the use of several concepts including (1) *management*, which covers unifying the data content and format [11, 12], automating the data preprocessing, e.g., microarray spot analysis [13], and publishing the data, e.g., by using online sites like MedLine and optimizing the data quality; (2) *documentation*, which aims at preprocessing the files in a way that allows their smooth usage and storage [14]; and (3) *metrics specification* that should be carefully selected based on the data analysis, in a way that mostly adheres with the studied information [15, 16].

Integration techniques such as unit normalization and statistical aggregation find common representation of related data gathered from different sources. The resulting integrated data forms a normalized input that is less biased than the data in its original form.

2.1.2 Feature Selection/Extraction

Data points can have several dimensions where each dimension, or a subset of dimensions, relates to a specific feature of the data. Often, when the dimensionality of the dataset is high, some of the dimensions turn out to be redundant or irrelevant to the classification problem. A *feature selection algorithm* aims at choosing the minimum number of features that have the highest discrimination power among the available features of a given dataset. Examples of features include hair color, height, and weight. Recursive feature elimination (RFE) and *relief* are examples of feature selection algorithms [17]. In contrast, *feature extraction* builds up synthetic features by combining and aggregating data in existing dimensions that do not necessarily belong to the original set of features.

Feature selection is important because it (1) helps to escape data over-fitting, (2) helps to achieve faster models, and (3) might allow a better understanding of the real parameters that produced the data. Feature selection algorithms can be classified into three classes: (1) filter methods, (2) wrapper methods, and (3) embedded methods (for more details refer to [18]). These feature selection algorithms are used for sequence analysis (content analysis and signal analysis) [19, 20], microarray analysis [21], mass spectra analysis [22], single-nucleotide polymorphism analysis [23, 24], and medical text mining [25].

2.1.3 Machine Learning

Given a certain dataset and possibly relevant expert information, machine learning techniques produce sets of rules that allow judging new data points. For example, such techniques can identify the patterns within the data provided by the high-throughput omics techniques [5]. Machine learning approaches include supervised methods, unsupervised methods, and semi-supervised methods (*see* also Chapter 16).

- *Supervised learning.* Supervised learning methods include algorithms and techniques that use a categorized subset of the dataset to train a computational model. Each point in the training subset is labeled with a known category, also known as a class. The points along with their class labels are passed to the computational model in the training phase where the model modifies its parameters to learn. Then, once the training phase is done, the computational model works as a classifier that takes an input point and returns its class.

Known algorithms belonging to this category include artificial neural networks, Bayesian classifiers, genetic algorithms, support vector machines (SVMs), k-nearest neighbor (KNN)-based classifiers, and others [26–31].

- *Clustering (unsupervised) learning.* When labels and classes of data points are not available or not trusted, techniques such as clustering algorithms exist that group *similar* points together. Clustering algorithms identify a structure that fits the training data points and allow for the use of the structure later to test new data. In the past, hierarchical clustering was firstly used on microarray data to find similar patterns of gene expression [32]. Then several other clustering techniques were used including simple ones such as the k-means algorithm and more advanced techniques such as self-organizing maps (SOM), SVMs, association rules, and general neural networks. GeneSpring and Spotfire use the above-listed algorithms for microarray analysis [9]. These algorithms return clustered data with an estimated accuracy. These clustered data can be imported to network analysis, modeling, simulation, and visualization tools [5].

New approaches using clustering algorithms were recently introduced like biclustering [33] and p-clustering [34]. Both work on microarray data (*see* **Note 1**). Greedy algorithms, spectral biclustering, column reordering, and 0–1 fractional programming provide other approaches to biclustering. In addition, biclustering may also be found in a supervised form depending on the application.

- *Semi-supervised learning.* Semi-supervised methods generally make use of both labeled and unlabeled data in the training process. They are useful in cases where part of the label is known but contains pairwise constraints; i.e., points a and b

are part of the same class. This aims to use all the available information to produce a more accurate model that represents the existing data for evaluation and prediction [8].

2.1.4 Visualizing

Networks or graphs represent a way of visualizing the relations among different components of the system model after applying the suitable algorithms to process the bioinformatics data. These networks include the visualization of metabolic network or pathway, protein network, and genetic or gene regulatory network. Tuning and cross-validation of a system (using k-fold, leave-one-out, and holdout cross validation) are methods used for improving the accuracy of the model.

2.2 Tools and Illustrations

Some tools are generic in the sense that they contain the different mining methodologies and allow users to make use of them optimally based on the application. Others can be classified as knowledge discovery in databases (KDD) and model visualization tools and interactive visualization environments for integrating data mining and visualization processes [35]. In addition, Weka (<http://www.cs.waikato.ac.nz/ml/weka>) [36] is one of the tools that contain a collection of machine learning algorithms to pre-process, classify, regress, cluster, and select features and visualize the mined bioinformatics data. Other computational bioinformatics tools that include a combination of some of the above-listed methods are:

- R (<http://www.r-project.org>).
- Cytoscape package [37].
- Octave (<http://www.gnu.org/software/octave/>).
- LibSVM and SVMlite, which are open-source packages used to implement SVM and can be used to distinguish between schizophrenia patients and controls [38].
- Matlab Arsenal: A Matlab toolbox that contains a large number of functions related to data clustering, feature selection, and extraction.

3 Applications in Psychiatric Research

The algorithms and tools summarized above are incorporated into different domains in psychiatric research. These include (1) analysis of gene associations with disease and implication of possible pathways, (2) evaluation of gene regulation and identification of co-regulated proteins, (3) assay of disease-associated protein profile changes along with protein network analysis, and (4) a holistic integration of all the previously enumerated factors with the ultimate phenotypic manifestations of the disease (Fig. 1).

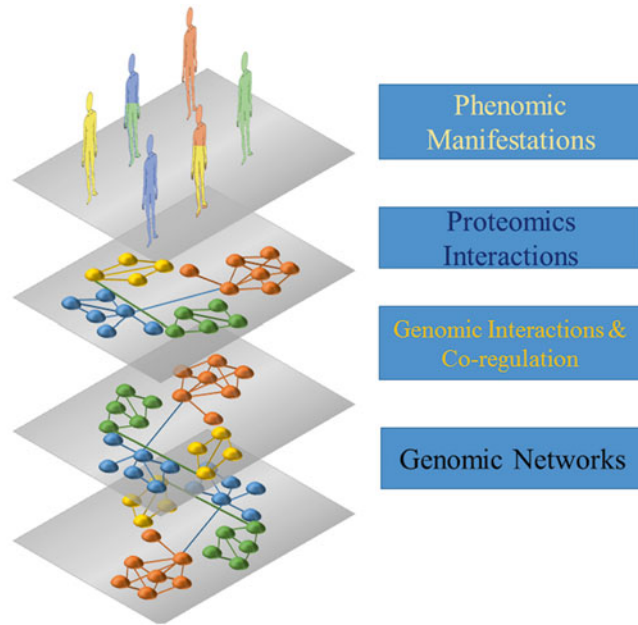


Fig. 1 Systems biology hierarchy. A proposed differential layer of investigation in a systems biology approach: at the very basic later interaction of genes whose polymorphism confers susceptibility to disease is first mapped; then the interaction and co-regulation of the different mRNA transcripts of the genes are mapped and correlated with the third-layer proteomics interactions. This final layer of protein–protein interaction network is considered to be the closest contributor to the phenotypic manifestations and variations

In this section, we illustrate the use of bioinformatics at different levels of pathway and network analysis as relevant to psychiatric disorders. The aim of this application is to investigate possible pathways implicated in the pathogenesis of the disease as well as identify disease-related network regulatory disturbances. Both the implicated pathways and the major nodes in the identified networks serve as putative targets for theragnostics applications.

3.1 Pathway Analysis in Psychiatric Genomics

Genome-wide association studies (GWAS) are large-scale genetic studies involving the study of large numbers of genes or SNPs across extensive populations with a particular phenotype compared to controls. This allows gene discovery associated with disease predisposition, severity, comorbidity, outcome, and relapse. Detected gene variations may be a single nucleotide (single-nucleotide polymorphisms or SNPs) or structural changes in the genome (copy number variants or CNVs). Along with the discovery of implicated SNPs and CNVs, GWAS analysis utilizes several tools and online databases of gene–gene interactions among involved loci [39]. This provides insight into the underlying pathophysiology of the disease.

Several tools are available for the use of investigators in this field such as the R & Bioconductor-associated packages (GWAStools, GenABEL, and others) or SNPranker [40–42]. These tools help analyze, display, clean, and stratify GWAS data and allow for pathway analysis based on online data on gene interactions (*see* Chapters 4 and 5). Special packages have also been dedicated for regional visualization of GWAS results, such as LocusZoom [43], the UCSC Genome Graphs (<http://genome.ucsc.edu/cgi-bin/hgGenome>) [44], and the Integrative Genomics Viewer or IGV (http://www.broadinstitute.org/igv/viewing_gwas). The latter can generate Manhattan plots from different GWAS formats [45]. The increasing use of these tools and their incorporation into psychiatric research have led to the association of different pathways with putative disease pathogenesis. For instance, the GWAS done by the Schizophrenia Psychiatric GWAS Consortium used both PLINK, an R-based GWAS tool, as well as Haploview for haplotype analysis and visualization of results and discovered the association of five new loci with schizophrenia [46]. A major discovery of the study was the implication of an SNP, rs1625579, in the intron of MIR137 gene which is a known regulator of neuronal development and for whom four target loci had also genome-wide significance. This was behind the implication of the MIR137 pathway of neuronal development in the pathogenesis of schizophrenia. Other significant pathways include calcium signaling in bipolar disorder; cholesterol metabolism and the innate immune response in Alzheimer disease; and postsynaptic signaling in schizophrenia and bipolar disorder [47]. A review of CNV studies followed by comprehensive pathway analysis have shown enrichment of specific genes in autism spectrum disorders related to pathways of cellular proliferation, projection and motility, GTPase/Ras signaling, neuronal synaptic complex genes, and ubiquitin degradation genes [48].

The main avenue of pathway analysis in genomics research depends on predetermined knowledge of gene function and gene interaction maps obtainable from online databases like Gene Ontology (<http://www.geneontology.org/>), GeneNet (<http://www.mgs.bionet.nsc.ru/mgs/gnw/genenet/>), and KEGG (<http://www.genome.jp/kegg/>). Assisted with pathway and network prediction tools like Ingenuity Pathway Analysis (IPA), those databases help to discover if disease-associated CNVs or SNPs can be associated with a physiological or a clinical phenotype (Fig. 2). For instance, Greenwood et al. used Collaborative Oncological Gene-environment Study (COGS) SNP chips to associate several SNPs with neurophysiological and neurocognitive endophenotypes in schizophrenia [49]. Results showed 47 SNP–endophenotype associations, and the involved genes were mapped into implicated pathways using IPA. Involved pathways in the pathogenesis of different endophenotypes include neurotransmitter

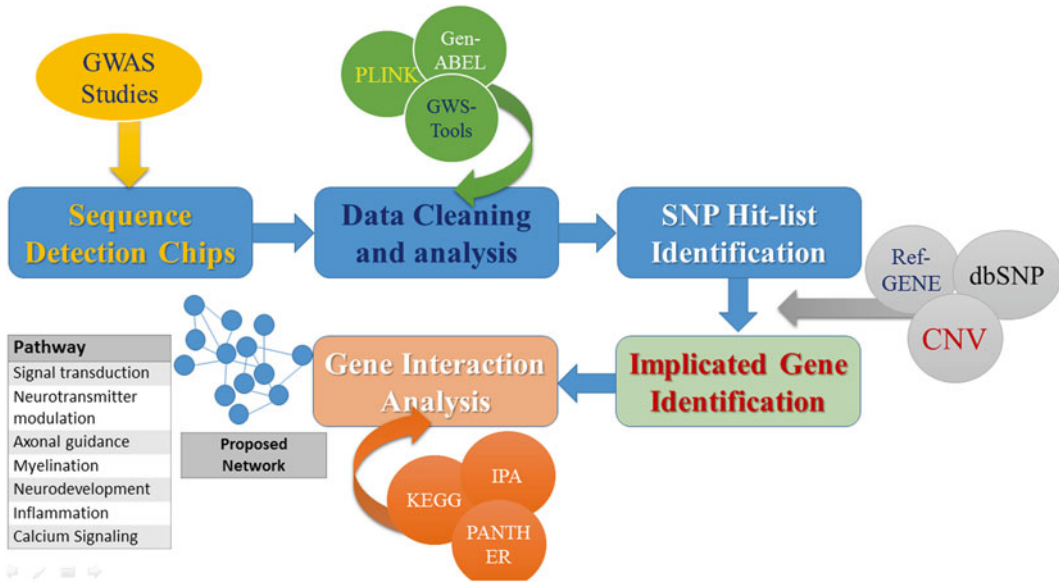


Fig. 2 Flow chart of genome–genome interaction analysis. Huge datasets obtained from genome-wide association studies (GWAS) using sequence detection chips. Then, high-throughput data processing is used to clean and analyze the data using available online databases. An SNP hit list is identified among the obtained datasets, and the relevant genes are recognized. Finally, genes are mapped by pathway and function using online ontology and gene interaction tools to determine pathways and networks associated with the disease

receptor signaling, cell signal transduction, amino acid metabolism, and axonal guidance. Similarly, Ayalew et al. used a translational convergent functional genomics to prioritize genes involved in schizophrenia pathogenesis and identified, using IPA, the involvement of brain development, myelination, cell adhesion, glutamate receptor signaling, G-protein-coupled receptor signaling, and cAMP-mediated signaling in the key pathophysiology of schizophrenia [50].

One major challenge when dealing with large genomics datasets is their high dimensionality, which complicates the identification of patterns and correlates within the data. To handle this problem, the HapMap project used *perfect proxy sets* or *co-sets*, which are sets of directly correlated SNPs, to elucidate interindividual differences allowing for reconstructing cellular interaction networks and their association with functional states [51]. Eventually, this opens a realm of possibilities for disease classification based on variations that lead to the same functional outcome. However, this bottom-up approach continues to have limited applicability in psychiatric research.

3.2 Pathway Analysis in Psychiatric Transcriptomics

In addition to SNPs and CNVs, a better insight into the gene involvement in psychiatric disease pathogenesis comes from the assessment of gene regulation that is measured by mRNA expression

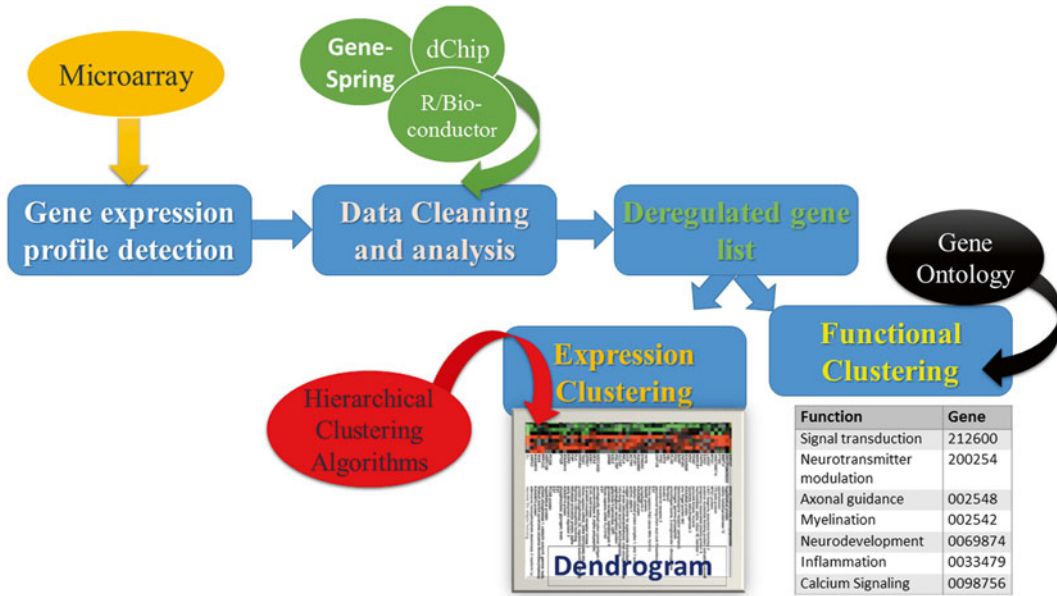


Fig. 3 Transcriptomics process overview. Gene expression profiles are obtained through microarray studies; then data cleaning and analysis with the help of online databases enable a list of dysregulated genes to be obtained. This list can be used to identify the pathways implicated in the disease process through functional clustering or to form a dendrogram of clusters of co-regulated genes

levels, namely, transcriptomics. Similar to genomics, the same clustering and data mining, cleaning, visualization, and prediction algorithms are employed. The goal of using such algorithms is to identify deregulated genes in each condition and the direction of deregulation, in addition to identifying clusters of co-regulated proteins and clusters of functionally related proteins. Eventually, a wealth of information can be obtained from these transcriptomics studies. Data management in transcriptomics follows a general sequence depicted in Fig. 3. First, tools like R-based bioconductor are utilized to process microarray data and obtain the differential expression profile. Then, genes can be clustered by biological function based on gene ontology or other gene-function databases or using tools that allow functional aggregation and network visualization like GenMAPP (<http://www.genmapp.org>). In addition, clustering based on expression levels can also be performed to allow visualization of co-regulated genes using a dendrogram (Fig. 3).

The work of Hakak et al. is used as an illustrative example [52]. The group first performed a microarray analysis comparing post-mortem dorsolateral prefrontal cortices of schizophrenia patients to control data. This yielded a set of 89 deregulated proteins that were mapped into different functions including myelination, plasticity, GABA signaling, signal transduction, and others. Following that, they used a hierarchical clustering algorithm with average

linkage methods to detect clusters of genes with similar expression profile. Results were displayed in a dendrogram and showed that genes involved in myelination are co-regulated, whereas they were found to be downregulated in schizophrenic patients. This suggested demyelination as a possible player in the pathogenesis of the disease. Sequeira et al. further reviewed transcriptional profiling in schizophrenia emphasizing all implicated pathways [53].

In addition, other bioinformatics tools were used for phenotype–transcriptome association studies in which transcriptomics-associated pathways are symmetrically correlated with phenotypic features. An example of this is the study done by Gormanns et al. on both depression and anxiety transcriptomes [54]. Data mining of available experimental microarray data was performed to build the disease transcriptome through extraction of annotations from Gene Expression Omnibus (GEO) repository. Then, MMTx library mapping with manual validation was used to map transcriptomic annotations to Unified Medical Language System (UMLS) concepts of anxiety and depression (*see Note 2*). The GEO datasets where the diseased and controls are significantly different were considered for further analysis resulting in a gene list with fold changes in expression. The genes were mapped into a protein list using pathway studio and into disease-related pathways using KEGG. Pathways were further enriched, the results showed six datasets matching anxiety-like phenotypes and five matching depression ones, and the relevant pathways were mapped.

In another example, Nakatani and colleagues performed wide genome expression analysis in bipolar disorder using samples from Brodmann’s area [55]. They used false discovery rate algorithms to analyze the microarray datasets while using GeneChip to assess direction of deregulation. Results were further validated by RT-PCR. In total, the study identified 84 differentially regulated genes with their functions and then mapped them using IPA software into three main networks: (1) cell growth and proliferation, (2) cell death, and (3) nervous system development [55].

3.3 Pathway Analysis in Proteomics

Proteomics is one of the most heavily investigated fields because of the work undertaken for biomarker discovery and disease pathway elucidation. In terms of relevance to psychiatry, proteomic studies start from assay of the differential protein expression between disease and control specimen using one of the two main techniques: 2D-differential in-gel electrophoresis (2D-DiGE) or liquid chromatography-mass spectrometric (LC-MS) analysis. Differentially expressed spots are detected, and MS is used to identify the proteins by mining MS databases like X! Tandem (<http://www.thegpm.org/TANDEM/>) or using tools like protein prospector (<http://prospector.ucsf.edu/prospector/mshome.htm>). Identified proteins are assigned into functional classes and interaction networks using several tools including PRotein Ontology (PRO—<http://www.obofoundry.org/cgi-bin/detail.cgi?id=protein>), IPA, MPPI (<http://mips.gsf.de/proj/ppi/>),

Reactome (<http://www.reactome.org/>), and others. Examples of such applications in psychiatric disorders are numerous. In the following, we highlight two studies using this approach.

Martins-de-Souza et al. used a nano-LC coupled to shotgun proteomics approach to detect differential expression of proteins in the prefrontal cortex of schizophrenic patients compared to controls. Acquired lists from the MS spectra were sent to the BioTools software package for analysis and searched against the NCBI database. The identified proteins were then clustered into groups based on biological functions using the human protein reference database (HPRD—<http://www.hprd.org>). Deregulated pathways subsequently identified included signal transduction, cell communication, cell growth maintenance, and energy metabolism [56].

In a similar approach, the proteomic profile of a mouse model was studied looking at the anxiety phenotype [57]. Proteins isolated from the synaptosome of mice cingulate cortices were detected by LC-tandem MS (LC-MS/MS). MS data were searched against International Protein Index mouse database using the BioWorks and SEQUEST software packages, KEGG was used to identify overrepresented pathways, and statistical pathway analysis was performed in R and visualized by Pathway Studio software. Among the different pathways, oxidative phosphorylation, metabolic processes, and fatty acid metabolism were overrepresented [57]. The proteomic findings were also correlated with the molecular processes involved in oxidative stress in a comparative approach described in the following section.

3.4 Comparative Omics Approaches

Comparative omics is an integrative approach that links high-throughput data from several layers of biological organizations, most commonly transcriptomics, proteomics, and metabolomics. Using this strategy, changes in regulatory pathways at each level are correlated with relevant changes at a different level in an investigation that aims to understand how different components of the central dogma contribute to the overall disease manifestation. This also serves as a means to validate findings at one level by demonstrating the corresponding interactive pathway at different levels. To illustrate this approach, we use the study done by Prbakaren et al. demonstrating the involvement of mitochondrial dysfunction and oxidative stress in the pathogenesis of schizophrenia [58]. Similar approaches to the ones previously described were used to analyze and interpret proteomics, transcriptomics, and metabolomics data. Results showed that half of the altered proteins obtained from prefrontal cortex samples were involved in oxidative stress and mitochondrial dysfunction. This protein alteration was at the same time associated with both cluster analysis of transcriptomic data showing similar association and relevant molecular metabolomics changes. Similarly, Filiou et al. used the same strategy to study an animal model of anxiety [57].

4 Conclusions

The understanding of interacting pathways in neuropsychiatric disorders is facilitated and made more accurate with the use of bioinformatics tools. A first step in the study of such disorders is the extraction and mathematical analysis of the massive amount of data engendered, a method referred to as *data mining*. Data mining methodologies follow a certain pattern, starting with a data cleaning step where the raw data are validated. This is followed by a data preprocessing step where the data are unified, automated, published, optimized, documented, and metrics specified. Finally, the integration step consists in normalizing the data, making them less biased than the original sets. In parallel, feature selection and extraction algorithms enable the elimination of any redundant or irrelevant feature of the disease and aggregate the data into different dimensions compared to the starting ones, respectively.

Different machine learning approaches can also be used. The *supervised learning approach* matches each data point to its corresponding class or category. The *unsupervised/clustering learning approach* classifies *similar* points that could not be individually categorized. *Semi-supervised learning approach* makes use of labeled and unlabeled data so that points falling under the same category but with pairwise constraints become apparent. However, all these methodologies and algorithms tactics should be assembled into graphs and networks to visualize better the relation among the different components of the system model. Most common networks include metabolic network or pathway, protein network, and genetic or gene regulatory network. In addition, some tools referred to as *generic* tools enfold data mining and visualization tools altogether.

Most importantly, these bioinformatics tools allow us to discover possible pathways implicated in the pathogenesis of psychiatric disorders and to identify disease-cross-related networks. Indeed, molecular and clinical psychiatry can be unveiled via bioinformatics tools that make use of data on psychiatric genomics, transcriptomics, proteomics, and comparative omics approaches. In *psychiatric genomics*, large-scale genetic analyses provided by GWAS along with disease-involved SNPs and CNVs give insight into pathways involving gene–gene interactions and so help in the understanding of the pathophysiology of the disease. In particular, LocusZoom, UCSC Genome Graphs, and IGV are dedicated for regional visualization of GWAS results. In addition, knowledge of gene function and gene interaction maps is obtained from online databases such as Gene Ontology, GeneNet, and KEGG. IPA is another bioinformatics tool that helps discover if disease-associated CNVs or SNPs can be associated with physiological or clinical psychiatric phenotypes. Still another promising tool to analyze large genomics data is the HapMap project. This uses *perfect proxy sets* or co-sets that are made of directly correlated SNPs, so that it becomes feasible to reconstruct cellular interaction networks and associate

them with functional states. Although this tool offers the possibility to classify disease variations with same functional outcome, its application remains limited in psychiatric research.

Psychiatric transcriptomics allows the assessment of gene regulation by measuring mRNA expression levels. Interestingly, gene identification in this category is done in a systematic way. First microarray data are processed via an R-based bioconductor. Second, genes are clustered by biological function via gene ontology databases and networks can then be visualized through GenMAPP. Specifically, a dendrogram can be used to visualize co-regulated genes. Still under the field of psychiatric transcriptomics, phenome–transcriptome association studies can be performed. In these studies, transcriptomics pathways are symmetrically correlated with phenotypic features of psychiatric disorders.

The applications of *psychiatric proteomics* are many, which is not surprising since proteomics provides a core approach for biomarker discovery and disease pathway elucidation. Psychiatric proteomics studies all follow a common trend, starting with the assay of differential protein expression between disease and control specimen. This is done using either the 2D-DiGE or the LC-MS technique. Once the differentially expressed spots/proteins are detected and identified by MS, they are assigned into functional classes and interaction networks using several tools including PRO, IPA, MPPI, and Reactome. Then they can be visualized through Pathway Studio software.

Finally, *psychiatric comparative omics* approaches link several layers of biological organizations, most commonly transcriptomics, proteomics, and metabolomics. Noteworthy, this approach serves to validate the findings of one level by demonstrating the occurrence of similar pathways at different levels. Comparative omics are also necessary to understand how different components of the central dogma contribute to the overall disease manifestation.

To sum up, the analysis of psychiatric disorders such as schizophrenia, autism, Alzheimer disease, and bipolar disorders among others is feasible and provides accurate results when applying bioinformatics tools. These novel approaches have the potential to allow a better pathway analysis and visualization of the pathophysiology of psychiatric disorders and, therefore, to help unravel and exploit the existing heterogeneity in biology of such conditions, potentially leading to a more effective targeted clinical interventions.

5 Notes

1. In the following are listed examples where clustering methods have used microarray gene expression data to cluster genes enabling each entry in the gene expression dataset to represent a gene feature:
 - (a) Row clustering was used to cluster genes having similar fluctuating behavior in all conditions [59]. This application

results in finding the tightly co-regulated genes and missing the weakly regulated ones.

- (b) Biclust is a method used to measure the similarity among a subset of genes and conditions. In brief, if the squared mean residue error value of a matrix is less than a threshold, then the set of genes is considered to be a biclust [60].
 - (c) Since biclust uses squared mean residue error, some gene sets having high similarity may also have high error value. The work in *P-clusters* solves this problem. It requires the following additional constraint to form a cluster. For any 2×2 submatrix, given two genes, and two conditions, respectively, x_{11} , x_{12} , y_{11} , and y_{12} , the following constraint should be satisfied: $|(x_{11} - x_{12}) - (y_{11} - y_{12})| \leq \delta$ where δ is a threshold value [34].
2. MetaMap Transfer (MMTx) is a tool used to identify the concepts, synonyms, and multi-word terms present in the Unified Medical Language System (UMLS). It parses the natural language format text into a list of computable concepts and their semantic types.

References

1. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature* 431:931–945
2. Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2000) GenBank. *Nucleic Acids Res* 28:15–18
3. Louie B, Mork P, Martin-Sanchez F et al (2007) Data integration and genomic medicine. *J Biomed Inform* 40:5–16
4. Luscombe NM, Greenbaum D, Gerstein M (2001) What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med* 40:346–358
5. Alawieh A, Zaraket FA, Li JL et al (2012) Systems biology, bioinformatics, and biomarkers in neuropsychiatry. *Front Neurosci* 6:187
6. Li MD (2010) Grand challenges and opportunities for molecular psychiatry research: a perspective. *Front Psychiatry* 1:2
7. Taurines R, Dudley E, Grassl J et al (2011) Proteomic research in psychiatry. *J Psychopharmacol* 25:151–196
8. Tovar D, Cornejo E, Xanthopoulos P et al (2012) Data mining in psychiatric research. *Methods Mol Biol* 829:593–603
9. Wang JT, Zaki MJ, Hannu TT et al (2005) Data mining in bioinformatics. In: Jain L, Wu X (eds) *Introduction to data mining in bioinformatics*. Springer, London, pp 3–8
10. Holloway AJ, van Laar RK, Tothill RW et al (2002) Options available—from start to finish—for obtaining data from DNA microarrays II. *Nat Genet* 32:481–489
11. Brazma A, Hingamp P, Quackenbush J et al (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat Genet* 29:365–371
12. Spellman PT, Miller M, Stewart J et al (2002) Design and implementation of microarray gene expression markup language (MAGE-ML). *Genome Biol* 3:Research0046
13. Bajcsy P (2004) Gridline: automatic grid alignment DNA microarray scans. *IEEE Trans Image Process* 13:15–25
14. Yandell MD, Majoros WH (2002) Genomics and natural language processing. *Nat Rev Genet* 3:601–610
15. Karr AF (2006) Exploratory data mining and data cleaning. *J Am Stat Assoc* 101:399
16. Fadlallah BH, Seth S, Keil A et al (2011) Robust EEG preprocessing for dependence-based condition discrimination. *Conf Proc IEEE Eng Med Biol Soc* 2011:1407–1410
17. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *JMLR* 3: 1157–1182
18. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics* 23:2507–2517

19. Liu H, Han H, Li J et al (2004) Using amino acid patterns to accurately predict translation initiation sites. *In Silico Biol* 4:255–269
20. Saeyns Y, Degroove S, Aeyels D et al (2004) Feature selection for splice site prediction: a new method using EDA-based feature ranking. *BMC Bioinformatics* 5:64
21. Ma S, Huang J (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics* 21:4356–4362
22. Resson HW, Varghese RS, Drake SK et al (2007) Peak selection from MALDI-TOF mass spectra using ant colony optimization. *Bioinformatics* 23:619–626
23. He J, Zelikovsky A (2006) MLR-tagging: informative SNP selection for unphased genotypes based on multiple linear regression. *Bioinformatics* 22:2558–2561
24. Wang Y, Makedon F, Pearlman J (2006) Tumor classification based on DNA copy number aberrations determined using SNP arrays. *Oncol Rep* 15:1057–1059
25. Han B, Obradovic Z, Hu ZZ et al (2006) Substring selection for biomedical document classification. *Bioinformatics* 22:2136–2142
26. Bishop CM (2006) *Pattern recognition and machine learning*. Springer, New York
27. Han J, Kamber M (2011) *Data mining: concepts and techniques*, 3rd edn. Morgan Kaufmann, San Francisco, CA
28. Hastie T, Tibshirani R, Friedman J et al (2005) *The elements of statistical learning: data mining, inference and prediction*. Math Intel 27(2):83–85
29. Lopresti D, Tomkins A (1997) Block edit models for approximate string matching. *Theor Comput Sci* 181:159–179
30. Mount DW (2004) *Bioinformatics: sequence and genome analysis*, 2nd edn. Cold Spring Harbour Laboratory Press, Cold Spring Harbour, NY
31. Witten IH, Frank E (2005) *Data mining: practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, CA
32. Eisen MB, Spellman PT, Brown PO et al (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95:14863–14868
33. Yang J, Wang W, Wang H et al (2002) δ -clusters: capturing subspace correlation in a large data set. *Data engineering, 2002*. In: *Proceedings 18th international conference, IEEE*, pp 517–528
34. Wang H, Wang W, Yang J et al (2002) Clustering by pattern similarity in large data sets. In: *Proceedings of the 2002 ACM SIGMOD international conference on management of data*, pp 394–405
35. Fayyad U, Wierse A, Grinstein G (2002) *Information visualization in data mining and knowledge discovery*. Morgan Kaufmann, San Francisco, CA
36. Frank E, Hall M, Trigg L et al (2004) *Data mining in bioinformatics using Weka*. *Bioinformatics* 20:2479–2481
37. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504
38. Diederich J, Al-Ajmi A, Yellowlees P (2007) E_x-ray: data mining and mental health. *Appl Soft Comput* 7:923–928
39. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26:445–455
40. Aulchenko YS, Ripke S, Isaacs A et al (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23:1294–1296
41. Gogarten SM, Bhargale T, Conomos MP et al (2012) GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 28:3329–3331
42. Merelli I, Calabria A, Cozzi P et al (2013) SNPranker 2.0: a gene-centric data mining tool for diseases associated SNP prioritization in GWAS. *BMC Bioinformatics* 14(Suppl 1):S9
43. Pruim RJ, Welch RP, Sanna S et al (2010) LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* 26:2336–2337
44. Meyer LR, Zweig AS, Hinrichs AS et al (2013) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* 41(Database issue):D64–D69
45. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26
46. Schizophrenia Psychiatric Genome-Wide Association Study (GWAS) Consortium (2011) Genome-wide association study identifies five new schizophrenia loci. *Nat Genet* 43:969–976
47. Collins AL, Sullivan PF (2013) Genome-wide association studies in psychiatry: what have we learned? *Br J Psychiatry* 202:1–4
48. Marshall CR, Scherer SW (2012) Detection and characterization of copy number variation in autism spectrum disorder. *Methods Mol Biol* 838:115–135

49. Greenwood TA, Lazzeroni LC, Murray SS (2011) Analysis of 94 candidate genes and 12 endophenotypes for schizophrenia from the Consortium on the Genetics of Schizophrenia. *Am J Psychiatry* 168:930–946
50. Ayalew M, Le-Niculescu H, Levey DF et al (2012) Convergent functional genomics of schizophrenia: from comprehensive understanding to genetic risk prediction. *Mol Psychiatry* 17:887–905
51. Jamshidi N, Palsson BO (2006) Systems biology of SNPs. *Mol Syst Biol* 2:38
52. Hakak Y, Walker JR, Li C et al (2001) Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia. *Proc Natl Acad Sci U S A* 98:4746–4751
53. Sequeira PA, Martin MV, Vawter MP (2012) The first decade and beyond of transcriptional profiling in schizophrenia. *Neurobiol Dis* 45:23–36
54. Gormanns P, Mueller NS, Ditzen C et al (2011) Phenome-transcriptome correlation unravels anxiety and depression related pathways. *J Psychiatr Res* 45:973–979
55. Nakatani N, Hattori E, Ohnishi T et al (2006) Genome-wide expression analysis detects eight genes with robust alterations specific to bipolar I disorder: relevance to neuronal network perturbation. *Hum Mol Genet* 15:1949–1962
56. Martins-de-Souza D, Gattaz WF, Schmitt A et al (2009) Prefrontal cortex shotgun proteome analysis reveals altered calcium homeostasis and immune system imbalance in schizophrenia. *Eur Arch Psychiatry Clin Neurosci* 259:151–163
57. Filiou MD, Zhang Y, Teplytska L et al (2011) Proteomics and metabolomics analysis of a trait anxiety mouse model reveals divergent mitochondrial pathways. *Biol Psychiatry* 70:1074–1082
58. Prabakaran S, Swatton JE, Ryan MM et al (2004) Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress. *Mol Psychiatry* 9:684–697, 643
59. Hartigan JA (1972) Direct clustering of a data matrix. *J Am Stat Assoc* 67:123–129
60. Cheng Y, Church GM (2000) Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol* 8:93–103

Chapter 10

Pathogen Genome Bioinformatics

Vitali Sintchenko and Michael P.V. Roper

Abstract

Recent advances in DNA sequencing technology have made the whole-genome sequencing of pathogens in a clinically relevant turn-around time both technically and economically feasible. The DNA sequencing of pathogens with epidemic potential offers new and exciting opportunities for high-resolution public health surveillance. This chapter outlines major methods and bioinformatics tools for pathogen genome characterization, the identification of infectious disease clusters, as well as for genomics-guided biosurveillance. Existing challenges are also considered.

Key words Disease clusters, Infectious disease bioinformatics, Microbial genomics, Public health surveillance

Abbreviations

ML	Maximum likelihood
NGS	Next-generation sequencing
PCA	Principal component analysis
SNPs	Single nucleotide polymorphisms
WGS	Whole-genome sequencing

1 Introduction

The number of microbial threats—in the form of newly identified pathogens, infections crossing the species barrier to people, diseases and vectors adapting to new environments, and microorganisms appearing in more virulent forms—has multiplied to an unprecedented degree in recent years. Furthermore, the epidemiology of well-known infectious diseases has been changing due to the globalization of trade, increased international travel and migration, and in response to immunization campaigns. This evolving epidemiology presents new challenges, both in terms of the understanding and monitoring of determinants of infections, and the

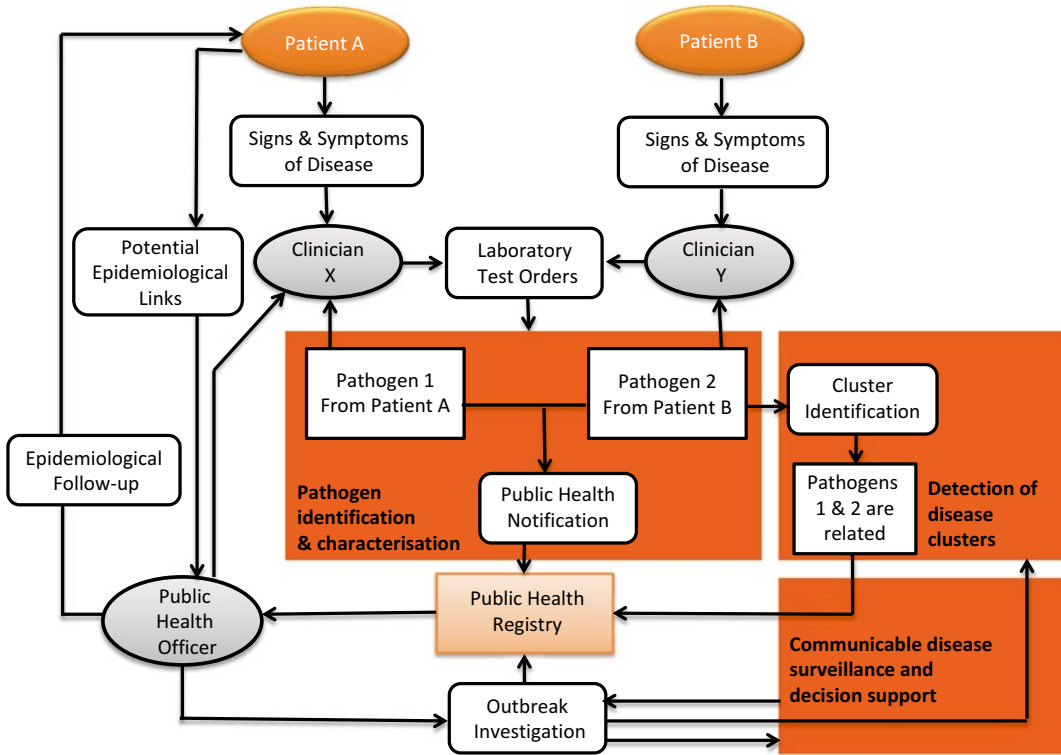


Fig. 1 Summary of three main elements of microbiology laboratory-based biosurveillance. In general, it starts with the initial presentation of patients to different clinicians and the referral of appropriate specimens to diagnostic laboratories

implementation of appropriate prevention measures. There is an urgent need to strengthen existing biosurveillance systems that remain vulnerable to the incomplete and delayed reporting of public health threats.

It is helpful to consider the flow of information between diagnostic laboratories, clinicians, and public health professionals in order to appreciate the added value of bioinformatics approaches to communicable disease management and control. Laboratories, clinicians, and public health officers are requestors and receivers of information and are involved in the analysis of multiple lines of evidence both independently and collectively (Fig. 1). The synthesis of these lines of evidence is enabled and supported by bioinformatics tools. Broadly, these operations can be divided into three consecutive stages: (1) pathogen identification and characterization; (2) detection of disease clusters; and (3) communicable disease monitoring and control (Fig. 1).

We believe that recent advances in DNA sequencing technology have made it technically and economically feasible to complete the whole-genome sequencing (WGS) of pathogens of public health significance in a clinically relevant timeframe [1, 2].

DNA sequencing offers important advantages over other methods of pathogen characterization. First, it provides a nearly universal solution with the potential of high throughput and quality. The process of genome sequencing is essentially the same regardless of the nature of a pathogen, and different microorganisms could be processed simultaneously in a single sequencing run. This means that WGS could allow economies of scale at the regional or national level. Furthermore, a single WGS run has the potential to replace multiple traditional tests carried out on the same isolate at a reference laboratory while providing equivalent or superior quality information [1, 3]. DNA sequences also represent an agnostic and likely *future-proof* data format amenable to exchange between laboratories and to comparison at national and international levels. Finally, the potential utility of WGS for public health surveillance has been supported by the rapid growth of public databases of reference genomes [2, 4]. Not surprisingly, researchers and public health professionals have turned their attention to genomics-guided approaches for biosurveillance [5]. Virologists have pioneered the use of WGS for pathogen characterization, targeting viral genomes small enough for WGS with traditional Sanger sequencing. This chapter, however, is focused on methods and tools for the genomics-guided biosurveillance of bacterial pathogens using DNA sequences.

The recovery of pathogens with epidemic potential requires notification to a public health registry which often leads to an investigation of epidemiological sources. In our scenario, pathogens 1 and 2 appeared to belong to the same species. When evidence suggests potential epidemiological links between patients A and B that are not apparent to different clinicians caring for these patients, an outbreak investigation is initiated and the laboratory is requested to undertake the clustering of pathogens. This assessment may progress to enhanced surveillance of the disease caused by the pathogen.

2 Materials

2.1 Software Tools Commonly Used and Freely Available for Pathogen Identification and Characterization

- Velvet (www.ebi.ac.uk/~zerbino/velvet/) and Spades (<http://bioinf.spbau.ru/spades/>) are de novo assembly programs.
- BWA and samtools (<http://samtools.sourceforge.net/>) are used for the alignment and calling of SNPs and small indels.
- Artemis (www.sanger.ac.uk/resources/software/artemis/) is a free genome browser and annotation tool that allows the visualization of sequence features.
- Prokka (www.vicbioinformatics.com/software/prokka.shtml) is an automatic microbial genome annotation tool.

2.2 Software Packages Commonly Used and Freely Available for Identifying Possible Infectious Disease Clusters by Genomic Comparison

- ACT [Artemis Comparison Tool] (www.sanger.ac.uk/resources/software/act/) is a user-friendly tool for the pairwise comparison of two or more DNA sequences. It can be used to identify and analyze regions of similarity and difference between genomes in the context of the entire sequence and their annotation.
- MEGA (www.megasoftware.net/) is an integrated package that estimates phylogenetic trees by a variety of methods, including Neighbor Joining, Maximum Parsimony, and Maximum Likelihood [6].
- BEAST (<http://beast.bio.ed.ac.uk/>) is a tool for the statistical phylogenetic analysis of molecular sequences in a Bayesian framework. It can be used to estimate phylogenies and test evolutionary hypotheses [7].
- RAXml (www.sfu.ca/biology/staff/dc/raxml/) is a tool for the statistical phylogenetic analysis of molecular sequences in the frequentist framework. It can be used to estimate phylogenies and test evolutionary hypotheses.
- Path-O-Gen (<http://tree.bio.ed.ac.uk/software/pathogen/>) is a program to estimate genome mutation rates by root-to-tip analysis. It can read and analyze contemporaneous trees (where all sequences have been collected at the same time) and dated-tip trees (where sequences have been collected at different times). It is useful for testing the molecular-clock hypothesis. It can also root the tree at the position that is likely to be the most compatible with the molecular clock assumption.
- PathSeq (www.broadinstitute.org/software/pathseq/) can be used for the analysis of the non-host portion of sequencing data. It enables the detection of both known and novel pathogens as well as any resident microflora [8].

2.3 Generic Software Packages Helpful for Analyzing Microbial Genomes

- Galaxy (<http://galaxyproject.org/>) is a free, extensible, open-source, web-based framework that seeks to package together a large number of commonly used informatics tools, including, but not limited to, some of those listed above.
- SplitsTree (www.splitstree.org/) is a tool that is primarily useful for the estimation of phylogenetic networks from molecular sequence data. For this purpose, the program implements the split decomposition, neighbor-net, consensus network, and super-networks methods.
- WEKA (The Waikato Environment for Knowledge Analysis, downloadable from www.cs.waikato.ac.nz/ml/weka/) is a collection of machine learning algorithms for data mining. It also offers tools for data preprocessing, classification, regression, clustering, association rules, and visualization.

3 Methods

The first steps in genomics guided biosurveillance and infectious disease control are laboratory based. DNA samples must be prepared and then sequenced. There are a number of choices of benchtop sequencing platforms, e.g., Illumina's MiSeq and Life Technology's Ion Torrent, each with its relative merits. Presently, these platforms generate libraries of reads (DNA sequences) that contain large numbers (typically $\sim 10^6$) of short (usually ≤ 500 bp) sequences of DNA of variable, but reasonably well-characterized, quality. These sequences can be represented in different file formats (*see Note 1*). The task of assembly has then been likened to accurately reconstructing a published Dickens novel based on millions of copies of overlapping sentence fragments (with spelling errors) possibly with missing pages. These libraries are then subjected to multiple analytical steps to identify and characterize the pathogen (Fig. 2) identify infectious disease clusters, and for genomics-guided biosurveillance (Fig. 1).

Ideally, all genomes would be constructed *de novo*, i.e., without the use of a reference genome to guide the assembly, as this introduces the assumption that the genome of an isolate has, in a sense, already been identified. Of course this is part of the task of genomics guided biosurveillance as we have characterized it. At this stage, however, this is a nontrivial task [9], so, for reasons of prac-

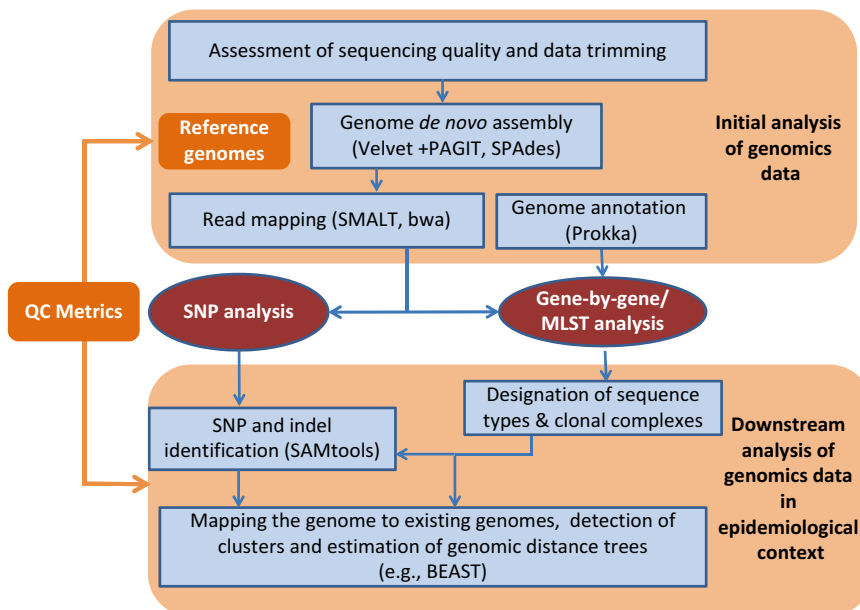


Fig. 2 Pathogen genome analysis workflow (suggested software in *brackets*). The workflow starts with obtaining sequencing data from an analyzer and assessing the data quality

tality, alignment to reference genomes is most commonly used to identify discriminatory features, e.g., SNPs, small indels, or gene alleles, useful for the stated purposes of this chapter [10].

Bioinformatics approaches for genome-wide analyses of pathogens are highly varied across the microbiology community, with an abundance of tools continually being developed, refined, and packaged together as software *pipelines*. There is an urgent need for national and international standards to be established due to the differing results produced by the various tools and the different ways that errors are accounted for at each analytic step.

Parts of the analytic process can be and should be automated—for the purposes of scientific reproducibility and economic efficiency. To do this, laboratories can either purchase pre-packaged commercial *pipelines* or implement in-house solutions, typically using open-source software. The relative merits of these two approaches are frequently argued and are outside the scope of this chapter (*see Note 2* for a brief discussion).

3.1 Pathogen Identification and Characterization

3.1.1 Phenotype-Independent Identification of Pathogens

The advances in WGS have enabled computational pathogen identification in biological samples independently of traditional microbial cultures. For example, sequence-based computational subtraction identifies novel pathogen-derived DNA sequences in infected tissues after subtracting human DNA sequences [8]. The method starts with a subtractive phase in which input reads are subtracted by alignment to human reference sequences. This is followed by an analytic phase in which the remaining reads are aligned to microbial reference sequences and/or assembled *de novo*.

WGS might be the ultimate tool in clinical microbiological typing. Freeware such as SAMtools can identify single nucleotide polymorphisms (SNPs) and some other mutations. The evolutionary history and *relatedness* of isolates can be estimated using tools such as Path-O-Gen (<http://tree.bio.ed.ac.uk/software/pathogen/>) and the other software packages listed in Subheading 2. Emerging evidence suggests that WGS-based identification and characterization of microbial pathogens improves monitoring for emerging clones or new pathogens and the resolution of laboratory-based surveillance [11, 12]. Specifically, this technology enhances the tracing of disease transmission in community and hospital settings through the identification of likely covert clusters as well as through the estimation of transmission events within putative outbreaks and by estimating specific source attribution with hypothesized geographical structure among related isolates. Recent proof-of-concept studies have demonstrated the superiority of WGS to current typing methods [13–15]. However, the vast majority of genomic data are currently medically not actionable because WGS technology is still maturing, and its cost-effectiveness for public health surveillance and laboratory workflow requires further assessment [1].

3.1.2 *Selecting Reference Genomes*

Reference bacterial genomes can be identified by querying available annotated genome sequences in the NCBI GenBank, typically using BLAST alignments and utilizing them to select the most appropriate reference sequence. It is important to note, however, that the genome sequences in the NCBI GenBank are not error free, as is evidenced by the relatively frequent issuing of updated versions for sequences.

3.2 *Detection of Disease Clusters*

The term *cluster* is context specific. For example, it can describe a set of potentially related sequences, or it can refer to patients with a disease grouped together by a common epidemiological link. There are two possible scenarios in which informatics tools are gainfully applied to identify disease clusters. The first is when the initial public health investigation points out an epidemiological link between patients, and the laboratory is requested to ascertain the similarity between microbial isolates obtained from these patients to lend credence to an existing epidemiological hypothesis. Alternatively, the laboratory may retrospectively examine microbial genotypes in order to identify clusters of genomically similar pathogens recovered from patients with a particular infectious disease. The recovery of such microorganisms aids in the identification of disease clusters. It provides quite strong evidence that the patients whose isolates are part of the genotyping cluster should be linked epidemiologically, e.g., share a common source of infection. Such analysis can be conducted prior to epidemiological follow-up. Ideally, genotyping clusters and clusters of patients defined by epidemiological links should coincide and overlap in time and space. However, this is not necessarily the case and certain assumptions must be satisfied for this to hold. The following subsections outline the main concepts and methods of assessing the similarity between microbial genomes. There are a number of methods available for determining the extent to which pathogens are genetically related. Arguably, this is best done by estimating a phylogeny from the available sequencing data. Nevertheless, the significant impact of genetic recombination on the evolution of bacteria should not be underestimated. Finally, even under the many simplifying assumptions made, phylogenetic analysis is very computationally intensive.

3.2.1 *Assessing Similarity Between Genes and Genomes*

In order to assess the similarity between genes or genomes, it is customary to place a numerical value on pairs of observations, which satisfies certain simple constraints so that we can sensibly use the term distance. The choice of observational unit is of course important as is the choice of how we measure distance or similarity. As a simplistic but concrete example, let us ask the question how similar two isolates are with respect to antibiotic drug resistance. We will take as our observational units the number of SNPs relative to a given reference genome at two loci of the genomes of two

isolates of a known pathogen as observations. We will let loci 1 be in a gene that codes for antibiotic drug resistance and loci 2 be in a stable area of the genome. We will denote these observations by x_1 and x_2 and let $x_1 = (0,0)$ and $x_2 = (1,1)$ where $(0,0)$ means that observation 1 has no SNPs at loci 1 and no SNPs at loci 2, and similarly for observation 2. For the question at hand, x_1 and x_2 should be considered to be identical and therefore have distance 0. Now, the most commonly used measure of distance is Euclidean distance. In our example, we have that the Euclidean distance between x_1 and x_2 is $\sqrt{(0-1)^2 + (0-1)^2} = \sqrt{2}$. Note that this distance gives equal weight (*significance*) to each component of each observation, and, as in the case in point, this may be inappropriate. Another frequently used measure of distance is the *city-block* or Manhattan distance. In our example, this is $|0-1| + |0-1| = 2$. Observe that the distances give different answers to our question both of which are clearly incorrect. This simple example serves to illustrate that the choice of units and distance function should be made in a way that is appropriate to the question at hand.

3.2.2 Clustering of Pathogens Using Similarity Metrics

There are various methods of cluster analysis. A popular technique is agglomerative hierarchical clustering, which we will briefly sketch out here. Such methods do not require the number of clusters to be determined in advance and employ various measures of distance between clusters, for example, minimum distance, maximum distance, or distance between centroids. These methods initially treat each single observation as a distinct cluster. The clusters are then aggregated during subsequent iterations, deriving clusters of increasingly larger size. The algorithm is stopped when a single cluster including all the observations is reached. The mergers can be graphically represented as a tree, indicating on one axis the distance between clusters (Fig. 3a, b) and on the other axis the terminal nodes of the (weighted) tree (a tree is a simple structure composed of nodes and branches with additional constraints that make it a special case of the more general object known as a graph, *see Note 3*). The interpretation of the nodes and branch lengths is context dependent. In the case of phylogenetic trees, the interpretation of internal nodes is typically of hypothetical ancestors, and the branch lengths represent an estimate of time.

The technique of cluster analysis is, however, a general one that can be applied to many problems of interest. Now, the simple example considered above already serves to illustrate that there are subtleties involved here as well. How many (weighted) trees are possible? If we just consider the number of nodes, then there is clearly just one possible tree. However, if we also take into account branch lengths, then under most reasonable models of length there are already an infinite number of trees even in this simple case (*see Note 3*). There are other possible complications that may arise, but we will not

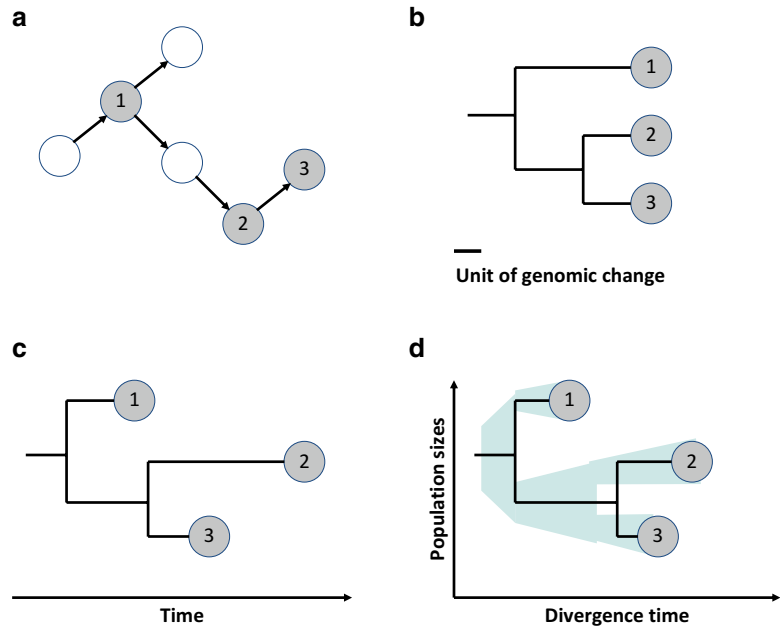


Fig. 3 Possible representation of relationships between pathogens within a putative outbreak. **(a)** A hypothetical set of six isolates from which only three were identified and analyzed (highlighted in *blue* and numbered from 1 to 3). **(b, c)** Estimated phylogenetic tree; panel **(b)** represents a traditional dendrogram and panel **(c)** illustrates a phylodynamic, or heterochronous sampling, view. **(d–a)** Phylogenetic/phylodynamic tree based on time-measured evolutionary history and representing posterior probabilities of divergence time and estimated population sizes (Colour figure online)

discuss them here. As above, the choice of a clustering method should be tailored to the specific question at hand. In addition, cluster analysis and the resulting dendrograms should be clearly distinguished from phylogenetic trees.

3.2.3 Visualizing Clusters as Trees

Rooted trees or dendrograms always provide a hierarchy of clusters. Phylogenetic trees have a particular interpretation. Statistical phylogenetics is based on explicit models of biological evolution which therefore allow for the more sophisticated analyses required to test hypotheses regarding evolutionary relationships between isolates. The evolutionary relationships of gene or protein sequences to their hypothesized ancestral sequences are represented by phylogenetic trees. Phylogenetic trees provide an estimate of the evolutionary relationship between isolates typically including estimates of their divergence times (*see Note 3*). The interior nodes of the tree represent hypothetical ancestors. Trees can be rooted, with a single ancestral organism implied, or unrooted, with no clear origin. For the purposes of the current

chapter, rooted trees are preferred. To produce a rooted tree, a known sequence can be added as an outlier, in order to anchor or root the tree. However, this introduces a number of assumptions that may not always be appropriate.

There has been recent progress in linking methods of phylogenetic analysis with epidemiological modeling. This new area is termed *phylodynamics*. The motivating idea is that if the evolutionary rate of change of a pathogen is rapid, then the contact structure should be able to be inferred from genetic information (Fig. 3). As is always the case in statistics, the sampling scheme and sample size is of great importance. BEAST and associated packages provide a flexible framework for such analyses [7]. Resulting trees can be spatially reconstructed with SPREAD [16]. Such visualization assists in inspection of key evolutionary changes in a geographical context and includes interactive exploration in the time dimension as well as with virtual globe software, e.g., Google Earth [16].

3.2.4 Selecting Methods for Analysis

A variety of algorithms are used to estimate the relationship between sequences from a number of observations. The most popular *distance estimation methods* are UPGMA (Unweighted Pair-Group Method with Arithmetic Mean) and Neighbor Joining. Both convert aligned sequences into a matrix of pair-wise distances and compute branching order and branch lengths. UPGMA has been used in microbial epidemiology, however, several assumptions of its algorithm make it inappropriate. For example, the assumption of a constant rate of evolution is likely to be incorrect. Neighbor Joining does not construct clusters but directly calculates distances to internal nodes. In contrast, *character-based methods* such as Maximum Parsimony, Maximum Likelihood (ML) and Bayesian Inference systematically compare characters within each column in multiple alignments using every data point, not just a distance matrix. In essence, the Maximum Parsimony method looks for the trees with the minimum number of changes. The ML method chooses the tree that maximizes the likelihood of observing the data and provides estimates of the likelihood of the resulting tree or trees (subject to the constraint that it is not always possible to actually maximize this quantity). Bayesian Inference produces a set of trees with posterior probabilities. Neighbor Joining and ML phylogenies are often investigated using MEGA. Node reliability is typically assessed using a technique known as bootstrapping. A variety of principled methods are available for assessing the relative merits of estimated phylogenies. These include the Akaike Information Criteria and Bayes Factors [17].

A significant limitation of these methods is computational cost as they are essentially computationally intractable. Phylogenetic estimation solves this challenge by employing a number of heuristics and other methods of approximation. Therefore, even the best

phylogeny estimations by maximum parsimony cannot guarantee the production of the true optimal solution, even when algorithms are run for a very long time.

3.2.5 Identifying Successful Clones of Pathogens from Phylogenies

The most successful clones of pathogens can be inferred from phylogenetic trees. Methods derived from Principal Component Analysis (PCA) (*see Note 4*) and model-based algorithms are commonly used to identify microbial population structures and to assign isolates to their population of origin. Both methods have limitations in their prior assumptions, in their capacity to handle large-scale genomic data and in interpreting analysis outputs.

Recently, network theory has been successfully applied to examine fine-scale population structures. Network methods subdivide the population into a network of nodes or community structures based on the density of the connections within and between different subgroups, which provides a means to identify and visualize the structure of the entire population. A median-joining network can be constructed with NETWORK (www.fluxus-engineering.com) and used to infer intraspecific phylogenies from SNP-based matrices where small genetic distances are expected.

3.3 Communicable Disease Surveillance and Decision Support

3.3.1 Linking Genotypic and Epidemiological Cluster

Concluding whether or not microorganisms are epidemiologically related remains a challenge. In practice, the relevant information about respective hosts, which may include such variables as age, sex, race, occupation, marital status, previous diseases, and contact structure, is often unavailable. Contact structure is clearly of great importance in determining epidemiological relatedness with respect to infectious diseases. It is important to model contact structure. Modeling of infectious disease dynamics is increasingly being done using social network theory [13]. Analysis of social network models can be much more difficult than of traditional deterministic models. However, this level of complexity is of importance on the short time scales, and in the small population sizes that are often of concern when considering infectious disease clusters.

The clustering of isolates has enabled a shift from a predominantly retrospective confirmation of epidemiological hypotheses to prospective laboratory based surveillance. Such prospective microbial-genomics guided surveillance and monitoring of infectious diseases demands the highest possible resolution and timeliness. Therefore, bioinformatics pipelines have to be designed and implemented to support these major requirements.

3.3.2 Estimating Transmission Chains and Deciphering the Temporal Dynamics of Bacterial Spread

The dynamics of bacterial spread can be explored using the models implemented in BEAST [7]. Constant bacterial population size, exponential growth, and molecular clock of genome-wide changes can be used as variables in these models. For each analysis, simulations should be run for 100 million generations with sampling every 10,000th generation (these figures are intended as a general guide only).

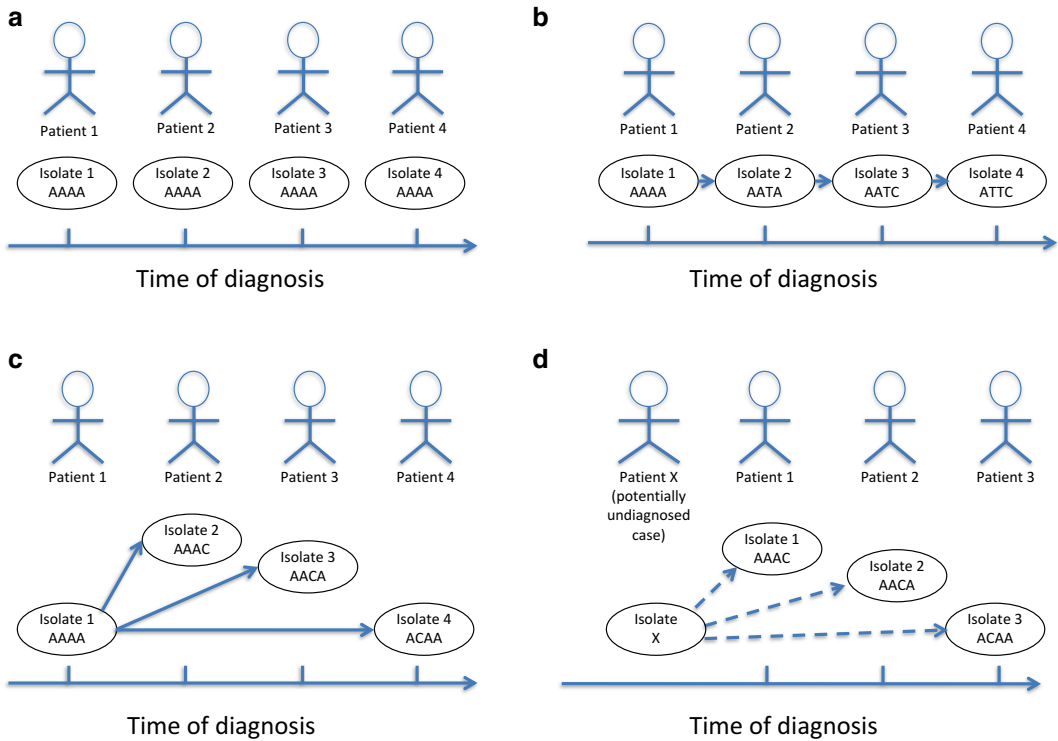


Fig. 4 Inferring the direction of transmission of a microorganism from accumulated mutations identified by WGS. Mutations are shown as SNPs in a hypothetical string of four nucleotides. **(a)** Four genomes are identical, and no direction can be inferred. **(b)** An apparent transmission chain from *left to right*, each patient accumulating a new mutation and passing the infection on to the next. **(c)** A possible central source case infects three secondary cases, each with a separate mutation not seen in other cases. **(d)** Three cases, each with a separate mutation not seen in other cases. For any one of these cases to have infected the other cases, it appears that two independent mutations would have had to occur at the same locus in separate individuals. The more likely explanation is an undiagnosed common source case with possible epidemiological links represented by *dashed lines*. Modified from [18]

Genomes of many pathogens appear to be fairly clonal and stable over time. These features suggest that patterns of accumulating SNPs can be used as a marker of microevolution within a clone. Furthermore, this *evolution by descent* offers the potential to use sequencing data as an indicator of the direction of transmission within an outbreak [18]. The directionality of transmission can be inferred through the comparison of genomic changes in isolates recovered from patients that belong to putative epidemiological clusters. Figure 4 illustrates this concept with four hypothetical short sequences representing four isolates of the same species from four hypothetical individual cases with a clinical disease and the potential for transmission. The approach has been validated for outbreaks of pulmonary tuberculosis when the topology of a phylogenetic tree suggested the existence of a common source of secondary cases and the secondary cases corresponded to the root

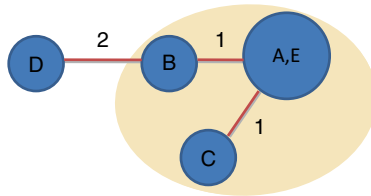
Step 1: Assess number of different alleles

Isolates	Allele 1	Allele 2	Allele 3	Allele 4	Allele 5
A	3	4	7	0	523
B	3	5	7	0	523
C	3	4	8	0	523
D	3	5	9	1	523
E	3	4	7	0	523

Step 2: Create a distance matrix

	A	B	C	D	E
A	0				
B	1	0			
C	1	2	0		
D	3	2	3	0	
E	0	1	1	3	0

Step 4: Assign clonal complexes



Step 3: Translate distance into tree

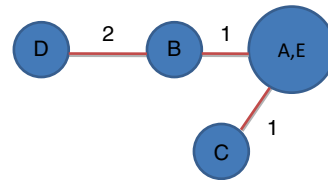


Fig. 5 Representation of microbial profiles of five isolates with a minimum spanning tree

of a phylogenetic tree, accurately predicting the existence of a common source case that was sequenced at a later date [13, 19]. However, this approach is based on several assumptions regarding the evolution of pathogens, nature of infectious diseases transmission, and technical capacity of detecting mutations.

3.3.3 Examining Microbial Subpopulations in Hospitals and Community Setting

The hypothesis of metapopulation structure, i.e., existence of different bacterial subpopulations in distinct geographical locations such as hospitals or communities, can be tested with a modified version of the Slatkin and Madison test [20]. The bacterial gene flow among different hospitals can be traced using the state changes and stasis tool (MacClade software), which counts the number of changes in a tree for each pair-wise character state and reconstructs the maximum parsimony of the ancestral characters. Subpopulations of pathogens are often visualized with spanning trees or connected graphs without cycles that reach out to or *span* all vertices. When represented as a connected graph with the least total weight, this estimated tree is called a *minimum spanning tree* (see **Note 3**). Figure 5 illustrates the process of the representation of a set of microbial profiles, e.g., obtained from a variable tandem repeat genomic typing, with a minimum spanning tree starting with the assessment of a number of different alleles and closing with the assignment of clonal complexes to nodes with the highest number of isolates. We argue that the Manhattan distance is a more appropriate option than Euclidian distance to measure the similarity of genomic profiles based on variable tandem repeats in multiple independent loci across a genome.

3.3.4 *Assessing the Cluster Definition and Its Performance*

Most spatial and temporal cluster definitions satisfy two important properties. First, they provide a unique way of clustering cases that is independent of the order in which the isolates are considered. This property guarantees that any two cases assigned to a cluster at a given time will remain in one cluster in the presence of additional cases. This makes it possible to search, retrospectively, for clusters (for given parameters N , t , and d) in historical data, compute the number of clusters, and determine how early they would have been detected, prospectively. In this way, one can adjust future values of N , t , and d according to prospective surveillance needs and the availability of public health resources. Algorithms that implement the working definitions of outbreaks often have three steps [21]:

1. Compute temporal and/or spatial distance of each new isolate with existing same-genotype isolates.
2. (a) If an existing isolate is found for which temporal and/or spatial distance is smaller or equal to t and/or d then the new isolate joins the set of this existing isolate. (b) If more than one isolate is found for which temporal and/or spatial distance is smaller or equal to t and/or d and they belong to different sets then the sets merge into one. (c) When no isolates have been found for which temporal and/or spatial distance is smaller or equal to t and/or d then the new isolate forms a new set.
3. A set becomes a cluster the moment it reaches N or more isolates.

The performance of these outbreak definitions in a prospective surveillance system can be tested by estimating how long it would take to detect each cluster in real time using a given outbreak definition. Following the algorithm described above, the detection date of an outbreak is simply the date at which a set of same genotype isolates that fulfil the appropriate spatiotemporal restrictions reaches N or more isolates and becomes a cluster (**step 3**) [21]. Figure 6 provides an example of the cluster definition performance. It compares three definitions, i.e., using 2, 4, and 6 isolates with indistinguishable genotypes recovered from patients who were getting ill in a comparable timeframe. The findings illustrate that these definitions perform similarly and offer the detection of at least 60 % of clusters of a disease within the first half of the cluster duration [21].

3.3.5 *Spatial and Temporal Clustering*

An alternative and more complex way to account for time and space dependencies within a set of disease counts (and potentially within a genotyping cluster) is to use a statistical method such as the space-time permutation scan statistic [22]. In this method, a cluster is defined as the region in space and time where the probability of an incident case occurring is higher inside than outside. Expected values are estimated from existing counts (for a given day and location the expected counts are taken to be proportional to

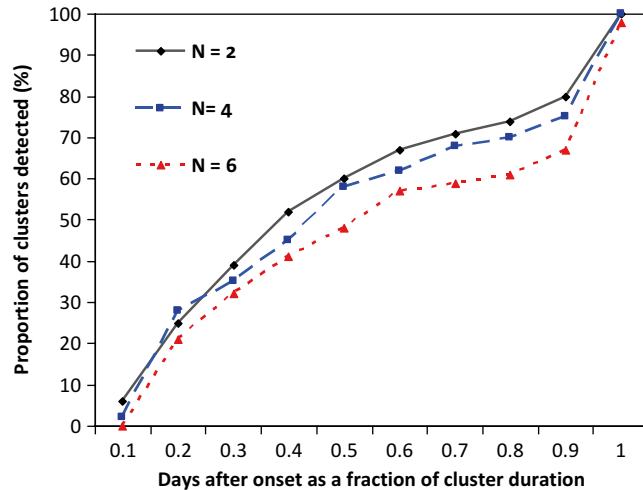


Fig. 6 Percentage of genotyping clusters as a function of their detection time (expressed as a fraction of cluster duration). Detection of spatiotemporal genotyping clusters with $(t, d) = (\text{no limit}, \text{no limit})$ for three different cluster size definitions (N): $N=2$, i.e., at least two cases with undistinguishable microbial genotypes (solid line), $N=4$ (dashed line), and $N=6$ (dotted line). Modified from Gallego et al. [21]

all the cases that occurred in that location multiplied by all the cases that occurred in that day).

Scan statistic has been one of the most popular methods for spatiotemporal analysis in both retrospective and prospective cases. Scan statistics were originally developed to test for spatial clusters. In the spatial case, scan statistics generally impose a circular window on the map under study and let the center of the circle move over the area so at different positions the window includes different sets of neighboring cases. The analysis is repeated for different sizes of circular windows. Conditioning on the observed total number of cases (N), the spatial scan statistic S is defined as the generalized maximum likelihood ratio over all possible circles Z . The algorithm maximizes the generalized likelihood ratio over all the circles and eventually identifies the circle that constitutes the most likely cluster. The method has been extended to a spatiotemporal one to enable prospective data analysis. In this scenario, instead of a circular window in two dimensions, the space-time scan statistic uses a *cylindrical window*. The base of the cylinder represents geographical data, while height represents time [22].

3.3.6 Data Transformation and Synthesis

The diversity and systematic heterogeneity of biological (in vitro and in vivo observations) and computational data require novel approaches to data synthesis. There are many strategies for data integration that employ statistical or mathematical models [23].

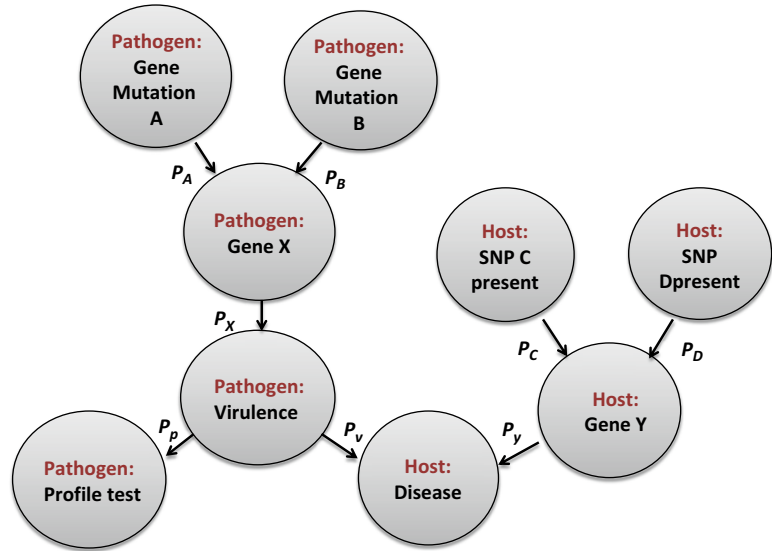


Fig. 7 A Bayesian network representing the relationship between infectious disease incidence (Disease), virulence of a pathogen, and changes in gene X and gene Y of the pathogen and a susceptible host (e.g., mutation or SNP), respectively. Edges represent probabilistic dependencies and the P notation defines conditional probability distributions between nodes

A well-known and useful tool in statistics is regression analysis. Bayesian statistics and Bayesian networks are becoming increasingly popular. There are many differences between the frequentist and Bayesian approaches to inference. In our view, Bayesian inference allows one to make predictions, including predictions about pathogen clusters and outbreaks, based both on the weight of evidence and a somewhat more explicit statement of the assumptions of a statistical model (Fig. 7). A related way of data integration is network inference. The premise of network inference is to quantify relationships between signals using some metric which can vary between models and algorithms. The use of several network inference methods can improve confidence of estimation by circumventing the biases inherent in any single algorithm or model. The application of network inference methods provides insights into multivariate structures and, by quantifying relationships between measured variables. These inferences can illuminate relationships or clusters not detectable by clustering of the raw data directly, especially when several different predefined network topologies segregate the same clusters [24].

To reduce the complexity of these models or the number of parameters, dimensionality reduction is employed to identify the key components that contain the maximum amount of information concerning the problem at hand. The most common form of dimensionality reduction is Principal Component Analysis (PCA) (*see Note 4*).

The identification of the key contributors to disease virulence or drug resistance enables the development of in vitro diagnostic and prognostic multivariate index assays (IVDMIA) which typically measure gene levels and mutations, often with complex machine learning algorithms, e.g., as implemented in WEKA, www.cs.waikato.ac.nz/ml/weka/. Such tests are composed of both a laboratory component and associated algorithms, used to score risk, the latter being an integral part to realizing the test's value.

4 Conclusions

There are three main challenges in communicable disease informatics: (1) technological, (2) statistical/mathematical, and (3) translational. First, sample sizes in current publications are often to enable pattern recognition which is statistically robust and consistent across different data collection methods. Second, high-dimensional data pose difficulties for statistical analysis and machine learning because of the large number of interrelated components, i.e., “small n , large p ” dilemma. *A priori* knowledge should be incorporated into the model in a principled manner to constrain the set of possible representations for the data and optimize a well-known trade-off between the small sample size and high model complexity. Despite these challenges, the application of WGS and informatics for pathogen detection and surveillance offers considerable potential for improving healthcare delivery (Table 1).

In addition, the requirements for the data processing, storage, and backup of WGS data outstrip the existing capacity of laboratory information systems. Many universities and vendors such as Amazon, Google, and Microsoft have created centralized super-computing facilities to support data-intensive analyses. The reliance on the university-based or commercial computer clusters provided might be appropriate to support biomedical research WGS experiments but their potential use for the processing of public health surveillance data must be carefully reviewed. For example, external cloud services may not comply with local data transfer security, patient privacy and confidentiality regulations. Another critical variable is the policy regarding the long-term storage and protection of clinical data in a commercial environment in which ownership is subject to change, mergers, or acquisition. Diagnostic and public health laboratories need to identify the jurisdictional laws and regulatory guidelines that oversee the transfer and storage of clinical and laboratory data as well as types of sequence data laboratories may be mandated to disclose to public health authorities.

Analysis of NGS data requires multidisciplinary teams of microbiologists, bioinformaticians, clinicians, and epidemiologists with substantial institutional support for resources and personnel.

Table 1
Main applications of WGS in pathogen genome informatics

Domains	Applications	Expected added value	References
Laboratory diagnosis and decision support	Identification of new pathogens	Improved timeliness of case detection; improved laboratory workflow	[1–3]
	Discovery of virulence mechanisms	Development of alternative therapeutics	[2, 25]
	Detection of drug-resistance markers	Optimized drug selection; reduced burden of drug resistance; improved patient outcomes	[3, 26, 27]
Laboratory-based public health surveillance	Detection of emerging clones	Monitoring of immune escape during clonal spread; high-resolution surveillance	[2, 12, 28]
	Tracing movements of mobile genetic elements between pathogens in clinical environments, e.g., new acquisitions by resident microflora by mutation or by spreading between patients	Improved hospital infection control practices; reduction in nosocomial transmissions	[11, 12]
	Identification of covert clusters of infections	Improved public health surveillance	[15, 28]
	Detection of disease outbreaks, ideally at point of first secondary case	Clinically and cost-effective targeted public health response	[15, 29]
	Tracing transmission events within outbreaks and determining directionality of transmission	Integration of genomics guided surveillance into communicable disease control and response; cost-effective contact tracing	[18, 19], [30–32]
	Source attribution with <i>molecular compass</i> of geographical population structure among related pathogens and their genomes	Cost-effective and timely response to public health threats	[2, 15, 33]

As WGS is applied to public health surveillance, standardizing quality metrics becomes critical. These metrics include standards for calibration, validation, and comparison among platforms; data reliability, robustness, and reproducibility, and the quality of assemblers. Like any technology, WGS has its advantages and limitations. Potential uncertainties and errors can be introduced into the sequence analysis by the sequencing machines, analytical algorithms, and residual errors in the reference data we align the new sequence against. Proficiency testing programs that cover both sequencing *wet laboratory* and analytical *dry laboratory* steps are urgently required.

5 Notes

1. Sequencing techniques generate large files containing thousands or millions of reads together with additional information such as read identifiers and descriptions. Different file formats have been introduced to efficiently manage this information. FASTQ is an extension of FASTA format. It stores a numeric quality score (PHRED) for every nucleotide in a sequence. Unfortunately, there is no uniform standard for encoding these quality scores and different PHRED-scales are in common use. SAM (Sequence Alignment Map) format is a file format for storing information about sequence alignments. The BAM file format is a binary representation of SAM, which was implemented to allow for efficient storage and processing of data. VCF (Variant Call Format) files have been introduced to store data about SNPs and small indels along with various types of metadata. The file format is simple and flexible.
2. The size of data generated by WGS instruments (>1 Gb per genome) requires additional computational resources for data analysis. Currently, there is no single platform offering data processing, database and data warehousing capabilities, and thus, institutions are required to establish their own data analysis pipelines or link together a variety of commercial and open-source software packages and data sets that contain information about microbial genomes of interest. Data exchange and online analyses are limited by the relatively low bandwidth and firewalls of existing laboratory and clinical networks. Not surprisingly, many laboratories have opted for in-house solutions instead of outsourcing. There is significant variability in the processes of storing sequencing and secondary data files. The most common approach of relying on external hard drives is not sustainable and the scalable capacity for systematic backup is needed. There is also a need for guidance for decisions regarding what types of WGS data should be stored and for how long, given the fact that the cost of storage will soon exceed the cost of data generation [34]. Whether to opt for commercial or open source solutions remains an open question. The enterprise software solutions are costly to laboratory budgets and it is not possible for large vendors to provide for all the needs of a laboratory. In-house solutions developed using open source software can lead to high labor costs, but are entirely flexible, and enable the control of sample and data processing and sharing. This is clearly critical for clinical diagnostic sequencing where process and quality control are of the utmost priority. It therefore seems likely that a mix of the two approaches will be appropriate for the foreseeable future.

3. The terminology regarding trees, graphs, and networks has been used somewhat inconsistently in the literature as it derives from various disciplines with different emphases. In particular, the distinction between a weighted graph and an unweighted graph is often important and has been dealt with briefly in the present chapter. An introduction to graph theory is found in the following reference [35].
4. Principal Component Analysis (PCA) is a technique for dimension reduction. It is invaluable as an exploratory tool and for data visualization. In essence, the idea is to reduce the number of dimensions of a high dimensional vector by obtaining a linear transformation which minimizes the *information loss* as characterized by a reduction in the *total variation* of the data set. The performance of PCA can be quantitatively evaluated using various metrics. PCA is an established technique and can be conducted in many software packages including the freely available R package (www.r-project.org).

Acknowledgments

VS was supported by the National Health and Medical Research Council Career Development Award.

References

1. Bertelli C, Greub G (2013) Rapid bacterial genome sequencing: methods and applications in clinical microbiology. *Clin Microbiol Infect* 19:803–813
2. Relman DA (2011) Microbial genomics and infectious diseases. *N Engl J Med* 365:347–357
3. Long SW, Williams D, Valson C et al (2013) A genomic day in the life of a clinical microbiology laboratory. *J Clin Microbiol* 51:1272–1277
4. Köser CU, Ellington MJ, Cartwright EJ (2012) Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog* 8:e1002824
5. Sintchenko V, Iredell JR, Gilbert GL (2007) Genomic profiling of pathogens for disease management and surveillance. *Nat Rev Microbiol* 5:464–470
6. Hall BG (2011) *Phylogenetic trees made easy*, 4th edn. Sinauer Associates, MA, USA, pp 61–138. ISBN 978-0-87893-606-9
7. Drummond AJ, Suchard MA, Xie D et al (2012) Bayesian phylogenetics with BEAUTi and the BEAST 1.7. *Mol Biol Evol* 29:1969–1973
8. Kostic AD, Ojesina AI, Pedomallu CS (2011) PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nat Biotechnol* 29:393–396
9. Earl D, Bradham K, John J et al (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res* 21:2224–2241
10. Harris SR, Török ME, Cartwright EJ et al (2013) Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks. *Nat Biotechnol* 31:592–594
11. Dunne WM Jr, Westblade LF, Ford B (2012) Next-generation and whole-genome sequencing in the diagnostic clinical microbiology laboratory. *Eur J Clin Microbiol Infect Dis* 31:1719–1726
12. Gilmour MW, Graham M, Reimer A et al (2013) Public health genomics and the new molecular epidemiology of bacterial pathogens. *Public Health Genomics* 16:25–30
13. Gardy JL, Johnston JC, Ho Sui SJ et al (2011) Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med* 364:730–739
14. Köser CU, Holden MT, Ellington MJ et al (2012) Rapid whole-genome sequencing for

- investigation of a neonatal MRSA outbreak. *N Engl J Med* 366:2267–2275
15. Underwood AP, Dallman T, Thomson NR et al (2013) Public health value of next-generation DNA sequencing of enterohaemorrhagic *Escherichia coli* isolates from an outbreak. *J Clin Microbiol* 51:232–237
 16. Bielejec F, Rambaut A, Suchard MA et al (2011) SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics* 27:2910–2912
 17. Yang Y (2005) Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* 92:937–950
 18. Walker TM, Monk P, Smith EG et al (2013) Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. *Clin Microbiol Infect* 19:796–802
 19. Walker TM, Ip CL, Harrell RH et al (2013) Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis* 13:137–146
 20. Switzer WM, Salemi M, Shanmugan V et al (2005) Ancient co-speciation of simian foamy viruses and primates. *Nature* 434:376–380
 21. Gallego B, Sintchenko V, Wang Q et al (2009) Biosurveillance of emerging biothreats using scalable genotype clustering. *J Biomed Inform* 42:66–73
 22. Kulldorff M, Heffernan R, Hartman J et al (2005) A space-time permutation scan statistic for disease outbreak detection. *PLoS Med* 2:e59
 23. Greene CS, Troyanskaya OG (2012) Chapter 2: data-driven view of disease biology. *PLoS Comput Biol* 8:e1002816
 24. Wagner JP, Wolf-Yadlin A, Sevecka M et al (2013) Receptor tyrosine kinases fall into distinct classes based on their inferred signalling networks. *Sci Signal* 6:ra58
 25. Bessen DE (2012) Population genomics: an investigative tool for epidemics. *Am J Pathol* 180:1358–1361
 26. Rolain JM, Diene SM, Kempf M et al (2013) Real-time sequencing to decipher the molecular mechanisms of resistance of a clinical pan-drug-resistant *Acinetobacter baumannii* isolate from Marseille, France. *Antimicrob Agents Chemother* 57:592–596
 27. Hornsey M, Loman N, Wareham DW et al (2011) Whole-genome comparison of two *Acinetobacter baumannii* isolates from a single patient, where resistance developed during tigecycline therapy. *J Antimicrob Chemother* 66:1499–1503
 28. Falush D (2009) Toward the use of genomics to study microevolutionary change in bacteria. *PLoS Genet* 5:e1000627
 29. Althomsons SP, Kammerer JS, Shang N et al (2012) Using routinely reported tuberculosis genotyping and surveillance data to predict tuberculosis outbreaks. *PLoS One* 7:e48754
 30. Roetzer A, Diel R, Kohl TA et al (2013) Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiology study. *PLoS Med* 10:e1001387
 31. Schürch AC, Kremer K, Daviena O et al (2010) High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Microbiol* 48:3403–3406
 32. Bryant J, Schürch AC, van Deutekom H et al (2013) Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. *BMC Infect Dis* 13:110
 33. Loman NJ, Constantinidou C, Christner M et al (2013) A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4. *JAMA* 309:1502–1510
 34. Gullapalli RR, Desai KV, Santana-Santos L et al (2012) Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. *J Pathol Inform* 3:40
 35. Cormen TH, Leiserson CE, Rivest RL et al (2009) Introduction to algorithmics, 3rd edn. McGraw-Hill, London

Setting Up Next-Generation Sequencing in the Medical Laboratory

Bing Yu

Abstract

The introduction of next-generation sequencing (NGS) technologies in research has proven to be very successful in the past 8 years. Now, there is considerable demand to apply these technologies for clinical diagnosis. The translation of research-to-clinical practice brings with it a unique set of challenges, particularly when it comes to setting up NGS in the medical laboratory. The practical issues related to infrastructure, selecting which NGS platform, and dealing with informatics requirements are discussed. Application of NGS for clinical diagnosis requires robust quality assurance at multiple levels including sample assessment, library preparation, template generation, and sequencing data which need to be generated, analyzed, and stored. The requirements for data generation, analysis, and storage are considerable.

Key words Genomic diagnosis, Medical genetics laboratory, Quality assurance

Abbreviations

ePCR Emulsion PCR
NGS Next-generation sequencing
qPCR Real-time/quantitative PCR

1 Introduction

Next-generation sequencing (NGS, also called second-generation (G2) sequencing) is so named relative to Sanger or first generation (G1) sequencing. NGS has three characteristics: (1) A need for clonal amplification of templates; (2) Sequencing is undertaken in a massively parallel way; and (3) Short read lengths (<700 bp) are generated [1]. The Genome Sequencer (GS) FLX was the first commercially available NGS platform introduced by Roche 454 Life Science in 2005 [2, 3]. Illumina® released the Solexa Genome Analyzer (GA) in 2006 [1, 4], followed by Life Technologies™ SOLiD™ (sequencing by Oligo Ligation Detection) by the end of 2007 [1, 5].

The NGS platforms can be categorized into two groups: (1) High-throughput instruments represented by HiSeq 2500 and SOLiD™ 5500 series, and (2) Rapid benchtop instruments (Ion Torrent™ / Proton™, MiSeq, and GS FLX/Junior) [6]. The Pacific Biosciences® PacBio RSII is also considered a rapid platform with potential benefits for the clinical laboratory [1, 7]. It represents third-generation (G3) sequencing methodology which does not require clonal amplification, captures the sequencing signals in real time, and has long reads, e.g., PacBio RSII >10 kb [1, 7].

NGS technologies have been extensively used in a range of research activities including (1) de novo genome sequencing of human samples, model organisms, and evolutionary informative species; (2) Re-sequencing for variation identification; (3) RNA-Seq for small RNA or transcriptome analyses; and (4) Characterizing cellular mechanisms such as chromatin/epigenomic modifications and spatial arrangement of cellular components [1]. The research applications of NGS have significantly improved our understanding of the human genome in the past 8 years.

Currently NGS technologies are moving from the research to the clinical laboratory [3], particularly the benchtop NGS instruments. Clinical applications of NGS are having an impact on health care delivered to patients [3]. The top three clinical indications are (1) diagnosis of Mendelian disorders (identification of disease-causing gene mutations), (2) molecular oncology analysis (a sequence-based companion diagnostic test), and (3) pharmacogenomics screening.

NGS limited to exome analysis can be used to address specific clinical questions and avoids overwhelming clinicians and patients with irrelevant or uninterpretable information obtained through whole genome sequencing. This approach has already had a significant impact on identifying the underlying genes and causal variants for a number of Mendelian disorders particularly those involving rare diseases [8].

Companion diagnostics based on a systematic NGS analysis are assuming greater importance for personalized (stratified) medicine to guide decisions on therapy leading to higher quality treatment options for patients while lowering health costs by avoiding therapies that are unlikely to work. Similarly, pharmacogenomics screening can improve drug efficacy and reduce the number of adverse drug reactions by providing clinicians with additional guidance on drug dosage based on the individual's genetic ability to metabolize or transport these drugs [3].

However, the goal of a rapid translation of research-to-clinical practice brings with it a unique set of challenges based around robust quality assurance to ensure the required reliability of clinical results. Some issues that address NGS applications in clinical genomic testing can be found in the CAP (College of American Pathologists) Checklist for NGS Laboratory Standards [9] and in

several reviews [10, 11]. This chapter focuses on a number of practical issues that should be considered when NGS is planned for implementation in the clinical laboratory.

2 Materials

2.1 Laboratory Infrastructure

Successful NGS sequencing depends on generating massively parallel reactions using either sequencing-by-synthesis or sequencing-by-ligation. These reactions can last for hours or even up to 12 days. Therefore, there will be specific infrastructural needs for rooms to host a NGS platform.

- *Temperature and humidity controls.* These two parameters must be controlled to ensure optimal sequencing reactions.
- *Vibration-free environment.* Primary data acquisition in most NGS platforms except the post-light ones such as the Ion Torrent/Proton relies on “super-density scanning.” For this it is necessary to have minimal vibration of the building structure, and any nearby sources for vibration such as centrifuges.
- *Uninterrupted power supply.* It is essential to have a secure and surge-free electricity supply for a NGS platform. Electricity should come from an uninterrupted power supply (UPS) source which is ideally backed up by a diesel generator since a sequencing reaction run can take up to 12 days.
- *Ultrapure water supply.* Access to type 1 ultrapure water (18.2 MΩ cm at 25 °C and TOC (total organic carbon) <10 parts per billion) is required particularly when using the post-light NGS platforms.
- *Excellent informatics network.* NGS can generate an unprecedented amount of data, which pose challenges for data analysis, management and storage [10]. An efficient transfer of such high volume data is necessary using sophisticated network connections between the data generation site to the analysis and storage servers. A 1 Gb (gigabyte = 1×10^9 bytes) network is essential for the NGS environment, with 10 Gb networks becoming more common using fiber-optic cables as the demands increase [12].
- The clinical laboratory should be designed not only for the optimal operation of a NGS platform, but also to ensure no potential for contamination or mix-up of samples at different stages in the process (*see Note 1*).

2.2 NGS Platform Selection for Clinical Laboratory

Selection of which NGS platform should be carefully considered based on its particular characteristics and the clinical needs [6]. Overall, the clinical laboratory requires versatile, robust, and affordable NGS platforms with user-friendly bioinformatics tools.

The latter, to some extent, will also be determined by the experience of the operator. Finally, excellent technical support and service is essential since delays because of machine downtime are not acceptable in patient care. Depending on the clinical indications, the turnaround time required will invariably influence the type of approach developed.

- *Minimal downtime.* The acquisition of two NGS platforms should be considered to ensure minimal disruption to the diagnostic service. These do not have to be the same platforms but should be interchangeable. For example, the Ion Torrent™ and Ion Proton™ platforms share the same chemistry and use the Ion Chef™ System for template preparation and chip loading. Alternatively, the rapid mode HiSeq 2500 can be the backup for the benchtop MiSeq instrument.
- *G3 platform.* G3 sequencing technologies are attractive for the clinical laboratory because of their shortened DNA preparation times, fast processing of sequencing signals in real time, and long read lengths [1, 7, 8]. However, a commercial Oxford Nanopore analyzer has yet to be released. The PacBio RSII analyzer presently has a relatively low throughput, high error rate for individual reads, high cost per Mb, and high capital cost [1, 7, 13].
- *Fast turnaround time.* It is essential in clinical NGS application to ensure an appropriate turnaround time for clinical decision making. For example, cancer somatic mutation profiling using NGS is being used for prioritization of target therapy. This type of test must be completed within 5 days from DNA extraction to reporting. High-throughput platforms such as the SOLiD™ 5500 series and HiSeq 2000 are unlikely to provide such fast turnaround results.
- *Flexibility.* This is required to ensure cost-effective and efficient applications in the clinical laboratory. Included here would be the batching of test samples when it is difficult to achieve a full run due to the rapid turnaround requirement. Different options for chip sizes or partial flow cells should be considered when selecting a platform based on the estimated sample numbers likely to be referred.
- *Starting template amount.* The DNA template requirement can be a significant limitation in cancer somatic mutation analysis. Formalin-fixed paraffin-embedded (FFPE) tissues are the most common source for DNA extraction in this scenario and will yield small amounts of what is often fragmented DNA. Cytology samples including fine needle aspiration and bronchial washing are even more challenging for providing sufficient DNA. PCR-based target amplification, particularly in a multiplex format, can be superior to capture-based method to overcome the limited template as well as formalin-mediated DNA damage.

2.3 NGS Accessories

Automated library preparation is preferred in the clinical laboratory to ensure reproducibility. It also reduces staff time and costs which comprise the laboratory's major budget component. Special instruments such as LifeTechnologies™ AB Library Builder™ and Beckman Coulter SPRIworks Fragment Library System are available. Alternatively, more flexible liquid handling robots including Tecan Freedom EVO®, Beckman Coulter Biomek® FXP, Agilent® Bravo™, or Caliper's Zephyr® Genomics Workstation can be introduced for automated library preparation.

The Covaris® system is a frequently used NGS accessory for shearing DNA with adjustable fragment size distributions. It uses acoustic wave energy (15–30 times higher than a sonicator) transmitted into a closed tube containing an aqueous DNA solution. This results in formation and collapse of air bubbles, which generate microscale water jets that cause physical shearing of the nucleic acid. The Covaris® enables rapid, reproducible, high-recovery, and unbiased DNA shearing. It is also used to declump post-emulsion PCR beads before chip loading in the SOLiD™ sequencing process.

The Agilent 2100 Bioanalyzer is a microfluidics-based platform for sizing, quantification, and quality control of NGS libraries. This accessory has many advantages over conventional techniques including improved data precision and reproducibility, short analysis times, and minimal sample consumption. The quality of DNA shearing, the range or distribution of DNA fragmentation before and after size selection and final quantification of a library can be assessed using this instrument. The LabChip® GX/GXII from Perkin Elmer is an alternative microfluidics instrument for sample assessment with higher sample throughput and less hands-on time.

The NanoDrop® Spectrophotometer and Qubit® Fluorometer are also useful for the quantitative assessment for DNA as well as beads generated. NanoDrop® can provide some indication in relation to impurities due to protein, peptides, and organic solvents. The Qubit® is a fluorescence-based quantification assay which is highly sensitive and accurate for double-stranded DNA without interference from RNA or nucleotides.

3 Methods

Validation, quality assurance, and quality control are essential for setting up NGS in the clinical laboratory (*see Note 2*) [10, 11]. Quality control measures should be implemented at different stages. Early detection and termination of a failed NGS test is required in the clinical laboratory. This is not only to avoid waste of expensive reagents but, perhaps more importantly, to reduce the risk for prolonged turnaround times.

3.1 *Sample Assessment*

The criteria for accepting or excluding a sample for NGS testing should be agreed on at the initial stage. This is important as NGS diagnostic applications presently have relatively higher costs and longer turnaround times compared with conventional DNA tests.

DNA extracted for any NGS test should be checked qualitatively and quantitatively. Minute amounts of DNA may only be available for analysis in circumstances like preimplantation genetic diagnosis or the use of circulating DNA for noninvasive prenatal diagnosis or cancer cell monitoring. Fine needle aspiration for cancer somatic mutation analysis is another example. Routine methods for DNA quantification such as spectrophotometry and fluorometer can be used, but these methods may be less suitable since the assay itself consumes a proportion of the available DNA. A multiplex PCR-assay which includes both identification and assessment targets would be ideal. This approach would provide the information for identity (useful for cross-checking when pooling samples) as well as its quality (amplifiable) and quantitation (available genome equivalent copies). Failure to exclude poor quality or insufficient DNA can significantly affect the sensitivity and specificity of NGS diagnosis and result in delay in the turnaround time.

In the diagnosis of Mendelian disorders, the collection of DNA samples from the index case *and* parents should be considered. Trio sample collection will enable more cost-effective data analysis and hence will achieve a reliable diagnostic result. It is also necessary to collect normal (non-tumor) genomic DNA to act as a control when a tumor sample is used for cancer somatic mutation analysis.

One should also consider whether two different samples from same individual are collected, e.g., a saliva sample collected in a physician's room and a blood sample from a collection center. The clinical NGS laboratory can use DNA extracted from both sources to exclude errors arising from sample misidentification or laboratory clerical errors. This will also help to verify the NGS results since there are many steps involved including sample pooling and barcoding [14].

3.2 *Library Preparation*

Fragment library preparation involves DNA shearing, adaptor/barcode ligation, size selection, amplification of ligated products, library purification, and final quantification. The initial DNA for library preparation has to undergo a target enrichment process. This can be achieved using different approaches including *PCR-based* (short multiplex or long range PCR products) or *capture-based* (fragmented genomic DNA) methods. Care should be taken to prevent any cross-contamination if the instruments are shared (*see Note 1*). It is a common practice to sequence a multiplex library with pooled DNA samples. Therefore, best practice needs to be implemented to ensure one specific barcode for every individual sample. Cross-checking by a coworker is required to prevent any sample swap during the barcode ligation [14].

Human exome sequencing allowing the capture of all relevant protein-coding targets is increasingly being used for diagnostic purpose [8]. Enrichment of DNA targets can be assessed by a real-time/quantitative PCR (qPCR) and be compared between pre- and post-capture using equal amounts of the templates. Excellent enrichment indicates a successful capture process. Super-multiplex PCR-based human exome enrichment (58 Mb target region in 12 pools of PCRs) is also available on the Ion Proton™ platform.

The Agilent 2100 Bioanalyzer is the method of choice to assess the quality and distribution of sheared DNA fragments. Effective size selection can narrow down the distribution peak, while successful ligation will right shift the distribution peak. Bioanalyzer results can be used to verify the purification quality, particularly how effective primers or primer dimers were removed. It also provides accurate quantification for equimolar pooling among different samples.

3.3 Template Generation

This step is unique to NGS since its aim is to generate millions of sequencing templates through clonal amplification. There are two methods for clonal amplification: emulsion PCR (ePCR) and bridge amplification (also called cluster generation). ePCR occurs within aqueous microdroplets separated by oil so that thousands of independent reactions can occur per microliter of volume. This process involves monoclonal amplification of individual DNA templates from a complex library pool. Multiple copies of a single DNA sequence can be generated within a water-in-oil emulsion droplet and are coated onto a single bead. Optimal monoclonal amplification without substantial multi-clone generation (two or more mixed signals from one bead) requires appropriate bead-to-fragment ratio. Precise quantification of the input library is critical for such optimal monoclonal amplification.

qPCR is the preferred method for determining the amount of amplifiable template in a library since it provides the high level of specificity and can accurately measure extremely low quantities of DNA. Consequently it allows the user to dilute libraries to very low concentrations for quantitation. The final optimal concentration of the library can be determined based on trial runs. Similarly, the quantification of template amount is also important in the library for bridge amplification. Overloading DNA fragments will cause the cluster density to be too high resulting in an overlapping or breaching of the signals.

In the sample pooling analysis, one should ensure the equimolar representation of multiple samples. Failure of equimolar pooling will cause a biased coverage and so insufficient coverage for some samples. Any potential cross-contamination should be prevented during clonal amplification (*see Note 1*).

3.4 Data Generation

The clinical laboratory must pay attention to base calling and raw data quality. This information is usually platform-dependent and should be monitored during each run. The expected signal intensity across a read should be evaluated to establish the normal performance ranges and expected decline in signal intensity. The raw data quality should pass the defined minimal criteria before the subsequent secondary or tertiary data processes are allowed to precede. Mapping quality can be used as a measure of the uncertainty that a read is mapped correctly to the genomic position. During validation, it should be shown that the test only provides sequence data that map to the specific regions interrogated in the test. Read mapping, coverage analysis, and variant calling with annotation will complement the quality profile for data generated (*see Note 2*).

3.5 Analysis of Data

NGS data analysis places significant demands in both computational and information technologies. Even when generating relatively limited data for clinical purposes, it is important to keep in mind that significant institutional support for resources and personnel is needed, e.g., to maintain and upgrade servers or to set up and apply tools for data analysis and management [12]. Clinical analysis of NGS data requires a multidisciplinary team of molecular geneticists, bioinformaticians, statisticians, pathologists, and clinicians. Various strategies for NGS data analysis are described in this book. The analysis pipeline needs to be validated including every subsequent software upgrade.

3.6 Storage of Data

The clinical laboratory will have to work within a number of legal obligations that will vary depending on local jurisdictional requirements. These will include obtaining the necessary accreditation to undertake clinical DNA testing as well as adhering to privacy legislation. The latter in particular is a challenge in terms of the large data sets that are generated (up to 200–300 Gb per run) [12] as well as the potential for detecting changes in DNA that are not directly relevant to the patient's current problem. *Incidental findings*, as these results are called, are discussed further in Chapter 12.

Of relevance to NGS is the storage of data sets many of which might need to be stored over long time periods to comply with legal and/or accreditation standards. VCF (variant information) file can be considered for long-term storage. The original genome sequence file (FASTQ format including quality information) or the alignment file (BAM format) should be retained for at least 1 year for potential reanalysis [14].

Storage can be considered under local and distant solutions. For the former, NGS sequencing data could be downloaded on an external hard disk, which can be kept securely and is easy to carry around. However, downloading large quantities of NGS data in the range of 20–200 Gb (Gb = 1×10^9 bytes) from the production

server and subsequently uploading the same data to the analysis server can introduce significant bit errors during the two transmissions. For example, a bit error ratio of 1 in 1,000,000 could have as many as 400,000 bit errors for 200 Gb of NGS data transferred. The external disk option could also undergo physical damage which is of particular concern as it is more difficult to ensure regular backup of data.

Dedicated local servers for NGS data analysis and storage are ideal as this would avoid the multiple transfers of large volume of data, and regular backups can be scheduled. The initial setup budget should include dedicated servers. If the local server is not available, an array disk system with a high storage capacity (6–12 Tb, terabyte = 1×10^{12} bytes) can be considered with the USB3 (Universal Serial Bus) connection or fast SATA3 (Serial Advance Technology Attachment) interface. This system can use RAID 5 (Redundant Array of Independent Disks Level 5), in which the data written to the system can be stripped across four disks in order to prevent data loss in case of a disk failure. The array disk system also supports “hot swap” in which a faulty disk can be replaced without shutting down the computer.

Ethical legal and social implications make remote storage (even analysis) of data using *cloud computing* less attractive in the *clinical* context. The reasons are as follows: (1) Privacy legislation will vary depending on where the cloud computing facility is located. (2) Data security is controlled by the *cloud* provider and, like privacy, will reflect the local jurisdictional requirements rather than those which the customer will need to address. Hopefully, these will be comparable but there might also be significant differences. (3) There is the potential to expose the data to interception during Internet transfer. In these circumstances, any problems that might emerge would be difficult to follow up from a legal perspective if another country or jurisdiction is involved.

4 Notes

1. A clinical laboratory utilizing NGS requires both state-of-the-art platforms and careful planning to prevent contamination becoming a source of error. Designated areas for pre-amplification, amplification, and post-amplification should be present, in line with the requirements of one international diagnostic accreditation standard (ISO 15189:2012). The airflow should be assessed and managed to reduce possible cross-contamination from post-amplification to pre-amplification areas.

Some accessory instruments such as Covaris® and Agilent 2100 Bioanalyzer may need to be duplicated in order to avoid the instrument being exposed to both genomic DNA and PCR products although the probability of cross-contamination is

relatively low at this stage. For example, DNA fragments are kept in a disposable and semi-closed microTUBE during the Covaris® process.

The locations for Life Technologies' EZ Bead™ System, Ion OneTouch™ and OneTouch™ ES Enrichment Station, or Illumina® cBOT Station should be considered carefully. The potential input for those clonal amplification instruments can be fragmented genomic DNA or PCR products. They should be located away from the conventional pre-PCR and PCR stations. The workflow in the clinical laboratory should be designed to move against any amplified products (even after only 6–12 cycle amplification) back to pre-PCR section, while the post-PCR area itself can affect the preparation for NGS template generation. As recommended by Life Technologies™, EZ Bead™ Emulsifier should be kept in a different room to the EZ Bead™ Amplifier or Enricher.

2. The performance characteristics of NGS need to be validated and documented before any clinical testing [10, 11]. These include the accuracy, precision, analytic sensitivity, analytic specificity, reportable range, and reference range.
 - *Accuracy.* For NGS this refers to the closeness of agreement between the sequencing results and the true value of accepted reference materials such as those generated by the US FDA from the Sequencing Quality Control project [15]. Other reference materials can be blood samples or cell lines with well-characterized data sets established. Control blood samples taken from a young adult are preferred as they avoid age-related variations [16], but these may not be readily renewable. Cell lines are superior in terms of renewability as well as sources of stable mutations and structural changes, but they per se may have rearrangements or loss of DNA.
 - *Coverage.* Sequencing errors in individual reads can be minimized via the analysis of multiple overlapping reads. The number of reads covering a given base position is defined as *depth of coverage*. *Average coverage* is the average number of overlapping reads within the regions of interest. Different diagnostic applications require different depths of coverage. For example, a heterogeneous sample such as tumor cells requires 500× to 1,000× average coverage to detect 1–5 % changes, while 50× average coverage is usually sufficient for a homogeneous sample, e.g., in the diagnosis of Mendelian diseases. The *uniformity of coverage* is a better parameter than the average coverage to describe the distribution or coverage across the region of interest that must be achieved to produce reliable sequencing results.

It can be expressed as 0.2× of the mean over 90 % of the total target regions. The uniformity of coverage should be monitored during diagnostic testing and compared to that established during validation.

- *Analytic sensitivity and specificity.* Analytic sensitivity is the likelihood that NGS will detect a sequence variation, if present, while analytic specificity is the probability that NGS will not detect a sequence variation, if not present. The *gold standard* data can be obtained from samples that have been characterized using Sanger sequencing or microarray analysis. Particular assays such as DNA mass spectrometry and real-time PCR can be used for the confirmation or exclusion of false positive and false negative results. The sensitivity of NGS diagnosis depends on the horizontal (how many target genes included) and vertical (depth) coverage of the genomic regions of interest [14]. The mutation detection limit can be identified using mutations present in known proportions in a cancer cell line that has been blended with different proportions of wild-type genomic DNA. Genomic DNA from a real frozen cancer tissue can also be used as the reference material for the assessment of NGS sensitivity and specificity. Several reference mutations can be characterized and verified by other methods such as DNA mass spectrometry, qPCR, or digital PCR.
- *Precision.* The clinical laboratory needs to demonstrate the precision, i.e., the degree to which repeated measurements give the same result—repeatedly (within-run precision) and reproducibility (between-run precision). Well-characterized reference materials can be used to monitor the intra- and inter-run variability.

NGS technologies continue to evolve rapidly, and any clinical application should be fully validated against the best available standards. As a minimum, the output from NGS (G2) should match what is obtained by Sanger sequencing [17]. The relevant sequence calling, mapping, and variant calling software need to be validated along with the test system validation. The clinical laboratory should establish a reportable range, e.g., multiple genes, exomes, or large genomic regions. It could maximize the diagnostic yield and minimize costs if a workflow can be established to deal with *disease essential* genes [14]. Ongoing verification and validation are necessary to demonstrate unchanged performance characteristics or to reestablish the characteristics when there are upgrades for the hardware and software, and changes in sequencing chemistries, reagents, or kits used for NGS diagnosis [9, 11].

Confirmation is recommended for any clinically relevant finding in NGS analysis by a different chemistry or a second method, particularly at the initial stages that NGS is applied for patient care. It is essential to enrol in external quality assurance programs and participate in proficiency testing or inter-laboratory sample exchange programs.

References

1. Liu L, Li Y, Li S et al (2012) Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012:251364
2. Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376–380
3. Martinez de Lecea MG, Rossbach M (2012) Translational genomics in personalized medicine – scientific challenges en route to clinical practice. *HUGO J* 6:1–9
4. Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456:53–59
5. Smith AM, Heisler LE, St Onge RP et al (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Res* 38:e142
6. 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
7. Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138
8. Rabbani B, Mahdih N, Hosomichi K et al (2012) Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J Hum Genet* 57:621–632
9. College of American Pathologists (2012) Next generation sequencing in Molecular pathology checklist (7.31.2012) – CAP Accreditation Program
10. Gargis AS, Kalman L, Berry MW et al (2012) Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol* 30:1033–1036
11. Lubin IM, Kalman L, Gargis AS (2013) Guidelines and approaches to compliance with regulatory and clinical standards: quality control procedures and quality assurance. In: Wong L (ed) *Next generation sequencing: translation to clinical diagnostics*. Springer, New York, pp 255–273
12. Sexton D (2012) Computational infrastructure and basic data analysis for high-throughput sequencing. In: Rodriguez-Ezpeleta N, Hackenberg M, Aransay AM (eds) *Bioinformatics for high throughput sequencing*. Springer, New York, pp 55–66
13. NGS Field Guide: Overview (2013) www.molecularecologist.com/next-gen-fieldguide-2013/. Accessed 4 Sept 2013
14. Weiss MM, Van der Zwaag B, Jongbloed JD et al (2013) Best practice guidelines for the use of next generation sequencing (NGS) applications in genome diagnostics: a national collaborative study of dutch genome diagnostic laboratories. *Hum Mutat* 34: 1313–1321
15. US Food and Drug Administration (2011) The third phase of the MAQC project – sequencing quality control (SEQC). www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/. Accessed 4 Sept 2013
16. Forsberg LA, Rasi C, Razzaghian HR et al (2012) Age-related somatic structural changes in the nuclear genome of human blood cells. *Am J Hum Genet* 90:217–228
17. Feliubadalo L, Lopez-Doriga A, Castellsague E et al (2012) Next-generation sequencing meets genetic diagnostics: development of a comprehensive workflow for the analysis of BRCA1 and BRCA2 genes. *Eur J Hum Genet* 21: 864–870

Chapter 12

Managing Incidental Findings in Exome Sequencing for Research

Marcus J. Hinchcliffe

Abstract

Exome sequencing for research has become available for broadly based genomic studies as well as smaller targeted investigations. New exome research projects being considered will intentionally process a large amount of common and rare DNA variation for the purpose of finding specific links between genotype and phenotype. However, the risks of uncovering a clinically relevant *incidental finding* are not uniform across projects but are highly dependent on the question being asked and exactly how it is intended to be answered.

Factors that influence the possibility of revealing a clinically relevant incidental DNA variation include the following: The overall design of the study and the number of participants involved, the mode of inheritance of the phenotype including whether the phenotype is likely to have a monogenic or a complex inheritance, whether the study is assessing a known list of genes or not, and whether the causative DNA variation is likely to be rare or common. Importantly, differing bioinformatics DNA variant filtering strategies strongly influence the odds of discovering an incidental finding. This chapter provides a framework for understanding and assessing the likelihood of discovering clinically relevant, incidental DNA variations that are not directly related to the question being addressed in a particular exome research project. It also outlines DNA variant filtering and functional informatics approaches that can investigate specific genomic questions while minimizing the risks of uncovering an incidental finding.

Key words Bioinformatics DNA variant filtering, Exome sequencing, Incidental finding, Next-generation sequencing

Abbreviations

GWAS Genome-wide association studies
NGS Next-generation sequencing
SNV Single-nucleotide variation

1 Introduction

A clinically relevant incidental DNA variation can be defined as a verified DNA variation that has a proven medically relevant phenotype not directly related to the condition being studied for research.

It is an unforeseen clinical finding relevant to the individual research participant involved. Therefore, it should be reported back to the participant and his/her doctor for follow-up. Thus, properly informed consent must explain the possibility of finding an incidental DNA variation.

Importantly, not all exome sequencing research projects have the same level of risk for uncovering an incidental finding. Understanding the factors affecting the likelihood of discovering an incidental finding is important for ensuring appropriate research ethics approval, for informed consent of the participants, and, for the researchers, to clarify their own study design and bioinformatics pipeline before embarking on a particular project.

Here described is a framework to assess the risks of discovering clinically relevant incidental DNA variations in particular research projects. Subheading 2 outlines a bioinformatics variant filtering process to minimize the risk of discovering an incidental DNA finding while answering the question that the research wishes to address. Subheading 2 also outlines a stepwise process for calculating the expected number of DNA variations for further investigation and the likelihood of finding a clinically relevant DNA variation.

There are three broad [1–3] and four narrow [4–7] factors listed here that influence the likelihood of uncovering an incidental finding in any research.

1.1 Study Design

A basic strategy of exome sequencing for research involves filtering out a very large number of DNA variations (25,000–30,000) sequenced from the 1.5 % of the genome that encompass the protein-coding section (*see Note 1*) to find the single (*or* maybe two or more for complex inherited traits) DNA variation(s) directly linked to the disease/phenotype under investigation. There are approximately 200,000 exons across 20,000 protein-coding genes (average of 10 exons per gene) that comprise the exome. While there is clear evidence that much of the nonprotein-coding portion of the genome is functional, protein-coding mutations have been consistently linked to Mendelian diseases probably because loss of their function leads to significantly greater alterations to cellular biology and so more penetrant phenotypes. This is reflected in the relatively higher evolutionary conservation of protein-coding exons compared with the noncoding portion of the genome. Hence, capturing and deep sequencing of the protein-coding exome is a worthwhile stratagem for gene and specific mutation discovery.

Beyond this general approach, there are specific study models that influence both the chance of finding the gene(s) of interest and the risk of also uncovering an incidental finding. Following is a list of some different study models and their respective incidental finding risks.

1.1.1 Exome of a Single Individual

A study with $n=1$ can leave the researcher vulnerable to uncovering an incidental finding as the entire list of rare variants might be examined to find the critical DNA variant. This method will work best for a rare condition that is likely to be monogenic and has been accurately phenotyped. It is an advantage to have either a list of genes to examine or a clear understanding of the biological system that is disrupted. It is also an advantage to have other relatives similarly affected and/or to have a recessive condition. This is a poor study design for a common condition or an inadequately characterized phenotype. An example of this type of approach can be seen in a whole-genome sequencing study for a rare form of Charcot–Marie–Tooth neuropathy [1].

1.1.2 Exomes of Multiple Unrelated Individuals for a Single Condition

This strategy is designed to detect genic overlap/intersection across multiple individuals with the same condition. Both the chance of discovering the gene(s) of interest and reducing the risk of uncovering an incidental finding are improved by enhancing the power of the study through an increase in the number of participants. If family segregation (linkage) analysis is also available subsequent to finding candidate DNA variation(s) this will also improve the chance of finding the gene(s) of interest and reduce the risk of uncovering an incidental finding. Accurate phenotyping is essential to avoid dilution of any signal from similar phenocopies. An example of this approach can be seen in a recent study of two families with familial episodic pain where genome-wide linkage scans with microsatellite markers were able to narrow down the region of interest to 7.8 Mb [2]. Later exome sequencing uncovered different missense mutations in the *SCN11A* gene in each of the families involved.

1.1.3 Exomes of Multiple Related Individuals for a Single Condition

The relationship distance between two directly related and similarly affected individuals is correlated with the power of the study. Every meiotic recombination event that has occurred between any two affected relatives (within a single family) that have been exome sequenced halves the number of DNA variants to sift through and halves the risk of uncovering an incidental finding (Fig. 1).

1.1.4 Exomes of Multiple Unrelated Individuals with Each Sequenced Individual Also Having a Pedigree Available for Subsequent Segregation Analysis, i.e., Combination of Subheadings 1.1.2 and 1.1.3

This is an ideal study design for gene discovery. Again, the more individuals available for analysis the greater becomes the power of the study and the lower the risk of identifying an incidental finding if the condition does not have a significant amount of genetic heterogeneity. On the other hand, it is possible to uncover a significant amount of genetic heterogeneity for a single condition with this study design if sufficient numbers of well-phenotyped families are available but the risk of uncovering an incidental finding increases with lesser genic overlap between affected individuals.

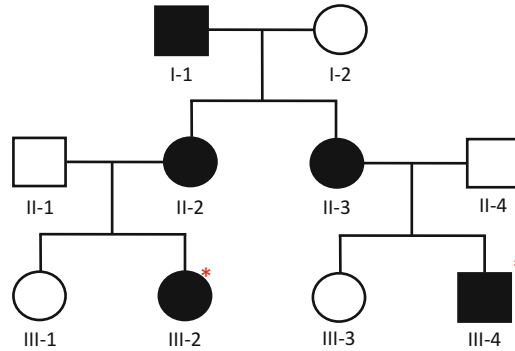


Fig. 1 Autosomal dominant pedigree with two affected participant's exomes analyzed. In this autosomal dominant condition with high-penetrance individuals III-2 and III-4 share 1/8th of their exome in common. Limiting the study's final DNA variant list to only the variants they have in common will reduce incidental finding risk to 1/8th and increase the focus on the likely candidate DNA variation (*see Note 2*)

1.1.5 Trio Exome Sequencing

Exome sequencing of an affected offspring and his/her two unaffected parents can efficiently identify either an autosomal recessive condition with both parents being heterozygous carriers with the affected offspring compound heterozygous or a de novo mutation in the offspring (*see Note 2*) for confounder of a possible de novo mutation—all de novo mutations must be validated by a second method in all members of the trio. This study design has an inherently low risk of uncovering an incidental finding as the bioinformatics filtering process will only examine rare homozygous mutations, compound heterozygous mutations, or de novo mutations. This necessarily eliminates the vast bulk of all DNA variations from the analysis. *See refs. 3, 4* for examples of exome trio studies that identified causal de novo mutations.

1.1.6 Somatic Mosaicism Detection by Exome Sequencing

It is possible to identify a somatic mosaic DNA variation by comparing exomes of affected with unaffected tissue. Cancer genotyping provides a useful model for this type of testing where a broad exome approach examines both the germline exome and subsequent somatic changes in the cancerous tissue. There is a high risk of detecting incidental findings associated with a familial basis for the cancer. This would have implications for blood relatives. A pure somatic mosaicism exome study by Lindhurst et al. [5] uncovered a rare activating mutation in *AKT1* as a common cause of Proteus syndrome. Bioinformatics filtering that focused on the differences between affected and normal tissue would have a very low risk of uncovering an incidental finding in this study design.

*1.1.7 Exome Sequencing
for Consanguinity/
Homozygosity Mapping
(for Recessive Disorders)*

This form of exome sequencing would specifically look for homozygosity of a rare DNA variant. If there is consanguinity in nearby generations then a relatively large number of rare homozygous variations may exist in an individual, thereby making the causal DNA variation difficult to track down. Homozygous variants that are difficult to interpret are likely to be found in this situation, and the incidental finding risk is potentially high but not easy to interpret. Conditions with a founder effect would come under this category of exome research. However, consanguinity may be distant, and the tracking down of a rare homozygous variant for a recessive condition becomes possible. The risks of an incidental finding in this case would depend on the degree of inbreeding of the study population.

**1.2 Inheritance
Pattern**

Part of the DNA variant filtering strategy will be dictated by the likely inheritance pattern of the phenotype under study. This will in turn affect the number of DNA variants to examine for their functional significance after the variant filtering is complete. Therefore, it influences the chance of incidental findings.

Careful scrutiny of the mode of inheritance from available pedigree(s) before embarking on an exome discovery project is prudent. The following is a list of some common inheritance patterns and their respective implications for the bioinformatics filtering strategy and hence the risk of finding an incidental finding.

*1.2.1 Mendelian
(Monogenic) Inheritance*

- Autosomal dominant inheritance will limit the exome analysis to heterozygous variants in chromosomes 1 to 22. X and Y chromosome DNA variants can be excluded from further analysis. This is the monogenic inheritance pattern with the greatest risk of uncovering an incidental finding.
- Autosomal recessive inheritance will necessarily limit the exome analysis to variants in genes that are either rare but homozygous or double heterozygous for a potentially function-altering DNA variant. This will significantly reduce the possibility of exposing an incidental finding.
- X-linked conditions will exclude from analysis over 90 % of the DNA variants that reside in the autosome. This will proportionally reduce the basic risk of finding an incidental finding by >90 %.
- Primary mitochondrial inheritance will be indicated by a maternal inheritance pattern although heteroplasmy and threshold effects can obscure this particular inheritance pattern. All autosome and sex chromosome-linked variants can be excluded in this case, significantly reducing the risk of an incidental finding. On the other hand, autosomally inherited,

nuclear encoded mitochondrial genes number >1,000, and from personal experience there is potentially a great degree of genetic heterogeneity possible for some specific mitochondrial related conditions. The risk of uncovering an incidental finding within a large cohort involving a potentially autosomal mitochondrial condition is real unless well managed through a bioinformatics functional analysis pipeline.

1.2.2 Complex Inheritance

Unravelling the sources of a complex inheritance pattern is a significant challenge for the genomic researcher. Exome sequencing can be both a blessing, because it simultaneously reveals an enormous amount of rare and common DNA variation, and a curse because the researcher is forced to scrutinize a substantial fraction of this. This “curse” also exposes the study to a greater risk of uncovering an incidental finding.

Multiple (>1) alleles can contribute to a quantitative trait in an additive way. They can also contribute in a non-quantitative combinatoric fashion; for example, epistasis and hypostasis can arise when specific rare and common digenic alleles interact in a common pathway. Both scenarios would require a significant number of well-phenotyped participant exomes to begin to decipher the causes of the inheritance pattern whether the condition is purely autosomal, sex linked, mitochondrial, or a combination of these.

Similarly, incomplete penetrance and variable expressivity also complicate exome interpretation potentially necessitating the analysis of a greater number of DNA variations, thus also exposing the study to a greater risk of uncovering an incidental finding.

1.3 Population Frequency of Specific DNA Variations

The debate about whether common or rare DNA variation is responsible for the majority of genetic disease is an important one and influences how the researcher approaches exome analysis. There are reasonable indications now that rare variation (<0.1 % in general population) is more important than was realized from the original HapMap-based common SNP investigations which formed the basis of many large genome-wide association studies (GWAS) [6–9]. Indeed, molecular genetic pathology (clinical) laboratories routinely see a broad spectrum of allelic heterogeneity in most studied disease genes with family-specific and novel mutations prevalent. Most published GWAS have been based on microarray SNP genotyping technology which interrogate common and uncommon but listed (*known*) SNPs. The fundamental (and debatable) assumption behind most GWAS has been that this non-rare DNA variation is commonly (statistically) responsible, or associated via linkage disequilibrium (LD) blocks, for common disease. While many broad associations have been linked to loci across the genome using the statistical power of often enormous GWAS patient cohorts, the results on the whole have not been predictive for the individual to the same extent as a rare segregating DNA variant.

Exome sequencing research unlocks this rare variation door in a way that microarray-based, huge cohort GWAS simply cannot.

A helpful plan for any exome-based DNA variation research is to focus on the novel and rare variation first and examine the more common variation secondarily, if necessary. This is also an important strategy to reduce the risk of spotlighting a gene unnecessarily that is known to be linked to disease.

Common DNA variation can be bioinformatically filtered out of further analysis using a DNA variant frequency filter based upon many different sources including the following:

- *The US National Heart, Lung, and Blood Institute (NHLBI) GO Exome Sequencing Project (ESP)*. At the time of writing, this resource had collated over 6,500 exomes (ESP6500 release) representing a population frequency call of >13,000 alleles per protein-coding nucleotide. This is an excellent source for filtering out common variation within protein-coding regions and splice sites. The only caveat being that the exomes in this population have a higher than normal frequency of heart, lung, and blood disorders. This should be taken into account if the exome research is examining a condition potentially linked to one of these biological processes in some way. An overview and access to this resource can be viewed at <http://evs.gs.washington.edu/EVS/>.
- *The 1000 Genome Project* (www.1000genomes.org/). This derives its DNA variation data from many broad racial groups and covers protein-coding, intronic, and intergenic regions. One of its main goals was to collate all variations of at least 1 % frequency in the populations studied. The data are accessible via links to dbSNP (www.ncbi.nlm.nih.gov/SNP/) and Ensembl (<http://ensembl.org/index.html>).
- *In-house*. A collection of control exomes derived from *exactly* the same exome sequencing and analysis pipeline as the study group being investigated is a valuable resource for both common and uncommon DNA variation. Most importantly, it is an essential resource for filtering out systematic false-positive DNA variant calls. The value of this filtering probably cannot be overstated as every next-generation sequencing (NGS) system and downstream exome DNA variant calling software algorithm *does* make idiosyncratic DNA variant calls that are novel, highly pathogenic looking, and completely false. These false calls are a function of a number of factors including the following:
 - The NGS raw sequencing accuracy: Most NGS systems will now call individual bases with a high degree of accuracy (>99.9 % per base accuracy), but as there are 6×10^9 bases in a diploid genome and approximately 9×10^7 bases in a diploid exome this small false-positive call rate becomes

important. Each NGS system has characteristic types of DNA sequence that the particular chemistry finds troublesome [10, 11].

- The read depth (average coverage): Some regions of exome have low read depth and as a consequence imperfectly call some DNA variation.
- The accuracy of the bioinformatics pipeline employed (both the initial basing calling and mapping steps as well as the final DNA variant calling step) does vary between software algorithms including whether a local realignment algorithm is employed or not.
- Paralogous and repetitive regions within the genome can cause systematic misalignments in the primary sequence mapping set.

It is essential to filter out DNA variant calls that incorrectly appear like possible incidental findings. A laboratory with many exomes, e.g., >100, already processed through the same pipeline is well placed to filter out efficiently these systematic false-positive calls.

1.4 “Hot Spot”

Genes

Genes differ in their respective tolerance to functional variation. For example, many immune-related genes necessarily contain a lot of variations. The HLA locus on chromosome 6 is a well-known “hot spot” of *rare* single-nucleotide variations (SNVs). Olfactory receptors are another group that recently acquired diminished evolutionary importance to the human race and are a common source of function-altering mutations. On the other end of the spectrum, some genes are completely intolerant of non-synonymous DNA variation. A recent paper by Petrovski et al. [12] systematically quantified the amount of variation in each gene using the ESP (6500 release) DNA variant data. A *Residual Variation Intolerance Score* for 16,900 of the 20,000 known protein-coding genes was derived, and the genes were ranked by a *Residual Variation Intolerance Score Percentile*. These data can be employed as a valuable filter whereby the researcher can initially limit the search to those genes found to have function-altering rare variation in genes that do not tolerate variation. Again, this strategy will also reduce the risk of uncovering an incidental finding by limiting the candidate gene list.

1.5 Racial Background of Participant and Control Exomes

The Human Reference Genome (Hg19) and ESP DNA variant dataset have a bias towards Caucasian genomic DNA sequence and variation, respectively. It is prudent to take this into account when collecting your control and affected cohorts for exome analysis. Two to three times as many DNA variants will remain after bioinformatics filtering from some non-Caucasian exomes.

1.6 Prevalence of Condition Under Study

It is worthwhile considering the background prevalence of the condition under study condition in both the research's affected group (which ideally is exclusively affected) *and* control group. A highly prevalent condition might involve the control exome group unintentionally.

1.7 Biological System(s) Involved in Study

Narrowing the research to the biological system(s) involved in the condition under study can considerably reduce the risk of uncovering an incidental finding. For example, if you are investigating a liver-specific problem then neuronal specific genes can be excluded, or if an inborn error of metabolism is suspected then the study will focus on metabolic enzymes and pathways and exclude unrelated genes. However, scientists must keep an open mind to an unexpected genetic cause. More specifically, if a restricted, methodically derived list of Human Genome Nomenclature Committee (HGNC) gene candidates for the research is available then the risk of uncovering a potential incidental finding can be largely eliminated. Finally, it is worthwhile noting that approximately two-thirds of all protein-coding genes have *not* had a specific genetic disease connected with them. This large fraction of genes cannot therefore be involved in any incidental finding back to the patient.

1.8 Practical and Ethical Considerations for the Return of Incidental Findings to Research Subjects

A large-scale analysis of human exomes for the frequency of pathogenic SNVs found that 3.4 % of European-ancestry ($n=500$) and 1.2 % of African-ancestry ($n=500$) exomes contained a high-penetrance medically actionable DNA variation in one of 114 genes selected by an expert panel to contain medically important genetic conditions [13]. They were assessed by reviewing the primary literature, although they were mostly (17 out of 22) initially identified by their listing in the Human Gene Mutation Database (HGMD). This may explain the bias towards European ancestry.

The American College of Medical Genetics and Genomics (ACMG) has provided recommendations for reporting incidental findings in *clinical* exomes. However, there have been ongoing discussions regarding a number of its recommendations [14]. Major concerns include (1) the non-certain clinical utility of some incidental findings and the possibility of doing more harm than good to the individual and relatives, (2) insufficient data on penetrance, (3) lack of resources to carry out specific reporting by certain clinical laboratories in genes they are not expert in, and (4) the lack of staffs to get through the workload.

Research laboratories are at a major disadvantage in that many are not involved in medical reporting of genetic findings often with little clinical genetics services available for pre- and post-test counselling. Confirmation of incidental findings in an accredited laboratory was highlighted as being potentially expensive with the source of funding uncertain [15]. However there is a consensus among

researchers that the return of incidental findings in some cases could be life saving and that there is a moral obligation to do so in these cases [15].

It is worthwhile highlighting that DNA sequencing and identification of DNA variation from initial sequence analysis is only an initial small step in the process of analyzing a *potentially* medically actionable finding. It is common for molecular genetic pathology laboratories to spend the majority of the time on any single DNA variation determining the confidence level of pathogenicity. This involves finding and assessing the primary and interpretative literature about the specific DNA variant, in silico analysis interpretation, background frequency rates, conservation and protein function studies related back to the individual amino acid/s involved, and family studies. At the end of this process, DNA variations may still be classified as *variants of unknown significance* despite the laboratory staffs being relative experts in the gene and its clinical implications.

For these collective reasons, it is recommended that the research exome investigation be geared towards specifically answering the targeted question or finding *the* gene of interest. This involves a well-structured exome study and analysis pipeline that has a minimal likelihood of inadvertently discovering a potential incidental finding. The method below describes typical sequential steps in exome research and identifies design factors in each that affect the chance of finding an answer to the research question and chance of uncovering an incidental finding.

2 Materials and Methods

Primary DNA sequence mapping of massively parallel DNA sequencing and variant calling are usually carried out on multi-nodal dedicated computers. Subsequent exome filtering may be carried out on a personal computer.

The following idealized analysis pipeline follows a series of sequential steps in the process of exome research for gene and DNA variation identification (Fig. 2). The exome research pipeline is split into five broad stages: (1) cohort production, (2) production of list of DNA variants, (3) filtering of DNA variants, (4) functional analysis of gene and mutation(s), and (5) segregation analysis. Not all steps are relevant for every type of exome research.

Step 1: Phenotyping. Accurate, objective phenotyping is essential for the *affected* cohort. Non-accurate alignment of individuals in the affected group will dilute the intersection gene signal (*see step 13*) and potentially expose the research to a larger probability of uncovering an incidental finding. The control group does not necessarily have to be well phenotyped if the condition under investigation is rare.

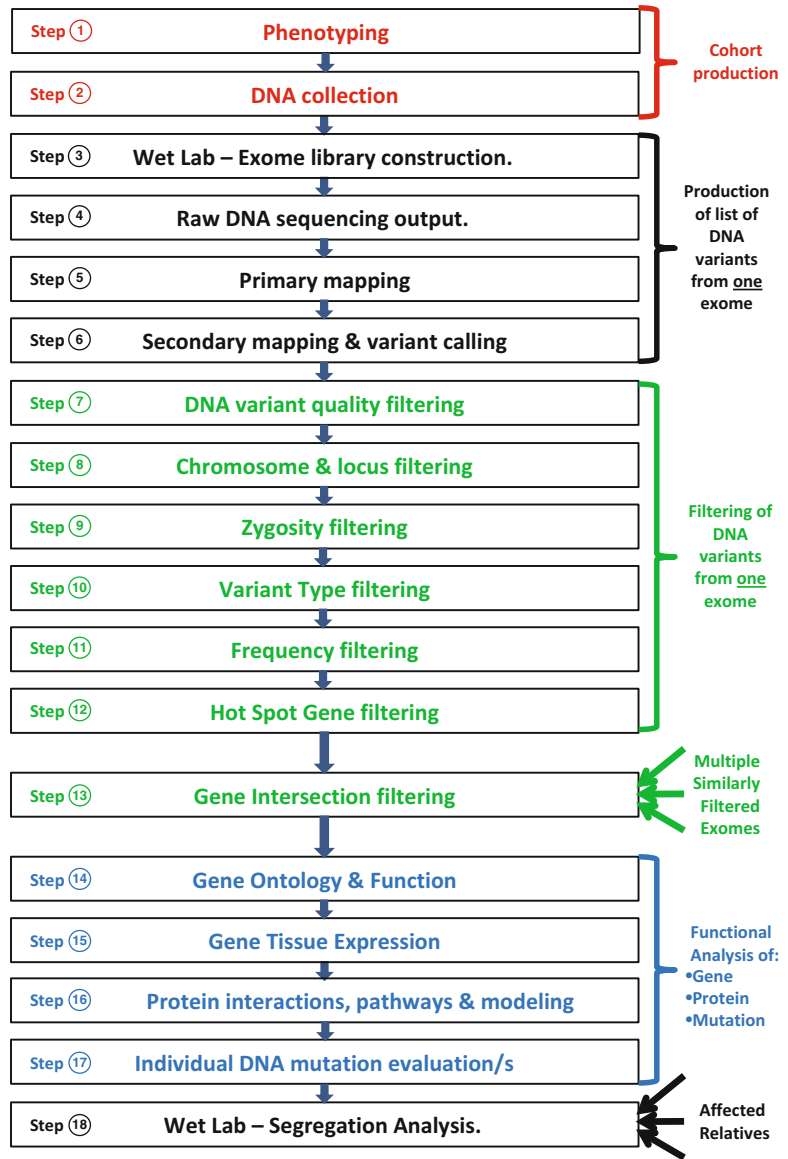


Fig. 2 Exome research pipeline (see explanation in text)

In general, the larger the size of the affected cohort, the greater the gene signal will be and the lower the chances will be of inadvertently uncovering an incidental finding. This effect is strongest with a Mendelian inheritance unless there is genetic heterogeneity involved in the disorder. The incidental finding risk will be inversely proportional to the number of affected participants if the disorder is monogenic. The incidental finding risk will be further increased by a more polygenic disorder. If a prior identified list of genes is to be investigated then the incidental finding risk may be minimal.

Step 2: DNA collection. From experience, both blood- and saliva-extracted DNA work well during the exome hybridization and capture process (*see Note 3*). This can allow the mailing of saliva samples and increase the size of the cohort.

Step 3: Exome library construction. An experienced laboratory is likely to produce a better technical NGS output. More efficient capture will produce greater on-target sequencing and greater depth of coverage, decreasing the false-positive DNA variant call rate and increasing the likelihood of finding the “needle in the haystack” causal DNA mutation. An average coverage of greater than $\times 40$ is required for calling most variants ($>90\%$). There is little value in increasing the coverage beyond $\times 100$.

Step 4: NGS output. Both of the major NGS systems currently in use produce high-quality raw DNA sequence. The author has experience with the SOLiD 5500 and in particular notes that its ligation-based chemistry provides a low rate of false-positive indels. False-positive indels can be bioinformatically filtered out if the researcher has access to a large normal control exome variant dataset derived from exactly the same sequencing and bioinformatics pipeline.

Step 5: Primary mapping. NGS short sequences are always mapped to the entire human reference genome (Hg19 at time of writing) first rather than the 1.5 % of the genome that an exome targets. This necessarily reduces the number of false-positive DNA variant calls due to misaligned paralogous sequences. Most software packages can also produce a mapping quality value that indicates the likelihood that a particular DNA variant was called from uniquely mappable sequence.

Step 6: Secondary mapping and variant calling. BED files (SOLiD systems) and Manifest files (Illumina systems) provide precise genomic coordinates to limit the variant calling to the exome capture regions (*see Note 4*). Local realignment algorithms are sometimes present in mapping software and can increase the accuracy of DNA variant calling, particularly in relation to small indels. Three types of variation can be sought:

- SNVs ($>95\%$ all variation), output is a .vcf file (variant calling file).
- Small insertions and deletions ($<5\%$ all variation), .vcf output.
- Large multi-exonic deletions and duplications (by normalized *relative* read depth).

Step 7: DNA variant quality filtering. Both a minimum mapping quality threshold and a minimum sequence read depth on both alleles for each variant including an acceptable allelic bias/ratio (*see Note 5*) will reduce the number of false-positive calls. Variants called by low (<3) specific sequence read *start site* counts can also be filtered to avoid PCR artefact (*see Note 6*). There is a significant

trade-off between sensitivity and specificity of DNA variant calling. If the DNA quality filtering is too low then too many false-positive DNA variant calls will be made, but if the DNA quality filtering is too stringent then it is possible to filter out the single DNA variant that the research is trying to find. Some empirical bioinformatics experimentation may be required before acceptable quality values are agreed on for the research.

Step 8: Chromosome and locus filtering. The risks of uncovering an incidental finding can be reduced by limiting the search to a specific chromosome or region previously identified by linkage (*see Note 4*). For example, if the inheritance pattern indicates an autosomal condition then filter the X and Y chromosomes. If the condition is X linked, filter chromosomes 1–22 and Y. If the condition is mitochondrial, filter everything except the 16 kb of the mitochondrial genome (*see Note 7*).

Step 9: Zygosity filtering. As outlined in Subheading 1, the inheritance pattern will dictate whether homozygous DNA variants can be filtered out. Be aware of potential consanguinity, inbreeding, and founder effects before filtering homozygous variants. Most recessive conditions will be compound heterozygous due to allelic heterogeneity unless the above inheritance modifiers are present.

Step 10: Variant-type filtering. Missense, nonsense, anti-nonsense, and canonical splice site mutations; frameshifts; in-frame deletions or insertions; and large deletions or insertions can be selected for. Keep in mind that synonymous amino acid DNA variants and deep intronic variants can be function altering, but as a first step these can be filtered out and later analyzed if the initial search is fruitless. An *initial scan* through candidate DNA variations for the condition under investigation can be limited to variations that are likely to be function altering such as nonsense and frameshift mutations near the N terminus.

Step 11: Frequency filtering. As summarized in Subheading 1, the frequency of a particular DNA variant in a background population can be the basis of a very good first-line filter when searching for the important single-DNA variation in research. A rare variant is generally more likely to be function altering than a common variant. This filter can be titrated up or down to include more or less candidate DNA variants depending on the question in mind (*see Table 1*). Looking at a shorter list of DNA variants by viewing only the novel changes first will reduce the risk of finding an incidental finding.

Step 12: “Hot spot” gene filtering. A “hot spot” gene list can be derived from the control exome data. The greater the size of the control dataset, the better the list of genes that commonly contain rare variation will be. The *Residual Variation Intolerance Score Percentile* from Petrovski [12] is a very useful ESP-derived rank of

Table 1

Number of variations by type and frequency from protein-coding regions (*per exome*) after NGS high-quality variation filtering up to the end of step 12 (derived from 70 exomes sequenced on a SOLiD 5500)

Average number of <i>potentially function altering</i> ^a variations seen <i>per individual exome</i> compared with the frequencies of those variations in the general population (value in bracket = 1 standard deviation)							
Frequency category (in general population)	Novel (Not seen in 1KGP or ESP)	<1 in 10,000 (Only 1 or 0 seen in ESP)	<1 in 1,000 (From ESP)	<1 in 200 (From 1KGP and ESP)	<1 in 100 (<1 %) (From 1 KGP and ESP)	<1 in 50 (<2 %) (From 1KGP and ESP)	<1 in 33 (<3 %) (From 1KGP and ESP)
Type of variation							
SNVs—heterozygous	76 (12)	93 (14)	145 (25)	259 (32)	330 (51)	401 (83)	437 (90)
Small indels ^b	16.5 (6.1)						
Large multi-exonic deletions ^c	1.98 (0.94)						

Note: From experience, there are two main causes of a significantly increased numbers of rare variants seen in an individual: (1) Non-Caucasian racial group as Hg19 is based largely on a Caucasian background and (2) sequencing chemistry artefact

1KGP 1000 Genome Project, ESP NHLBI Exome Sequencing Project (6500 exomes or >13,000 allele count for most *protein-coding bases* in database at the time of writing)

^aPotentially function altering = missense, nonsense, anti-nonsense, canonical splice site mutations; frameshifts; in-frame deletions or insertions; and large deletions or insertions. This table *excludes* synonymous mutations and nonprotein-coding mutations (which are both sometimes function altering)

^bPopulation frequencies of small indels can be sought from a control exome set that has been derived from the same DNA sequencing and bioinformatics pipeline. Individual NGS platforms and pipelines may perform differently, calling small indels at different rates. Small indel frequency datasets are not readily accessible from the 1KGP or the ESP. The value listed represents the average number of novel small insertions or deletions seen after bioinformatics filtering. From experience, a number of small indels in an individual exome with many standard deviations higher than the average of 16.5, e.g., >100, can actually indicate a mismatch repair gene defect, and therefore an incidental finding involving a possible increased familial cancer risk should be investigated

^cNo frequency data available for multi-exonic deletions and duplications from 1KGP or ESP

each gene's tolerance of function-altering variation and could quickly highlight a single-DNA variation residing in a gene that does not tolerate variation.

At the end of the DNA variant filtering process, the number of variants that can be expected to remain *per exome* is listed in Table 1. This gives the researcher a raw level of expected DNA variation for a single individual within a research project. As can be seen from the table, the number of DNA variants per individual is strongly influenced by the variation frequency filter (**step 11**). The DNA variant list can be <100 per individual if only high-quality, heterozygous, novel, or rare DNA variations (<1 in 10,000 alleles) are not filtered. Hence, it is important to consider this particular filter carefully.

Step 13: Gene intersection filtering. This is the first step in the exome analysis pipeline where multiple exome variant lists can be combined to potentially find an answer to the research question.

If the research's *affected* cohort has >1 individual then searching for genic overlap is a powerful way of finding a new gene/phenotype relationship. It is the level of the *gene* that this sorting should occur on, as allelic heterogeneity is common. Be aware of possible pleiotropy as a single gene can be involved in more than one type of related or unrelated disease depending on the particular exact DNA variation and other environmental and epistatic factors.

The larger the number of affected exomes at this stage, the greater will be the chance of finding an answer to the research question and the risk of uncovering an incidental finding will be less.

Step 14: Gene ontology and function. A candidate gene's function can be initially searched at the GeneCards website (www.genecards.org/index.shtml). An Entrez gene summary on function can be found here together with a host of accompanying information. BioGPS (<http://biogps.org/#goto=search>) allows the researcher to look up data on gene function and expression. Links to gene ontology terms including molecular function, biological process, and cellular component controlled ontological nomenclature can give helpful initial hints to a gene's relevance to the research question. PubMed (www.ncbi.nlm.nih.gov/pubmed/) is a well-known medical publication retrieval database for searching and examining more closely the known function of the gene(s).

Step 15: Gene tissue expression. Obviously, a gene should be expressed in the tissue type of interest. Both BioGPS and GeneCards (see above) also give good-quality tissue expression data and graphs from different sources.

Step 16: Protein interactions, pathways, and modeling. String (<http://string-db.org>) is a portal for the access of evidence-based protein interactions based on a number of lines of support including experiments, databases, text mining, and homology modeling.

The protein data bank (<http://www.pdb.org>) gives access to protein function and 3D structure modeling from the protein workshop. Individual amino acid locations within atomic structure-determined proteins can be accessed from here.

Step 17: Evaluation of individual DNA variants. This is a complicated and potentially difficult step in the whole process. Determining the functional effect of a specific DNA variant relies on multiple lines of evidence as briefly outlined in Subheading 1 and include the following:

- Mutation databases (general or locus specific) which often list DNA variants that do not have a good line of evidence to suggest that they are indeed function altering (see Chapter 15).

- In silico programs such as SIFT, Polyphen, and Grantham deviation/variation which rely on amino acid conservation information and consequently have significant idiosyncratic sensitivity and specificity issues (these in silico programs should not be used, in my opinion, as a front-line filter in exome research as they run the risk of over-calling many variants' effect and even incorrectly filtering the very DNA variation sought in the entire research project) (*see* Chapters 13 and 14).
- Published articles (both primary and interpretive).
- Frequency in the general population.

The final **step 18** outlined next is highly recommended if family phenotyping and DNA collections are available.

Step 18: Segregation analysis. As demonstrated in Subheading 1 (Fig. 1), linkage of a candidate causal DNA variation with other affected family members is a potent way of demonstrating causality. The utility of a segregation analysis is exponentially (power of 2) powerful depending on the number and relationship distance of affected relatives to determine a DNA variant's linkage to a phenotype. However, be mindful of linked chromosomal regions that if close may segregate with meiotic divisions with the true causal DNA variant. Non-affected family members are also useful if there is a low degree of non-penetrance in the condition under study. This is also one of the most useful ways to focus on the causal DNA variant and exclude other variants from analysis, thereby reducing the risk of uncovering an incidental finding in research exome analysis.

3 Incidental Findings in the Medical Laboratory

Various options to deal with incidental findings in the *research* environment have been discussed. Most are also applicable in the *medical* testing laboratory which increasingly will generate incidental findings as the evolution from genetics to genomics continues. A key consideration in medical testing is what the patient understands about the genomics test being undertaken, i.e., the consent process, as this type of testing will not usually be overseen by a research ethics committee. Consent is presently the subject of much debate particularly when genomics-based approaches move beyond targeted diagnostic-type tests to include broader (screening) tests. The latter is more likely to generate incidental findings of relevance to the health and well-being of the person being tested as well as family members.

4 Notes

1. Most exome capture using hybridization kits from different suppliers are based on the consensus protein-coding sequences of the human reference genome which is about 38 Mb in size. However, many kits now also include extra regions of the genome that are known to be transcribed but not involve proteins. They include many defined noncoding RNAs (ncRNAs) including microRNAs, small nucleolar RNAs, small nuclear RNAs, long ncRNAs, and others. The list of ncRNAs is large and growing. Some kits have over 55 Mb of target sequence as a consequence. Analyzing the functional significance of DNA variations of ncRNAs is challenging, but a good study design would involve a number of unrelated participant exomes and the DNA variant filtering strategy should aim to show locus proximity of rare variants across a number of participants in a particular ncRNA. Functional information is difficult to determine for ncRNAs, so the greater the number of participants in the study, the greater the probability of showing statistical significance of a link.
2. Exome sequencing depth of coverage is not uniform across all target regions. Homologous and repetitive regions as well as regions of low sequence complexity and extreme GC or AT content can affect the capture hybridization or the accuracy for primary mapping of the short sequence reads to the genome. For these reasons there are certain exons that frequently have poor coverage and therefore do not call all DNA variants present. Always bear in mind that exome studies may simply not sequence or call a particular causal DNA variant despite excellent study design.
3. While DNA extraction from saliva kits is known to contain a level of bacterial DNA contamination, the exome capture process and primary mapping of sequences to a human reference genome eliminate nonspecific sequences from generating false-positive variant calls. Nevertheless, it is advisable that the participant giving a saliva sample not eat meat products prior to donation!
4. BED or manifest files can specify limited genomic coordinates that limit the DNA variant calling to much more narrow genomic regions involved in genic pathways known to be associated with the disease in question. This can potentially eliminate or greatly reduce the risk of uncovering an incidental finding by not calling DNA variants outside the regions of interest.

5. True DNA variant calls should have a reasonably even distribution of allele counts on the reference and variant bases. A rule of thumb for SNVs is that if either allele has greater than triple the read count of the other the DNA variant may be a false-positive call. All important DNA variants including incidental findings must be validated by a second method such as Sanger sequencing for SNVs and small indels or Gap PCR for large deletions. Small indels do have a greater allelic bias away from the variant allele due to capture hybridization bias. An increased capture probe density (kits differ on this parameter) will decrease this allelic bias.
6. Allele start sites refer to an exact nucleotide coordinate on which a single DNA sequence begins. It is important to have more than one start site on both the forward and reverse sequence reads as the multiple PCR steps during library construction can falsely increase the number of reads calling a particular variant. This artefact will be apparent by viewing a highly repeated start site in one read direction on the Broad Institute's IGV software.
7. Although most exome capture kits do not have any mitochondrial genome-specific probes, the mitochondrial genome outnumbers the nuclear genome hugely so that off-target capture of the mitochondrial genome is usually present and can be seen if a mitochondrial chromosome-specific BED or manifest file is used in **step 6**.

References

1. Lupski JR, Reid JG, Gonzaga-Jauregui C et al (2010) Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med* 362:1181–1191
2. Zhang XY, Wen J, Yang W et al (2013) Gain-of-Function Mutations in SCN11A Cause Familial Episodic Pain. *Am J Hum Genet* 93:957–966
3. Suls A, Jaehn JA, Kecskes A et al (2013) De Novo Loss-of-Function Mutations in CHD2 Cause a Fever-Sensitive Myoclonic Epileptic Encephalopathy Sharing Features with Dravet Syndrome. *Am J Hum Genet* 93:967–975
4. O'Roak BJ, Deriziotis P, Lee C et al (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. *Nat Genet* 43:585–589
5. Lindhurst MJ, Sapp JC, Teer JK et al (2011) A mosaic activating mutation in AKT1 associated with the Proteus syndrome. *N Engl J Med* 365:611–619
6. Schork NJ, Murray SS, Frazer KA et al (2009) Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19: 212–219
7. Do R, Kathiresan S, Abecasis GR (2012) Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* 21:R1–9
8. Tennessen JA, Bigham AW, O'Connor TD et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69
9. Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336:740–743
10. Liu L, Li Y, Li S et al (2012) Comparison of next-generation sequencing systems. *J Biomed Biotechnol* 2012:251364
11. 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
12. Petrovski S, Wang Q, Heinzen EL et al (2013) Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9:e1003709

13. Dorschner MO, Amendola LM, Turner EH et al (2013) Actionable, pathogenic incidental findings in 1,000 participants' exomes. *Am J Hum Genet* 93:631–640
14. Green RC, Berg JS, Grody WW et al (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med* 15:565–574
15. Klitzman R, Appelbaum PS, Fyer A et al (2013) Researchers' views on return of incidental genomic research results: qualitative and quantitative findings. *Genet Med* 15:888–895

Chapter 13

Approaches for Classifying DNA Variants Found by Sanger Sequencing in a Medical Genetics Laboratory

Pak Leng Cheong and Melody Caramins

Abstract

Diagnostic applications of DNA sequencing technologies present a powerful tool for the clinical management of patients. Applications range from better diagnostic classification to identification of therapeutic options, prediction of drug response and toxicity, and carrier testing. Although the advent of massively parallel sequencing technologies has increased the complexity of clinical interpretation of sequence variants by an order of magnitude, the annotation and interpretation of the clinical effects of identified genomic variants remain a challenge regardless of the sequencing technologies used to identify them. Here, we survey methodologies which assist in the diagnostic classification of DNA variants and propose a practical decision analytic protocol to assist in the classification of sequencing variants in a clinical setting. The methods include database queries, software tools for protein consequence, evolutionary conservation and pathogenicity prediction, familial segregation, case-control studies, and literature review. These methods are deliberately pragmatic as diagnostic constraints of clinically useful turnaround times generally preclude obtaining evidence from in vivo or in vitro functional experiments for variant assessment. Clinical considerations require that variant classification is stringent and rigorous, as misinterpretation may lead to inappropriate clinical consequences; thus, multiple parameters and lines of evidence are considered to determine potential biological significance.

Key words Clinical annotation, Databases, Diagnostics, Pathogenicity prediction, Sequencing, Variants

Abbreviations

HGMD	Human Gene Mutation Database
NCBI	National Center for Biotechnology Information
NG	Genomic
NM	mRNA
NP	Protein from RefSeq database
VUS	Variant of unknown significance

1 Introduction

The increasingly higher throughput and lower cost of sequencing technologies have facilitated routine mutation identification and characterization in medical laboratories. The ready availability of sequence data has expanded the incorporation of these results into clinical care and decision making. Consequently, the discovery of variants of unknown significance (VUS) is a common occurrence in clinical sequencing, regardless of the sequencing methodology utilized. The focus of a genetic diagnosis in response to a clinical question increasingly revolves around the key challenge of accurate, reliable, and reproducible variant annotation and interpretation.

A review of the concepts of analytical validity, clinical utility, and clinical validity is useful in this context. *Analytic validity* is defined as the process by which the performance of a test system is measured and assessed and often involves addressing inherent issues of quality control, robustness, accuracy, reliability, efficiency, and traceability. In this chapter, we assume that identified DNA variants have been detected in an analytically valid manner and do not directly address this aspect. *Clinical validity* refers to the accuracy with which a test predicts the presence or the absence of the phenotype or, stated as a question, *how accurately does the sequencing result predict the clinical phenotype*. *Clinical utility* of a sequencing test is the capacity of the result to rule a diagnosis in or out and thus make a decision to adopt or to reject a therapeutic course of action possible. Or *does the sequencing result allow the recommendation of a clinical course of action*? Both clinical validity and clinical utility are of great importance when interpreting variants in a diagnostic environment.

The context of Sanger sequencing variant annotation and interpretation usually involves addressing a specific biological hypothesis, which can often be phrased as the following: “Could the patients’ signs and symptoms be the result of the detected variant(s) in this particular gene(s)?” Generally, Sanger sequencing will consider only a handful of genes to address this question, and therefore the hypothesis is tested only once or a handful of times. This contrasts with whole-exome or whole-genome sequencing, where multiple testing returns much larger numbers of variants and therefore requires greater interpretive caution due to the increased likelihood of a type I (false positive) error or the risk of increasing a type II error (false negative). Therefore, it is important that patients and physicians who order DNA genetic tests are aware of these limitations.

The diagnostic environment is also necessarily pragmatic; the need for clinically useful turnaround times precludes the ability to develop functional assays to assess directly biological effects of variants. Almost all assessments must be made more or less bioinformatically (in silico) by reference to literature databases, mutation

databases, and variant prediction software. Impressive research efforts such as the Duke University Task Force for Neonatal Genomics, where functional characterization of variants occurs in near real time, thus enabling results to be returned in a time frame which is useful in a neonatal intensive care setting, offer an interesting glimpse to the future.

In this chapter, we refer to many published international best practice guidelines on variant interpretation and classification. The reader is also encouraged to seek further information by consulting local best practice guidelines and interpretation standards, where available.

2 Materials

In order to illustrate clearly the approaches taken, we will use variants in the *CPOX*, *LDLR*, and *BRCA2* genes as examples in a process utilizing resources which can generally be classified into three groups: (1) databases (gene/locus specific and more generic), (2) browsers, and (3) tools.

Most of the resources are available online (*see* Table 1). Each step in variant assessment may use one or more of these. The lack

Table 1
Web resources used in variant classification

Tool	URL address
Align GVDG	http://agvgd.iarc.fr/
IARC breast cancer database	http://brca.iarc.fr/PRIORS/index.php
Breast Cancer Information Core	http://research.nhgri.nih.gov/bic/
COSMIC	www.sanger.ac.uk/genetics/CGP/cosmic/
Gene Ontology	www.geneontology.org
HGMD	www.biobase-international.com/product/hgmd
HGVS recommendations for the description of sequence variants	www.hgvs.org/mutnomen/recs.html
MutPred	http://mutpred.mutdb.org/about.html
NCBI Gene	www.ncbi.nlm.nih.gov/gene
CD-Search on NCBI Conserved Domains Database	www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
NCBI dbSNP	www.ncbi.nlm.nih.gov/projects/SNP/
PFAM	http://pfam.sanger.ac.uk

(continued)

Table 1
(continued)

Tool	URL address
PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/
SIFT	http://sift.jcvi.org
SIFT BLink	http://sift.jcvi.org/www/SIFT_BLink_submit.html
SMART	http://smart.embl.de
SNPeffect	http://snpeffect.switchlab.org
SNPs&GO	http://snps.uib.es/snps-and-go/
T-COFFEE regular	www.igs.cnrs-mrs.fr/Tcoffee/tcoffee.cgi/index.cgi?stage1=1 &daction=TCOFFEE::Regular
UCSC Genome Browser	http://genome.ucsc.edu/
UniProt (for Swiss-Prot protein code)	www.uniprot.org
UCL <i>LDLR</i> FH database	www.ucl.ac.uk/ldlr/LOVDv.1.1.0/index.php?select_db=LDLR

of standardization and differences in database update and review dates can sometimes be a source of interpretive conflict. In order to overcome this, the Human Variome Project (HVP) is considering the option of accreditation of databases for use in the clinical setting in the future.

3 Methods

There are several steps in establishing the potential clinical and biological significance of a given variant. The initial step in this process hinges on accurate annotation. This facilitates other downstream steps, including in silico predictions, obtaining population frequency data, and literature searches for functional information. A variant in the *CPOX* gene will be used as an example to illustrate these steps, followed by a comparative discussion on how this variant and two other variants in *LDLR* and *BRCA2* are classified based on the evidence of pathogenicity in a clinical setting.

3.1 Annotation of Variant and Visualization of Genomic Context

One of the first annotation steps typically involves contextualizing a particular genomic position within the sequence of known gene/s, transcript/s, or regulatory regions. This facilitates the process of prediction of likely variant effects (if any) on the resulting protein.

The current standard nomenclature system used in the annotation of variants has been developed by the Human Genome Variation Society (HGVS) for the description of genetic variants [1]. Older nomenclature systems may still be in historical use, referred

to in the literature or in databases; this should be noted during review by the user with appropriate caution. Annotation of variants from Sanger sequencing traces can be undertaken using commercially available packages (Mutation Surveyor® from Softgenetics is one such example) or manually by using freely available software. In either instance, a reference sequence is generally a key initial requirement.

It is also important to note at this stage that available software (both commercial and free) will frequently come with a disclaimer that the product should only be used for research purposes and not clinical decision making. This should be recognized as it will mean that a formal evaluation is required to validate the software prior to its use for clinical testing.

Curated reference sequences can be obtained by searching for the gene name, e.g., *CPOX* for coproporphyrinogen-III oxidase, at the National Center for Biotechnology Information (NCBI) Gene website. Genomic, mRNA, and protein sequence accessions are listed with the prefixes NG, NM, and NP, respectively, under the *NCBI Reference Sequences* (RefSeq) section (Fig. 1). For mRNAs,

NCBI Reference Sequences (RefSeq)

[RefSeqs maintained independently of Annotated Genomes](#)

These reference sequences exist independently of genome builds. [Explain](#)

Genomic

NG_015994.1 RefSeqGene

Range	5001..19166
Download	GenBank , FASTA , Sequence Viewer (Graphics)

mRNA and Protein(s)

[NM_000097.5](#) → [NP_000088.3](#) **coproporphyrinogen-III oxidase, mitochondrial precursor**

[See proteins identical to NP_000088.3](#)

Status: REVIEWED

Source sequence(s)	AK290140 , BC017210 , BC023551
Consensus CDS	CCDS2932.1
UniProtKB/Swiss-Prot	P36551
Related	ENSP00000264193 , OTTHUMP00000217368 , ENST00000264193 , OTTHUMT00000358900

Conserved Domains (1) [summary](#)

	pfam01218	Coprogen_oxidase; Coproporphyrinogen III oxidase
	Location:150 – 453	
	Blast Score: 1341	

Fig. 1 A view on NCBI Reference Sequences (RefSeq). The link to Sequence Viewer (Graphics) is *arrowed*

Homo sapiens coproporphyrinogen oxidase (CPOX), RefSeqGene on chromosome 3

NCBI Reference Sequence: NG_015994.1
[GenBank](#) [FASTA](#)

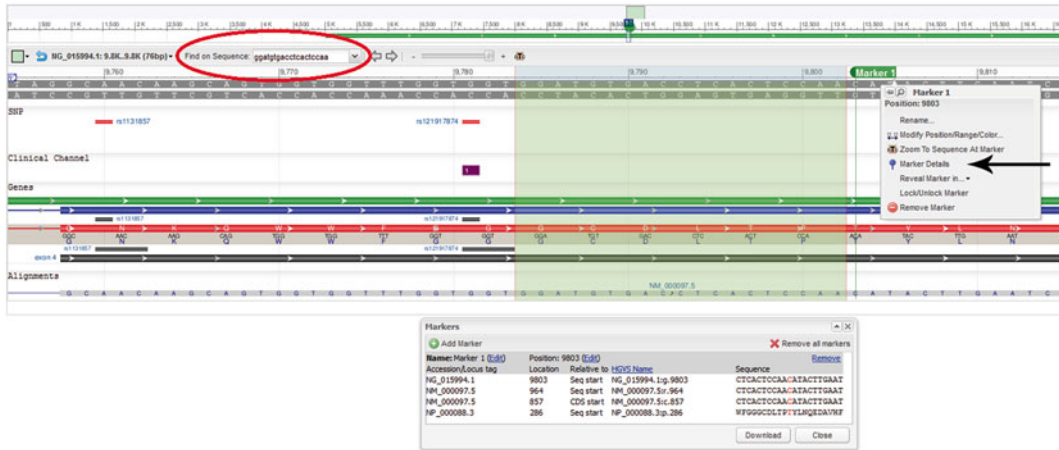


Fig. 2 Searching for sequence and HGVS nomenclature on NCBI Sequence Viewer

the longest transcript is usually used as the reference transcript, although in some cases (for example, the *FECH* gene), the longest transcript may not be the predominant transcript in vivo. Arriving at that site page, the GenBank or FASTA format genomic reference sequences for the gene of interest can be downloaded under the NCBI Reference Sequences tab. A graphical view of the gene is also available by clicking the link “Sequencer Viewer (Graphics).”

One method of locating a variant from the reference sequence involves searching the surrounding sequence on the “Find on Sequence” panel in the Sequencer Viewer (circled at the top of Fig. 2) by entering the sequence into this search box. In our example, for the sequence string containing the C>A change in exon 4 of the *CPOX* gene (...GGATGTGACCTCACTCCAA(C/A)ATACTTGAA...), enter the immediately adjacent sequence “GGATGTGACCTCACTCCAA” in the Sequence Viewer search box. This will pinpoint the region adjacent to the C>A change, and a marker can then be created at the SNP by right clicking on the nucleotide and selecting “Set New Marker At Position.” Once the marker is set (Marker 1 in this case), move the cursor to the “Marker 1” label and select “Marker Details” (highlighted by arrow in Fig. 2) to show the HGVS nomenclature of this variant in the fourth column. The amino acid change (ACA>AAA; Thr>Lys) can also be deduced. HGVS nomenclature for this variant is NM_000097.5:c.857C>A or NP_000088.3:p.Thr286Lys depending on whether the coding DNA sequence or amino acid change is emphasized.

Once annotation is completed, the variant needs to be evaluated further according to its sequence context and location. In the *CPOX* example above, there is a non-synonymous variant within the coding region (Table 2).

Table 2
For coding variants the following should also be considered when evaluating biological effects

Type of variant and potential effects	Additional points to consider
Frameshift—insertion or deletion	Is there an alteration in reading frame?
Missense—stop-gain/nonsense (substitution resulting in a stop codon)	In some instances nonsense mutations may not have functional significance, such as the p.Lys3326Ter variant in <i>BRCA2</i> , arising from an NM_000059.3:c.9976A>T substitution which results in a stop codon and loss of the final 93 amino acids of the <i>BRCA2</i> protein. This variant has a reported allele frequency of 0.8 % in some populations and is not considered to be clinically significant [5]
Missense—substitution resulting in loss of stop codon	For example Hb Constant Spring (p.Ter143Gln in <i>HBA2</i>) resulting from a stop-loss mutation leading to a lengthened peptide [6]
Insertion/deletion not causing a frameshift	Caution is advised, e.g., in the <i>LDLR</i> c.2397_2405delCGTCTTCCT in-frame deletion. This deletion is interesting in that it has no or little effect per se in vitro but becomes functional when found in cis in combination with a non-synonymous variant p.Asn543His [7]
Non-synonymous single-nucleotide variant (SNV) or synonymous SNV	Synonymous SNV can sometimes be pathogenic by affecting splicing. As an example, a critical yet translationally silent C>T variant at position 6 in <i>SMN2</i> exon 7 compromises its splicing, causing most of the <i>SMN2</i> mRNA (~80 %) to lack exon 7 (<i>SMNΔ7</i>). The resulting unstable molecule is rapidly degraded, leaving patients with SMN deficiency, the degree of which correlates with clinical severity of spinal muscular atrophy [8]

3.2 *In Silico Analysis of Annotated Missense Variants*

There are several considerations in assessing whether a missense variant is likely to be pathogenic or nonpathogenic.

3.2.1 *Location of Variant in Relation to the Transcript*

When evaluating biological significance, the following list presents some general considerations. In the absence of other evidence, these are listed from more to less likely predictive of functional effects:

1. Coding region variants—See Table 2.
2. Invariant splice sites—Coding nucleotides close to the exon-intron boundaries may not only affect amino acid sequence but also splicing; the first two nucleotides at the start (donor site) or the end (acceptor site) of introns are invariable in 98.71 % of genes—these are called canonical dinucleotides, and mutation to these nucleotides is invariably associated with alternative splicing effects [2].
3. 5' or 3' untranslated region (UTR) of a transcript—Variants in these regions *may* have influence on gene expression.
4. Noncoding exon—In some genes not all exons are transcribed, and this alternative transcription may be tissue specific.

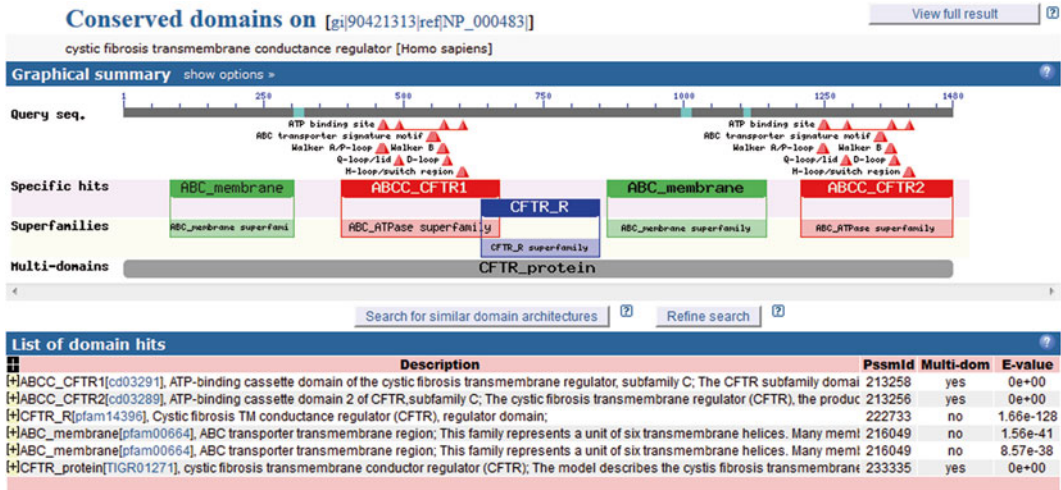


Fig. 3 Conserved protein domain of CFTR on NCBI Conserved Domains Database

For example, exon 1 and part of exon 2 of *HMBS* are transcribed in non-erythroid tissues but not in erythroid tissue [3].

- Intronic—Intronic variants outside the canonical splice site can also affect splicing. As an example, the variant NM_000140.3:c.315-48T>C, in intron 3 of *FECH*, promotes the use of a cryptic acceptor site, resulting in an aberrant transcript with a premature stop codon [4].
- Upstream or downstream of transcript start site: An *arbitrary* distance of 500–1,000 bp may be used by some laboratories.

3.2.2 Functional Domains

An important consideration for predicting the functional significance of a variant is its position within known protein domains, such as those defined by PFAM [9], SMART [10, 11], or other domain classification approaches. Conserved residues within functional domains are an indicator of negative evolutionary selection and so considered to provide some indirect evidence that changes will affect protein function. Conserved Domains Database (CDD) on NCBI [12–14] is a curated database that incorporates information from such sources. The CD-Search tool can align a given accession, GenoInfo Identifier (GI) number, or FASTA protein sequence to known domains in the database. An example using cystic fibrosis transmembrane conductance regulator (CFTR) protein sequence is provided in Fig. 3. The two transmembrane domains (ABC_membrane), two ATPase domains (ABCC_CFTR1 and 2), and R domain are shown with conserved amino acids within these domains marked in triangles. It is possible to identify whether variants of interest lie within these domains and if the amino acid of interest is conserved.

3.2.3 *In Silico Prediction Strategies*

Where no experimental data are available, the effect of a missense variant on protein function may be predicted by using various *in silico* tools. These predictive tools are generally based on two principles outlined below and should be used very cautiously, especially if predictions are not supported by additional evidence. Examples include the following:

- *Analyses based on evolutionary conservation of the nucleotide or amino acid.* As highlighted, negative evolutionary selection is an important indicator of functional significance of residues. By aligning nucleotide and amino acid sequences of orthologs from different species, the divergence of these residues and implication for putative functional effects may be deduced. Examples of these tools include phyloP, phastCons, SIFT, Align GVGD, Mutation Assessor, PANTHER, and MAPP (detailed discussion in Tavtigian et al. [15]).
- *Structural and biophysical property-based analyses.* This involves an analysis of the difference in biophysical properties between the reference and variant amino acid and predicting the probability that the resulting change will significantly affect protein structure. Examples of these tools include PolyPhen-2, SNPeffect (incorporating FoldX which is a protein stability-based prediction), and LS-SNP/PDB.

Some tools (for example, Align GVGD and SNPs&GO) use a combination of both strategies and may include supervised machine learning to improve their predictions (Table 3).

3.2.4 *Examples of In Silico Prediction Tools*

As mentioned above, the following *in silico* prediction tools utilize three broad strategies/methodologies: (1) evolutionary conservation, (2) structural and biophysical properties, and (3) machine learning.

- *PhyloP and phastCons.* PhyloP (phylogenetic *P*-value) and phastCons are two phylogenetic scoring systems which quantitatively measure evolutionary conservation [16, 17]. PhyloP scores are based on a measure of conservation at the level of individual nucleotides and are calculated as $-\log P$ -values, where a positive score indicates conservation. PhastCons relies on identifying elements (“runs”) of conserved sites, with scores ranging between 0 and 1, representing the probability of negative selection. These scores are integrated into the University of California Santa Cruz (UCSC) Genome Browser under the conservation track (Fig. 4). PhyloP and phastCons scores can easily be visualized by changing the settings in the Conservation Track to include them (Fig. 5).
- *SIFT.* Another commonly used tool based on sequence homology is Sort Intolerant From Tolerant (SIFT) [18–21], which comes with a user protocol [22]. One advantage of SIFT is

Table 3
Examples of some classifications used for DNA variants

ACMG classification	IARC classification (with probability of the variant being pathogenic)
Variant previously reported and is a recognized cause of the disorder	Class 1—not pathogenic (<0.1 %)
Variant previously unreported and is expected to cause the disorder	Class 2—likely not pathogenic (likelihood of pathogenicity 0.1–5 %)
Variant previously unreported and may or may not be causative of the disorder	Class 3—uncertain (5–94.9 %)
Variant previously unreported and is probably not causative of disease	Class 4—likely pathogenic (likelihood of pathogenicity 95–99 %)
Variant previously reported and is a recognized natural variant	Class 5—pathogenic (>99 %)
Variant is not known or expected to be causative of disease but is found to be associated with a clinical presentation, e.g., variants associated to particular disease from genome-wide association studies or modifier genes	

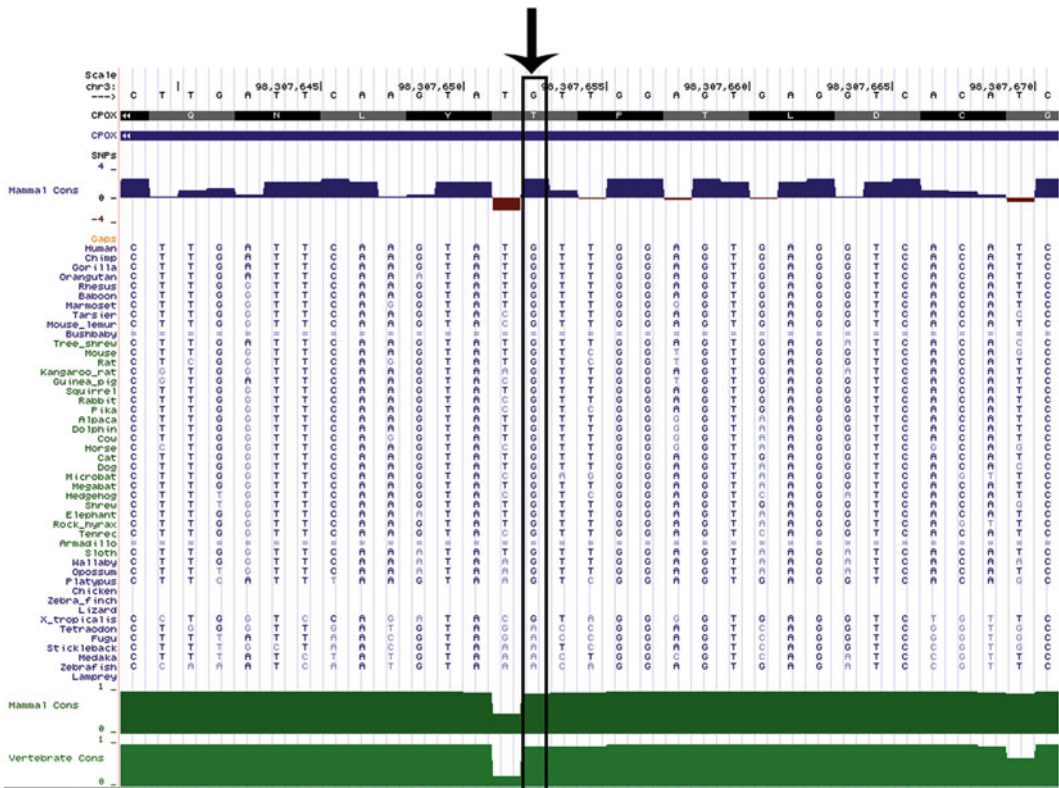


Fig. 4 PhyloP and phastCons scores in UCSC Genome Browser. The nucleotide change leading to CPOX p.Thr286Lys (arrowed) is highly conserved as indicated by phyloP (blue bar at top) and phastCons (green bars at the bottom). This is in contrast with the adjacent nucleotide where the negative phyloP (in red) and low phastCons score indicate accelerated evolution (Color figure online)

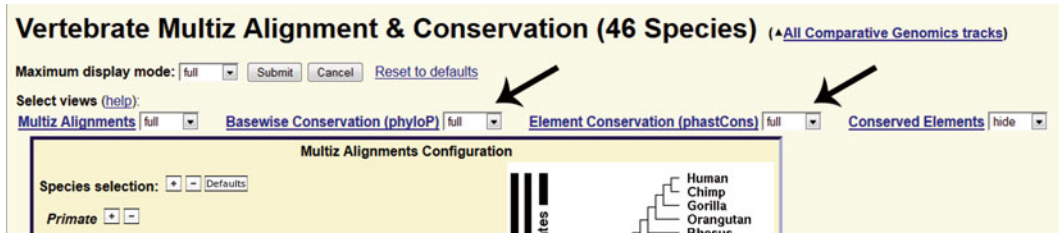


Fig. 5 Conservation track settings in UCSC Genome Browser. Click on “Conservation” under Comparative Genomics in the UCSC Genome Browser to adjust conservation track settings. Select “full” for Basewise Conservation (phyloP) and Element Conservation (phastCons) to display these scores in the Genome Browser (arrows)

that it accepts various input formats (e.g., Ensembl protein transcript ID, NCBI GI number, protein FASTA sequence, or RefSeq ID from dbSNP if it is a known SNP) for single-protein or -batch analyses. Users can either allow SIFT to build a multiple sequence alignment or they can submit their own. A SIFT score based on normalized probability of all 20 amino acids appearing in that particular position is calculated, and SIFT will “call” the variation damaging if the score lies below a threshold (predefined at 0.05). It also returns a median sequence conservation score which ranges from 0 (all 20 amino acid substitutions have been observed in multiple sequence alignment at the position) to 4.32 (where only one amino acid is observed at that position). Although a score of >3.25 would ordinarily indicate high conservation, if too few organisms are considered in the alignment, this may simply be reflective of this lack of diversity. Ideally, representation should be as broad as possible, including species from all vertebrate groups, e.g., Mammalia, Primates, Aves, Amphibia, and Reptilia. In our example, we use SIFT BLink, a rapid version of SIFT, as it runs analysis with pre-computed multiple sequence alignment from BLAST search. The GI number for CPOX (41393599) was retrieved from NCBI Protein. Submitting a query for the variant of interest (T286K), SIFT BLink predicted the variant to be tolerated with a score of 0.20 (median sequence conservation score 2.94, with 83 sequences aligned at this position).

- *Align GVGD*. Align GVGD [23] combines biophysical characteristics and multiple protein sequence alignments to predict the pathogenicity of variants. Align GVGD calculates the Grantham variation (GV, variation in the biophysical properties of all amino acids at a particular position in the multiple protein sequence alignment) and Grantham deviation (the deviation in biophysical properties of the altered amino acid from the reference). These scores are based on Grantham scores which measure the volume, polarity, and side chain composition of amino acids [24]. The two scores are combined

to provide a classification of pathogenicity likelihood, ranging from C0 (less likely to be deleterious) to C65 (most likely to be deleterious).

To perform Align GVGD:

- Download FASTA protein sequences for alignment. For CPOX, reference CPOX protein sequences were downloaded from NCBI. There are six NP accessions, i.e., non-predicted protein sequences as prefixed by XP, from *Homo sapiens*, *Sus scrofa*, *Mus musculus*, *Danio rerio*, *Rattus norvegicus*, and *Bos taurus*.
- Submit the downloaded FASTA file to T-COFFEE regular [25], a multiple sequence alignment tool (see **Note 1**).
- The multiple sequence alignment is uploaded in FASTA format onto Align GVGD. Enter “T286K” for Substitutions list, and submit the job.

Align GVGD classified the p.Thr286Lys variant as less likely to be deleterious with GV of 144.57 and GD of 8.11. The classification (C0) remained unchanged even when all CPOX protein sequences including predicted sequences were used for alignment.

- *SNPeffect*. SNPeffect [26] analyzes the structural effect of variants on protein using various algorithms (TANGO, WALTZ, LIMBO, and FoldX). Some pre-computed variants are available for search on their database. Using the CPOX p.Thr286Lys example, SNPeffect showed that the variant had no effect on aggregation tendency, amyloid propensity, or chaperone binding. Structural analysis using FoldX however predicted that the Thr-to-Lys change would result in a difference in free energy and hence reduction in protein stability (Fig. 6).

FoldX prediction is only provided if there is a homologous structural model available. Users can use partial protein sequence, e.g., a particular domain only, to broaden the homology search.

- *MutPred*. MutPred [27] predicts the effect of amino acid changes on (1) protein structure and dynamics, e.g., secondary structure and transmembrane helix; (2) predicted functional properties, e.g., catalytic residues and glycosylation sites; and (3) evolutionary information (based on SIFT). Input requirements include the FASTA sequence of the wild-type protein and the amino acid change. Two scores are returned—a general score to predict whether the variant is deleterious and *P*-values of the top five properties that may be altered as a result. The calling algorithm is based on machine learning with

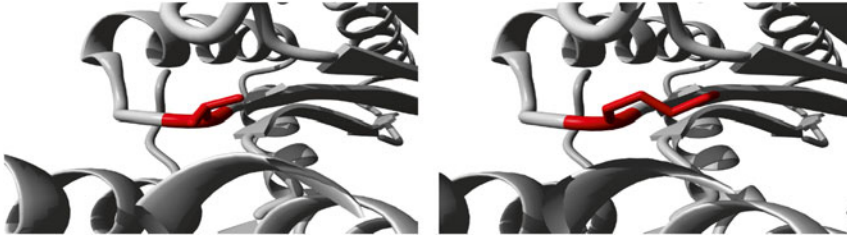


Fig. 6 FoldX prediction on the CPOX p.Thr286Lys variant from SNPeffct. The empirical protein design force-field FoldX is used to calculate the difference in free energy of the mutation: $\Delta\Delta G$ (delta delta G). If the mutation destabilizes the structure, $\Delta\Delta G$ is increased, whereas stabilizing mutations decrease the $\Delta\Delta G$. Since the FoldX error margin is around 0.5 kcal/mol, changes in this range are considered insignificant. 2aex has 100.00 % homology with the submitted sequence. This pdb is then used to get some more information on the structural effect. The mutation from THR to LYS at position 286 results in a $\Delta\Delta G$ of 3.29 kcal/mol. This implies that the mutation reduces the protein stability. Molecular visualization of the WT (*left*) and variant (*right*) amino acid. The residues colored in *red* represent the wild-type (THR) and variant residue (LYS)

Prediction Site Results								
Mutation	Probability of Actionable Hypothesis	Actionable Hypotheses	Molecular Mechanism	Substrate	Non-Actionable Hypotheses	Top 5 Substrates	Exact PFM Match	Neighbor PFM Match
Thr286Lys	0.555	<ul style="list-style-type: none"> Loss of protein stability (P = 0.228) Gain of substrate binding (P = 0.000) Gain of substrate release (P = 0.000) 	<ul style="list-style-type: none"> Confident Hypotheses 		<ul style="list-style-type: none"> Gain of protein stability (P = 0.228) Loss of substrate binding (P = 0.000) Loss of substrate release (P = 0.000) Gain of substrate release (P = 0.000) Loss of protein stability (P = 0.228) 			

Fig. 7 Output from MutPred. In the p.Thr286Lys example, MutPred returned a general score of 0.555 with three actionable hypotheses, suggesting some evidence that the variant may affect function

a random forest classifier, using data from HGMD and Swiss-Prot as a training set. Depending on the two scores the potential alteration in molecular mechanism is categorized into actionable, confident, or very confident hypotheses (Fig. 7).

- *PolyPhen-2*. PolyPhen-2 uses machine learning to select optimally 11 sequence- and structure-based predictive features for assessment of pathogenicity [28]. Various input formats are allowed. Two datasets (HumDiv and HumVar), both retrieved from UniProt, were used to train the algorithm (*see* supplementary material in ref. 28 for details of the two datasets). PolyPhen-2 returns a probabilistic score on whether the variant is likely to be damaging or not on both datasets. The scoring system for PolyPhen-2 is complex. For example, the HumVar-trained PolyPhen-2 score is more conservative as the HumDiv dataset assumes all non-synonymous SNPs with no disease annotation as benign, meaning that variants with mild effect

will be considered benign. On the other hand, the HumDiv-trained PolyPhen-2 score would be more suitable for association discovery albeit a higher false-positive rate. This difference can be highlighted in our example CPOX p.Thr286Lys, where it is *probably damaging* with a score of 0.0995 on HumDiv, but only *possibly damaging* with a score of 0.733 on HumVar.

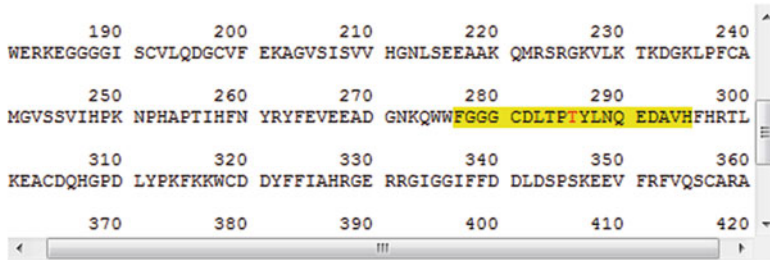
- **SNPs&GO.** SNPs&GO [29, 30] is an algorithm that makes use of Gene Ontology (GO) functional annotation. It uses support vector machines to incorporate the amino acid change, surrounding protein sequence environment, residue conservation (all of which form the basis of the algorithm PhD-SNP developed by the same laboratory group [31]), multiple sequence alignment (based on another algorithm PANTHER), and GO terms. For an explanation on the actual calculation of GO analysis *see* Kaminker et al. [32]. To use SNPs&GO, find the Swiss-Prot code for the protein of interest from the UniProt website, e.g., for CPOX it is HEM6_HUMAN. If Swiss-Prot code is not available, the FASTA sequence may be utilized but the associated GO terms will need to be entered manually (*see* **Note 2**). GO term associated to a protein can be searched on the Gene Ontology website. Enter the variant, and submit the job. The results will show predictions from SNPs&GO and the two related algorithms (PhD-SNP and PANTHER) (Fig. 8). A variant will be predicted as disease causing (second column) if the probability is above the default setting of 0.5. Measure of the quality of this binary classification (disease or neutral) is provided as a reliability index (RI), which correlates to the accuracy and the Matthews correlation coefficient (MCC; *see* ref. 33 for explanation on using MCC in evaluating the accuracy of predictions). In our example of CPOX p.Thr286Lys, SNPs&GO and PhD-SNP predicted the variant to be disease causing, while PANTHER predicted it to be neutral. The lower probability and reliability index assigned by SNPs&GO as a final score in this instance is reflective of the differing information provided by the PhD-SNP and PANTHER inputs.

3.2.5 Evaluation of In Silico Prediction Tools

There is currently no single consensus method for assessing variant pathogenicity using in silico prediction tools, and no tool alone is universally acknowledged as providing the most accurate prediction in all circumstances. The National Genetics Reference Laboratory (NGRL) at Manchester, UK, has published an evaluation of some prediction tools [33]. This report recommended a consensus approach when the best in silico tool for the particular gene of interest is unknown. The combination of three commonly used in silico prediction tools (PolyPhen-2, SIFT, and Align GVGD) was shown to have inferior prediction, often because the predictions

SNPs&GO Predicting disease associated variations using GO terms

Sequence File: HEM6_HUMAN.seq
 Alignment File: HEM6_HUMAN.seq.blast
 GO-terms File: HEM6_HUMAN.seq.go
 Output File: output.txt



Mutation	Prediction	RI	Probability	Method
T286K	Disease	8	0.902	PhD-SNP: F[T]=5% F[K]=0% Nali=512 PANTHER: F[T]=6% F[K]=3% SNPs&GO
	Neutral	6	0.223	
	Disease	4	0.696	

Fig. 8 SNPs&GO output for the CPOX variant p.Thr286Lys

from these tools are contradictory. These three tools are also those accessed directly through Alamut® (Interactive Biosoftware), a commercially available software package commonly used in many diagnostic laboratories. The combination of in silico tools that provided the most accurate predictions for the four genes investigated in the report (*BRCA1*, *BRCA2*, *MLH1*, and *MSH2*) included MutPred, SNPs&GO, and MAPP. MutPred and SNPs&GO were also shown to have the best predictions in over 40,000 pathogenic and neutral variants tested in another study [34].

Evolutionary conservation-based tools are highly sensitive to input multiple sequence alignments. The NGRl report demonstrated that pathogenicity prediction could change substantially depending on input alignment. Align GVGD provides curated alignment for several cancer susceptibility genes. If a laboratory is performing regular assessment of particular genes, building an in-house alignment for these genes should be considered.

The NGRl report did not recommend the use of protein stability-based methods such as FoldX in variant effect prediction due to the variability of tolerance to stability change between proteins.


Reference SNP(refSNP) Cluster Report: rs5925		
RefSNP	Allele	HGVS Names
Organism: human (Homo sapiens)	Variation Class: SNV: single nucleotide variation	NC_000019.9.g.11230881T>C NG_009060.1.g.35825T>C
Molecule Type: Genomic	RefSNP Alleles: C/T	NM_000527.4.c.1959T>C
Created/Updated in build: 52/137	Allele Origin:	NM_001195798.1.c.1959T>C
Map to Genome Build: 37.4	Ancestral Allele: C	NM_001195799.1.c.1836T>C
Validation Status: 	Clinical Channel: unknown	NM_001195800.1.c.1455T>C
Citation: PubMed	Clinical Significance: NA	NM_001195802.1.c.1596T>C
	MAF/MinorAlleleCount: C=0.332/724	NM_001195803.1.c.1578T>C
	MAF Source: 1000 Genomes	NP_000518.1.p.Val653= NP_001182727.1.p.Val653= NP_001182728.1.p.Val612= NP_001182729.1.p.Val485= NP_001182731.1.p.Val532= NP_001182732.1.p.Val526= NT_011295.11.g.2493683T>C

Fig. 9 Summary of SNP information on NCBI dbSNP. Validation status is depicted by various *symbols*. Click on the “Validation Status” link for description on these symbols

3.3 Population Frequencies

Large-scale sequencing projects such as the HapMap project, 1000 Genomes Project, and Exome Sequencing Project from the National Heart, Lung, and Blood Institute (NHLBI-ESP) provide frequency information on polymorphisms that allow inference on pathogenicity. Although common polymorphisms are unlikely to be deleterious, this does not exclude the possibility of milder or modifying effects on protein function.

When accessing these data it is important to (1) understand the source, phenotype, and ethnicity of selected samples in these projects as these will influence variant frequency and (2) be aware of the validation status of reported SNPs. SNPs are considered validated on dbSNP when at least one of the submissions is obtained by experimental methods, the submission contains frequency information, e.g., data from HapMap or 1000 Genomes Project, or there are multiple independent observations [35]. SNPs that have not been validated may be false positives.

The NCBI dbSNP database collates frequency information from various sources. To view SNP summary on dbSNP, users can search for reported SNPs using the HGVS name (under “Search by ID on All Assemblies”). Alternatively, reported SNPs will be highlighted in red under the SNP track in the NCBI Sequence Viewer (see above). The SNP summary shows the minor allelic frequency (MAF) count using data from the 1000 Genomes cohort. Different levels of validation are also shown in symbols (Fig. 9).

Ethnic based population frequencies can be obtained under the “Population Diversity” section (Fig. 10). In the example of rs5925, the allelic frequencies in European (CEU), Asian (HCB and JPT), sub-Saharan African (YRI), and other ethnicities obtained from the HapMap project are shown (red box). Where available, data from NHLBI-ESP are also provided (green box).

ss#	Sample Ascertainment				Genotype Detail				Alleles	
	Population	Individual Group	Chrom. Sample Cnt.	Source	C/C	C/T	T/T	HWP	C	T
ss107937053	ABECASIS_CLINICAL_PANEL		748	AF					0.354	0.646
ss142653712	ENSEMBL_Venter		2	IG	1.000				1.000	
	ENSEMBL_celera		2	IG	1.000				1.000	
ss171497730	PGP		2	IG		1.000			0.500	0.500
ss19402919	CEPH		184	AF					0.370	0.630
HapMap-CEU	European		226	IG	0.204	0.460	0.336	0.527	0.434	0.566
HapMap-HCB	Asian		86	IG	0.070	0.326	0.605	0.584	0.233	0.767
HapMap-JPT	Asian		172	IG	0.023	0.360	0.616	0.317	0.203	0.797
HapMap-YRI	Sub-Saharan African		226	IG	0.018	0.186	0.796	0.584	0.111	0.889
HAPMAP-ASW			98	IG	0.020	0.367	0.612	0.371	0.204	0.796
HAPMAP-CHB	Asian		82	IG	0.073	0.146	0.780	0.010	0.146	0.854
HAPMAP-CHD			168	IG	0.060	0.333	0.607	0.752	0.226	0.774
HAPMAP-GIH			176	IG	0.250	0.455	0.295	0.439	0.477	0.523
HAPMAP-LWK			172	IG	0.035	0.267	0.698	0.752	0.169	0.831
HAPMAP-MEX			100	IG	0.320	0.480	0.200	1.000	0.560	0.440
HAPMAP-MKK			282	IG	0.014	0.312	0.674	0.251	0.170	0.830
HAPMAP-TSI			170	IG	0.141	0.565	0.294	0.150	0.424	0.576
ss34248622	ESP_Cohort_Populations		4550	GF	0.139	0.420	0.441	0.001	0.349	0.651

Fig. 10 Allelic frequency of rs5925 in different ethnicities on dbSNP

The collection of in-house data where no dbSNP entry is available is encouraged. An example is a promoter variant c.-24C>G in *HBA2* or *HBA1* that is commonly found in patients investigated for alpha thalassemia. This variant has been found in association with other known causative *HBA2/HBA1* variants and is also seen on its own in association with a normal phenotype. Therefore this variant would be considered benign in terms of function.

3.4 Family Studies

Family studies provide information about whether a variant of interest is inherited or de novo, about its pattern of inheritance (dominant/recessive/paternal/maternal), and whether the variant co-segregates with the phenotype. This can be especially useful when interpreting variants private to a family which may not be described in the literature, a situation which is frequently present.

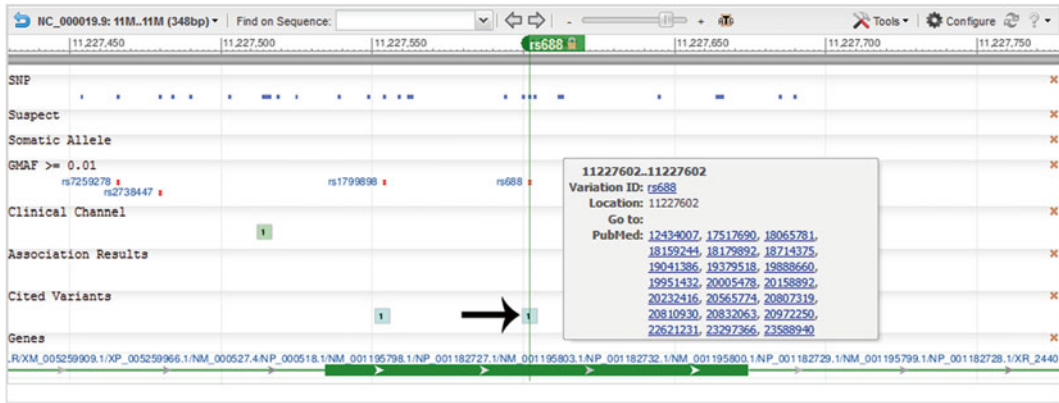


Fig. 11 PubMed IDs for articles related to a specific SNP can be viewed under “Cited Variants” (arrowed)

3.5 Literature Search for Published Evidence of Biological Effect (Functional Studies, etc.)

Searching the published literature for functional evidence of a particular variant is essential when attempting to establish potential effects. In dbSNP, PubMed IDs for articles related to a reported SNP are listed under the “Cited Variants” track where available (Fig. 11). Locus-specific databases such as the University College of London *LDLR* familial hypercholesterolemia database for the *LDLR* gene [36], the Breast Cancer Information Core (BIC), and the International Agency for Research on Cancer (IARC) breast cancer database for *BRCA1* and *BRCA2* are also useful. The UK Clinical Molecular Genetics Society (CMGS) guidelines for interpretation and reporting of unclassified variants states under section 4.1 that consulting locus-specific databases when reporting unclassified variants is *essential*, although curatorial rigor of all databases is a consideration [37]. It is envisaged that in the future, larger curated databases such as the Human Gene Mutation Database (HGMD) and Catalogue of Somatic Mutations in Cancer (COSMIC), which also provide links to reference journal articles, may become comprehensive enough to include information available in locus-specific databases and may thus supplant this requirement.

In all instances, the *quality* of source data must be carefully evaluated. It is important to establish whether the evidence presented has been based on *in silico*, *in vitro*, or *in vivo* studies, with robustness of reported results verified by checking for reproducibility by independent groups and/or involving different populations. Too frequently a variant is reported as having been independently found a number of times, but, on closer inspection, the multiple observations are actually based on the one original publication. It may also be useful to search for information other than journal articles such as conference abstracts via search engines, although non-peer-reviewed data are of limited use in a clinical setting.

3.6 Underlying Biological Knowledge

Understanding the underlying molecular mechanisms of disease and their consequences in phenotype causation is essential for variant interpretation. For example, in dominant conditions (such as some inherited cancer syndromes) where homozygous/compound heterozygous allelic loss may be embryonic lethal, co-occurrence of a VUS with another known pathogenic mutation in trans can indicate that the variant is unlikely to be pathogenic.

In other instances, such as in the molecular diagnosis of thalassemia, laboratory phenotypic data may be helpful. In these instances it is important to correlate genotyping results with the phenotype to ascertain whether further studies are required. For example, a hematological profile may be less or more severe than predicted on the basis of a known hemoglobin beta gene mutation. This may be indicative of gene–gene interactions such as a deletion, single-base change, or even a duplication of the hemoglobin alpha globin gene.

3.7 Classification Models for Variants Based on Evidence of Pathogenicity

The classification of variant pathogenicity is a complex task that requires professional judgment based on the collective evidence from all the aspects discussed and considered in this chapter. The American College of Medical Genetics (ACMG) has published an approach, detailed in a decision flow chart, on variant classification ([38]; Fig. 12). In both research and medical diagnostics, not all variants will have sufficient information for an unequivocal determination, and not all evidence will have the same strength. As a general principle, variants demonstrated to have biological effects with in vivo and/or in vitro evidence are more convincing than those suggested solely on the basis of in silico predictive effects. Even an in vitro environment, although indicative, may not always be a true reflection of in vivo effects as complex biological interactions cannot be assessed. Cassa et al. [39] have shown that 8.5 % variants classified as disease causing in the manually curated HGMD database are found in asymptomatic individuals. Conversely, a disease phenotype may be due to quantitative or pleiotropic effects of variants beyond the gene of interest. This is especially the case as more and more genotype–phenotype associations are unveiled by genomic scale researches. Interdisciplinary consortia such as evidence-based Network for the Interpretation of Germline Mutant Alleles (ENIGMA) are increasingly being formed to harmonize variant interpretation for some clinically important genes [40].

There are various classification systems to categorize variants in the clinical context. The ACMG have proposed a six-category classification [38], and for cancer susceptibility genes the IARC Unclassified Genetic Variants Working Group has suggested a five-class classification defined by the probability of a variant being pathogenic [2, 41] (Table 3).

With limited information, there are likely to be a large proportion of variants classified in the *uncertain* or the *likely pathogenic/likely nonpathogenic* categories, complicating genetic

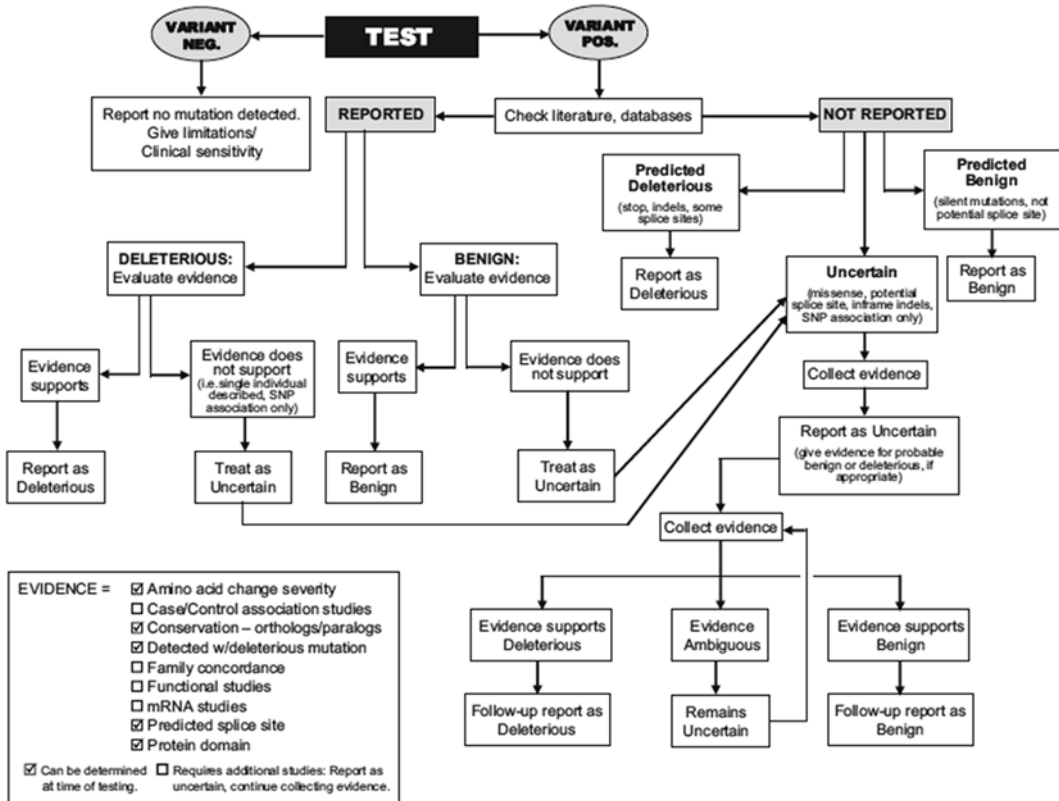


Fig. 12 ACMG flow chart on variant classification and reporting. Adapted from [45] with permission from Nature Publishing Group

counseling, potentially causing ongoing uncertainty for patients, and requiring follow-up studies. Diagnostic laboratories may be justifiably reluctant to allocate the variant into the *likely pathogenic* or the *expected to cause the disorder* categories without overwhelming supportive evidence, particularly if clinical stakes are high. In many instances, calculation of an exact posttest probability of disease for the variant may not be possible. These considerations are highlighted below.

3.8 Case Studies

Following are comparisons of three cases and the findings utilizing the methods described above (summarized in Table 4).

Case 1. CPOX variant (NM_000097.5:c.857C>A, NP_000088.3:p.Thr286Lys).

This variant was detected in two family members who had a biochemical diagnosis of hereditary coproporphria. The nucleotide and amino acid sequences are highly conserved. In silico evidence using SNPs&GO, MutPred, and Align GVG D was inconsistent. FoldX did suggest reduced stability, although as mentioned above it is not recommended for prediction. The variant is listed on

Table 4
Approaches used to classify variants in the three case studies

	CPOX p.Thr286Lys	LDLR p.Glu101Lys	BRCA2 p.Asn289His
Nucleotide and amino acid conservation (PhyloP and phast Cons)	Conserved	Conserved	Not conserved
Protein domain	No association to functional domain	Calcium-binding site	No association to functional domain
SNPs&GO (probability of being disease causing)	0.696	0.965	0.966
MutPred (probability of being pathogenic)	0.555	0.937	0.163
Align GVGD	C0	C0–C55	C0
SIFT	Tolerated (score = 0.20)	Affect protein function (score = 0)	Tolerated (score = 0.12)
PolyPhen-2 (HumVar)	Possibly damaging (0.733)	Probably damaging (0.985)	Benign (0.075)
dbSNP entry	N/A	rs144172724	rs766173
Population frequency	No frequency information	No frequency information	Found in 5.8 % in population, up to 20 % in Han Chinese
Literature and functional studies	Reported in one patient with hereditary coproporphyrinuria. No functional studies available	Reported in multiple populations with familial hypercholesterolemia. Functional studies showed 15–30 % of normal LDLR activity in homozygous state	Associated with decreased risk to breast cancer. No functional studies available
Variant classification in database	Not available	Disease-causing mutation (HGMD)	Disease-associated polymorphism (HGMD)
Classification	May be pathogenic	Pathogenic	Likely not pathogenic

HGMD database and had been reported once in another patient with hereditary coproporphyrinuria [42]. No functional studies were available. Due to the nature of variable penetrance in porphyria and lack of functional studies, the variant would be classified as *may or may not be causative of the disorder*.

Case 2. LDLR variant (NM_000527.4:c.301G>A (NP_000518.1: p.Glu101Lys)).

This variant was found in a patient with clinical familial hypercholesterolemia. The nucleotide and amino acid sequences are highly conserved. The variant is within a calcium-binding site on NCBI CDD. It is reported in the locus-specific variation database and had been described in various populations [36]. The variant is also known as FH Lancashire or E80K (using a different transcript as reference). Align GVDG prediction ranged from C0 to C55 depending on the input of multiple protein sequence alignments. Manual curation of the sequence alignment (i.e., removing predicted/hypothetical protein or unrelated protein sequences) would see the classification changing from C0 to C55. This highlights the importance of carefully selected alignment. SIFT predicted the variant to be not tolerated. SNPs&GO and MutPred predicted a high probability for the variant to be disease causing. Functional studies showed that LDLR activity was 15–30 % of normal in a homozygous individual [43]. The above evidence was considered to be sufficient to classify this variant as *pathogenic*.

Case 3. BRCA2 variant (NM_000059.3:c.865A>C (NP000050.2:p.Asn289His)).

This variant was detected in a Chinese patient referred for familial breast cancer testing. The variant is listed as a validated SNP (rs766173) on dbSNP with a minor allele frequency of 5.8 % in 1000 Genomes Project. However, the population frequency of this variant is up to 20 % in Han Chinese. It is a polymorphic SNP where a different nucleotide change (c.865A>G, p.Asn289Asp) is also found. The variant is reported in the HGMD database, and there is one article linked to the variant, reporting the variant to be associated with *decreased* risk of breast cancer [44]. In silico studies showed that the variant is not conserved based on PhyloP and phastCons, and it is not associated with any functional domain. Align GVDG using the built-in sequence alignment for BRCA2 indicated that the variant is not likely to affect function (class C0). MutPred predicted the variant to be benign with the probability of it being deleterious at 0.163. However, SNPs&GO called it a disease-causing variant with a high RI of 9 (probability 0.966). Given the high population frequency especially in Han Chinese, it is unlikely that the variant is pathogenic, at least in the Chinese population. Results from in silico studies are inconsistent and therefore inconclusive. There was no positive association of this variant to breast cancer at the moment of reporting. The variant was therefore classified as *likely not pathogenic*.

4 Notes

1. Alternatively, one can perform multiple sequence alignment by SIFT BLink as described previously. SIFT will perform PSI-BLAST, and the FASTA file of the multiple protein sequence alignment can be downloaded. Beware that it may contain

unrelated proteins/predicted proteins which may need to be removed manually.

2. If using FASTA sequence instead of Swiss-Prot code, SNPs&GO will *only* call the CPOX variant disease associated if one includes the two GO terms associated with CPOX (GO:0006779 and GO:0004109) while entering information in SNPs&GO. It will be called neutral if you do not!

References

1. den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 15:7–12
2. Bures M, Seledtsov IA, Solovyev VV (2000) Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res* 28:4364–4375
3. Chretien S, Dubart A, Beaupain D et al (1988) Alternative transcription and splicing of the human porphobilinogen deaminase gene result either in tissue-specific or in housekeeping expression. *Proc Natl Acad Sci U S A* 85:6–10
4. Gouya L, Puy H, Robreau AM et al (2002) The penetrance of dominant erythropoietic protoporphyria is modulated by expression of wild-type FECH. *Nat Genet* 30:27–28
5. Wu K, Hinson SR, Ohashi A et al (2005) Functional evaluation and cancer risk assessment of BRCA2 unclassified variants. *Cancer Res* 65:417–426
6. Clegg JB, Weatherall DJ (1974) Hemoglobin constant spring, and unusual alpha-chain variant involved in the etiology of hemoglobin H disease. *Ann N Y Acad Sci* 232:168–178
7. Jensen HK, Jensen TG, Faergeman O et al (1997) Two mutations in the same low-density lipoprotein receptor allele act in synergy to reduce receptor function in heterozygous familial hypercholesterolemia. *Hum Mutat* 9:437–444
8. Kashima T, Manley JL (2003) A negative element in SMN2 exon 7 inhibits splicing in spinal muscular atrophy. *Nat Genet* 34:460–463
9. Punta M, Coghill PC, Eberhardt RY et al (2012) The Pfam protein families database. *Nucleic Acids Res* 40:D290–D301
10. Schultz J, Milpetz F, Bork P et al (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* 95:5857–5864
11. Letunic I, Doerks T, Bork P (2012) SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res* 40:D302–D305
12. Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* 32:W327–W331
13. Marchler-Bauer A, Anderson JB, Chitsaz F et al (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* 37:D205–D210
14. Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F et al (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* 39:D225–D229
15. Tavtigian SV, Greenblatt MS, Lesueur F et al (2008) In silico analysis of missense substitutions using sequence-alignment based methods. *Hum Mutat* 29:1327–1336
16. Pollard KS, Hubisz MJ, Rosenbloom KR et al (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20:110–121
17. Siepel A, Bejerano G, Pedersen JS et al (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15:1034–1050
18. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. *Genome Res* 11:863–874
19. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. *Genome Res* 12:436–446
20. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31:3812–3814
21. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 7:61–80
22. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073–1081
23. Mathe E, Olivier M, Kato S et al (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res* 34:1317–1325

24. Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185:862–864
25. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302: 205–217
26. De Baets G, Van Durme J, Reumers J et al (2012) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res* 40: D935–D939
27. Li B, Krishnan VG, Mort ME, Xin F et al (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 25:2744–2750
28. Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249
29. Calabrese R, Capriotti E, Fariselli P et al (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* 30: 1237–1244
30. Capriotti E, Altman RB (2011) Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics* 12(Suppl 4):S3
31. Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* 22: 2729–2734
32. Kaminker JS, Zhang Y, Waugh A et al (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. *Cancer Res* 67:465–473
33. Williams S (2012) Analysis of in silico tools for evaluating missense variants. http://www.ngri.org.uk/Manchester/sites/default/files/publications/Add-To-Menu/Missense_Prediction_Tool_Report.pdf. Accessed 13 June 2013
34. Thusberg J, Olatubosun A, Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat* 32:358–368
35. Kitts A, Sherry S (2002) The single nucleotide polymorphism database (dbSNP) of nucleotide sequence variation. In: *The NCBI handbook* [Internet]. <http://www.ncbi.nlm.nih.gov/books/NBK21088/>. Accessed 10 June 2013
36. Leigh SE, Foster AH, Whittall RA et al (2008) Update and analysis of the University College London low density lipoprotein receptor familial hypercholesterolemia database. *Ann Hum Genet* 72:485–498
37. Bell J, Bodmer D, Sistermans E et al (2007) Practice guidelines for the Interpretation and reporting of unclassified variants (UVs) in clinical molecular genetics. <http://www.cmgs.org/BPGs/pdfs%20current%20bpgs/UV%20GUIDELINES%20ratified.pdf>. Accessed 16 June 2013
38. Richards CS, Bale S, Bellissimo DB et al (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. *Genet Med* 10:294–300
39. Cassa CA, Tong MY, Jordan DM (2013) Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum Mutat* 34: 1216–1220
40. Spurdle AB, Healey S, Devereau A et al (2012) ENIGMA-evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. *Hum Mutat* 33:2–7
41. Plon SE, Eccles DM, Easton D et al (2008) Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. *Hum Mutat* 29:1282–1291
42. Whatley SD, Mason NG, Woolf JR et al (2009) Diagnostic strategies for autosomal dominant acute porphyrias: retrospective analysis of 467 unrelated patients referred for mutational analysis of the HMBS, CPOX, or PPOX gene. *Clin Chem* 55:1406–1414
43. Hobbs HH, Brown MS, Goldstein JL (1992) Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. *Hum Mutat* 1:445–466
44. Bhatti P, Struewing JP, Alexander BH et al (2008) Polymorphisms in DNA repair genes, ionizing radiation exposure and risk of breast cancer in U.S. radiologic technologists. *Int J Cancer* 122:177–182
45. Richards CS, Bale S, Bellissimo DB et al (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations. *Genet Med* 10:294–300

Chapter 14

Designing Algorithms for Determining Significance of DNA Missense Changes

Sivakumar Gowrisankar and Matthew S. Lebo

Abstract

Humans differ from each other in their genomes by <1 %. This determines the difference in susceptibility to disease, phenotypes, and traits. Predominantly, when looking for causal disease mutations, protein-coding sequences are screened first since those have the highest probability of affecting the function of a protein. Recent technological advances have seen a rise in the number of experiments being conducted to study a variety of diseases from monogenic to complex traits. Several computational approaches have been developed to extract putative functional missense variants. In this chapter we review some of these approaches and describe a standard step-by-step procedure that can be used to classify variants for the purpose of clinical care. We also provide two examples demonstrating this approach, one for a patient with a dilated cardiomyopathy diagnosis, and the other for a patient with an unknown etiology undergoing whole-genome sequencing (WGS).

Key words Missense variants, Variant classification, Variants of unknown significance

Abbreviations

ESP	Exome Sequencing Project
HGMD	Human Gene Mutation Database
LOF	Loss of Function
OMIM	Online Mendelian Inheritance in Man
SNV	Single Nucleotide Variation
VUS	Variant of Unknown Significance
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

1 Introduction

Single nucleotide variations (SNVs) are one of the most common forms of alterations in the human genome and account for 99 % of the differences between people [1]. Missense variants, also known



Fig. 1 A hypothetical protein with a T>C missense variant causing a valine to alanine amino-acid change

as non-synonymous SNVs, are a class of variants that overlap the coding region of the genome and alter the amino-acid sequence of the encoded protein (Fig. 1). Missense SNVs are known to be responsible for a wide variety of adverse phenotypes and diseases. However, it is estimated that only 13–25 % of the known missense mutations are deleterious and the remainder are benign [2, 3].

In research, several studies have been conducted to study systematically various diseases and phenotypes to identify the causal genes and their pathogenic mutations [4, 5]. These studies employ techniques such as genome-wide association studies (GWAS) [6, 7] and more recently high-throughput sequencing known as next-generation sequencing (NGS) [8, 9]. Briefly, GWAS work under the hypothesis that the causal variants linked to disease can be captured by detecting adequate number of variants that are in linkage disequilibrium (LD). However, in most experiments the causal gene let alone the causal variant can be difficult to identify. The advent of NGS has seen a tremendous increase in the number of rare coding variants detected and associated with complex traits in case versus control studies. With this comes greater complexity in separating functional from benign variants.

In a clinical molecular diagnostic setting, the traditional mode of testing a patient for a known disease-associated gene and variant has shifted towards screening for a large number of disease-associated genes and variants. Although this has improved the clinical sensitivity of diagnostic tests, the number of variants whose significance cannot be determined (called VUS or variants of unknown significance) has also increased. In addition several clinical laboratories have begun offering whole-exome and whole-genome sequencing (WES and WGS) tests [10]. This raises the bar on the interpretation of new DNA variants as the genes they overlap with may not have a definitive proof of association with a disease.

Different kinds of information are currently being employed to classify variants by significance. These can be broadly categorized into genetics-based evidence, functional evidence, and computational predictions (also *see* **Note 1**).

- In genetics-based evidence one of the most commonly used metric is the population frequency of the variant allele. This refers to the proportion of any given population that carries

the variant allele in question. If the variant allele frequency (also called as allele frequency) is greater than the prevalence of the disease in the population—assuming complete penetrance—then it can be inferred that this variant is not associated with the disease.

- In functional evidence the gene and protein expression are analyzed either *in vivo* or *in vitro* to assess the effect of a mutation. Specifically mouse models have been used for characterization of specific diseases and phenotypes. When using mouse models as evidence for assessing clinical significance of a mutation care must be taken to ensure the phenotypes and diseases tested are related to the ones at hand.
- Computational prediction tools are available particularly to predict the functional significance of coding mutations. A number of tools exist, and many of them share a number of features that go into their prediction model. These include nucleotide and amino-acid conservation, location of the mutation on the coded-protein's 3D structure, physicochemical properties of the protein itself, and so on. In spite of using similar features many tools have different underlying models to predict pathogenicity [11] (also *see* **Note 2**).

There are also databases and Web servers developed that pre-classify publically available variants from sources such as dbSNP and ESP (Exome Sequencing Project). These databases essentially provide a collection of information that includes several of the ones discussed above. However, before using these sources care must be taken to ensure the databases and Web servers are up-to-date.

2 Materials

2.1 Annotation Tools

Variants need to be annotated with information such as conservation, allelic frequency, functional models, and so on. Several free and commercial tools exist that cater to this purpose. One such frequently used and freely available tool for academic institutions is ANNOVAR [12]. ANNOVAR accepts as input a text-based tab-delimited input file with the first five columns as chromosome number, start position, end position, wild-type allele, and variant allele. The rest of the columns can be used for any annotations and will be reproduced in the output.

Another tool that is commercially available for annotation is Alamut (Interactive Biosoftware, LLC. Rouen, France). Alamut is a powerful variant annotation and visualization tool that can be used to populate a variety of fields. One of the advantages of Alamut is that it can be used to obtain annotations for the same variant from different transcripts. Table 1 shows a list of currently used variant annotation tools.

Table 1
Variant annotation tools

Tool name	Description	Reference/URL
VEP	Variant Effect Predictor—annotations with predicted effect on proteins	[13]
ANNOVAR	Annotate variants with a number of optional data	[12]
VAT	Variant Annotation Tool—cloud-based tool for variant annotation	http://vat.gersteinlab.org/
GenomeTrax™	Commercial tool to identify pathogenic variants in silico	www.biobase-international.com/product/genome-trax
SeattleSeq	Variant annotation tool	[14]
Alamut	Commercial software for variant annotation and visualization	www.interactive-biosoftware.com/software/alamut/
FAVR	Filtering and annotating of variants that are rare	[15]

2.2 Allele Frequency Databases

A handful of consortia and various other efforts have sought to identify and determine the allele frequency for naturally occurring variants in various populations. These include the ESP (<http://evs.gs.washington.edu/EVS/>), the 1000 genomes project [16], and the PanSNPdb (The Pan-Asian SNP Genotyping Database) [17], ClinSeq [18]. The ESP has sequenced >6,500 African American and European American normal populations and subsequently identified common and rare variants that occur in them. Similarly, the 1000 genomes project has sequencing and variant data for >24 different populations. Another good source of common variants is dbSNP [19]. dbSNP data are an accumulation of variant submissions from multiple different laboratories and consortia. Hence, the quality of the downloaded data needs to be assessed before its use. These variants can either be downloaded in bulk or obtained through a Web interface.

2.3 Computational Functional Impact Prediction Tools

Several in silico functional prediction tools are available as mentioned earlier to assess the functional significance of the missense variant in question. More than often these tools use a similar set of features such as nucleotide conservation, amino-acid conservation, physicochemical properties, overlapping with known domains, among other things to make a prediction on the putative effect of the variant on protein function. Users are encouraged to read other detailed descriptions published [20, 21]. Table 2 shows some of the most commonly used tools.

Table 2
Commonly used in silico analysis tools (see also Chapter 13)

Tool name	URL	Reference
PolyPhen2	http://genetics.bwh.harvard.edu/pph2/	[22]
SIFT	http://sift.bii.a-star.edu.sg/	[23, 24]
MutationTaster2	www.mutationtaster.org/	[25]
SNAP	www.rostlab.org/services/SNAP/	[26]
MutationAssessor	http://mutationassessor.org/	[27]
MAPP	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	[28]

Table 3
Commonly used DNA variant and mutation databases

Tool name	Description	Reference
OMIM	Database of inherited mutations especially in humans	http://omim.org/
HGMD	Human Gene Mutation Database	[29]
ClinVar	Free archive of human variations and phenotypes with supporting evidence	[30]
LOVD	Leiden Open Variation Database	[31]
GET-Evidence	Collaborative database of human variants, traits, and diseases	http://evidence.personalgenomes.org/about
PharmGKB	Tool to investigate effect of genetic variation on drug response	[32]
HGVS—database and other tools	Database of databases organized by various categories	www.hgvs.org/dblist/dblist.html

2.4 Variants in Clinically Relevant Databases

Many databases that store causal variants associated with clinically significant phenotypes exist. For example OMIM, Online Mendelian Inheritance in Man (<http://omim.org/>) is a database of human genes and associated phenotypes that also has information about causal variants. This information is gleaned from the literature and updated on a daily basis. It must be noted that this database primarily focuses on inherited diseases and their causal variants. Another database of significance is HGMD, Human Gene Mutation Database [29] that catalogs all known mutations and associated diseases from the literature. For HGMD, one might need to obtain a license depending on the type of institution and the intended use. Table 3 shows some of the most commonly used databases.

2.5 Literature Search Currently, there are limited sources that have accurately and thoroughly gathered variants of significance from literature. Literature searches are essential for identifying the primary sources for family, case-control, and functional studies. Searches using PubMed and search engines such as Google are most effective, provided care is taken to include all possible nomenclature for the variant (also *see Note 3*).

3 Methods

While there are no accepted universal standards to filter and categorize variants, there are steps that are used widely by a number of laboratories. The method described in this chapter is one of the approaches that our laboratory generally follows for filtering variants detected by NGS. In addition, for reporting variants to patients we follow a more elaborate set of rules that are detailed in ref. 33. Although we have focused on NGS, the same approach can be used with other related technologies.

3.1 Filtering Strategies

Depending on the size of the genomic region sequenced the number of variants detected can range from a few hundreds to a few million [34]. Identifying variants of interest that might be functionally significant can be like searching for a needle in a haystack. Nevertheless, a variety of *filtering* mechanisms can be used to reduce the number of variants to a list of manageable size.

3.1.1 Variant Call Quality

Most of the tools in use today to call variants include a set of standard metrics as part of the output. These range from depth of coverage, mapping quality, and strandedness. Depth of coverage or simply coverage is the number of reads overlapping the variant of interest. Usually for germ-line-based analysis a coverage of $\geq 20\times$ indicates good quality [35], while for somatic mutations a higher depth $>100\times$ might be preferred to detect low percent tumor or heterogeneity. When analyzing whole-exome or whole-genome data, a lower coverage might be used as threshold to filter out likely false-positive variants. This might reduce the likelihood of incurring a false-negative variant [36].

3.1.2 Population Allele Frequency

When a variant present in a patient is commonly found in the general population then that variant most likely is not clinically significant. In general, an allele frequency of $\geq 5\%$ points to a variant being benign [37, 38]. This cutoff can be further reduced when assessing the significance of variants pertaining to a specific disease with known prevalence, age of onset, and penetrance (also *see Note 4*). For example for assessing hypertrophic cardiomyopathy related genes a variant may be considered benign if its allele frequency is $>0.3\%$ [33].

3.1.3 Presence of Variant in Clinically Relevant Databases

A variant present in a patient may be classified pathogenic if it is already implicated with a disease in any of the clinically relevant databases mentioned earlier. However, two caveats need to be noted.

- False-positive assertions exist in both commercial and public databases. These incorrectly implicate a variant with a disease. Therefore, it is always recommended to read the actual publication(s) that link(s) a variant to a disease. Even if the results of the publications are correctly cited in the database, it remains possible that conclusions in them are incorrect.
- In spite of having the variant in the database, it is possible that the variant is associated with a different phenotype or disease than what is present in the patient. A variant that might cause a particular disease does not necessarily cause the other.

3.1.4 Segregation Patterns

If family-trio or other extended pedigrees are available, it is best to check for segregation of the variant with the affected. For example if the proband does not share a variant with other affected family members then that variant may not be the cause of the phenotype or disease.

In cases where the inheritance pattern of a disease is known, one might look for specific types of variants. For dominant inheritance a single heterozygous variant is enough to cause the disease [39]. For recessive mode of inheritance, two compound heterozygous variants might be expected, whereas in the case of consanguineous marriage a homozygous variant might be expected [40]. In rare cases when the parents are not affected a de novo mutation might be the cause of the disease [41]. A pedigree or detailed family information is not always available but remains one of the most powerful ways of confirming the causality of a variant with a phenotype.

3.1.5 Other General Considerations

- *Conserved domains:* Certain segments or domains of the protein may have greater functional significance than others. These domains are often conserved across multiple species. A missense variant overlapping this domain will change the amino-acid structure and therefore might alter the domain's functionality. For example, the RS domain in the gene *RBM20* is commonly mutated in dilated cardiomyopathy [42].
- *Conservation:* A missense variant will cause a change in the amino-acid sequence. However, it is possible that the position of the amino acid is not evolutionarily conserved across multiple species indicating that it is not functionally significant. Tools such as the UCSC genome browser [43] or Alamut can be used to do this. On the other hand it is possible to download all this information from UCSC or NCBI and use it in custom tools for high-throughput processing.

- In silico prediction tools: The most commonly used tools among others include SIFT, PolyPhen2, and MutationTaster. Though these tools use a common set of features the algorithms designed to make the predictions differ significantly. Therefore, it is desirable to use more than one prediction tool. These tools are used in a research setting, but for *clinical* assessment of a variant they are used with great care as they have not been validated for this purpose (*see* **Note 5**).

3.1.6 Special Considerations for Mendelian Phenotypes

We have discussed earlier that a variant previously associated with disease in Mendelian disease databases is a sign of pathogenicity of the variant. Sometimes, however, a slightly different variant may have been implicated in the disease. Although the exact same variant was not observed earlier it indicates that the location of the variant is functionally important. This suggests that the current variant is also significant.

Missense variants can also give rise to a splicing variant (Fig. 2) depending on its location. Many splicing variants have been implicated in diseases [44]. It is important to check if a variant will affect the splicing event through in silico tools such as NNSPLICE [45] and GeneSplicer [46]. For clinical testing, the same cautions apply as described above for in silico prediction tools.

It is also possible to obtain an estimation of the frequency of occurrence of missense variants in specific genes of interest. One could then relate the occurrence of a pathogenic missense variant to the likelihood of that type of variant occurring in a gene. For example, missense variants are common in the *TTN* gene (dilated cardiomyopathy) while relatively rare in the *PTPN11* gene (Noonan syndrome). Therefore, a missense variant in *PTPN11* has a higher *a priori* likelihood of being pathogenic than a missense variant in *TTN*.

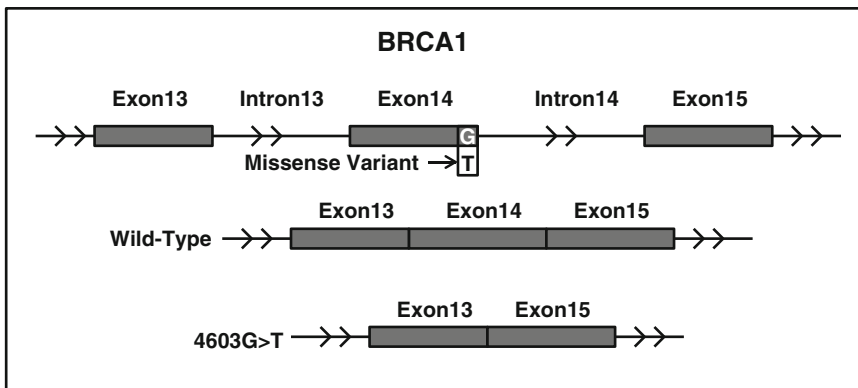


Fig. 2 A 4603G>T missense variant at the splicing junction in *BRCA1* causing exon 14 skipping [47]

3.1.7 *Special Considerations for Complex Traits*

For complex diseases such as diabetes it is hypothesized that a collection of variants represent cumulative risk factors [48, 49] as opposed to a single variant causing a disease in the Mendelian disorders. In these cases the problem becomes the identification of risk alleles rather than causal alleles. Often, patient cohorts are needed to predict these alleles followed by model creation to predict risk for a new patient in acquiring the complex trait. The readers are encouraged to read publications that give a much detailed account of these issues [50].

3.1.8 *Special Considerations for Somatic Mutations*

Analyzing cancer mutations is a broader topic. Here, we leave a brief description of what is involved in such an analysis. The readers are encouraged to refer to other materials for detailed information. Somatic mutations are generally called accurately with the presence of tumor-normal sample pairs. Variant calling is typically done using tools such as VarScan2 [51], Strelka [52], and MuTect [53] which are developed specifically for cancer-related data. These variants can then be annotated with matches to cancer-specific databases such as COSMIC [54]. As with Mendelian databases care must be taken to ensure the validity of the entries by going through publications referenced in the database. For research-based studies the variant can be further annotated and filtered as part of a clinical trial or, via its effect if there are drugs that target the specific mutation and/or related pathway(s).

3.2 *Examples*

3.2.1 *Filtering Variants Obtained from a Dilated Cardiomyopathy-Specific Gene Panel*

The panel covers a total of 51 genes covering only coding regions and can be developed in what is called *targeted* NGS. What follows is a step-by-step breakdown on the number variants remaining at each filtering step.

1. Total Variants: 142.
2. Variants after quality filtering: 140.
3. Variants after reference sequence error filtration: 132.
4. Variants after filtration by population allele frequency: 6.
5. Manual review (segregation, literature) which produces the final breakdown into:
 - (a) One likely benign variant.
 - (b) Four variants of unknown significance (VUS).
 - (c) One pathogenic variant.

3.2.2 *Filtering Variants Obtained from a Patient with Unknown Etiology Using WGS*

1. Total variants: 5,131,222 (SNVs and small indels).
2. Filter by quality: 3,227,455 substitutions and 418,331 indels.
3. Filter by coding region: 20,240 coding sequence (CDS) or splice variants (also *see* **Note 6**).

4. Filter by allele frequency: 616 Rare CDS/splice Variants.
 - (a) *Present in clinical databases*: 25 “Disease-Causing” Variants.
 - Manual review—two pathogenic.
 - (b) *Novel/rare loss-of-function (LOF) variants*: 189 LOF Variants.
 - Variants present in medically relevant genes (11 variants).
 - Manual review (two variants).

4 Notes

1. In the absence of pedigree/segregation information, any computationally based method assessing the effect of a missense variant is as accurate as the annotations that go into it. It is possible that some disease-causing mutations may not have been studied in detail yet. As a result computational methods to assess significance will be inconclusive.
2. There can be important differences between *functional* significance, *medical* significance, and *population* significance for any allele. Briefly, a variant can be functionally significant, meaning that it might be affecting the protein function, but may not be medically significant, meaning that it may not give rise to a clinically relevant trait. Similarly, a variant allele on a conserved location might not necessarily have a medically significant effect [11].
3. When using allele frequency for variant filtration, care must be taken that the cutoff is fairly conservative based on disease prevalence, penetrance, and age of onset. Prevalence gives an estimate of how frequently the disease occurs in the population. If the population frequency is more than the prevalence of the disease, then the variant is likely benign. In addition, the penetrance of the disease, defined as the percentage of people harboring the mutation that are symptomatic, must be determined. Finally, age of onset defines the age at which individuals harboring a mutation become symptomatic. Once a base-line threshold is set based on prevalence, the allele frequency threshold should be further increased or decreased based on penetrance and age of onset.
4. Not all functional prediction tools work well with all types of proteins, as these methods may have been trained on specific types of proteins, e.g., structural, and therefore are not generalizable. For example PolyPhen2 is primarily trained on globular proteins and hence tend to work better with those kinds. Sometimes a different version of an existing tool is created for specific types of protein [55].
5. At times extensive searches of the literature are needed to interpret accurately the significance of a DNA variant. This can be time-consuming but is essential.

6. The search for causal mutation typically begins with screening the coding missense variants. However, if the search is inconclusive or negative, it is generally recommended to look for other types of mutations including synonymous, stop loss, stop gain, splicing, and noncoding.

References

1. Kruglyak L, Nickerson DA (2001) Variation is the spice of life. *Nat Genet* 27:234–236
2. Fu W, O'Connor TD, Jun G et al (2013) Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216–220
3. Yue P, Moulton J (2006) Identification and analysis of deleterious human SNPs. *J Mol Biol* 356:1263–1274
4. Ariyaratnam R, Casas JP, Whittaker J et al (2007) Genetics of ischaemic stroke among persons of non-European descent: a meta-analysis of eight genes involving approximately 32,500 individuals. *PLoS Med* 4:e131
5. Lopes LR, Rahman MS, Elliott PM (2013) A systematic review and meta-analysis of genotype-phenotype associations in patients with hypertrophic cardiomyopathy caused by sarcomeric protein mutations. *Heart* 99:1800–1811
6. McCarthy MI, Zeggini E (2009) Genome-wide association studies in type 2 diabetes. *Curr Diabetes Rep* 9:164–171
7. O'Seaghdha CM, Fox CS (2012) Genome-wide association studies of chronic kidney disease: what have we learned? *Nat Rev Nephrol* 8:89–99
8. Bolze A, Byun M, McDonald D et al (2010) Whole-exome-sequencing-based discovery of human FADD deficiency. *Am J Hum Genet* 87:873–881
9. Foroud T (2013) Whole exome sequencing of intracranial aneurysm. *Stroke* 44:S26–S28
10. Karow J (2011) Baylor Whole Genome Laboratory launches clinical exome sequencing test. <http://genomeweb.com/print/988726>
11. Sunyaev SR (2012) Inferring causality and functional significance of human coding DNA variants. *Hum Mol Genet* 21:R10–R17
12. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164
13. McLaren W, Pritchard B, Rios D et al (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26:2069–2070
14. Ng SB, Turner EH, Robertson PD et al (2009) Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 461:272–276
15. Pope BJ, Nguyen-Dumont T, Odefrey F et al (2013) FAVR (Filtering and Annotation of Variants that are Rare): methods to facilitate the analysis of rare germline genetic variants from massively parallel sequencing datasets. *BMC Bioinformatics* 14:65
16. 1000 Genome Project Consortium, Abecasis, G.R., Auton, A. et al (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56–65
17. Ngamphiw C, Assawamakin A, Xu S et al (2011) PanSNPdb: the Pan-Asian SNP genotyping database. *PLoS One* 6:e21451
18. Biesecker LG, Mullikin JC, Facio FM et al (2009) The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. *Genome Res* 19:1665–1674
19. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29:308–311
20. Tennessen JA, Bigham AW, O'Connor TD et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64–69
21. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat Rev Genet* 12:628–40
22. Adzhubei IA, Schmidt S, Peshkin L et al (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7:248–249
23. Hu J, Ng PC (2013) SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. *PLoS One* 8:e77940
24. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 4:1073–1081
25. Schwarz JM, Rodelsperger C, Schuelke M et al (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 7:575–576

26. Johnson AD, Handsaker RE, Pulit SL et al (2008) SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 24:2938–2939
27. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 39:e118
28. Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 15:978–986
29. Stenson PD, Mort M, Ball EV et al (2013) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* [Epub ahead of print]
30. Landrum MJ, Lee JM, Riley GR et al (2013) ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* [Epub ahead of print]
31. Fokkema IF, Taschner PE, Schaafsma GC et al (2011) LOVD v. 2.0: the next generation in gene variant databases. *Hum Mutat* 32: 557–563
32. Whirl-Carrillo M, McDonagh EM, Hebert JM et al (2012) Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 92:414–417
33. Duzkale H, Shen J, McLaughlin H et al (2013) A systematic approach to assessing the clinical significance of genetic variants. *Clin Genet* 84:453–63
34. Shendure J, Lieberman Aiden E (2012) The expanding scope of DNA sequencing. *Nat Biotechnol* 30:1084–1094
35. Gowrisankar S, Lemer-Ellis JP, Cox S et al (2010) Evaluation of second-generation sequencing of 19 dilated cardiomyopathy genes for clinical applications. *J Mol Diagn* 12:818–827
36. Kitzman JO, Snyder MW, Ventura M et al (2012) Noninvasive whole-genome sequencing of a human fetus. *Sci Transl Med* 4: 137ra76
37. Johnson GC, Esposito L, Barratt BJ et al (2001) Haplotype tagging for the identification of common disease genes. *Nat Genet* 29:233–237
38. Raychaudhuri S (2011) Mapping rare and common causal alleles for complex human diseases. *Cell* 147:57–69
39. Kottgen M (2007) TRPP2 and autosomal dominant polycystic kidney disease. *Biochim Biophys Acta* 1772:836–850
40. Myerowitz R (1997) Tay-Sachs disease-causing mutations and neutral polymorphisms in the Hex A gene. *Hum Mutat* 9:195–208
41. Simons C, Wolf NI, McNeil N et al (2013) A de novo mutation in the beta-tubulin gene TUBB4A results in the leukoencephalopathy hypomyelination with atrophy of the basal ganglia and cerebellum. *Am J Hum Genet* 92:767–773
42. Brauch KM, Karst ML, Herron KJ et al (2009) Mutations in ribonucleic acid binding protein gene cause familial dilated cardiomyopathy. *J Am Coll Cardiol* 54:930–941
43. Karolchik D, Barber GP, Casper J et al (2013) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* [Epub ahead of print]
44. Fan X, Tang L (2013) Aberrant and alternative splicing in skeletal system disease. *Gene* 528:21–26
45. Reese MG, Eeckman FH, Kulp D et al (1997) Improved splice site detection in genic. *J Comput Biol* 4:311–323
46. Pertea M, Lin X, Salzberg SL (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res* 29:1185–1190
47. Yang Y, Swaminathan S, Martin BK et al (2003) Aberrant splicing induced by missense mutations in BRCA1: clues from a humanized mouse model. *Hum Mol Genet* 12:2121–2131
48. Hakonarson H, Grant SF, Bradfield JP et al (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* 448:591–594
49. Sladek R, Rocheleau G, Rung J et al (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445: 881–885
50. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
51. Koboldt DC, Zhang Q, Larson DE et al (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22:568–576
52. Saunders CT, Wong WS, Swamy S et al (2012) Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28:1811–1817
53. Cibulskis K, Lawrence MS, Carter SL et al (2013) Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* 31:213–219
54. Forbes SA, Bindal N, Bamford S et al (2010) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res* 39:D945–D950
55. Jordan DM, Kiezun A, Baxter SM et al (2011) Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. *Am J Hum Genet* 88:183–192

Chapter 15

DNA Variant Databases: Current State and Future Directions

John-Paul Plazzer and Finlay Macrae

Abstract

In this chapter we aim to provide an overview of DNA variant databases, commonly known as Locus-Specific Databases (LSDBs), or Gene-Disease Specific Databases (GDSDBs), but the term *variant database* will be used for simplicity. We restrict this overview to germ-line variants, particularly as related to Mendelian diseases, which are diseases caused by a variant in a single gene. Common difficulties associated with variant databases and some proposed solutions are reviewed. Finally, systems where technical solutions have been implemented are discussed. This work will be useful for anyone wishing to establish their own variant database, or to learn about the global picture of variant databases, and the technical challenges to be overcome.

Key words Disease, Gene, Ontology, Phenotype, Standards, Variant database

Abbreviations

HGVS Human Genome Variation Society
HPO Human Phenotype Ontology
HVP Human Variome Project

1 Introduction

A genetic variant is a difference found in a DNA sequence as compared to a reference sequence. A variant database stores variants that are found during clinical diagnostic testing, for the purpose of sharing information. There are two main reasons for sharing variant information. One is to assist with the interpretation of variant pathogenicity in relation to a clinical phenotype. The other is to enable research, either on specific genes or variants, or across multiple genes and diseases [1].

A variant database stores information about variants which have been collected from one or more sources. There are three main sources of information in variant databases: (1) directly from

laboratory or clinical data, (2) indirectly by extraction from the published literature, or (3) a combination of both. The type of information collected is the gene and variant, patient demographics, and patient phenotype or disease. Additional data can be added to enhance the information and assist with clinical decision making, including *in silico* predictions, pathology information, *in vitro*/functional assays, and family history.

Variant databases can be as simple as a spreadsheet or HTML table, to more complex relational databases. They are generally divided as *variant-specific* (each variant has a record with a summary of patient or phenotype information) or *patient-specific* (each patient has a record and variants are stored against the patient). Patient-specific databases should be distinguished from patient registries which can house similar information, but in an identified form for clinical management or ethically approved research purposes.

Variant databases are often started by a small group of researchers or clinical geneticists interested in collating variant information and sharing it with the wider community. Alternatively, individual researchers may start a database on their own initiative, if the database for their gene(s) of interest does not yet exist. At the other end of the scale are larger organizations or consortiums that create databases in their fields of expertise [2]. Many varieties of databases exist, at different levels of organization—local, regional, national, and global. There are databases grouped by gene and disease and supported by organizations such as the International Society for Gastrointestinal Hereditary Tumours (InSiGHT). Variant databases are also divided by their commercial, research, or clinical focus. These databases all have one thing in common, that is, to share information about genetic variations and their role in disease. A comprehensive list of variant databases is available online: <http://www.hgvs.org/dblist/glsdb.html>.

2 Current State of Mutation Databases

Review articles have reported on the nature and quality of variant databases [2, 3]. The reviews cover the degree to which variant databases are supported by expert curation, the types of information contained within a database, and the amount of data that is available. The 2010 review by Mitropoulou et al. found only 13 % of variant databases had information for all three of the most important categories of information: (1) variant pathogenicity, (2) reference sequence, and (3) appropriate nomenclature for variant names.

Errors in reporting of genetic variants in published literature are also a recognized problem, and these errors can find their way into variant databases [4]. The best way to prevent errors in variant descriptions is to use Mutalyzer software, as shown in Subheading 4.3 on Standards. Other errors are less likely to be

detected and fixed, and submitters need to be relied upon to provide accurate information.

Another major issue is how to define adequate database structures. Most databases are now implemented through a relational database system [3]. Established databases usually have a table for variants, and commonly a second table for patient and phenotype information. However, this structure is not ideal if phenotypes and disease can change over time in a patient, or if treatments alter the course of disease development. A properly designed relational database will have a third table for phenotype, so that new phenotype records can be added for each patient as needed. This is recognized and implemented by more advanced databases. Defining fields to handle patient phenotype, family history, patient ethnicity and in vitro test results is also challenging. We will look at a potential solution for the issue of patient phenotypes.

Data for specific genes and diseases is shared by clinicians or laboratory diagnostic staff to variant databases. Historically, this has not always occurred, and unpublished variant information is held in private systems of diagnostic laboratories, clinic-based patient records or other registries [5]. Automated technologies, like Next-generation sequencing, should make sharing easier than in the past as information will be generated in a digital format and processed by software pipelines. Other factors that prevent sharing include concerns over patient privacy and consent. Guidelines concerning the ethical sharing of patient information have been formulated [6]. The most important rule is to remove identifying information from data submissions.

3 Types of Variant Databases

3.1 *Universal Databases*

Database systems designed to handle multiple genes and diseases can be called *universal* variant databases. There are two ways this is accomplished, generally divided into *locus-specific* versus *centralized* approaches [2]:

1. *Locus-specific*. Individual database tables are created and tailored for each disease or disease grouping. Therefore, such a system can be viewed as a collection of variant databases, rather than a unified whole, albeit running on the same software platform. These universal variant databases collate variants from many genes and diseases into a single system. They are specifically designed to include detailed clinical data for any disease. A major advantage of this approach is the curator for a particular variant database can maintain close supervision over the database, informed by his or her dedicated expertise relating to the gene and associated diseases. The best example is the Leiden Open Variation Database (LOVD) (<http://www.lovd.nl>) described in Subheading 3.2.

2. *Centralized.* A consistent database structure is used for all genes and diseases. This is a more standard relational database system with well-defined fields and data types. However, this approach loses the flexibility of a decentralized database in favor of centralized control. The main limitation of centralized variant databases is a decrease in disease-specific information, with a focus on more generalized disease terms. ClinVar (<http://www.ncbi.nlm.nih.gov/clinvar>), a new system developed by the National Center for Biotechnology Information (NCBI) is an excellent example of a centralized database. However, as well as accepting submissions from diagnostic laboratories, data from some locus-specific variant databases is shared with ClinVar, showing the mutually beneficial roles of both centralized and non-centralized databases. Another example of a centralized variant database is the Human Gene Mutation Database (HGMD), a commercially orientated system which stores information extracted from published literature.

There are circumstances where one approach is more suitable than the other, and both have value to the clinical genetics community. Universal databases can also provide automated integration with external databases and systems, such as the University of California Santa Cruz (UCSC) genome browser. However, any use nonstandard data limits full integration with external systems. Universal systems may also provide free hosting services for the databases and are recommended over in-house databases as they already incorporate tools for analyzing genetic variant information and are interconnected with other systems.

3.2 Leiden Open Variation Database

Leiden Open Variation Database (LOVD) could be considered the de facto standard for variant database systems. It was created to generate automatically locus-specific variant databases for any gene or disease [7]. These are then hosted by organizations on their own servers, or alternatively the LOVD developers may provide free hosting. LOVD has recently been upgraded to version 3 (LOVDv3) which includes significant enhancements to its architecture. An important feature LOVDv3 introduced is a more sophisticated patient phenotype model. This allows phenotype information to be recorded over time as a disease is diagnosed, treated, or as it progresses in a patient. LOVD also handles reference sequences automatically, and integrates with Mutalyzer and other external systems.

3.3 Universal Mutation Database

The Universal Mutation Database (UMD) (<http://www.umd.be>) system also comprises multiple genes and diseases. It incorporates tools to enable the analysis of genotype-phenotype relations across genes and diseases, a benefit for having a unified architecture [8]. The UMD central portal enables access to 37 genes

involved with cancer and non-cancer genetic diseases. Phenotypes, disease symptoms, and other clinical features can be selected and filtered by gene, regardless of whether the phenotype selected is normally associated with the gene.

3.4 Federated Databases

A potential solution to the multiplicity of incompatible databases is the notion of federated databases. This is a top-down solution that requires cooperation between the developers of databases. Database federation entails a system that can query multiple databases and return a unified response. This approach can also solve the multiple-database access problems which face clinicians and geneticists. However, for such a system to work it has to deal with issues of incompatible database systems, nonstandard disease fields, and different data formats. This approach works best when different groups adopt the same underlying database software. A federated approach has been applied in the case of FinDis (Subheading 5.3).

4 Standards

In order to accomplish efficient sharing of data and access to it, standards are necessary (*see Note 1*). They enable researchers and bioinformaticians to create systems without having to reinvent the wheel each time. Furthermore, the ability to interrogate many different databases using a common interface would have scientific value, and improve clinical decision making processes. Standards would make this all possible. However, there is a general lack of standards for variant databases, except for variant nomenclature and reference sequences [9].

Standards for variant databases can be grouped into the following categories:

- Curation (Variant descriptions, reference sequences, data sharing).
- Disease and phenotype ontologies.
- Database systems and software applications.

Standards and guidelines exist which cater to some of these categories, though they may need refinement to be fully acceptable. There is often more than one standard to choose from, and also the task of ensuring that the standards are widely adopted.

4.1 The Human Variome Project

The Human Variome Project (HVP) (<http://www.humanvariomeproject.org>) has recognized the limitations and restrictions to the free sharing of data, and has developed a global strategy for data sharing. The HVP operates through two Councils—a disease/gene specific council and a country specific council, corresponding to different ways databases are organized. It is incorporated as a nonprofit organization, has a Board with elected

directors with limited terms of office. It has high level UNESCO status as an official NGO partner, and is in negotiation with the World Health Organization for similar status. Its mission is simple: *To facilitate the documentation and sharing of all genetic variation worldwide*. It aims to accomplish this through the development of standards and guidelines that can be adopted by any group or person to improve the sharing of genetic information [9].

4.2 Minimal Database Fields

Some guidelines exist which define a minimal set of fields for variant databases. The absolute minimum is the HGVS description of the variant, in combination with a reference sequence and the submitter name or identifier. Additional preferred information includes patient gender, patient ethnicity, sample ID, patient phenotype, variant pathogenicity or classification, and a publication identifier if the variant has been published [9].

4.3 Human Genome Variation Society Nomenclature

The Human Genome Variation Society (HGVS) has designed a nomenclature to describe variations at the DNA, RNA, and protein level which has been widely adopted [10]. This is a suitable way to denote variants in databases. Importantly, a free Web application called Mutalyzer (<https://mutalyzer.nl/>) makes checking variant nomenclature easy [11]. This tool checks that a variant is correctly formatted according to the HGVS standard. It can also check a batch of variants from a file, and has a Web service component which can be called automatically from other systems. Mutalyzer's other functions include converting between genomic and reference sequence coordinates.

4.4 Locus Reference Genomic

Having a standard nomenclature description for genetic variants would be worthless without standard reference sequences for genes and proteins. Historically, this was difficult to achieve as the amount of DNA sequence data grew over time. One factor which hindered standardization was the use of ad hoc reference sequences with inconsistent nucleotides, amino acids, or even exon segments. Furthermore, when reference sequences were available, incorrect reporting of the proper versions for genomic, cDNA, and protein sequences had led to ambiguous variant descriptions.

Locus Reference Genomic (LRG) (<http://www.lrg-sequence.org/>) is a new scheme for standardizing reference sequences, which are not subject to change [12]. It encompasses all the sequence transcripts necessary for complete coverage of a gene and protein, and can handle legacy numbering formats. LRGs are ideal for variant databases and reporting of genetic variants in general. LRGs for genes or genomic regions are created only on request by individuals or groups, but are made freely available. The LRG Web site also provides a Web service for programmatic access to the reference sequences and annotated information.

4.5 Phenotype Standards

The most difficult aspect of variant databases is storing diverse disease terms and related phenotypic descriptions. It is complicated by the fact that submitters often adhere to their own terminologies. This causes a wide range of terms to be stored in a database that in some cases mean the same thing but are not easy to compare. Standard terminologies to describe diseases and associated phenotypes would allow databases to be linked, and enable computational processing and analysis of genotype–phenotype associations. A number of systems exist for describing disease including the Systematized Nomenclature of Medicine (SNOMED), Online Mendelian Inheritance in Man (OMIM), International Classification of Diseases (ICD), Medical Subject Headings (MeSH), Human Phenotype Ontology (HPO), and Disease Ontology (DO) [13]. However, a detailed examination of their suitability for variant databases is beyond the scope of this work, and we will review HPO only. In general, variant databases require a standard which allows submitters to describe diseases in a flexible way, while still using well-defined terms.

The Human Phenotype Ontology (HPO) provides standardized disease terms based on a hierarchical representation of human anatomy and phenotypes, at various levels of abstraction [14]. It is specifically aimed at the human genetics community. The HPO is suitable for variant databases given that different submitters may provide different details when describing the same disease or symptoms. Due to its hierarchical design, HPO can handle ambiguous terms that are submitted or found in biomedical literature. It also enables automated reasoning to be performed over the terms due to their relationships defined in the hierarchy.

Variant databases will need to be converted to use these standards. This may require translating existing fields to match the new standards. As many standards, including HPO, are not complete, database curators may need to develop the ontology terms in collaboration with the ontology designers.

5 Examples of Variant Databases and Technical Solutions

We describe now a few examples of variant databases that aim to provide technical solutions to the common challenges we have encountered. These examples focus on multiple genes and diseases within a disease class, which have more support in terms of funding and academic interest than would be the case with more isolated databases (*see Note 2*).

5.1 Neurogenetics

The field of neurogenetics provides a microcosm of the challenges involved in the systematic collection of variants across the human genome. There are at least 2,400 neurogenetic disorders listed in

the Orphanet database [15]. The range and diversity of the disorders, many with overlapping phenotypes, have resulted in over 1,200 databases covering almost 1,000 genes, as reviewed by Sobrido et al. [16]. The work of Köhler et al. has produced HPO terms for use in variant databases in the neurogenetics field. This was accomplished by automated processing, using existing terms listed in Orphanet. With the HPO terms in place, researchers were then able to perform computational analysis over the neurogenetics phenotypes and measure similarities between diseases. This is one example of the utility of the HPO and its applicability to variant databases.

The Alzheimer Disease & Frontotemporal Dementia Mutation (AD&FDM) Database has ten genes associated with neurodegenerative phenotypes, with fields for patient phenotype, variant frequency, functional study data, patient ethnicity, and age of onset in addition to the variant field [17]. A feature in this database is integration with the UCSC Genome Browser. It uses a system called PhenCode [18] to share data with the Genome Browser, where it may be viewed alongside useful functional and evolutionary information. This means anyone looking at relevant genes via the Genome Browser can see the variants from the AD&FDM database listed. Phencode is not in dynamic linkage with the variant database, but requires upload to a separate MySQL database. This requires some extra effort to ensure that the databases are kept in sync. However, it serves a useful and important role in bridging the gap between variant databases and centralized databases.

5.2 Immunodeficiency

Similarly, Immunodeficiency Databases (IDBases) also has links with UCSC genome browser. IDBases is a group of databases covering hundreds of genes associated with immunodeficiency diseases. This focus allows it to create an extensive database for specific genes, with well-defined fields. The advantage of combining databases for genes that are associated with a disease or syndrome is that searching for variants is streamlined for clinicians or genetic counsellors. The ultimate aim is to have all genes and diseases that are not (necessarily) related in a unified system. This will allow for so called *disease families* to be understood [14].

5.3 FinDis: A Federated National Database

The Finnish Disease Heritage Database (FinDis) (<http://www.findis.org>) records variant information for diseases that are more prevalent in Finland compared to other countries [19]. A federated system was implemented so that already existing databases could be incorporated into the FinDis network. It works using software designed to interrogate LOVD (versions 2 and 3) and other databases and presents the results in a unified portal. The software which was developed for this task is freely available on GitHub (<http://www.github.com/findis-db>). It is envisioned that other

countries could create similar systems using this software, as Nodes for the Human Variome Project.

5.4 InSiGHT

The need to centralize variant data relating to the mismatch repair gene syndromes (Lynch Syndrome) was recognized by InSiGHT in 2008. Three databases merged—the original database curated from Finland which collected the early results of mutation detection, a database of the published literature curated from Canada, and a database of functional assays from the Netherlands. The merged database is now housed on a LOVD platform and is publically available through the InSiGHT Web site (<http://www.insight-group.org>). Strengths of the InSiGHT approach, which is widely recognized as a leading example of variant databasing, is its governance through a committee which reports to InSiGHT's democratically elected Council, the incorporation of InSiGHT to minimize any medicolegal threats engendered by publication of pathogenicity assignments on the database (which if incorrect, may lead to adverse health outcomes), funding for and appointment of a full time curator, and most particularly the work of its active Variant Interpretation Committee (VIC). This expert panel is systematically addressing all variants submitted to the database for pathogenicity assignment on a 5 class system. The VIC meets by teleconference every few months and matches data around variants (published and more powerfully, unpublished sourced through calls to the InSiGHT membership) against gene-specific classification guidelines which have been refined during the work and experience of the committee. The committee currently has over 40 members from all continents. Deliberations of the Committee are published through the InSiGHT database. InSiGHT has utilized microattribution to provide improved acknowledgment to submitters of variant information. This allows the submitters of small unpublishable parcels of information to be acknowledged in a form that can be incorporated into their publication record; publishing houses are beginning to recognize the process of microattribution [20].

6 Conclusion

The field of variant databases has matured over recent years. Its importance is even more pronounced in the era of Next-generation sequencing as the scientific community grapples with the challenge of interpreting the enormous quantities of sequencing data now emerging. More assistance, funding, and recognition are needed for these databases and groups that support them such as the HVP. Readers are invited to contact the authors of this chapter to engage in this challenge, for example, by offering support for curation relating to genes and diseases of their interest (*see Note 3*).

7 Notes

1. There are already some guidelines, standards and database systems available which facilitate improved sharing of genetic information. They are important for implementing solutions that benefit everyone, and are the result of many years of accumulated experience.
2. Variant database systems and related applications are often open-source—it would be preferable to use these if possible to avoid reinventing the wheel, and improve upon them where necessary and share the updated code with the wider community.
3. There are many genes and diseases that do not have adequate database curation—this is an opportunity for enterprising individuals to become involved in the clinical informatics field.

References

1. Greenblatt MS, Brody LC, Foulkes WD et al (2008) Locus-specific databases and recommendations to strengthen their contribution to the classification of variants in cancer susceptibility genes. *Hum Mutat* 29:1273–1281
2. Claustres M, Horaitis O, Vanevski M et al (2002) Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. *Genome Res* 12:680–688
3. Mitropoulou C, Webb AJ, Mitropoulos K et al (2010) Locus-specific database domain and data content analysis: evolution and content maturation toward clinical use. *Hum Mutat* 31:1109–1116
4. Cotton RG, Auerbach AD, Beckmann JS et al (2008) Recommendations for locus-specific databases and their curation. *Hum Mutat* 29:2–5
5. Cotton RG, Al Aqeel AI, Al-Mulla F et al (2009) Capturing all disease-causing mutations for clinical and research use: toward an effortless system for the Human Variome Project. *Genet Med* 11:843–849
6. Povey S, Al Aqeel AI, Cambon-Thomsen A et al (2010) Practical guidelines addressing ethical issues pertaining to the curation of human locus-specific variation databases (LSDBs). *Hum Mutat* 31:1179–1184
7. Fokkema IF, Taschner PE, Schaafsma GC et al (2011) LOVD v. 2.0: the next generation in gene variant databases. *Hum Mutat* 32:557–563
8. Bérout C, Hamroun D, Collod-Bérout G et al (2005) UMD (Universal Mutation Database): 2005 update. *Hum Mutat* 26:184–191
9. Kohonen-Corish MR, Al-Aama JY, Auerbach AD et al (2010) How to catch all those mutations—the report of the third Human Variome Project meeting, UNESCO Paris, May 2010. *Hum Mutat* 31:1374–1381
10. den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Hum Mutat* 20:7–12
11. Wildeman M, van Ophuizen E, den Dunnen JT et al (2008) Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. *Hum Mutat* 29:6–13
12. Dagleish R, Flicek P, Cunningham F et al (2010) Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Med* 2:24
13. Schriml LM, Arze C, Nadendla S et al (2012) Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 40(D1):D940–D946
14. Robinson PN, Mundlos S (2010) The human phenotype ontology. *Clin Genet* 77:525–534
15. Köhler S, Doelken SC, Rath A et al (2012) Ontological phenotype standards for neurogenetics. *Hum Mutat* 33:1333–1339
16. Sobrido MJ, Cacheiro P, Carracedo A et al (2012) Databases for neurogenetics: introduction, overview, and challenges. *Hum Mutat* 33:1311–1314
17. Cruts M, Theuns J, Van Broeckhoven C (2012) Locus-specific mutation databases for neurodegenerative brain diseases. *Hum Mutat* 33:1340–1344

18. Giardine B, Riemer C, Hefferon T (2007) PhenCode: connecting ENCODE data with mutations and phenotype. *Hum Mutat* 28: 554–562
19. Polvi A, Linturi H, Varilo T et al (2013) The Finnish Disease Heritage database (FinDis) update—a database for the genes mutated in the Finnish Disease Heritage brought to the next-generation sequencing era. *Hum Mutat* 34:1458–1466
20. Thompson BA, Spurdle AB, Plazzer JP et al (2014) Application of a five-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants lodged on the InSiGHT locus-specific database. *Nat Genet* 46:107–115

Chapter 16

Natural Language Processing in Biomedicine: A Unified System Architecture Overview

Son Doan, Mike Conway, Tu Minh Phuong, and Lucila Ohno-Machado

Abstract

In contemporary electronic medical records much of the clinically important data—signs and symptoms, symptom severity, disease status, etc.—are not provided in structured data fields but rather are encoded in clinician-generated narrative text. Natural language processing (NLP) provides a means of unlocking this important data source for applications in clinical decision support, quality assurance, and public health. This chapter provides an overview of representative NLP systems in biomedicine based on a unified architectural view. A general architecture in an NLP system consists of two main components: *background knowledge* that includes biomedical knowledge resources and *a framework* that integrates NLP tools to process text. Systems differ in both components, which we review briefly. Additionally, the challenge facing current research efforts in biomedical NLP includes the paucity of large, publicly available annotated corpora, although initiatives that facilitate data sharing, system evaluation, and collaborative work between researchers in clinical NLP are starting to emerge.

Key words Biomedicine, Electronic medical record, Machine learning method, Natural language processing, Rule-based learning method, System architecture, Unified Medical Language System

Abbreviations

BNF	Backus–Naur form
cTAKES	Clinical Text Analysis and Knowledge Extraction System
EMR	Electronic medical record
GATE	General Architecture for Text Engineering
LSP	Linguistic String Project
MedLee	Medical Language Extraction and Encoding System
MLP	Medical language processor
NER	Named entity recognition
NLP	Natural language processing
POS	Part of speech
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System

1 Introduction

In contemporary electronic medical records (EMRs) most of the clinically important data—signs and symptoms, symptom severity, disease status, etc.—are not provided in structured data fields but are rather encoded in clinician-generated narrative text. Natural language processing (NLP) provides a means of unlocking this important data source, converting unstructured text to structured, actionable data for use in applications for clinical decision support, quality assurance, and public health surveillance. There are currently many NLP systems that have been successfully applied to biomedical text. It is not our goal to review all of them, but rather to provide an overview of how the field evolved from producing monolithic software built on platforms that were available at the time they were developed to contemporary component-based systems built on top of general frameworks. More importantly, the performance of these systems is tightly associated with their “ingredients,” i.e., modules that are used to form its background knowledge, and how these modules are combined on top of the general framework. We highlight certain systems because of their landmark status as well as on the diversity of components and frameworks on which they are based.

The Linguistic String Project (LSP) was an early project starting in 1965 that focused on medical language processing [1]. The project created a new schema for representing clinical text and a dictionary of medical terms in addition to addressing several key clinical NLP problems such as de-identification, parsing, mapping, and normalization. The system’s methodology and architecture have substantially influenced many subsequent clinical NLP systems.

One of the main requirements for developing clinical NLP systems is a suitable biomedical knowledge resource. The Unified Medical Language System (UMLS) [2], initiated in 1986 by the National Library of Medicine, is the most widely used knowledge resource in clinical NLP. The UMLS contains controlled vocabularies of biomedical concepts and provides mappings across those vocabularies.

With the development of machine learning, NLP techniques, and open-source software, tools have been developed and are now available in open source, e.g., NLTK (<http://www.nltk.org>), Mallet (<http://mallet.cs.umass.edu/>), Lingpipe (<http://alias-i.com/lingpipe/>), and OpenNLP (<http://opennlp.apache.org/>). These tools can help biomedical researchers reuse and adapt NLP tools efficiently in biomedicine. Several software frameworks that facilitate the integration of different tools into a single pipeline have been developed, such as General Architecture for Text Engineering (GATE, <http://gate.ac.uk/>) and Unstructured

Information Management Architecture (UIMA, <http://uima.apache.org/>). Given the success of IBM's Watson in the 2011 Jeopardy challenge, the UIMA framework, which was used for real-time content analysis in Watson, has now been applied widely by the biomedical NLP community. The highly recognized open-source system clinical Text Analysis and Knowledge Extraction System (cTAKES) was the first clinical NLP system to use the UIMA framework to integrate NLP components and is rapidly evolving.

In this chapter, we provide an overview of NLP systems from a unified perspective focused on system architecture. There are already comprehensive reviews and tutorials about NLP in biomedicine. Spyns provided an overview of pre-1996 biomedical NLP systems [3], while Demner-Fushman et al. more recently reviewed and summarized NLP methods and systems for clinical decision support [4]. The use of NLP in medicine has been comprehensively reviewed by Friedman [5], Nadkarni et al. [6], and more recently by Friedman and Elhadad [7]. The review in this chapter differs from previous work in that it emphasizes the historical development of landmark clinical NLP systems and presents each system in light of a unified system architecture.

We consider that each NLP system in biomedicine contains two main components: biomedical background knowledge and a framework that integrates NLP tools. In the rest of this chapter, we first outline our model architecture for NLP systems in biomedicine, before going on to review and summarize representative NLP systems, starting with an early NLP system, LSP-MLP, and closing our discussion with the presentation of a more recent system, cTAKES. Finally, we discuss challenges as well as trends in the development of current and future biomedical NLP systems.

2 Materials

2.1 A General Architecture of an NLP System in Biomedicine

We start from a discussion by Friedman and Elhadad [8] in which NLP and its various components are illustrated, as reproduced in Fig. 1. NLP aspects can be classified into two parts in the figure: the left part contains trained corpora, domain model, domain knowledge, and linguistic knowledge; the right part contains methods, tools, systems, and applications. From the viewpoint of system architecture, we consider a general architecture in which an NLP system contains two main components: *background knowledge*, which corresponds to the left part of the figure, and a *framework* that integrates NLP tools and modules, which corresponds to the right part of the figure. Our view of a general architecture is depicted in Fig. 2. Below we describe the two main components and their roles in biomedical NLP systems.

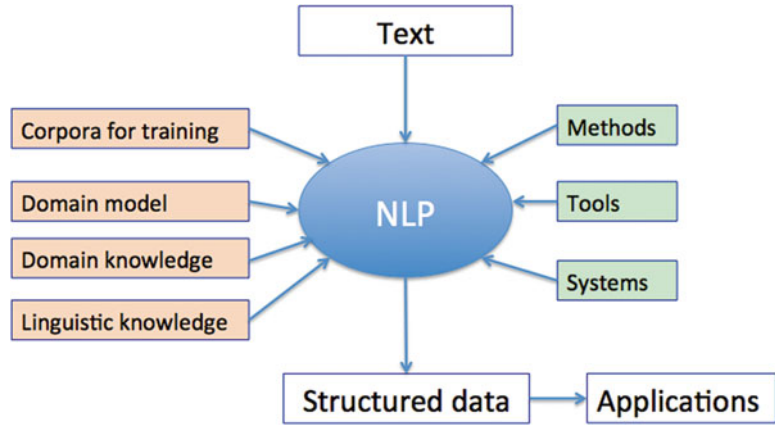


Fig. 1 Aspects of clinical NLP systems as described by Friedman and Elhadad [8]. The rectangles on the left side represent background knowledge, and the components on the right side represent the framework, i.e., algorithms and tools. Background knowledge and framework are the main components of an NLP system

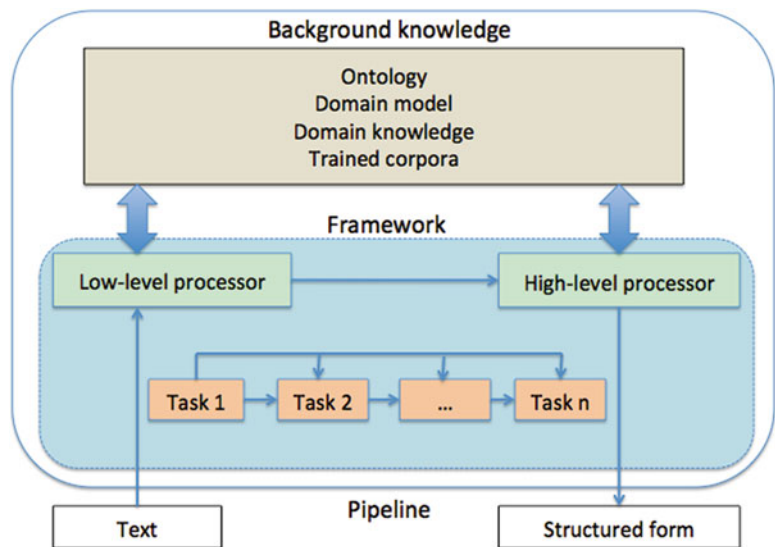


Fig. 2 The general architecture of a clinical NLP system contains two main components: background knowledge and framework. Background contains ontologies, a domain model, domain knowledge, and trained corpora. Framework includes a low-level processor for tasks such as tokenization and part-of-speech tagging. A high-level processor is used for tasks such as named entity recognition and relation extraction. Tasks or modules in the framework can be dependent or independent and are organized sequentially or hierarchically

2.1.1 *Background Knowledge for NLP in Biomedicine: The Unified Medical Language System*

As mentioned in the introduction, biomedical knowledge is an important component in building clinical NLP systems. Domain knowledge and linguistic knowledge are key elements. Earlier systems such as LSP-MLP built their own medical vocabulary and tools due to the lack of easily available resources at that time. The creation of the UMLS, which began development in 1986, substantially benefited clinical NLP systems. The UMLS contains three main components: the Metathesaurus, the Semantic Network, and the SPECIALIST lexicon. For practical purposes, the UMLS can be considered as an ontology of biomedical concepts and their relations. Each UMLS component is briefly summarized below.

- The UMLS's *Metathesaurus* currently contains over one million biomedical concepts and five millions concept names originating from over 150 controlled vocabularies in the biomedical sciences, such as ICD-10, MeSH, SNOMED CT, and RxNorm.
- The UMLS *Semantic Network* provides a consistent categorization of all concepts represented in the UMLS Metathesaurus. It reduces the complexity of the Metathesaurus by grouping concepts according to semantic types. Currently, it contains 135 broad categories and 54 relationships among categories. For example, the category *Disease or Syndrome* has a relationship "associated_with" with the category *Finding*, and the category *Hormone* has a relationship "affects" with the category *Disease or Syndrome* in the semantic network.
- The UMLS *SPECIALIST lexicon* contains syntactic, morphological, and spelling information for biomedical terms [9]. Currently, it contains over 200,000 terms and is used by the UMLS lexical tools for NLP tasks.

Background knowledge also includes domain models and trained corpora, which are used to deal with specific domains such as radiology reports, pathology reports, and discharge summaries. Annotated corpora are manually marked up by human annotators and used to train machine learning linguistic classifiers as well as to evaluate rule-based systems.

2.1.2 *NLP Tools and Integrated Frameworks*

There are two main approaches for building NLP tools. The first is rule based, which mainly uses dictionary lookup and rules. The second uses a machine learning approach that relies on annotated corpora to train learning algorithms. Early systems often used rule-based approach since they were relatively easy to design and implement. Currently, with the development of robust statistical machine learning methods and an increasing number of annotated corpora, many clinical NLP systems have moved away from relying exclusively on rule-based methods, although there is still a high cost in generating new annotated training data, which are still required to

account for differences in tasks, types of documents, as well as their provenance. As shown in many clinical NLP challenges, machine learning methods often achieve better results than rule-based methods. However, rule-based methods are somewhat easier to customize and adapt to a new domain. Most contemporary NLP systems are hybrid, i.e., built from a combination of rule-based and machine learning methods [8].

Figure 2 shows how NLP tools can be integrated into a pipeline built on top of a particular framework. By framework we mean a software platform for the control and management of components such as loading, unloading, and handling components of the pipeline. Components within a framework can be embedded and linked together or used as plug-ins. For NLP systems in biomedicine, the framework can be divided into two levels: low-level and high-level processors. *Low-level processors* perform foundational tasks in NLP such as sentence boundary detection, section tagging, part-of-speech (POS) tagging, and noun phrase chunking. *High-level processors* perform semantic level processing such as named entity recognition (NER), e.g., diseases/disorders, sign/symptoms, medications, relation extraction, and timeline extraction.

The framework can be integrated into the NLP system itself or it can leverage available general architectures with the two most widely used being GATE (<http://gate.ac.uk/>) and UIMA (<http://uima.apache.org/>). Both consist of open-source software.

GATE, written in Java, was originally developed at the University of Sheffield in 1995 and is widely used in the NLP community. It includes basic NLP tools for low-level processing such as tokenizers, sentence splitters, POS taggers, packaged in a wrapper called Collection of REusable Objects for Language Engineering (CREOLE), and a high-level processor for NER packaged in an information extraction system called ANNIE. It can integrate available NLP tools and machine learning software such as Weka (<http://www.cs.waikato.ac.nz/ml/weka/>), RASP (<http://www.sussex.ac.uk/Users/johnca/rasp/>), SVM Light (<http://svmlight.joachims.org/>), and LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>). Several clinical NLP systems have used GATE as their framework. They include HITEx (which will be in the next section) and caTIES (<http://caties.cabig.upmc.edu/>) for cancer text information extraction.

UIMA, written in Java/C++, was originally developed by IBM and is part of the Apache Software Foundation software since 2006. Its motivation is *to foster reuse of analysis components and to reduce duplication of analysis development. The pluggable architecture of UIMA allows to easily plug in your own analysis components and combine them together with others* (<http://uima.apache.org/doc-uima-why.html>).

The framework is best known as the foundation of IBM's 2011 Jeopardy challenge Watson system. UIMA's functionalities are similar to GATE but are more general since UIMA can be used for analysis of audio and video data in addition to text. There are several clinical NLP systems that use the UIMA framework such as cTAKES (described in the next section), MedKAT/P <http://ohnlp.sourceforge.net/MedKATp/> for extracting cancer-specific characteristics from text, and MedEx [10, 11] (Java version, <http://code.google.com/p/medex-uima/>) for medication extraction.

2.1.3 System Selection

In order to give a unified view of system architecture, we selected representative NLP systems in this review based on their historical importance and influence in the biomedical NLP community.

We first chose two widely influential landmark clinical NLP systems: LSP-MLP and Medical Language Extraction and Encoding (MedLEE). LSP-MLP is a pioneering project and has greatly influenced subsequent NLP systems. MedLEE is a system that is currently widely used in clinical NLP communities. We then selected a specific-purpose system called SymText, which was designed for radiology report processing. SymTex began development in the 1990s and is still in active use today. We also briefly review MetaMap, a widely used tool in the biomedical NLP community. We chose two systems based on GATE and UIMA: HITEx and cTAKES, respectively. Summaries of characteristic features of the clinical NLP systems reviewed in this chapter are presented in Table 1.

3 Methods (Systems)

3.1 Linguistic String Project: Medical Language Processor

The LSP (<http://www.cs.nyu.edu/cs/projects/lsp/>) was developed in 1965 at New York University by Sager et al. [1, 12]. It is one of the earliest research and development projects in computer processing of natural language. The development of LSP was based on the linguistic theory of Zellig Harris: linguistic string theory, transformation analysis, and sublanguage grammar [13–15]. It mainly focused on medical language processing, including the sublanguage of clinical reporting, radiograph reports, and hospital discharge summaries. The LSP approach used a parsing program to identify the syntactic relations among words in a sentence. The project strongly influenced subsequent clinical NLP projects. The LSP's system was called the Medical Language Processor (MLP).

The core component of MLP is a parser. The authors first developed a general NLP parser for the general English language domain, including English grammar and lexicon, and then they extended the system to the sublanguage of biomedicine by adding a medical lexicon and corresponding grammar. Below we summarize the main components of MLP.

Table 1
Summary of characteristic features of clinical NLP systems

System	Programming language	Creator	Framework	Open/closed source and license	Background knowledge resource	Clinical domain or source of information	Encoding
LSP-MLP	Fortran C++	New York University		Software provided by Medical Language Processing LLC corporation	Developed its own medical lexicons and terminologies	Progress notes, clinical notes, X-ray reports, discharge summary	SNOMED
MedLEE	Prolog	Columbia University		Closed source commercialized by Columbia University and Health Fidelity Inc.	Developed its own medical lexicons (MED) and terminologies	Radiology, mammography, discharge summary	UMLS's CUI
SPRUS/SymText/ MPLUS	LISP, C++	University of Utah		Closed source	UMLS	Radiology concepts from findings in radiology reports	ICD-9
MetaMap	Perl, C, Java, Prolog	National Library of Medicine		Not open source but available free under UMLS Metathesaurus License Agreement	UMLS	Biomedical text Candidate and mapping concepts from UMLS	UMLS's CUI
HITEx	Java	Harvard University	GATE	Open-source i2b2 software license	UMLS	Clinical narrative family history concept, temporal concepts, smoking status, principal diagnosis, comorbidity, negation	UMLS's CUI
cTAKES	Java	Mayo clinic and IBM	UIMA	Open-source Apache 2.0	UMLS+ trained models	Discharge summary, clinical note, clinical named entities (diseases/disorders, signs/symptoms, anatomical sites, procedures, medications), relation, co-reference, smoking status classifier, side effect annotator	UMLS's CUI and RxNorm

3.1.1 Background Knowledge

- Lexicons: MLP developed lexicons for both general English language and medical knowledge. In the lexicon, each word has an associated POS and grammatical and medical “attributes” called subclasses. The lexicon has 60 possible verb objects and 50 medical subclasses. It also had lists of predefined prepositions, abbreviations, and doses. These attributes are used throughout the processing to guide the parsing and to resolve ambiguities. Predefined lists consist of:
 - Standard numbers, times, and dates.
 - Medical terms.
 - Dose strings.
 - Organism terms.
 - Geographic nouns.
 - Patient nouns.
 - Institution/ward/service nouns.
 - Physician/staff nouns.
- Grammar: The grammar is written in Backus–Naur Form (BNF). It finds grammatical structures in clinical text and contains the following components:
 - BNF: The context-free component.
 - The RESTR (restriction) contains procedures written in the MLP’s “Restriction Language.” Those procedures test the parse tree for the presence or the absence of particular features.
 - The LISTS contains lists used in procedures other than RESTR.

3.1.2 Pipeline

- The *preprocessor* breaks input text into sentences. Then, the preprocessor identifies possible spelling errors, abbreviations, and all forms for names of patients, staffs, facilities, and administrative and geographic areas for de-identification. Numbers, units, and dates are transformed into ANSI standard format.
- The *MLP parser* uses a top-down, context-free grammar-based parser. The system generates multiple parses of ambiguous sentences guided by a BNF grammar. The parser was originally written in FORTRAN and then partly converted into Prolog [16]. Today it is written in C++. The MLP system is now publicly available through the Web site provided by Medical Language Processing, LLC—a Colorado corporation (<http://mlp-xml.sourceforge.net/>).

The parser proceeds from left to right through the sentence and top to bottom through the BNF definitions. Once the parser associates a terminal symbol of the parse tree, the attributes of the word can be tested by a restriction, for example, the agreement of subject and verb. The following steps are involved in the processing of text:

- *Selection* passes or rejects a parse based on subtrees.
- *Transformation* decomposes sentences into their basic canonical sentences.
- *Regularization* connects basic canonical sentences by conjunctions.
- *Information format* maps the syntactic parse trees into medical information structures. MLP considers 11 information structures related to patients such as patients, family, medication, treatments, and laboratory test.

Finally, the output is written into two formats: tab-delimited and XML format.

LSP-MLP was used for processing clinical narratives in English, and it was also extended into other languages such as French, German, and Dutch [1]. It has been used to map clinical text into SNOMED codes [17, 18]. LSP-MLP was designed for information retrieval from clinical text; hence, there were no reports evaluating mapping. The performance in information retrieval tasks indicated 92.5 % recall and 98.6 % precision [18]. With its complete structures, LSP-MLP provided an early successful example for the development of subsequent NLP systems.

3.2 MedLEE

The MedLEE system was developed by Friedman et al. at Columbia University [19, 20] in 1994. It was first designed for radiology reports and was then extended to other domains such as discharge summaries. The system was written in Quintus Prolog. MedLEE contains two main components: (1) a knowledge base including medical concepts and (2) a natural language processor. MedLee was the first NLP system used as part of a system for actual patient care, and some systems in which it was embedded have been shown to improve care [21, 22]. It was commercialized in 2008. The architecture of MedLEE is depicted in Fig. 3.

3.2.1 Background Knowledge

MedLEE's knowledge base is called the Medical Entities Dictionary (MED) [20], which contains a knowledge base of medical concepts and their taxonomic and semantic relations. Each concept in MED is assigned to an identifier. The MED originally contained over 34,000 concepts.

3.2.2 Pipeline

The natural language processor has three phases of processing as follows.

- *Phase 1: Parsing.* Identifies the structures of the text through the use of a grammar. It contains three main components: a set of grammar rules, semantic patterns, and lexicon.
 - Grammar rules: MedLEE uses a BNF grammar which originally contained 350 grammar rules.

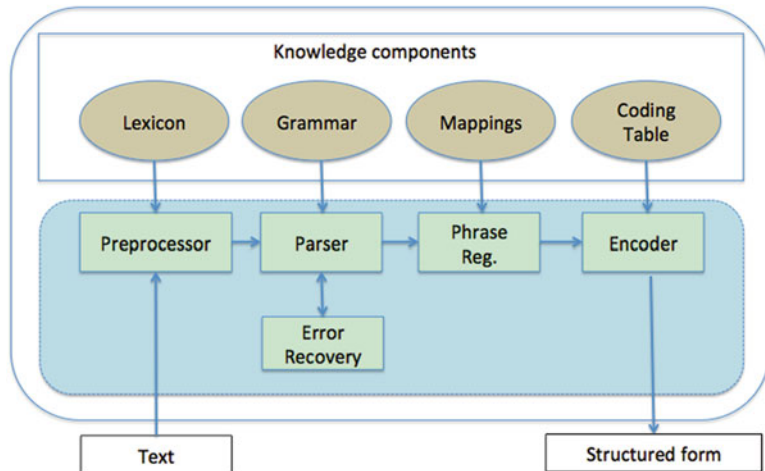


Fig. 3 Architecture of MedLEE, where the background knowledge contains components for the lexicon, grammar, mappings, and coding tables. The low-level processor is a preprocessor, and the high-level processor consists of modules for parsing, error recovery, phrase regularization, and encoding

- Semantic classes: MedLEE considers sentences that contain semantic patterns connected by conjunctions. Semantic patterns can be a word or a phrase and/or belong to a semantic class. Examples of semantic classes are Bodyloc, Cfinding, and Disease. MedLEE also considers negation as a semantic pattern in its grammar.
- Lexicon: The semantic lexicon originally contained both single words (1,700) and phrases (1,400).
- *Phase 2: Phrase regularization.* This module regularizes the output forms of phrases that are not contiguous. This is a critical step that further reduces the variety that occurs in natural language. The method is automatically applied by processing all phrasal lexical entries that begin with the symbol phrase. Phrase is used to specify that a phrase may occur in a non-contiguous variant form.
- *Phase 3: Encoding.* This step maps the regularized structured forms to controlled vocabulary concepts. This process is accomplished using a knowledge base containing synonymous terms. The synonym knowledge base consists of associations between standard output forms and a controlled vocabulary. At the end of this stage of processing, the only values that remain are unique controlled vocabulary concepts.

The output of MedLEE is represented as a formal model of clinical information in the domain of interest such as radiology. It has been extended to map extracted concepts into UMLS codes [23], and its architecture was also extended to build an information

extraction system for molecular pathways from journal articles [24]. Evaluation on 150 random sentences from clinical documents achieved 0.77 recall and 0.83 precision compared to 0.69–0.91 recall and 0.61–0.91 precision for seven domain experts performing the same tasks [23].

3.3 *SPRUS/ SymText/ MPLUS*

SPRUS/SymText/MPLUS [25–28] was developed in 1994 by Haug et al. at the University of Utah. It has been implemented using common LISP, the Common Lisp Object System (CLOS), and C++. The original system was called SPRUS, and it evolved into Symbolic Text Processor (SymText), Natural Language Understanding System (NLUS), and the latest version of system, MPLUS (M++). The system was specifically designed for processing chest radiograph reports.

3.3.1 *Background Knowledge*

- SPECIALIST lexicon from UMLS, a synonyms database, POS lexicon.
- An Augmented Transition Network (ATN) grammar, a transformational rule base, and a set of resolution strategy rules.
- Knowledge bases also contain belief network node structures, values, and training cases for each context. The context was predefined such as events in chest radiology reports.

3.3.2 *Pipeline*

SymText consists of three primary modules for the analysis and interpretation of sentences [27].

- First, a structural analyzer generates an initial structural interpretation of a sentence.
- Second, a transformational module transforms the initial structure according to the targeted semantic contexts.
- Third, a resolution module semantically resolves the conceptualizations of the text according to its structure. Encoded data are the system's outputs.

SymText's outputs contain three semantic concepts: finding, disease, and appliances (devices).

The distinct feature of SymText when compared to other systems is that it uses belief networks to represent biomedical domain knowledge and discover relationships between nodes within parse trees. SymText has been used in several applications such as mapping chief complaints into ICD-9 codes [29] and extracting pneumonia-related findings from chest radiograph reports [30, 31].

Evaluation using 292 chest radiograph reports to identify pneumonia-related concepts showed that the system achieved 0.94 recall, 0.78 precision, and 0.84 specificity, outperforming lay persons [31]. MPLUS was evaluated for the extraction of American College of Radiology utilization review codes from 600 head CT reports. The system achieved 0.87 recall, 0.98 specificity, and 0.85 precision in identifying reports as positive, i.e., containing brain findings [28].

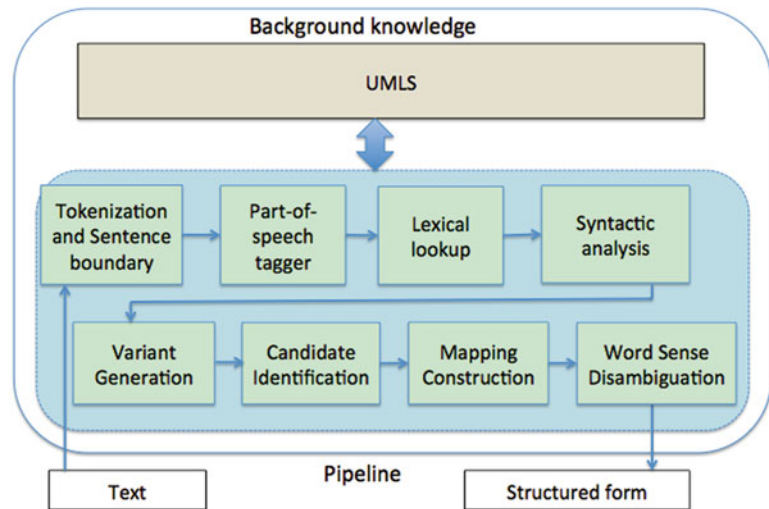


Fig. 4 Architecture of the MetaMap system, modified from the original [33], where background knowledge is based on UMLS and different modules represent the pipeline

3.4 MetaMap

MetaMap (<http://metamap.nlm.nih.gov/>) [32, 33] was originally developed in 1994 by Aronson at the National Library of Medicine. It was created for mapping the biomedical literature to concepts in the UMLS Metathesaurus [2]. It has been widely used for processing clinical text [34–36]. The tool uses a variety of linguistic processes to map from text to Concept Unique Identifiers (CUI) in the UMLS. It is written in Perl, C, Java, and Prolog. The architecture of MetaMap is depicted in Fig. 4.

3.4.1 Background Knowledge

The UMLS is used as the knowledge resource.

3.4.2 Pipeline

The most recent version of the system, as described by Aronson and Lang [33], has a two-stage architecture:

- Lexical/syntactic processing:
 - Tokenization (including sentence splitting and acronym expansion).
 - POS tagging.
 - Lexical lookup that uses the UMLS SPECIALIST lexicon.
 - Syntactic analysis that generates phrases for further processing.
- Phrasal processing:
 - A table lookup is used to identify variants of phrase words.
 - Candidate identification identifies and ranks strings from the UMLS that match phrasal terms.

- Mapping to text through selection, combination, and mapping of candidates to the text.
- Word sense disambiguation selects senses consistent with the surrounding text.

MetaMap's output can be provided in XML format, MetaMap Output (MMO), or human-readable (HR) formats. Since its initial development MetaMap has been used in a variety of clinical text processing tasks. For example, Shah et al. [34] used it to extract the cause of death from EMRs, while Meystre et al. [35] used it to extract medication information from the clinical record. Pakhomov et al. [36] used MetaMap to extract health-related quality of life indicators from diabetes patients described in physician notes. Recently, Doan et al. [37] used MetaMap for phenotype mapping in the PhenDisco system, a new information retrieval system for the National Center for Biotechnology Information's database of genotypes and phenotypes (dbGaP <http://www.ncbi.nlm.nih.gov/gap>).

The MetaMap tool is highly configurable, consisting of advanced features such as *negation detection* (using the NegEx algorithm described in Chapman et al. [38]) and *word sense disambiguation*. Although not open source, the software is freely available from the National Library of Medicine as a stand-alone command-line tool implemented primarily in Prolog. In addition to the Prolog version of MetaMap, a Web-based interface is available that facilitates simple queries and also batch processing of text. Furthermore, a Java implementation of MetaMap, MMTx, is available although this version is no longer under active development.

MetaMap was used by the NLM team in the 2009 i2b2 challenge on medication extraction. It achieved an F-score of 0.803, with precision 0.784 and recall 0.823. Although it ranked fourth in the challenge, it had the highest recall among participating teams [39, 40]. Another system that used MetaMap, Textinator, developed by Meystre et al. was also among the top ten in that competition [35, 40].

3.5 HITex

Health Information Text Extraction (HITex, http://www.i2b2.org/software/projects/hitex/hitex_manual.html) is an open-source NLP system (under i2b2 software license) developed at Brigham and Women's Hospital and Harvard Medical School. It was built based on the GATE framework. The system leverages a set of NLP modules known as CREOLE in GATE for low-level processing, such as sentence splitting and POS tagging. Other components for high-level processor, such as a UMLS mapper and classifier, were developed as plug-in components and are easily handled for loading/reloading. The architecture of HITex is depicted in Fig. 5.

3.5.1 Background Knowledge

HITex uses UMLS for background knowledge. It has trained corpora for several tasks such as building a classifier for smoking status.

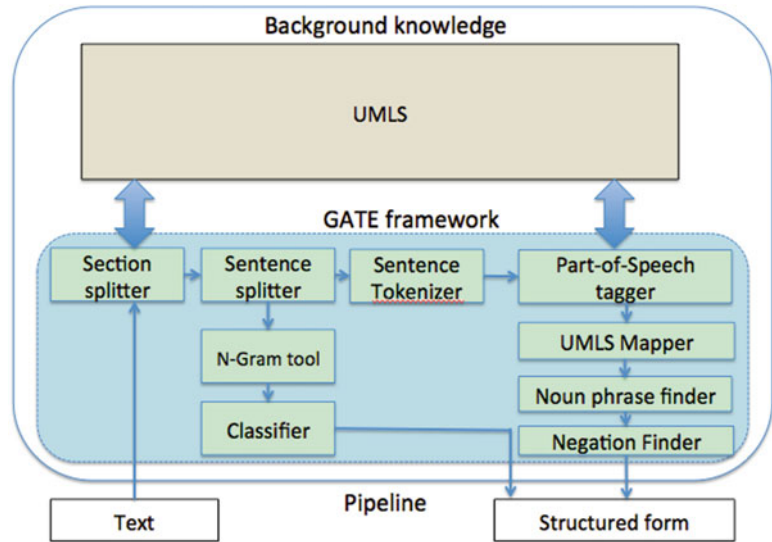


Fig. 5 Architecture of HITEx system, simplified from the original publication by Zeng et al. [41]

3.5.2 Pipeline

HITEx contains the following modules integrated in the GATE framework.

- The *section splitter/filter* splits clinical reports into sections and assigns them to section headers. There are over 1,000 section headers in HITEx. Then it filters sections based on selection criteria such as section names.
- The *sentence splitter* breaks sections into sentences. It is based on regular based rules.
- The *sentence tokenizer* breaks sentences into words; it uses an extensive set of regular expressions that define both token delimiters and special cases.
- The *POS tagger* assigns POS tags to each token in the sentence. This module is a rule-based POS tagger as a plug-in for the GATE framework.
- The *noun phrase finder* groups POS-tagged words into the noun phrases using the set of rules and the lexicon. This module is a plug-in for the GATE framework.
- The *UMLS mapper* associates the strings of text to UMLS concepts. It uses a UMLS dictionary lookup: it first attempts to find exact matches, and when exact matches are not found it stems, normalizes, and truncates the string.
- The *negation finder* assigns the negation modifier to existing UMLS concepts. It used the NegEx algorithm [38].
- The *N-Gram tool* extracts n-word text fragments along with their frequency from a collection of text.

- The *classifier* takes a smoking-related sentence to determine the smoking status of a patient. It determines one of the following classes: *current smoker*, *never smoked*, *denies smoking*, *past smoker*, or *not mentioned*.

The system has been used for the extraction of family history from 150 discharge summaries, with accuracies of 0.82 for principal diagnosis, 0.87 for comorbidity, and 0.90 for smoking status extraction, when excluding cases labeled *insufficient data* in the gold standard [41, 42].

3.6 cTAKES

The cTAKES (<http://ctakes.apache.org/>) system [43], initiated by a Mayo-IBM collaboration in 2000, was first released as an open-source toolkit in 2009 by Savova et al. It is an open-source software system under the Apache v2.0 license and is widely used by multiple institutions. The system leverages NLP tools from OpenNLP [44] with trained clinical data from the Mayo Clinic. It is the first clinical NLP system to adopt UIMA as its framework.

3.6.1 Background Knowledge

cTAKES used trained corpora from Mayo clinic data and other sources, utilizing the UMLS as the main background knowledge. Trained corpora were used for low-level processing such as sentence splitting and tokenizing. The UMLS was used for NER lookup.

3.6.2 Pipeline

cTAKES employs a number of rule-based and machine learning methods. The system can take inputs in plain text or in XML format. It initially included these basic components:

- The *sentence boundary detector* extends OpenNLP's supervised maximum entropy sentence detection tool.
- The *tokenizer* breaks sentences into tokens and applies rules to create tokens for date, time, fraction, measurement, person title, range, and roman numerals.
- The *normalizer* maps multiple mentions of the same word that do not have the same string in the input data. It leverages the SPECIALIST NLP tools (<http://www.specialist.nlm.nih.gov/>) from the National Library of Medicine.
- The *POS tagger* and the *shallow parser* are wrappers around OpenNLP's modules.
- The *NER* uses a dictionary lookup based on noun phrase matching. The dictionary resource is from UMLS. It maps words into UMLS semantic types including diseases/disorders, signs/symptoms, procedure, anatomy, and medications. After being mapped into semantic types, name entities are also mapped into UMLS's CUIs.

cTAKES incorporates the NegEx algorithm [38] for detecting negation from clinical text. Since UIMA is a framework that can easily adapt to new modules, cTAKES integrates other modules such as an assertion module, a dependency parser, a constituency parser, a semantic role labeller, a co-reference resolver, a relation extractor, and a smoker status classifier.

There has been considerable focus on the evaluation of cTAKES core preprocessing modules. The sentence boundary detector achieved an accuracy of 0.949, while tokenizer accuracy was also very high at 0.949. Both POS tagger and shallow parsing performed well, achieving accuracies of 0.936 and 0.924, respectively. For NER, the system achieved a 0.715 F-score for exact and a 0.824 F-score for overlapping span [43].

cTAKES was first applied to phenotype extraction studies [43] and then was extended to identify document-level patient smoking status [45] and patient-level summarization in the first i2b2 challenge [46]. The system was used to generate features for a state-of-the-art system in the 2010 i2b2 challenge on relation extraction of medical problems, tests, and treatments [47].

4 Conclusions

We have provided an overview of several clinical NLP systems under a unified architectural view. Background knowledge plays a crucial role in any clinical NLP task, and currently the UMLS is a major background knowledge component of most systems. Rule-based approaches utilizing the UMLS are still dominant in many clinical NLP systems. Rule-based NLP systems have historically achieved very good performance within specific domains and document types such as radiology reports and discharge summaries. One of the main reasons for using a rule-based approach is that rules are relatively easy to customize and adapt to new domains as well as to different types of clinical text.

Earlier NLP systems such as LSP-MLP and MedLEE comprise “hard coded” system modules that do not facilitate reuse. The development of general frameworks such as GATE and UIMA allows sub-tasks or modules to be developed independently and integrated easily into the framework. Machine learning algorithms have been shown to benefit significantly NLP sub-tasks such as NER. Therefore, they can serve as independent modules to be integrated into a framework to improve a sub-task in a clinical NLP system. The combination of machine learning and rule-based approaches in a single hybrid NLP system often achieves better performance than systems based on a single approach. In recent years, a clear trend has developed towards creating reusable NLP modules within open-source frameworks like GATE and UIMA.

The main limitation of machine learning when compared to rule-based approaches is that rule-based systems do not require significant amounts of expensive, manually annotated training data machine learning algorithms typically do. This problem is exacerbated in the biomedical domain, where suitably qualified annotators can be both hard to find and prohibitively expensive [48, 49].

There is an increasing trend towards building community-wide resources and tools for clinical NLP. There have been several shared tasks that bring researchers in clinical NLP together to solve, evaluate, and compare different methods. Additionally, there are shared computing resources that aggregate several NLP tools to facilitate the work of researchers, such as the NLP environment in iDASH [50]. The Online Registry of Biomedical Informatics Tools (ORBIT <http://orbit.nlm.nih.gov>) project is another platform allowing sharing and collaborating for biomedical researchers in order to create and maintain a software registry, in addition to knowledge bases and data sets.

A unified overview of a few exemplary NLP systems has been presented from the architectural perspective that all these systems have two important components: background knowledge and a computational framework. How these components are constructed and integrated into pipelines for biomedical NLP is a critical determinant for their performance. Applications that benefit from biomedical NLP systems, such as EMR linking to genomic information [51], are likely to have great utilization in the next few years.

Acknowledgements

S.D. and L.O.M. were funded in part by NIH grants U54HL108460 and UH3HL108785.

References

1. Sager N, Friedman C, Lyman M (1987) Medical language processing: computer management of narrative data. Addison-Wesley, Reading, MA
2. Lindberg DA, Humphreys BL, McCray AT (1993) The Unified Medical Language System. *Methods Inf Med* 32:281–291
3. Spyns P (1996) Natural language processing in medicine: an overview. *Methods Inf Med* 35: 285–301
4. Demner-Fushman D, Chapman WW, McDonald CJ (2009) What can natural language processing do for clinical decision support? *J Biomed Inform* 42:760–772
5. Friedman C (2005) Semantic text parsing for patient records. In: Chun H, Fuller S, Friedman C et al (eds) Knowledge management and data mining in biomedicine. Springer, New York, pp 423–448
6. Nadkarni PM, Ohno-Machado L, Chapman WW (2011) Natural language processing: an introduction. *J Am Med Inform Assoc* 18: 544–551
7. Friedman C, Elhadad N (2014) Natural language processing in health care and biomedicine. In: Shortliffe EH, Cimino J (eds) Biomedical informatics; computer applications in health care and biomedicine. Springer, London, pp 255–284
8. Friedman C, Rindfleisch TC, Corn M (2013) Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *J Biomed Inform* 46:765–773

9. McCray AT, Srinivasan S, Browne AC (1994) Lexical methods for managing variation in biomedical terminologies. *Proc Annu Symp Comput Appl Med Care* 1994:235–239
10. Xu H, Stenner SP, Doan S et al (2010) MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 17: 19–24
11. Doan S, Bastarache L, Klimkowski S et al (2010) Integrating existing natural language processing tools for medication extraction from discharge summaries. *J Am Med Inform Assoc* 17:528–531
12. Sager N, Lyman M, Bucknall C et al (1994) Natural language processing and the representation of clinical data. *J Am Med Inform Assoc* 1:142–160
13. Harris Z (1968) *Mathematical structures of language*. Wiley, New York
14. Harris Z (1982) *A Grammar of English on mathematical principles*. Wiley, Australia
15. Harris Z (1991) *A theory of language and information: a mathematical approach*. Clarendon, Oxford
16. Hirschman L, Puder K (1985) Restriction grammar: a Prolog implementation. In: Warren D, van Canegham M (eds) *Logic programming and its applications*. Ablex Publishing Corporation, Norwood, NJ, pp 244–261
17. Sager N, Lyman M, Nhàn NT et al (1994) Automatic encoding into SNOMED III: a preliminary investigation. *Proc Annu Symp Comput Appl Med Care* 1994:230–234
18. Sager N, Lyman M, Nhàn NT et al (1995) Medical language processing: applications to patient data representation and automatic encoding. *Methods Inf Med* 34:140–146
19. Friedman C, Alderson PO, Austin JH et al (1994) A general natural-language processor for clinical radiology. *J Am Med Inform Assoc* 1:161–174
20. Friedman C, Cimino JJ, Johnson SB (1994) A schema for representing medical language applied to clinical radiology. *J Am Med Inform Assoc* 1:233–248
21. Knirsch CA, Jain NL, Pablos-Mendez A et al (1998) Respiratory isolation of tuberculosis patients using clinical guidelines and an automated clinical decision support system. *Infect Control Hosp Epidemiol* 19:94–100
22. Friedman C, Hripcsak G (1999) Natural language processing and its future in medicine. *Acad Med* 74:890–895
23. Friedman C, Shagina L, Lussier Y et al (2004) Automated encoding of clinical documents based on natural language processing. *J Am Med Inform Assoc* 11:392–402
24. Friedman C, Kra P, Yu H et al (2001) GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics* 17:S74–S82
25. Haug P, Koehler S, Lau LM et al (1994) A natural language understanding system combining syntactic and semantic techniques. *Proc Annu Symp Comput Appl Med Care* 1994: 247–251
26. Haug PJ, Koehler S, Lau LM et al (1995) Experience with a mixed semantic/syntactic parser. *Proc Annu Symp Comput Appl Med Care* 1995:284–288
27. Koehler S (1998) *SymText: a natural language understanding system for encoding free text medical data*. Doctor Dissertation, University of Utah. ISBN:0-591-82476-0
28. Christensen LM, Haug PJ, Fiszman M (2002) MPLUS: a probabilistic medical language understanding system. In: *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, vol 3, pp 29–36
29. Haug PJ, Christensen L, Gundersen M et al (1997) A natural language parsing system for encoding admitting diagnoses. *Proc AMIA Annu Fall Symp* 1997:814–818
30. Fiszman M, Chapman WW, Evans SR et al (1999) Automatic identification of pneumonia related concepts on chest x-ray reports. *Proc AMIA Symp* 1999:67–71
31. Fiszman M, Chapman WW, Aronsky D et al (2000) Automatic detection of acute bacterial pneumonia from chest X-ray reports. *J Am Med Inform Assoc* 7:593–604
32. Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001: 17–21
33. Aronson AR, Lang F-M (2010) An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 17: 229–236
34. Shah PK, Perez-Iratxeta C, Bork P et al (2003) Information extraction from full-text scientific articles: where are the keywords? *BMC Bioinformatics* 4:20
35. Meystre SM, Thibault J, Shen S et al (2010) Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *J Am Med Inform Assoc* 17:559–562
36. Pakhomov S, Shah N, Hanson P et al (2008) Automatic quality of life prediction using electronic medical records. *AMIA Annu Symp Proc* 2008:545–549
37. Doan S, Lin K-W, Conway M et al (2014) PhenDisco: phenotype diversity system for the

- database of genotypes and phenotypes. *J Am Med Inform Assoc* 21:31–36
38. Chapman WW, Bridewell W, Hanbury P et al (2001) A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 34:301–310
 39. Mork JG, Bodenreider O, Demner-Fushman D et al (2010) Extracting Rx information from clinical narrative. *J Am Med Inform Assoc* 17: 536–539
 40. Uzuner O, Solti I, Cadag E (2010) Extracting medication information from clinical text. *J Am Med Inform Assoc* 17:514–518
 41. Zeng QT, Goryachev S, Weiss S et al (2006) Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med Inform Decis Mak* 6:30
 42. Goryachev S, Sordo M, Zeng QT (2006) A suite of natural language processing tools developed for the I2B2 project. *AMIA Annu Symp Proc* 2006:931
 43. Savova GK, Masanz JJ, Ogren PV et al (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 17:507–513
 44. Apache Software Foundation OpenNLP. <http://opennlp.apache.org/>
 45. Savova GK, Ogren PV, Duffy PH et al (2008) Mayo clinic NLP system for patient smoking status identification. *J Am Med Inform Assoc* 15:25–28
 46. Sohn S, Savova GK (2009) Mayo clinic smoking status classification system: extensions and improvements. *AMIA Annu Symp Proc* 2009: 619–623
 47. de Bruijn B, Cherry C, Kiritchenko S et al (2011) Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010. *J Am Med Inform Assoc* 18:557–562
 48. Albright D, Lanfranchi A, Fredriksen A et al (2012) Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc* 20:922–930
 49. Chapman WW, Nadkarni PM, Hirschman L et al (2011) Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *J Am Med Inform Assoc* 18:540–543
 50. Ohno-Machado L, Bafna V, Boxwala AA et al (2012) iDASH: integrating data for analysis, anonymization, and sharing. *J Am Med Inform Assoc* 19:196–201
 51. Denny JC (2012) Chapter 13: mining electronic health records in the genomics era. *PLoS Comput Biol* 8:e1002823

Chapter 17

Candidate Gene Discovery and Prioritization in Rare Diseases

Anil G. Jegga

Abstract

A rare or orphan disorder is any disease that affects a small percentage of the population. Most genes and pathways underlying these disorders remain unknown. High-throughput techniques are frequently applied to detect disease candidate genes. The speed and affordability of sequencing following recent technological advances while advantageous are accompanied by the problem of data deluge. Furthermore, experimental validation of disease candidate genes is both time-consuming and expensive. Therefore, several computational approaches have been developed to identify the most promising candidates for follow-up studies. Based on the *guilt by association* principle, most of these approaches use prior knowledge about a disease of interest to discover and rank novel candidate genes. In this chapter, a brief overview of some of the in silico strategies for candidate gene prioritization is provided. To demonstrate their utility in rare disease research, a Web-based computational suite of tools that use integrated heterogeneous data sources for ranking disease candidate genes is used to demonstrate how to run typical queries using this system.

Key words Gene prioritization, Gene ranking, Test set, Training/seed gene set, Orphan disease, Rare disease

Abbreviations

HSP	Hereditary spastic paraparesis
NCL	Neuronal ceroid lipofuscinosis
OD	Orphan disease
OMIM	Online Mendelian Inheritance in Man
PPIN	Protein–protein interaction networks

1 Introduction

A rare or orphan disease (OD) is any disease that affects a small percentage of the population. Most of the known ODs appear early in life and are genetic. Hence, they are present throughout the life of an affected individual. A large number involve children

and about 30 % of affected children die before the age of five. In the USA, the Rare Disease Act of 2002 defines an OD as any disease or condition that affects fewer than 200,000 persons in the USA. Although the incidence of individual ODs may be small, cumulatively, the 8,000 known ODs affect about 25 million Americans, or about 10 % of the US population [1]. Further, about 250 new ODs and conditions are described each year [2]. While some of the ODs like cystic fibrosis are well known and well studied, the majority continue to be understudied mainly because a large number affect very few individuals.

Despite the advances in genome-wide techniques such as linkage analysis and association studies, the selected disease loci are usually chromosomal regions and do not represent the precise location of a *gene*. A *locus* thus identified typically contains several hundred candidate genes. For example, in the OMIM database, >900 rare disorders are described that have been mapped to one or more such gene map loci and are classified as having an *unknown molecular basis* (OMIM IDs prefixed with “#”). The prioritization of the positional candidate genes in these rare disorder loci constitutes an important step to facilitate gene identification for further experimental studies. To this effect, as shown in Table 1, several candidate gene prioritization methods have been developed [3–13] (additional references in [14]). While most of these computational approaches are based on the assumption that similar phenotypes are caused by genes with comparable or related functions [4, 13, 15–17], they differ by the data sources utilized and the algorithms used for calculating similarity [18]. Some of these state-of-art approaches (e.g., ENDEAVOUR [7, 18] and ToppGene [12, 13]) use an extensive set of gene features and data sources in computing similarities between known and candidate sets of genes for a disease.

An alternate set of approaches adopt similar or modified algorithms used to analyze social and Web networks for disease gene identification and ranking because biological networks have been found to be comparable to communication and social networks through commonalities such as scale-freeness and small-world properties [19]. These network-based approaches predominantly use protein–protein interaction networks (PPIN) and the candidate genes are typically ranked based on their connectivity to known disease genes (described as a *training* or *seed* set). While PPINs have been used widely to identify novel disease candidate genes [20–24], several recent studies [22, 23, 25–27] report also using them for candidate gene prioritization.

Recent technological advances in whole-genome or exome sequencing are increasingly used to search for Mendelian disease genes in an unbiased manner [28]. Of the ODs with a known causal gene mutation, about 70 % are monogenic [29] and according to the current version of OMIM, there are about 5,000 monogenic ODs and for half of these the underlying genes remain unknown.

Table 1
List of some of the bioinformatics approaches and tools to rank human disease candidate genes^a

Approach	URL	Data types used	Training set (input)
Genes2Diseases [48, 49]	http://www.ogic.ca/projects/g2d_2/	Sequence, Gene Ontology (GO), literature mining	Phenotype GO terms Known genes
BITOLA [50]	http://www.mf.uni-lj.si/bitola/	Literature mining	Concept
GeneSeeker [51, 52]	http://www.cmbi.ru.nl/GeneSeeker/	Expression, phenotype, literature mining	N/A
GFINDER [53, 54]	http://www.bioinformatics.polimi.it/GFINDER/	Expression, phenotype	N/A
ToppGene [13]	http://toppgene.cchmc.org	Mouse phenotype, expression, GO, pathways, literature mining	Known genes
ToppNet [27]	http://toppgene.cchmc.org	Protein interactions	Known genes
Endeavour [7]	http://www.esat.kuleuven.be/endeavour	Sequence, expression, GO, pathways, literature mining	Known genes
Gene Weaver [55]	http://www.GeneWeaver.org	Variety of gene annotations	Known genes
TargetMine [56]	http://targetmine.nibio.go.jp	Gene annotations and protein interactions	Known genes
ProphNet	http://genome2.ugr.es/prophnet	Gene annotations and protein interactions	Known genes
BioGraph [57]	http://www.biograph.be	Gene annotations and protein interactions	Known genes or keywords
PosMed [58]	http://biosparql.org/PosMed	Gene annotations and protein interactions	Known genes or keywords

^aThe *first column* has the source or the name of the tool (including reference, if available) while the *second column* has the URL of the corresponding Web application. At the time of writing this manuscript, all the URLs were functional. The *third column* shows the list of genomic annotation types/features used by each of the methods for candidate gene ranking. The *last column* has details of the training or the input data, if used (*Note: This list is extensive, but not exhaustive; reference [14] provides an additional list of tools*)

The OD causal gene identification thus represents the first step to a better understanding of the pathophysiological mechanisms underlying ODs, which in turn can lead to developing effective therapeutic interventions. While massively parallel DNA sequencing technologies have rendered the whole-genome re-sequencing of individual humans increasingly practical, the associated expenses are still a hurdle. Since ~85 % of the known genetic causes for Mendelian disorders affect the protein-coding exonic regions [30], an alternative approach involves the targeted re-sequencing of all protein-coding subsequences (exome sequencing), potentially

overcoming the financial hurdle [31, 32]. Thus, whole-exome sequencing for identifying causative mutations in ODs is likely to become the most commonly used tool for OD gene identification [28]. However, exome sequencing has some limitations: (1) The cause of a disease could be a noncoding variation or a large indel or structural genomic variant, all of which are missed by exomic sequencing, and (2) Some variants may not be identified because of lack of sequence coverage across the variant or due to technical errors, e.g., bioinformatics variant calling issues [28].

In the following sections two examples from recently published studies [33, 34] are presented showing how the bioinformatics-based disease-network analysis approach is used along with exome sequencing, to identify and prioritize novel orphan disease causal variants.

In the first study, Erlich et al. [34] used exome re-sequencing experiments along with bioinformatics-based approaches successfully to prioritize OD candidate genes illustrating the potential of combining genomic variant and gene level information to identify and rank novel causal variants. Three different candidate gene prioritization tools (Endeavour [7], ToppGene [13], and SUSPECTS [11]) were used to prioritize the most likely candidate gene for hereditary spastic paraparesis (HSP). Briefly, a familial case of HSP was first analyzed through whole-exome sequencing and four largest homozygous regions (containing 44 genes) were identified as potential HSP loci. This list was further narrowed using several filters. For example, a gene was considered as potentially causative if it contained at least one variant that is either under purifying selection or not inherited from the parents or absent in dbSNP or the 1,000 Genomes Project data. Because the majority of the known OD variants as mentioned earlier affect coding sequences, the authors also checked whether the variant was non-synonymous. After the multistep filtering step, 15 candidate genes were identified which were then subjected to computational ranking or prioritization using three methods (Endeavour [7], ToppGene [13], and SUSPECTS [11]), each of which uses different types of data and algorithms for prioritization. As a training set, a list of 11 seed genes associated with a pure type of HSP was compiled through literature mining. Interestingly, the top-ranking gene from all the three bioinformatics approaches was *KIF1A*. Subsequent Sanger sequencing confirmed that *KIF1A* indeed was the causative variant for HSP. The same example will be used to demonstrate the use of computational approaches to rank disease candidate genes.

In a second study, Benitez et al. [33] used Endeavour [7] and ToppGene [13] to rank the neuronal ceroid lipofuscinosis (NCL) candidate variant genes identified by exome sequencing. Known causal genes for other NCLs along with genes that are associated with phenotypically similar disorders were used as a training set. Interestingly, the three variants identified by exome sequencing

(*PDCD6IP*, *DNAJC5*, and *LIPJ*) were in the top five genes in the combined analysis using ToppGene and Endeavour suggesting that they may be functionally or structurally related with NCLs encoded genes and constituting true causative variants for adult NCL.

2 Materials

Since the application presented here is Web-based, a computer with Internet connection and a compatible Web browser is needed. The system described here (ToppGene: <http://toppgene.cchmc.org>) is tested regularly on a number of browsers and operating systems and should not have compatibility issues. For any *guilt by association*-based disease gene prioritization approach, a training set representing known knowledge in the form of genes associated/related to disease of interest is critical.

3 Methods

The methods described here, and the screenshots used to illustrate them, are correct for the servers/databases at the time of writing (September 2013). From time to time, interfaces and query/search options may be redeveloped in response to users' feedback and details may change.

The ToppGene application from the ToppGene Suite [13] will be used for ranking candidate genes in the orphan disease HSP.

3.1 ToppGene: Functional Annotations-Based Candidate Gene Ranking

ToppGene takes into account multiple layers of data to generate a signature from a list of user-specified training genes representing contemporary knowledge, e.g., known disease associated/causative genes. Based on this signature, the program then ranks user-uploaded list of new genes (*test* set). The backend knowledgebase of ToppGene consists of 17 gene feature types compiled from different publicly available resources. These include disease-dependent and disease-independent information such as known disease-genes, previous linkage regions, association studies, human and mouse phenotypes, known drug-targets, and microarray expression results, gene regulatory regions (transcription factor target genes and microRNA targets), protein domains, protein interactions, pathways, biological processes, literature co-citations, and so on. Each of these sources is used in an integrated manner to prioritize disease candidate genes.

As part of the workflow, a representative profile of the training genes (functional enrichment) using 17 different features (as listed above) is generated first. From the functional enrichment profile of the training genes, over-representative terms are identified. The test set genes are then compared to these overrepresented terms for all

categorical annotations and the average vector for the expression values. For each of the test set genes, a similarity score to the training profile for each of the 17 features is then derived and summarized (17 similarity scores). For computing similarity measures of categorical annotations, e.g., GO annotations, ToppGene uses a fuzzy-based similarity measure (*see* Popescu et al. [35] for additional details). In case of numeric annotations, e.g., microarray expression values, the similarity score is calculated as the Pearson correlation of the two expression vectors of the two genes. The 17 similarity scores are combined into an overall score using statistical meta-analysis and a *p-value* of each annotation of a test gene G is derived by random sampling of the whole genome. For additional algorithmic details, validation, and comparison with other related applications, readers are referred to previously published studies [12, 13].

3.2 Identifying and Ranking Disease Genes for Hereditary Spastic Paraparesis (HSP)

The disease HSP and a recently reported pathogenic mutation in *KIF1A* provide a model to illustrate the use of ToppGene [34]. The goal is to demonstrate the effectiveness of integrated functional annotation-based approaches in ranking novel disease candidate genes. The following sections describe the workflow and the results.

3.2.1 Compiling Training Set and Test Sets Genes for HSP

The training and test gene sets for HSP were derived from [34] (Tables 2 and 3). The training set was expanded further using additional HSP-associated genes reported in NCBI's Clinical Variations database. Some of the resources which are commonly used to compile known disease-associated genes are OMIM [36], the Genetic Association Database [37], GWAS [38], and the Comparative Toxicogenomics Database [39]. The latter database also integrates diseases biomarkers derived from literature and specialized database mining. The test set genes can come from a variety of approaches including sequencing, neighboring genes on the chromosome, protein interactome, or even the entire genome. Where no sequencing data are available, computational approaches can be used to compile test sets or, in some cases, the entire set of coding genes can be used for finding the most likely candidate genes for a disease. The test set or candidate genes in such cases can be compiled from mining protein interactomes and/or functional linkage networks. Briefly, for each of the training set genes (known disease causal gene), their interacting partners (both from the protein interactome and functional networks) can be extracted to generate a test set. The protein interactome data can be downloaded from the NCBI (<ftp://ftp.ncbi.nih.gov/gene/GeneRIF/interactions.gz>) while for functional networks, users can use either (1) Functional Linkage Network (FLN) [40] or (2) STRING [41]. The ToppGene Suite has an application which allows mining and ranking the protein interactome for novel candidate genes (*see* Subheading 3.2.3).

Table 2**A list of HSP-associated genes used as the training set for ranking HSP-candidate genes**

Entrez Gene ID	Gene symbol	Gene name
51062	<i>ATL1</i>	Atlastin GTPase 1
26580	<i>BSCCL2</i>	Berardinelli–Seip congenital lipodystrophy 2 (seipin)
22948	<i>CCT5</i>	Chaperonin containing TCP1, subunit 5 (epsilon)
9420	<i>CYP7B1</i>	Cytochrome P450, family 7, subfamily B, polypeptide 1
57165	<i>GJC2</i>	Gap junction protein, gamma 2, 47 kDa
3329	<i>HSPD1</i>	Heat shock 60 kDa protein 1 (chaperonin)
9897	<i>KIAA0196</i>	KIAA0196
3798	<i>KIF5A</i>	Kinesin family member 5A
3897	<i>LICAM</i>	L1 cell adhesion molecule
123606	<i>NIPA1</i>	Non imprinted in Prader–Willi/Angelman syndrome 1
5354	<i>PLP1</i>	Proteolipid protein 1
10908	<i>PNPLA6</i>	Patatin-like phospholipase domain containing 6
65055	<i>REEP1</i>	Receptor accessory protein 1
26278	<i>SACS</i>	Spastic ataxia of Charlevoix-Saguenay (sacsin)
6683	<i>SPAST</i>	Spastin
80208	<i>SPG11</i>	Spastic paraplegia 11 (autosomal recessive)
23111	<i>SPG20</i>	Spastic paraplegia 20 (Troyer syndrome)
51324	<i>SPG21</i>	Spastic paraplegia 21 (autosomal recessive, Mast syndrome)
6687	<i>SPG7</i>	Spastic paraplegia 7 (pure and complicated autosomal recessive)
23503	<i>ZFYVE26</i>	Zinc finger, FYVE domain containing 26
118813	<i>ZFYVE27</i>	Zinc finger, FYVE domain containing 27

3.2.2 Prioritization of HSP Candidate Genes

Now that you have the training and test set genes for HSP ready (Tables 2 and 3), proceed with the ranking of 14 HSP test set genes using ToppGene [13].

1. From the ToppGene Suite homepage (<http://toppgene.cchmc.org>) click on the second link (“ToppGene: Candidate gene prioritization”) (Fig. 1).
2. On the following page (“ToppGene: Candidate gene prioritization”), enter either gene symbols or Entrez gene IDs from the training and test set genes (21 and 14 genes respectively in this case; *see* Tables 2 and 3), and click “Submit query” (Figs. 2 and 3).

Table 3
List of ToppGene ranked HSP candidate genes

ToppGene rank	Entrez gene ID	Gene symbol	Gene name
1	547	<i>KIF1A</i>	Kinesin family member 1A
2	728294	<i>D2HGDH</i>	D-2-hydroxyglutarate dehydrogenase
3	23192	<i>ATG4B</i>	Autophagy related 4B, cysteine peptidase
4	3069	<i>HDLBP</i>	High density lipoprotein binding protein
5	23178	<i>PASK</i>	PAS domain containing serine/threonine kinase
6	84289	<i>ING5</i>	Inhibitor of growth family, member 5
7	25992	<i>SNED1</i>	Sushi, nidogen and EGF-like domains 1
8	1841	<i>DTYMK</i>	Deoxythymidylate kinase (thymidylate kinase)
9	389090	<i>OR6B2</i>	Olfactory receptor, family 6, subfamily B, member 2
10	150681	<i>OR6B3</i>	Olfactory receptor, family 6, subfamily B, member 3
11	653437	<i>AQP12B</i>	Aquaporin 12B
12	375318	<i>AQP12A</i>	Aquaporin 12A
13	51078	<i>THAP4</i>	THAP domain containing 4
14	643905	<i>PRR21</i>	Proline rich 21

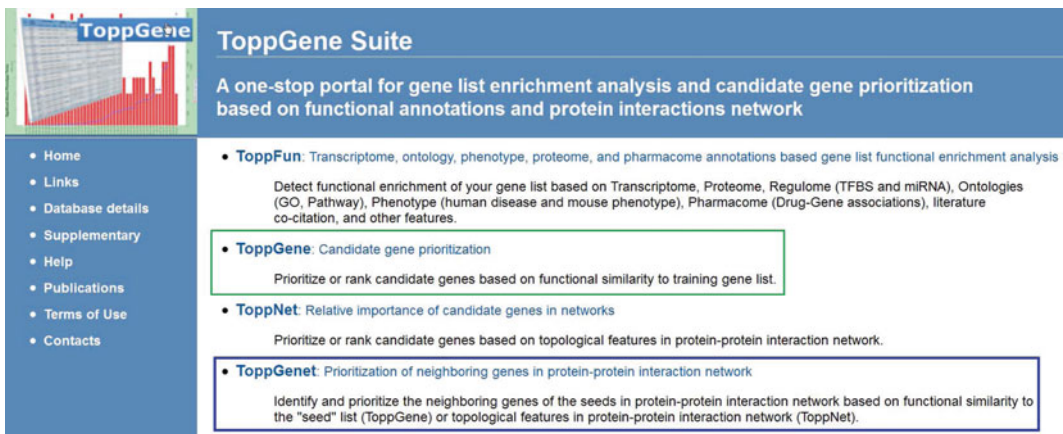


Fig. 1 The ToppGene Suite home page

3. Select the appropriate statistical parameters. For the “Training Parameters” and “Test Parameter” you can use the default parameters, i.e., Bonferroni correction and *p*-value cutoff set to 0.05; see “Help” section for more details. Under the “Test parameter”, the “random sampling size” option is for select-

Select your gene identifier type, paste your training and test gene sets below or select example sets, then submit.

Example gene sets: [HGNC Symbol](#) [Entrez ID](#)
(click on "HGNC Symbol" or "Entrez ID" to use the example training and test set of genes)

Symbol Types: HGNC Symbol Entrez ID

Training Gene Set:

ATL1
BSCL2
CCT5
CYP7B1
GJC2
HSPD1
KIAA0196
KIF5A
L1CAM
NIPA1
PLP1
PNPLA6
REEP1
SACS
SPAST
SPG11
SPG20
SPG21
SPG7
ZFYVE26
ZFYVE27

Test gene set:

547
728294
23192
3069
23178
84289
25992
1841
389090
150681
653437
375318
51078
643905

Clear Submit Query

Fig. 2 The ToppGene entry page for launching gene prioritization showing the lists of training and test set genes for HSP

Training set (21 / 21)			Test set (14 / 14)		
Entered	Human Symbol	Gene ID	Entered	Human Symbol	Gene ID
ATL1	ATL1	51062	547	KIF1A	547
BSCL2	BSCL2	26580	728294	D2HGDH	728294
CCT5	CCT5	22948	23192	ATG4B	23192
CYP7B1	CYP7B1	9420	3069	HDLBP	3069
GJC2	GJC2	57165	23178	PASK	23178
HSPD1	HSPD1	3329	84289	ING5	84289
KIAA0196	KIAA0196	9897	25992	SNED1	25992
KIF5A	KIF5A	3798	1841	DTYMK	1841
L1CAM	L1CAM	3897	389090	OR6B2	389090
NIPA1	NIPA1	123606	150681	OR6B3	150681
PLP1	PLP1	5354	653437	AQP12B	653437
PNPLA6	PNPLA6	10908	375318	AQP12A	375318
REEP1	REEP1	65055	51078	THAP4	51078
SACS	SACS	26278	643905	PRR21	643905
SPAST	SPAST	6683			
SPG11	SPG11	80208			
SPG20	SPG20	23111			
SPG21	SPG21	51324			
SPG7	SPG7	6687			
ZFYVE26	ZFYVE26	23503			
ZFYVE27	ZFYVE27	118813			

Fig. 3 Training and test set genes for HSP

Training parameters

Feature	Correction	p-Value cutoff	Gene Limits	
<input checked="" type="checkbox"/> All	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> GO: Molecular Function	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> GO: Biological Process	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> GO: Cellular Component	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Human Phenotype	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Mouse Phenotype	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Domain	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Pathway	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Pubmed	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Interaction	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Cytoband	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Transcription Factor Binding Site	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Gene Family	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Coexpression	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Coexpression Atlas	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Computational	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> MicroRNA	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Drug	Bonferroni	0.05	1	$\leq n \leq 1500$
<input checked="" type="checkbox"/> Disease	Bonferroni	0.05	1	$\leq n \leq 1500$

Test parameter

Random sampling size: 1500 (6% of genome)
 Min. feature count: 2

Home Modify Query Start prioritization

Fig. 4 The TopGene entry page for selecting prioritization parameters

ing the background gene set from the genome for computing the *p*-value while the “Min. feature count” represents the number of features that need to be considered for prioritization. The default options are 6 % of the genome (or 1,500 genes from a total of 25,000 genes) for random sampling size, and feature count is 2 (Fig. 4).

4. If your gene lists contain alternate symbols or duplicates or obsolete symbols, they are ignored or presented with the option to resolve them and add them back to your input list. Additionally, if there are common genes between training and test sets, i.e., test set genes which are also found in training set, these will be removed from the test set and no ranks will be assigned to them. After selecting the appropriate statistical parameters (training and test) click on the “Start prioritization” button.
5. Once the analysis is complete, the first half of the results page shows the enrichment results for the training set (Fig. 5).
6. The prioritized list of test set genes sorted according to their ranks based on the *p*-values are displayed in the lower half (Fig. 6). Each column indicates the features used to compute similarity between training and test sets.

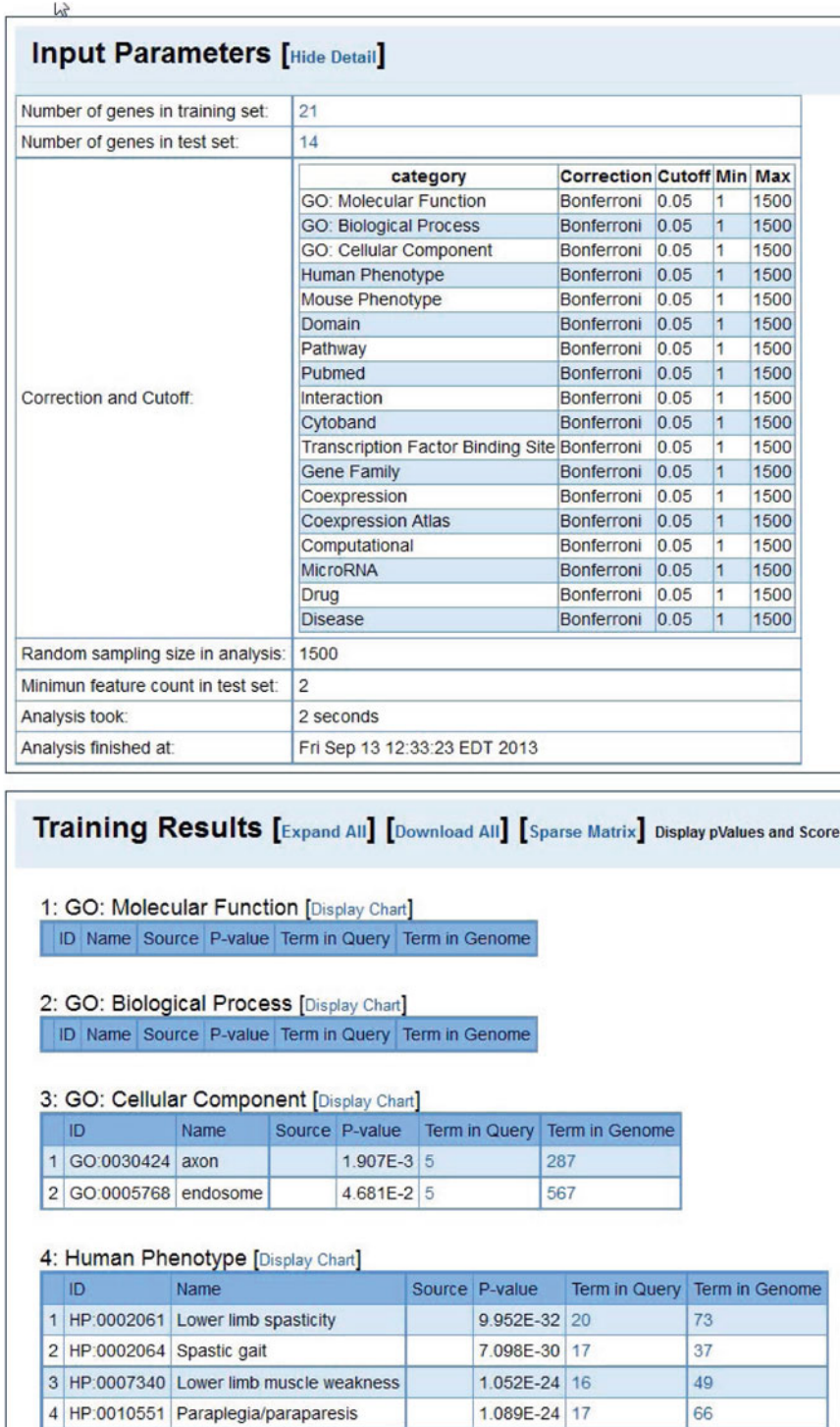


Fig. 5 Result of prioritization showing the input parameters and a partial view of the training set enrichment results

Test Results [Hide Detail] [Download] [Show Network]													
Rank (net)	Gene Symbol	Gene ID	GO: Molecular Function		GO: Biological Process		GO: Cellular Component		Human Phenotype		Mouse Phenotype		
			Score	pValue	Score	pValue	Score	pValue	Score	pValue	Score	pValue	
1 <input type="checkbox"/>	KIF1A	547	0.000E0	4.997E-1	0.000E0	4.997E-1	5.783E-2	1.524E-1	1.000E0	6.540E-4	7.961E-1	3.270E-3	
2 <input type="checkbox"/>	PRR21	643905							1.000E0	6.540E-4			
3 <input type="checkbox"/>	ATG4B	23192	0.000E0	4.997E-1	0.000E0	4.997E-1	0.000E0	6.272E-1	1.000E0	6.540E-4	6.160E-1	1.243E-2	
4 <input type="checkbox"/>	ING5	84289	0.000E0	4.997E-1	0.000E0	4.997E-1	5.783E-2	1.524E-1	1.000E0	6.540E-4			
5 <input type="checkbox"/>	PASK	23178	0.000E0	4.997E-1	0.000E0	4.997E-1	1.551E-1	1.962E-2	1.000E0	6.540E-4	0.000E0	5.311E-1	
6 <input type="checkbox"/>	DTYMK	1841	0.000E0	4.997E-1	0.000E0	4.997E-1	9.730E-2	6.736E-2	1.000E0	6.540E-4			
7 <input type="checkbox"/>	OR6B3	150681	0.000E0	4.997E-1			0.000E0	6.272E-1	1.000E0	6.540E-4			
8 <input type="checkbox"/>	THAP4	51078	0.000E0	4.997E-1					1.000E0	6.540E-4			
9 <input type="checkbox"/>	SNED1	25992	0.000E0	4.997E-1	0.000E0	4.997E-1	0.000E0	6.272E-1	1.000E0	6.540E-4			
10 <input type="checkbox"/>	OR6B2	389090	0.000E0	4.997E-1			0.000E0	6.272E-1	1.000E0	6.540E-4			
11 <input type="checkbox"/>	HDLBP	3069	0.000E0	4.997E-1	0.000E0	4.997E-1	5.783E-2	1.524E-1	1.000E0	6.540E-4			
12 <input type="checkbox"/>	AQP12A	375318	0.000E0	4.997E-1			0.000E0	6.272E-1	1.000E0	6.540E-4	0.000E0	5.311E-1	
13 <input type="checkbox"/>	AQP12B	653437	0.000E0	4.997E-1			0.000E0	6.272E-1	1.000E0	6.540E-4	0.000E0	5.311E-1	
14 <input type="checkbox"/>	D2HGDH	728294	0.000E0	4.997E-1	0.000E0	4.997E-1	9.730E-2	6.736E-2	7.170E-1	3.466E-2			

Fig. 6 Result of prioritization showing the ranked list of test set genes

- Additional details about the ranked genes can be obtained in both graphical and tabular format. For this, select the gene(s) you are interested in from the “Rank” column (Fig. 7), and click on the “Show Network” link. This will lead to the “Common Terms for selected genes and Training Set” page where you can see the details on how a test set gene is connected to the training set (Fig. 8). The network view can be downloaded as an XGMML file and can be imported into Cytoscape [42] for visualization and further analysis.
- The prioritized list can be downloaded as a table. Of the 14 HSP candidate genes, *KIF1A* is ranked at the top (Table 3). Occasionally the ranks may vary a little because for every run, a different set of random genes are selected for computing the statistical significance.

3.2.3 Mining HSP Interactome for Novel Candidate Genes

The ToppGeNet application from the ToppGene Suite will be used for finding and ranking novel candidate genes for HSP. ToppGeNet allows the user to rank the interacting partners (direct or indirect) of known disease genes for their likelihood of causing a disease. Here, given a training set of known disease genes, the test set is

Rank (net)	Gene Symbol	Gene ID
1 <input checked="" type="checkbox"/>	KIF1A	547
2 <input type="checkbox"/>	PRR21	643905
3 <input type="checkbox"/>	ATG4B	23192
4 <input type="checkbox"/>	ING5	84289
5 <input type="checkbox"/>	PASK	23178
6 <input type="checkbox"/>	DTYMK	1841
7 <input type="checkbox"/>	OR6B3	150681
8 <input type="checkbox"/>	THAP4	51078
9 <input type="checkbox"/>	SNED1	25992
10 <input type="checkbox"/>	OR6B2	389090
11 <input type="checkbox"/>	HDLBP	3069
12 <input type="checkbox"/>	AQP12A	375318
13 <input type="checkbox"/>	AQP12B	653437
14 <input type="checkbox"/>	D2HGDH	728294

Fig. 7 List of top ranked genes for HSP showing *KIF1A* ranked at the top

Feature	ID	Name	Gene
Human Phenotype	HP-0002530	Urinary bladder sphincter dysfunction	ATL1 HSPD1 KAA0196 KIF1A KIF5A NRP1 REEP1 SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0002169	Clonus	CCT5 KIF1A KIF5A NRP1 REEP1 SPG11 SPQ20 ZFYVE26 ZFYVE27
Human Phenotype	HP-0002064	Spastic gait	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 REEP1 SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0003026	Abnormality of the ankles	KIF1A KIF5A REEP1 SPG11 SPQ20 ZFYVE27
Human Phenotype	HP-0007340	Lower limb muscle weakness	ATL1 BSCL2 CYP7B1 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0002127	Upper motor neuron abnormality	CCT5 KIF1A KIF5A NRP1 REEP1 SPG11 SPQ20 ZFYVE26 ZFYVE27
Human Phenotype	HP-0002091	Lower limb spasticity	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0000690	Limb muscle weakness	ATL1 BSCL2 CYP7B1 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0003487	Babinski sign	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0002460	Distal muscle weakness	ATL1 BSCL2 CYP7B1 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0002492	Abnormality of the corticospinal tract	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0003676	Progressive disorder	ATL1 CYP7B1 HSPD1 KAA0196 KIF1A KIF5A NRP1 SPAST SPG11 ZFYVE26
Human Phenotype	HP-0003450	Abnormality of the motor neurons	BSCL2 CCT5 GJC2 KIF1A KIF5A NRP1 PNPLA6 REEP1 SACS SPG11 SPQ20 ZFYVE26 ZFYVE27
Human Phenotype	HP-0000009	Functional abnormality of the bladder	ATL1 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPAST SPG11 SPQ7 ZFYVE26
Human Phenotype	HP-0000014	Abnormality of the bladder	ATL1 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPAST SPG11 SPQ7 ZFYVE26
Human Phenotype	HP-0000630	Peripheral neuropathy	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 REEP1 SACS SPAST SPG11 SPQ7 ZFYVE26
Human Phenotype	HP-0003302	Amyotrophy	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPG11 SPQ20 ZFYVE26
Human Phenotype	HP-0001347	Hyperreflexia	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0007367	Atrophy/Degeneration affecting the central nervous system	ATL1 CCT5 KAA0196 KIF1A L1CAM NRP1 PLP1 PNPLA6 SPAST SPG11 SPQ20 SPQ7
Human Phenotype	HP-0001256	Abnormality of pyramidal motor function	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0000759	Abnormality of the peripheral nervous system	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 REEP1 SACS SPAST SPG11 SPQ7 ZFYVE26
Human Phenotype	HP-0002062	Abnormality of the pyramidal tracts	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0001257	Spasticity	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0003579	Pace of progression	ATL1 BSCL2 CCT5 CYP7B1 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0001324	Muscle weakness	ATL1 BSCL2 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0002014	Abnormality of the lower limb	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0000079	Abnormality of the urinary system	ATL1 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPAST SPG11 SPQ7 ZFYVE26
Human Phenotype	HP-0002813	Abnormality of limb bone morphology	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0003011	Abnormality of the musculature	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 PNPLA6 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Human Phenotype	HP-0000004	Onset and clinical course	ATL1 BSCL2 CCT5 CYP7B1 GJC2 HSPD1 KAA0196 KIF1A KIF5A NRP1 PLP1 REEP1 SACS SPAST SPG11 SPQ20 SPQ7 ZFYVE26 ZFYVE27
Mouse Phenotype	MP-0002229	neurodegeneration	GJC2 KIF1A KIF5A PLP1 PNPLA6 SPQ7
Mouse Phenotype	MP-0002882	abnormal neuron morphology	GJC2 KIF1A KIF5A L1CAM PLP1 PNPLA6 SPAST SPQ20 SPQ7
Mouse Phenotype	MP-0002966	abnormal motor coordination/movement	ATL1 GJC2 KIF1A KIF5A L1CAM PLP1 PNPLA6 REEP1 SACS SPQ20 SPQ7
Coexpression	dev_lower_uro	dev lower_uro_neuro_e14.5_BladPervGanglion_Sor10_top-relative-expression-ranks_500	ATL1 BSCL2 KIF1A KIF5A L1CAM PLP1
Atlas	neur_e14.5_BladPervGanglion_Sor10_500		ATL1 BSCL2 KIF1A KIF5A L1CAM PLP1

Fig. 8 Browsing details of the top ranked gene. *Red boxed* gene (*KIF1A*) is the target gene and the remaining are the training set genes. The “Feature” and “Name” columns show the details of functional annotations shared between the ranked test set gene and the training set (Color figure online)

generated by mining the protein interactome and compiling the genes interacting either directly or indirectly (based on user input) with the training set genes. The test set genes can then be ranked using either ToppGene (functional annotation-based method) or ToppNet (PPIN-based method).

1. From the ToppGene Suite homepage (<http://toppgene.cchmc.org>) click on the fourth link (“ToppGenet: Prioritization of neighboring genes in protein-protein interaction network”) (Fig. 1).
2. On the following page, copy and paste the list of known HSP genes (from Table 2) in the box “Set of seeds”. Keep the “Distance to seeds” as 1 which means the immediate interacting partners of known HSP genes will be considered as the test set. Select “Functional annotation based” for the “Prioritization method” and click “Submit Query” (Fig. 9).
3. The next page you can see the list of 21 training set genes and 210 genes comprising the test set. The test set represents genes whose encoding proteins are direct interactants of training set genes encoded proteins. Of the 210 test set genes, 195 only are considered for ranking as the 15 training set overlapping genes

ToppGenet: Prioritization of neighboring genes in Protein-Protein Interaction Network

Select your gene identifier type, paste your sets below or select example set, then submit.

Entry Type:

Example gene sets:
(click on "HGNC Symbol" or "Entrez ID" to use the example "seeds")

Set of seeds:

ATL1
 BSCL2
 CCT5
 CYP7B1
 GJC2
 HSPD1
 KIAA0196
 KIP5A
 LICAM
 NIPAL
 PLP1
 PNPLA6
 REEP1
 SACS
 SPAST
 SPG11
 SPG20
 SPG21
 SPG7
 ZFYVE26
 ZFYVE27

Distance to seeds:

Prioritization method:

Fig. 9 The ToppGeNet entry page showing the list of HSP genes for interactome ranking (immediate interacting partners of HSP genes encoded proteins) using functional annotation based method (ToppGene)

are excluded from test set. The remaining steps are as described earlier (*see* **steps 3** through **8** under Subheading **3.2.2**).

4. Based on ToppGene ranking, the top five candidate genes for HSP are *KLC1*, *CANX*, *PPP2R2B*, *KIF5C*, and *MYC*.

3.3 Limitations

Candidate gene ranking approaches based on functional annotation similarity have the following limitations:

- Most of these approaches rank test set genes based on user-specified training set. Although this is an improvement over the “favorite” gene method, it should be noted that they are based on the assumption that novel disease genes yet to be discovered will be consistent with what is already known about a disease and/or its genetic basis which may not always be true. Nevertheless, having a “good” or representative training set is critical. The training set may not necessarily be always a set of known causal genes but can be an implicated pathway or biological process or even a list of symptoms (or phenotype). Prior knowledge can sometimes be also inferred from related or similar diseases. This similarity can be either similar manifestation or symptoms or similar molecular mechanisms of related or similar diseases.
- Second, selecting an appropriate approach is also important and frequently depends on the disease type and the molecular mechanism that causes it. For example, using protein–protein interaction data for identifying novel candidates may be useful when a disease is known to be caused by the disruption of a larger protein complex. On the other hand, using a protein interaction network may not be totally justified for a disease known to be caused by aberrant regulatory mechanisms. In such cases, either using gene regulatory networks and/or high-throughput gene expression data may be more appropriate [14].
- The accuracy of the prioritization is heavily dependent not only on the training set but also on the accuracy and completeness of the original literature-mining or database-derived annotations. In other words, functional annotation based candidate gene prioritization approaches tend to be biased towards well studied or better annotated genes. For instance, a novel, “true” disease gene can be missed if it lacks sufficient annotations.
- Although it has been speculated that complex traits result more often from noncoding regulatory variants than from coding sequence variants [43–45], current disease gene identification and prioritization approaches predominantly are gene-centric. However, interpreting the consequences of noncoding sequence variants is relatively complex because the relationships among promoter, intergenic, or noncoding sequence

variations, gene expression levels, and trait phenotypes are less well understood than the relationship between coding DNA sequence and protein function.

- Finally, different gene prioritization methods use different algorithms to integrate, mine, and rank, and objectively, there is no one methodology that is better than the others for all data inputs [46]. Several previous studies have shown that the computational approaches for disease gene ranking are largely complementary [10, 33, 34, 47]. To increase the robustness of prioritization analysis, we therefore recommend using at least two algorithmically different ranking approaches, e.g., functional annotation-based and network topology-based approaches.

References

1. Rados C (2003) Orphan products: hope for people with rare diseases. *FDA Consum* 37:10–15
2. Wastfelt M, Fadeel B, Henter JI (2006) A journey of hope: lessons learned from studies on rare diseases and orphan drugs. *J Intern Med* 260:1–10
3. Freudenberg J, Propping P (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics* 18 Suppl 2S: 110–115
4. Turner FS, Clutterbuck DR, Semple CA (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol* 4:R75
5. Tiffin N, Kelso JF, Powell AR et al (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res* 33:1544–1552
6. Adie EA, Adams RR, Evans KL et al (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics* 6:55
7. Aerts S, Lambrechts D, Maity S et al (2006) Gene prioritization through genomic data fusion. *Nat Biotechnol* 24:537–544
8. Thornblad TA, Elliott KS, Jowett J et al (2007) Prioritization of positional candidate genes using multiple web-based software tools. *Twin Res Hum Genet* 10:861–870
9. Zhu M, Zhao S (2007) Candidate gene identification approach: progress and challenges. *Int J Biol Sci* 3:420–427
10. Tiffin N, Adie E, Turner F et al (2006) Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res* 34:3067–3081
11. Adie EA, Adams RR, Evans KL et al (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics* 22:773–774
12. Chen J, Xu H, Aronow BJ et al (2007) Improved human disease candidate gene prioritization using mouse phenotype. *BMC Bioinformatics* 8:392
13. Chen J, Bardes EE, Aronow BJ et al (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res* 37(Web Server issue):W305–W311
14. Piro RM, Di Cunto F (2012) Computational approaches to disease-gene prediction: rationale, classification and successes. *FEBS J* 279:678–696
15. Goh KI, Cusick ME, Valle D et al (2007) The human disease network. *Proc Natl Acad Sci U S A* 104:8685–8690
16. Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. *Nature* 409:853–855
17. Smith NG, Eyre-Walker A (2003) Human disease genes: patterns and predictions. *Gene* 318:169–175
18. Tranchevent LC, Barriot R, Yu S (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Res* 36(Web Server issue):W377–W384
19. Junker BH, Koschutski D, Schreiber F (2006) Exploration of biological network centralities with CentiBiN. *BMC Bioinformatics* 7:219
20. George RA, Liu JY, Feng LL et al (2006) Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res* 34:e130
21. Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* 8:333–346

22. Kohler S, Bauer S, Horn D et al (2008) Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet* 82:949–958
23. Wu X, Jiang R, Zhang MQ et al (2008) Network-based global inference of human disease genes. *Mol Syst Biol* 4:189
24. Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein-protein interaction network. *Bioinformatics* 22:2800–2805
25. Chen JY, Shen C, Sivachenko AY (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pac Symp Biocomput* 367–378
26. Ortutay C, Vihinen M (2009) Identification of candidate disease genes by integrating Gene Ontologies and protein-interaction networks: case study of primary immunodeficiencies. *Nucleic Acids Res* 37:622–628
27. Chen J, Aronow BJ, Jegga AG (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics* 10:73
28. Gilissen C, Hoischen A, Brunner HG et al (2012) Disease gene identification strategies for exome sequencing. *Eur J Hum Genet* 20:490–497
29. Zhang M, Zhu C, Jacomy A et al (2011) The orphan disease networks. *Am J Hum Genet* 88:755–766
30. 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
31. Bainbridge MN, Wiszniewski W, Murdock DR et al (2011) Whole-genome sequencing for optimized patient management. *Sci Transl Med* 3:87re3
32. Kingsmore SF, Saunders CJ (2011) Deep sequencing of patient genomes for disease diagnosis: when will it become routine? *Sci Transl Med* 3:87ps23
33. Benitez BA, Alvarado D, Cai Y et al (2011) Exome-sequencing confirms DNAJC5 mutations as cause of adult neuronal ceroid-lipofuscinosis. *PLoS One* 6:e26741
34. Erlich Y, Edvardson S, Hodges E et al (2011) Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res* 21:658–664
35. Popescu M, Keller JM, Mitchell JA (2006) Fuzzy measures on the Gene Ontology for gene product similarity. *IEEE/ACM Trans Comput Biol Bioinform* 3:263–274
36. Hamosh A, Scott A, Amberger J et al (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* 33:D514–D517
37. Becker KG, Barnes KC, Bright TJ et al (2004) The genetic association database. *Nat Genet* 36:431–432
38. Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* 106:9362–9367
39. Davis AP, Murphy CG, Saraceni-Richards CA et al (2009) Comparative Toxicogenomics Database: a knowledgebase and discovery tool for chemical-gene-disease networks. *Nucleic Acids Res* 37(Database issue):D786–D792
40. Linghu B, Snitkin ES, Hu Z et al (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol* 10:R91
41. Szklarczyk D, Franceschini A, Kuhn M et al (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* 39:D561–D568
42. Shannon P, Markiel A, Ozier O et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–2504
43. King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. *Science* 188:107–116
44. Korstanje R, Paigen B (2002) From QTL to gene: the harvest begins. *Nat Genet* 31:235–236
45. Mackay TF (2001) Quantitative trait loci in *Drosophila*. *Nat Rev Genet* 2:11–20
46. Bromberg Y (2013) Chapter 15: disease gene prioritization. *PLoS Comput Biol* 9:e1002902
47. Navlakha S, Kingsford C (2010) The power of protein interaction networks for associating genes with diseases. *Bioinformatics* 26:1057–1063
48. Perez-Iratxeta C, Bork P, Andrade MA (2002) Association of genes to genetically inherited diseases using data mining. *Nat Genet* 31:316–319
49. Perez-Iratxeta C, Wjst M, Bork P et al (2005) G2D: a tool for mining genes associated with disease. *BMC Genet* 6:45
50. Hristovski D, Peterlin B, Mitchell JA et al (2005) Using literature-based discovery to identify disease candidate genes. *Int J Med Inform* 74:289–298

51. van Driel MA, Cuelenaere K, Kemmeren PP et al (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur J Hum Genet* 11:57–63
52. van Driel MA, Cuelenaere K, Kemmeren PP et al (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res* 33(Web Server issue):W758–W761
53. Masseroli M, Galati O, Pinciroli F (2005) GFINDER: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res* 33(Web Server issue):W717–W723
54. Masseroli M, Martucci D, Pinciroli F (2004) GFINDER: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res* 32(Web Server issue):W293–W300
55. Baker EJ, Jay JJ, Bubier JA et al (2012) GeneWeaver: a web-based system for integrative functional genomics. *Nucleic Acids Res* 40(Database issue):D1067–D1076
56. Chen YA, Tripathi LP, Mizuguchi K (2011) TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. *PLoS One* 6:e17844
57. Liekens AM, De Knijf J, Daelemans W et al (2011) BioGraph: unsupervised biomedical knowledge discovery via automated hypothesis generation. *Genome Biol* 12:R57
58. Makita Y, Kobayashi N, Yoshida Y et al (2013) PosMed: ranking genes and bioresources based on Semantic Web Association Study. *Nucleic Acids Res* 41(Web Server issue):W109–W114

Computer-Aided Drug Designing

Mohini Gore and Neetin S. Desai

Abstract

Computer-aided drug designing has emerged as a cost-effective and rapid tool for the discovery of newer therapeutic agents. Several algorithms have been developed to analyze protein structure and function, to identify interacting ligands, active site residues, and to study protein–ligand interactions, which can eventually lead to the identification of new drugs. In silico drug designing involves identification of the target protein which is responsible for the development of the disease under study. The three-dimensional structure of the protein can be predicted using homology modeling, while molecular docking is applied to study the interaction of a drug molecule with the protein. The best orientation of the ligand-protein docked structure which has overall minimum energy needs to be obtained. In silico methods can be used to identify potential drugs for various diseases. Thus, computer-aided drug designing has become an indispensable and integral part of the drug discovery process.

Key words Computer-aided drug designing, Docking, Homology modeling, Virtual screening

1 Introduction

The discovery and development of new drug is a long and complicated process. It is estimated that a typical drug discovery process initiating from lead identification to clinical trials can take up to 14 years with a cost of 800 million US dollars [1]. The traditional approach involves random screening of chemical compounds obtained from nature or synthesized in the laboratories involving long design cycle and higher cost. Computer-aided drug design involves structure-based drug design using in silico approaches which has made the drug discovery process cost-effective and much faster [2]. In silico drug design plays a significant role in all stages of drug development. By selecting only a potent lead molecule it may avert the late stage clinical failures, thus reducing the cost by a significant amount [3].

Proteins are organic molecules, which are an essential part of metabolic reactions of a cell. The majority of the cell's chemical reactions and structural components entail proteins crucial for

appropriate cell functions. Therefore, if their function is impaired, the consequences can lead to a number of diseases [4]. Studying protein structure, function, and interactions within and between cells is vital for drug discovery as any impairment in this function can result in malfunction and disease. Therefore, categorizing protein as a drug target may result in developing better lead molecule for drug development. A drug target is a molecule where the drug interacts to bring about the desired change. The goal of drug designing is to develop a drug that is highly specific to a particular target and affects it in a desired way, so as to interfere with the disease process. Computer-based simulation is being used extensively to model drug–target interactions to guide drug discovery.

In simpler terms, drug discovery involves search for lead molecules which could alter a diseased pathway. As a part of the discovery process, one or more molecular as well as cellular processes that occur in the affected cells of a diseased tissue or organ need to be altered. Computer-aided drug designing is being used extensively to establish potential drugs for the treatment and containment of various diseases. *In silico* drug designing for a particular disease involves identification of the protein molecule causing the disease in question. The three-dimensional structure of a protein molecule can be obtained by homology modeling using its amino acid sequence. Furthermore it has been observed that during evolution, the protein structure is more stable than the sequence, so that similar sequences have identical structures [5]. Moreover, using the protein sequence alignment and template structure, the structure of a protein molecule can be predicted. The next step involves identification of the drug molecule which can interact with the protein molecule. The interaction of the structures of protein molecule and drug molecule can be studied using docking methods [6]. The finest orientation of the ligand which forms a complex with overall minimum energy needs to be determined.

2 Materials

2.1 Identification of Target Protein

Proteins are responsible for almost all essential functions of a cell including metabolism, regulation, and development. Impairment in the structure or function of a protein can result in disease. Computer-aided drug designing can be used for drug discovery by identifying drug molecules which can interact with the protein which is implicated in causing the disease. Hence, first the causative protein molecule for the disease under study needs to be identified. If the three-dimensional structure of the protein is not known, it can be predicted by using homology modeling. The sequence of the protein under study can be obtained by performing a search on NCBI's Entrez database. Normally, the protein sequences are used in FASTA format.

2.1.1 Sequence Analysis

In computer-aided drug designing, one can use the genetic (DNA or RNA) sequence or the protein's amino acid sequence of several organisms or species. With this information one can determine the evolutionary relationships of a species by finding identical sequences in biological databases. There are two types of sequence analysis methods to determine similarity among biological sequences [7].

- *Pairwise sequence alignment.* This predicts regions of similarity that may indicate functional, structural as well as evolutionary relationships between two protein or nucleic acid sequences.
- *Multiple sequence alignment.* Often applied to align three or more biological sequences of similar length, homology and the evolutionary relationships between the sequences studied from its output.

2.2 Homology Modeling

In silico methods can be used to predict a protein structure from its amino acid sequence. Homology modeling involves prediction of the three-dimensional structure of a protein from the known structures of one or more related proteins, which are used as templates. Several automated programs are available for homology or comparative modeling of protein three-dimensional structures. The models obtained can be subjected to structural validation by assessment of the Ramachandran plot.

2.2.1 Identification of Templates

The structures of the templates to be used for homology modeling have to be searched in protein structure databases such as the Protein Data Bank (PDB). The target sequence (query) is used for searching proteins with sequence similarity. There are a number of options available on the World Wide Web (WWW) to perform the sequence similarity study. A few examples of software programs are given in Table 1.

2.2.2 Selection of the Templates

One or more of the template identification methods can be employed to obtain a list of potential templates. It is necessary to select the appropriate templates for protein modeling so that there is a higher overall sequence similarity between the target and template sequences, i.e., higher percentage of identical residues, lower number and shorter gap lengths in the alignment.

2.2.3 Three-Dimensional Structure Modeling

Once the templates have been selected, comparative modeling is done to build a three-dimensional protein model. Several software programs and servers are available for protein modeling. Some examples are mentioned in Table 1.

2.2.4 Model Validation

Different software or servers used for comparative modeling will provide a number of models. These models need to be evaluated and the best one identified. The stereochemical quality of the protein structure, like bond length, the phi/psi angles, etc., has to be

Table 1
Software and servers for homology modeling and docking

Name	URL ^a	References
<i>Homology modeling</i>		
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	[8]
HHpred	http://toolkit.tuebingen.mpg.de/hhpred	[9]
ModBase	http://modbase.compbio.ucsf.edu/	[10]
GeneSilico Metaserver	https://genesilico.pl	[11]
Predict protein	http://www.predictprotein.org/	[12]
UCLA-DOE	http://fold.doe-mbi.ucla.edu/	[13]
COMA	http://www.ibt.lt/en/coma.html	[14]
COMPASS	http://prodata.swmed.edu/compass	[15]
PMP	http://www.proteinmodelportal.org/	[16]
MODELLER	http://salilab.org/modeller/	[17–20]
I-TASSER	http://zhanglab.ccmb.med.umich.edu	[21–23]
Swiss model	http://swissmodel.expasy.org/	[24–26]
Phyre	http://www.sbg.bio.ic.ac.uk/phyre2/html	[27]
<i>Docking</i>		
Autodock	http://autodock.scripps.edu/	[28]
DOCK	http://dock.compbio.ucsf.edu/	[29]
FlexX	http://www.biosolveit.de/flexx/	[30]
GOLD	http://www.ccdc.cam.ac.uk/gold/	[31]
ICM	http://www.molsoft.com/docking.html	[32]
SwissDock	http://swissdock.vital-it.ch/	[33]
Hex	http://hexserver.loria.fr	[34]

^aAccessed 24 May 2013

assessed. One way of doing this is by generating Ramachandran plots [35]. This method evaluates the correctness of structural coordinates based on standard deviations in phi and psi angle pairs for residues in a protein. Some tools available for generating Ramachandran plots are RAMPAGE server [36], Molprobability [37, 38], PROCHECK [39, 40], STAN server [41], ERRAT [42], Verify 3D [43], etc.

2.2.5 Chemical Structure of the Drug Molecule

Small molecule databases embody a major resource for the study of biochemical interactions. A variety of repositories of biological molecules and their physicochemical properties are available. These include databases of known chemical compounds, drugs, carbohydrates, enzymes, reactants, natural products, and natural-product-derived compounds. PubChem (<http://pubchem.ncbi.nlm.nih.gov/>) provides information on the biological activities of more than 40 million small molecules and 19 million unique structures [1]. The chemical structures of the drug molecules to be used for docking studies can be drawn using software like ACD/Chemsketch, MarvinSketch, ChemWriter, ChemPen, etc. If another protein molecule is to be used for docking studies, its

structure can be predicted by homology modeling. Virtual screening for suitable ligands has emerged as a rapid and inexpensive method for the identification of lead compounds. Virtual screening enables systematic evaluation of large chemical libraries in a very short span of time to identify potential lead compounds. The ligand should have the maximum interaction with the target protein, i.e., it should have the lowest value of free energy of interaction. Interactions can be steric, hydrophobic, electrostatic, or hydrogen bonding (H-bonding).

Ligand–protein target interactions occur at the atomic level. For a detailed understanding of how and why ligands bind to receptors, one should consider the biochemical and biophysical properties of both the drug compound as well as the target at an atomic scale. For instance the Swiss-PDB tool can determine important physicochemical parameters, such as hydrophobicity and polarity that are the key factors in drug-protein binding [7].

2.3 Docking

Docking is a process, in which two molecules fit or dock together in a three-dimensional space. Docking explores ways in which a target protein molecule and another molecule, such as a drug fit together. Several docking software programs are available namely, interactive protein docking and molecular superposition programs. A few examples are given in Table 1. Docking programs such as Autodock, GOLD, DOCK, ICM, etc. are based on the principle of achieving an optimized conformation of both protein and ligand, where the energy of the overall system is minimized (*see Note 1*).

3 Methods

3.1 Identification of the Target Protein

Selection and validation of the target protein is a crucial step for drug designing. The repositories of protein structures and sequences are rapidly growing and serve as a useful source for the identification of potential protein targets for drug designing. For drug discovery, the function of a protein should be known. Protein function annotation is available in the protein databases. Once the causative protein for the disease under study has been identified, its sequence can be retrieved using NCBI's Entrez database for identification of its structure using homology modeling.

3.2 Homology Modeling

3.2.1 Identification of Templates

The first step in homology modeling involves searching for template structures using the target protein sequence as a query. This is normally achieved by comparing the target sequence with the sequence of each of the structures in the database. There are many servers that allow databases to be searched on the WWW. Examples of these are listed in Table 1. Some servers search directly against the PDB. Threading programs and fold-recognition WWW servers can also be used to find the maximum number of possible templates.

3.2.2 Selection of Templates

The list of templates obtained by the template search is then used to select the appropriate ones for homology modeling. Templates with higher overall sequence similarity are chosen. Model accuracy increases with the use of several templates. In cases where the target-template sequence identity is above 40 %, the alignment of the target and template sequences for model prediction is relatively simple and optimal. If the target-template sequence identity is lower than 40 %, the alignment generally has gaps and so the numbers of misaligned residues need to be minimized. Sequence-structure alignment, or fold recognition, is crucial to obtain an optimal model. Generally, the higher the sequence identity the more accurate is the model.

3.2.3 Three-Dimensional Structure Modeling

After the target-template alignment has been built, a three-dimensional model for the target protein can be constructed. The sequence of the unknown structure is threaded into the sequence of the known structure (template) and the fitness of the sequences for that structure is examined. There are several software programs and servers available for this purpose (Table 1). Some modeling methods employ rigid-body assembly, in which models are constructed from a few core regions, loops and sidechains obtained from related structures. A few methods involve modeling that addresses spatial restraints. In this method distance geometry or optimization techniques to satisfy spatial restraints derived from the alignment of the target sequence with the template structures are used [19, 44]. A commonly used program called MODELLER applies this method for automated model construction. The user has to input the sequences of the target and the templates. The MODELLER program then calculates models based on multiple sequence alignment between target and template proteins, to distinguish between highly conserved residues from less conserved ones. Next, the model needs to be optimized using structural, stereochemical and energy calculations.

3.2.4 Model Validation

Once the homology model has been constructed, it needs to be validated for reasonable bond length, bond angles, torsion angles, etc. The conformational rotations can then be verified using Ramachandran Plot which shows whether the distribution of the backbone bond angles is optimum. A plot for the two torsional variables Φ (phi) and Ψ (psi) indicating energetically allowed combinations of the two backbone torsional angles adjacent to the α carbon is known as Ramachandran plot. Software packages available to generate Ramachandran plots for model validation are described in Subheading 2.2.4. The overall quality of a model can be derived from the Ramachandran plot by computing the percentage of residues in the most favorable regions and the percentage of residues in the unfavorable regions. The models with maximum number of residues in the favorable region and the least

number of residues in the disallowed region are considered to be the finest models and are used for further docking studies.

3.2.5 Chemical Structure of the Drug Molecule

Drug discovery involves study of the interaction between a protein molecule and a ligand. The chemical structure of the ligand can be drawn using a variety of software packages available. The chemical structures need to be drawn correctly and clearly using consistent long lengths and angles. The drawing programs should be able to deduce additional information about the compound for further use in docking studies. Several software programs are available for this purpose, e.g., ACD/ChemSketch. This program has other uses such as calculation of molecular properties, 2D and 3D structure cleaning, structure naming, etc. Chemical structures drawn using ACD/ChemSketch software can be saved as a .MOL file and then be converted to .PDB file using Argus laboratory software [45].

3.3 Docking

Docking studies involve the binding of two interacting molecules, with the aim to fit them into favorable conformations. Docking algorithms search all the possible conformations of the ligand–receptor molecule interaction. The scoring function of the docking algorithms evaluates the different conformations and returns an energy value for them. The lower the energy value the better is the conformation. The orientation of the ligand with respect to the protein molecule is verified to ensure that there are no unacceptable steric interactions between the ligand and the protein molecules. For acceptable orientation, an interaction energy is calculated which represents the score for the docking. Docking of the protein molecule and the receptor can be done using docking programs (Table 1). The receptor and ligand molecules have to be uploaded in PDB format. For example, the docked structures can be examined using PatchDock, which gives a geometric shape complementarity scores [46–48] (*see Note 2*).

In silico drug designing is being extensively used for drug discovery. Several inhibitors including human lipoxigenase inhibitors, kinase inhibitors and cannabinoid CB2 receptor agonists have been discovered using virtual screening with homology models [1]. Several successful drugs have been developed using the drug design method and some of them have been marketed. Examples include: imatinib, which is used to treat hematologic cancers like chronic myelogenous (myeloid) leukemia (CML); ritonavir is a protease inhibitor used as an antiretroviral agent for treating HIV/AIDS [49].

4 Notes

1. Homology modeling and docking are complex tasks, because most molecules being flexible can adopt a number of different conformations of similar energy. During docking, the

characteristics of not only the receptor and the ligand but also the surrounding solvent in which these are present *in vivo* have to be taken into account. Although researchers try to attain models that are as precise as possible, several approximations need to be made in practice. Researchers have worked on this problem and have come up with algorithms which take this conformational flexibility into account during docking. Increased computing power can help to enhance the speed of virtual docking and, at the same time, maintain the required accuracy.

2. Docking studies can be automated or manual. The method to be used depends on the docking problem. Docking of single ligands can be performed by using automated methods. However, for the docking of a database of ligands, manual docking methods are preferred since the molecules with the maximum structural similarity with the receptor can be selected for better results. Manual docking methods also enable more exhaustive control of the molecules and their orientations.

References

1. Song CM, Lim SJ, Tong JC (2009) Recent advances in computer-aided drug design. *Brief Bioinform* 10:579–591
2. Sahoo BM, Dinda SC, Ravi Kumar BVV et al (2012) Computational approaches for drug design and discovery process. *Curr Pharma Res* 2:600–611
3. Bharath EN, Manjula SN, Vijaychand A (2011) *In silico* drug design-tool for overcoming the innovation deficit in the drug discovery process. *Int J Pharm Pharm Sci* 3:8–12
4. Reynaud E (2010) Protein misfolding and degenerative diseases. *Nat Educ* 3:28
5. Krieger E, Nabuurs SB, Vriend G (2003) Homology modeling. In: Bourne PE, Weissig H (eds) *Structural bioinformatics*, 1st edn. Wiley-Liss, Inc., New York, pp 507–521
6. Desai NS, Gore M (2012) Computer aided drug designing using phytochemicals-bacoside A3 and myricetin and nitric oxide donors-s-nitroso-n-acetylpenicillamine and nitroglycerin as a potential treatment of pancreatic cancer. *J Comput Sci Syst Biol* 5:001–008
7. Casey R (2005) Bioinformatics in computer-aided drug design. *BeyeNETWORK*. <http://www.b-eye-network.com/print/852>. Accessed 24 May 2013
8. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
9. Biegert A, Mayer C, Remmert M et al (2006) The MPI bioinformatics toolkit for protein sequence analysis. *Nucleic Acids Res* 34:W335–W339
10. Pieper W, Webb BM, Barkan DT et al (2011) ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res* 39:D465–D474
11. Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server. *Nucleic Acids Res* 31:3305–3307
12. Rost B (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. *Proc Int Conf Intell Syst Mol Biol* 3:314–321
13. Ficher D, Eisenberg D (1996) Fold recognition using sequence-derived predictions. *Protein Sci* 5:947–955
14. Margelevičius M, Venclovas Č (2010) Detection of distant evolutionary relationships between protein families using theory of sequence profile-profile comparison. *BMC Bioinformatics* 11:89
15. Sadreyev RI, Grishin NV (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* 326:317–336
16. Arnold K, Kiefer F, Kopp J et al (2009) The protein model portal. *J Struct Funct Genomics* 10:1–8
17. Eswar N, Webb B, Marti-Renom MA et al (2006) Comparative protein structure modeling with MODELLER. *Curr Protoc Bioinformatics* 15:5.6.1–5.6.30
18. Marti-Renom MA, Stuart AC, Fiser A et al (2000) Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* 29:291–325

19. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol* 234:779–815
20. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. *Protein Sci* 9:1753–1773
21. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9:40
22. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* 5:725–738
23. Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res* 40(Web Server issue):W471–W477. doi:10.1093/nar/gks372
24. Arnold K, Bordoli L, Kopp J et al (2006) The SWISS-MODEL Workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 22:195–201
25. Kiefer F, Arnold K, Kunzli M et al (2009) The SWISS-MODEL repository and associated resources. *Nucleic Acids Res* 37:D387–D392
26. Peitsch MC (1995) Protein modeling by E-mail. *Nat Biotechnol* 13:658–660
27. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the web: a case study using the Phyre server. *Nat Protoc* 4: 363–371
28. Morris GM, Huey R, Lindstrom W et al (2009) Autodock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J Comput Chem* 16:2785–2791
29. Kuntz ID, Blaney JM, Oatley SJ et al (1982) A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 161:269–288
30. Rarey M, Kramer B, Lengauer T et al (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 261:470–489
31. Jones G, Willett P, Glen RC (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J Mol Biol* 245:43–53
32. Abagyan R, Totrov M, Kuznetsov D (1994) ICM—a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J Comput Chem* 15: 488–506
33. Grosdidier A, Zoete V, Michielin O (2011) SwissDock, a protein-small molecule docking web service based on EADock DSS. *Nucleic Acids Res* 39:W270–W277
34. Ritchie DW, Venkatraman V (2010) Ultra-Fast FFT protein docking on graphics processors. *Bioinformatics* 26:2398–2405
35. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7:95–99
36. Lovell SC, Davis IW, Arendall WB 3rd et al (2003) Structure validation by C α geometry: Φ , ψ and C β deviation. *Proteins* 50:437–450
37. Chen VB, Arendall WB 3rd, Headd JJ et al (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66:12–21
38. Davis IW, Leaver-Fay A, Chen VB et al (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res* 35:W375–W383
39. Laskowski RA, MacArthur MW, Moss DS et al (1993) PROCHECK—a program to check the stereochemical quality of protein structures. *J Appl Cryst* 26:283–291
40. Laskowski RA, Rullmann JAC, MacArthur MW et al (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR* 8:477–496
41. Kleywegt GJ, Jones TA (1996) Phi/psi-chology: Ramachandran revisited. *Structure* 4:1395–1400
42. Colovos C, Yeates TO (1993) Verification of protein structures: patterns of non-bonded atomic interactions. *Protein Sci* 2:1511–1519
43. Lüthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. *Nature* 356:83–85
44. Srinivasan S, March CJ, Sudarsanam S (1993) An automated method for modeling proteins on known templates using distance geometry. *Protein Sci* 2:227–289
45. Thompson MA (1995) A QM/MM molecular dynamics study of the potential of mean force for the association of k⁺ with dimethylether in aqueous solution. *J Am Chem Soc* 117: 11341–11344
46. Duhovny D, Nussinov R, Wolfson HJ (2002) Efficient unbound docking of rigid molecules. In: Gusfield D, Guigo R (eds) *Proceedings of the 2nd workshop on algorithms in bioinformatics*, Rome, Italy, Springer Verlag, pp 185–200
47. Schneidman-Duhovny D, Inbar Y, Polak V et al (2003) Taking geometry to its edge: fast unbound rigid (and hinge-bent) docking. *Proteins* 52:107–112
48. Schneidman-Duhovny D, Inbar Y, Nussinov R et al (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* 33:W363–W367
49. Mandal S, Moudgil M, Mandal SK (2009) Rational drug design. *Eur J Pharmacol* 625: 90–100

INDEX

A

Algorithms

- bioinformatics.....158–161
- DNA sequencing.....4, 20, 182
- variant classification.....218, 252, 254, 255

Annotation

- conservation.....253
- DNA variant.....19, 227, 254
- genome.....5, 104, 175, 297

Association analysis

- algorithm.....3–4
- genome wide association study.....3, 47, 71
- single nucleotide polymorphism.....71

B

Bayesian analysis.....90, 160, 176

Bioinformatics

- algorithm.....158, 251–261
- analytic.....17–29
- bioinformatician.....18, 41, 189, 202, 267
- computational biology.....18
- next-generation sequencing (NGS).....4, 18, 84, 189, 195–206
- production.....17–29
- storage.....197, 202–203

C

Cancer

- germline.....87, 90, 92, 210
- somatic cell.....95, 198, 200

Chromosome microarray (CMA)

- array comparative genomic hybridization (aCGH).....122
- array single nucleotide polymorphism.....122
- balanced/unbalanced.....118, 139, 150, 151
- copy number variant (CNV).....118–125, 128–137, 140, 144–148, 150–152
- database.....119, 128–133
- de novo.....121, 130, 133, 134, 139, 150
- International Standards for Cytogenomics Analysis (ISCA).....119, 125, 133, 145, 152
- loss of heterozygosity (LOH).....127, 138, 151
- quality control (QC).....122, 123

variant of unknown significance

(VOUS).....129, 131–135, 145

Clinical annotation

- cluster analysis.....34, 40, 167, 180, 181

Comparative genomic hybridization CGH.

See Chromosome microarray (CMA)

Computational biology. *See* Bioinformatics

Computer aided drug designing

- docking.....314, 316, 317, 319
- 3-D protein structure.....314, 315, 317
- drug discovery.....313, 314, 317, 319
- drug–target interactions.....314
- homology modeling.....314–319

Copy number variant/alteration (CNV/CNA).....12, 14,

19, 25, 92–93, 118–125, 128–137, 140, 144–148, 150–152, 162–164, 168

D

Database

- browser.....8, 128, 133, 175, 229, 236, 237, 270, 299
- centralized.....266, 270
- curation.....111, 248, 267, 271, 272
- locus-specific.....221, 244, 265, 266
- mutation.....8, 22, 215, 221, 255, 264–267
- relational.....264–266
- standards.....128, 267
- variant.....13, 263–272

Data imputation

- genome wide association study (GWAS).....76
- single nucleotide polymorphism (SNP).....76, 77

Data mining.....158, 161, 165,

166, 168, 176

Deletion

- chromosome.....124, 125, 129, 136, 138, 147

Diagnostics

- comparative genomic hybridization (CGH).....122
- DNA (genetic) test.....2, 13, 66, 200, 228

Disease

- complex.....7, 67, 259
- genetic.....18, 212, 215, 267
- Mendelian.....196, 200, 204, 208, 258, 259, 296, 297
- somatic.....198, 200, 210

DNA

- mutation 83–89, 95, 218
- sequence
 - next-generation..... 17–29, 83–96
 - reference 4–5, 9, 11, 14–15
 - Sanger.....5, 9, 13, 27, 84, 175, 205, 224, 227–249, 298
 - targeted..... 8–9, 19, 87–88
- variant.....209–214, 216, 218–224, 227–249, 252, 255, 260, 263–272

Duplication

- chromosome118, 124, 125, 131, 136–138
- gene 92, 119, 136

E

Electronic medical record (EMR)

- coding.....70
- natural language processing 70, 276
- phenotype..... 66, 67, 70

Epistasis

- biological72
- statistical.....72

Exome

- next-generation sequencing 9, 18, 19, 25, 87–89, 196, 218, 220
- orphan (rare) disease.....297, 298

F

Filtering

- DNA variant.....211, 220, 223
- negative filtering.....27
- quality score.....12
- read trimming.....23, 24
- variant call format.....21

G

Gene

- candidate 6, 8–9, 13, 15, 54, 221, 295–310
- discovery..... 162, 209, 295–310
- mapping.....2, 3, 5, 8, 14, 15
- mutation 22, 196, 215, 244, 245, 255, 296
- orphan (rare)..... 295, 298, 299

Gene–gene interactions 162, 168, 245

Gene mapping

- candidate6, 8–9
- discovery.....2–3
- next-generation sequencing 4, 5, 8, 9

Gene prioritization (ranking)296–301, 303, 309, 310

- orphan (rare) disease.....298, 299

Genome

- algorithms.....4, 7
- NCBI..... 110, 130, 179, 231
- reference sequence21, 24, 26, 178, 231, 268

Genome-wide association study (GWAS)

- case-control48, 49, 54, 56, 58, 60
- Hardy Weinberg equilibrium.....52, 54
- imputation57–59
- linkage disequilibrium55, 58
- meta-analysis66, 75–76
- minor allele frequency.....67–68
- missing heritability66, 71–73
- population stratification..... 48, 53–56, 65, 71
- principal component analysis (PCA)..... 54–56, 71
- quality control (QC)..... 51–53, 65, 73, 77
- single nucleotide polymorphism 48, 63, 65, 67–68, 162

Genomic diagnosis

- complex disease.....83–96
- incidental finding..... 134, 145, 146, 207–224
- next-generation sequencing (NGS).....196–198
- variant of unknown significance (VUS).....228

Genomics

-13, 18, 19, 83–96, 156, 162–165, 168, 175, 177, 190, 199, 222, 237
- genotype138, 142, 143, 146, 150

H

High-throughput sequencing

- targeted..... 8–9, 95, 259
- whole exome sequencing (WES)..... 8–10, 19, 21, 24, 28, 228, 252, 298
- whole genome sequencing (WGS)..... 8, 10–15, 19, 21, 28, 87, 88, 93, 95, 174, 175, 178, 184, 189–191, 196, 209, 228, 252, 259–260, 296

Homology modeling

-221, 314–319
- comparative genomics.....315

I

Imputation

- genome wide association study (GWAS) 49, 57–59, 76–77
- genotype7, 76–78
- haplotype7, 76, 77
- polymorphisms57

Incidental finding

- consent.....208, 222
- filtering 210, 211, 216–223
- inheritance211–212, 219
- phenotype 207–209, 211, 212, 221
- risk..... 208–215, 219, 223
- variant database 208–211, 218–219, 221–222

Infectious disease. *See* Pathogen

- Insertion deletion (indel).....3, 6, 9, 12, 13, 19, 25, 26, 28, 89–92, 96, 175, 178, 191, 218, 220, 224, 259, 298

In silico

-95, 107, 121, 216, 222, 228, 230, 233–242, 244–246, 248, 254, 255, 258, 264, 313–315, 319

In vitro.....187, 189, 233, 244, 245, 253, 264, 265
In vivo..... 32, 38, 187, 232, 244, 245, 253, 320

L

Linkage disequilibrium (LD) 5–7, 55–58, 65, 66, 68, 73–77, 212, 252
 single nucleotide polymorphism.....6, 68
Loss of heterozygosity (LOH). *See also* Somatic mutation
 cancer..... 127, 138, 151

M

Machine learning method/algorithm
 cross validation38, 43
 data set.....37, 176
 statistical model.....37
 supervised/unsupervised (clustering)160
Mapping. *See* Gene mapping
Mass spectrometry (MS). *See* Metabolomics
Metabolites
 Human Metabolome Database.....36
 phenotype..... 31, 32, 34, 48, 49
 profiling.....31, 33
Metabolomics
 analyte36–40
 liquid chromatography (LC).....33, 39
 mass spectrometry (MS).....35, 36
 nuclear magnetic resonance (NMR).....36
 targeted.....32–39
 untargeted.....34–36, 39–40
Microarray. *See also* Chromosome microarray (CMA)
 comparative genomic hybridization (CGH).....122
 single nucleotide polymorphism (SNP).....122
Microbial genomics. *See* Pathogen
MicroRNA (miRNA)
 miRBase 107–109, 111, 112
 miRNA target.....110, 115
 non-coding RNA (ncRNA).....99, 100, 104, 106, 107, 111, 113
Missense variants. *See* Mutation detection; Variant
Missing heritability
 complex disease.....64, 66
 genome wide association study
 (GWAS)..... 64, 66, 71–73
Multiple testing..... 49, 74, 228
Mutation detection
 animal model14
 annotation.....5
 comparative genomics.....5
 database7
 family studies.....243–244
 literature 271
 pathogenic245, 271
 pathogenicity prediction241
 population frequency 212–214, 242–243, 247

in silico 95, 228, 233, 258
variant of unknown significance (VUS)..... 245, 259
in vitro 189, 253, 265
in vivo 187, 253

N

Natural language processing (NLP)
 architecture275–292
 biomedicine275–292
 clinical decision support275, 277
 electronic medical record (EMR) 276, 288, 292
 knowledge resources 276–279, 282–288, 290
 machine learning276, 279, 280, 290, 291
 tools 276–281, 290
 unified medical language system.....276, 279
Next-generation sequencing (NGS)
 bioinformatics (production, analytic, storage).....83–96
 clinical *vs.* research applications.....196
 high throughput 8, 84, 198
 massively parallel195, 197
 quality measures..... 86, 196, 199
 targeted gene sequencing.....87–89
 whole exome sequencing (WES).....9–10
 whole genome sequencing (WGS).....10–14

O

Omics
 comparative167
 genomics..... 86, 162–164
 metabolomics.....167
 phenomics.....158
 proteomics166–167
 transcriptomics164–166
Orphan (rare) disease
 biological networks296
 candidate gene295–310
 locus.....296
 prioritization.....295–310

P

Pathogen
 bio-surveillance..... 174, 175, 177
 cluster174–176, 179–183, 186–187
 infectious disease176, 177, 179, 183, 188
 phylogenies 176, 179, 181–183
 public health 174, 175, 189, 190
 whole genome sequencing (WGS)..... 174, 175, 178, 184, 189–191
Pathway
 analysis..... 41, 158–167, 169
 data mining.....158, 161, 165, 166, 168
 discovery 162, 163, 166, 169
 interactive161, 167
 machine learning 160–161, 168

Phenomics158

Phenotype 1–15, 31, 32, 34, 48, 49, 59,
65–67, 69–70, 72–75, 77, 118, 119, 130, 132,
134, 137, 145–147, 162, 163, 166–168, 178,
207–209, 211, 212, 221, 222, 228, 242, 243, 245,
252, 253, 255, 257, 258, 263, 265–270, 288, 291,
296, 297, 299, 309

gene mapping2, 3, 5, 8, 14, 15

Polymorphisms

restriction fragment length polymorphism
(RFLP)2, 13

single nucleotide polymorphism (SNP)..... 2, 48, 63, 65,
67–68, 122, 159, 162, 178

Population stratification48, 53–56, 60, 65, 71

genome wide association study (GWAS)47–60

Principal component analysis (PCA)..... 54–56, 71,
183, 188, 192

Proteomics

biomarker discovery 158, 166, 169

pathway discovery158

Public health surveillance. *See* Pathogen

Q

Quality assurance

coverage5, 10, 202

reference genome3, 5, 10

replication65

Quality control (QC)

filtering 22, 23, 65, 73

validation199

R

Rare disease. *See* Orphan disease

S

Sequencing. *See* Next-generation sequencing (NGS)

Single nucleotide polymorphism (SNP)

annotation239, 240

association study 63, 162, 164

dbSNP 13, 22, 26, 67, 87, 213, 229,
237, 242–244, 247, 248, 253, 254, 298

GWAS65, 67, 164

linkage disequilibrium6, 55, 65, 66, 68, 212

population frequency212, 242–243

synonymous/non-synonymous26, 239

variant 10, 239

Single nucleotide variation (SNV). *See* Single nucleotide
polymorphism (SNP)

Somatic mutation

copy number alteration/variant (CNV/CNA)92, 93

germline mutation89

heterogeneity 84, 85, 94, 256

ploidy84–86, 94

structural rearrangement 84, 88, 92–94

tumor/cancer 84–87, 89, 94, 200, 256, 259

variant caller software88–90

Standards

quality analysis (QA)101–102

quality control (QC) 19, 22–24, 127

statistical analysis34, 36–40

Structural rearrangements12, 14, 84, 88, 89, 92–94

Supervised analysis. *See* Machine learning

System architecture275–292

T

Transcriptomics 158, 164–169

U

Unified medical language system. *See* Natural language
processing

Unsupervised analysis. *See* Machine learning

V

Variant. *See also* Single nucleotide polymorphism (SNP)

annotation228, 253, 254

causal2, 4, 6–9, 11, 13, 68, 196, 252, 255, 298

classification229, 245–247

database 13, 214, 218,
263–272

Variant of unknown significance (VUS/VOUS)

chromosome microarray 129, 131–135, 145

DNA variant/mutation228

genomic diagnosis228

somatic mutation245, 259

W

Whole exome sequencing (WES) 8, 19, 21, 24,
28, 208, 298

Whole genome sequence(ing) (WGS) 8, 10–15,
18, 21, 28, 87, 93, 95, 174, 175, 178, 184,
189–191, 196, 209, 228, 252, 259–260.
See also Next-generation sequencing (NGS)