The OMICs

Applications in Neuroscience

Edited by
Giovanni Coppola

OXFORD

THE OMICs

THE OMICs

Applications in Neuroscience

EDITED BY

GIOVANNI COPPOLA

Director, Center for Informatics and Personalized Genomics
Semel Institute for Neuroscience and Human Behavior
Departments of Psychiatry & Neurology
David Geffen School of Medicine
University of California, Los Angeles

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide.

Oxford New York
Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in
Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trademark of Oxford University Press
in the UK and certain other countries.

Published in the United States of America by
Oxford University Press
198 Madison Avenue, New York, NY 10016

Library of Congress Cataloging–in–Publication Data
OMICS (2014)
The OMICs : applications in neuroscience / edited by Giovanni Coppola.
p. ; cm.
Includes bibliographical references.
ISBN 978–0–19–985545–2 (alk. paper)
I. Coppola, Giovanni, 1971– editor of compilation. II. Title.
[DNLM: 1. Brain—physiology. 2. Genomics—methods. 3. Drug Discovery.
4. Translational Medical Research. QU 460]
QP376
612.8′2—dc23
2013024509

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

CONTENTS

Contributors vii

PART ONE: *DNA*

1. Medical DNA Sequencing in Neuroscience *03*
   KAROLA REHNSTRÖM, ARVID SULS, AND AARNO PALOTIE
2. Epigenomics: An Overview *27*
   KEVIN HUANG AND GUOPING FAN
3. The Role of Epigenomics in Genetically Identical Individuals *42*
   ZACHARY A. KAMINSKY

PART TWO: *RNA*

4. Transcriptomics *63*
   T. GRANT BELGARD AND DANIEL H. GESCHWIND
5. Decoding Alternative mRNAs in the "Omics" Age *73*
   YUAN YUAN AND DONNY D. LICATALOSI
6. Transcriptomics: From Differential Expression to Coexpression *85*
   MICHAEL C. OLDHAM
7. High-Throughput RNA Interference as a Tool for Discovery in Neuroscience *114*
   LISA P. ELIA AND STEVEN FINKBEINER
8. The Genetics of Gene Expression: Multiple Layers and Multiple Players *132*
   AMANDA J. MYERS

PART THREE: *PROTEIN*

9. Proteomics *155*
   JONATHAN C. TRINIDAD, RALF SCHOEPFER, AND A. L. BURLINGAME
10. Focused Plasma Proteomics for the Study of Brain Aging and Neurodegeneration *183*
    PHILIPP A. JAEGER, SAUL A. VILLEDA, DANIELA BERDNIK, MARKUS BRITSCHGI, AND TONY WYSS-CORAY

PART FOUR: *CELLS AND CONNECTIONS*

11. Cellomics: Characterization of Neural Subtypes by High-Throughput Methods and Transgenic Mouse Models *195*
    JOSEPH DOUGHERTY
12. Neuroscience and Metabolomics *220*
    REZA M. SALEK
13. Brain Connectomics in Man and Mouse *232*
    ARTHUR W. TOGA, KRISTI CLARK, HONG WEI DONG, HOURI HINTIRYAN, PAUL M. THOMPSON, AND JOHN D. VAN HORN
14. Optogenetics *254*
    RICHIE E. KOHMAN, HUA-AN TSENG, AND XUE HAN
15. Characterizing the Gut Microbiome: Role in Brain–Gut Function *265*
    GERARD CLARKE, PAUL W. O'TOOLE, JOHN F. CRYAN, AND TIMOTHY G. DINAN

PART FIVE: *THERAPEUTICS*

16. OMICs in Drug Discovery: From Small Molecule Leads to Clinical Candidates *291*
    B. MICHAEL SILBER
17. Pharmacogenomics *315*
    STEVEN P. HAMILTON

PART SIX: *OMICsOME: INTEGRATION OF OMICs DATA*

18. Multidimensional Databases of Model Organisms *333*
    KHYOBENI MOZHUI AND ROBERT W. WILLIAMS
19. Network Biology and Molecular Medicine in the Postgenomic Era: The Systems Pathobiology of Network Medicine *345*
    STEPHEN Y. CHAN AND JOSEPH LOSCALZO

Links to Helpful Resources *357*

Index *361*

CONTRIBUTORS

T. Grant Belgard
Department of Psychiatry
David Geffen School of Medicine
University of California, Los Angeles
Los Angeles, CA

Daniela Berdnik
Department of Neurology and Neurological Sciences
Stanford University School of Medicine
Stanford, CA

Markus Britschgi
F. Hoffmann-La Roche AG
pRED, Pharma Research & Early Development, DTA Neuroscience
Basel, Switzerland

A. L. Burlingame
Department of Pharmaceutical Chemistry
University of California, San Francisco
San Francisco, CA

Stephen Y. Chan
Division of Cardiovascular Medicine
Department of Medicine
Brigham and Women's Hospital
Harvard Medical School
Boston, MA

Kristi Clark
The Institute for Neuroimaging and Informatics (INI) and Laboratory of Neuro Imaging (LONI)
Keck School of Medicine of USC
University of Southern California
Los Angeles, CA

Gerard Clarke
Department of Psychiatry and Alimentary Pharmabiotic Centre
University College Cork
Cork, Ireland

John F. Cryan
Department of Anatomy and Neuroscience & Alimentary Pharmabiotic Centre
University College Cork
Cork, Ireland

Timothy G. Dinan
Department of Psychiatry and Alimentary Pharmabiotic Centre
University College Cork
Cork, Ireland

Hong Wei Dong
The Institute for Neuroimaging and Informatics (INI) and Laboratory of Neuro Imaging (LONI)
Keck School of Medicine of USC
University of Southern California
Los Angeles, CA

Joseph Dougherty
Department of Genetics & Department of Psychiatry
Washington University School of Medicine in St. Louis
St. Louis, MO

Lisa P. Elia
Gladstone Institute of Neurological Disease and
Taube-Koret Center for Huntington's Disease Research
University of California, San Francisco
San Francisco, CA

Guoping Fan
Department of Human Genetics
David Geffen School of Medicine
University of California, Los Angeles
Los Angeles, CA

Steven Finkbeiner
Gladstone Institute of Neurological Disease and
Departments of Neurology and Physiology
University of California, San Francisco
San Francisco, CA

Daniel H. Geschwind
Departments of Psychiatry, Neurology and Human Genetics
David Geffen School of Medicine
University of California, Los Angeles
Los Angeles, CA

Steven P. Hamilton
Department of Psychiatry and Institute for Human Genetics
University of California, San Francisco
San Francisco, CA

Xue Han, Ph.D.
Assistant Professor
Biomedical Engineering Department and
Joint Professor
Department of Pharmacology and Experimental Therapeutics Member
Photonics Center Boston University
Boston, MA

Houri Hintiryan
The Institute for Neuroimaging and Informatics (INI) and Laboratory of Neuro Imaging (LONI)
Keck School of Medicine of USC
University of Southern California
Los Angeles, CA

Kevin Huang
Department of Human Genetics
David Geffen School of Medicine
University of California, Los Angeles
Los Angeles, CA

Philipp A. Jaeger
Departments of Bioengineering and Medicine
University of California San Diego
La Jolla, CA

Zachary A. Kaminsky
Johns Hopkins University School of Medicine
Department of Psychiatry
Baltimore, MD

Richie E. Kohman
Biomedical Engineering Department
Boston University
Boston, MA

Donny D. Licatalosi
Center for RNA Molecular Biology
Case Western Reserve University
Cleveland, OH

Joseph Loscalzo
Department of Medicine
Brigham and Women's Hospital
Harvard Medical School
Boston, MA

Khyobeni Mozhui
Department of Preventive Medicine
University of Tennessee Health
Memphis, TN.

Amanda J. Myers
Laboratory of Functional Neurogenomics
Department of Psychiatry & Behavioral Sciences
Program in Neuroscience
Interdepartmental Program in Human Genetics and Genomics
Center on Aging
University of Miami Miller School of Medicine
Miami, FL

Michael C. Oldham
Department of Neurology
The Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research
University of California, San Francisco
San Francisco, CA

Paul W. O'Toole
School of Microbiology and Alimentary Pharmabiotic Centre
University College Cork
Cork, Ireland

Aarno Palotie
Wellcome Trust Sanger Institute
Hinxton, United Kingdom and
Institute for Molecular Medicine
University of Helsinki
Helsinki, Finland and
Program for Human and Population Genetics
The Broad Institute of MIT and Harvard
Cambridge, MA

Karola Rehnström
Wellcome Trust Sanger Institute
Hinxton, United Kingdom and
Institute for Molecular Medicine
University of Helsinki
Helsinki, Finland

Reza M. Salek
Department of Biochemistry and Cambridge Systems Biology Centre
University of Cambridge
Cambridge CB2 1GA, UK

Ralf Schoepfer
Laboratory for Molecular Pharmacology
NPP (Pharmacology)
University College London
London, United Kingdom

B. Michael Silber
Department of Bioengineering and Therapeutic Sciences
Schools of Medicine and Pharmacy
University of California, San Francisco
San Francisco, CA

Arvid Suls
VIB-Department of Molecular Genetics and
University of Antwerp
Antwerpen, Belgium

Paul M. Thompson
The Institute for Neuroimaging and Informatics (INI) and Laboratory of Neuro Imaging (LONI)
Keck School of Medicine of USC
University of Southern California
Los Angeles, CA

Arthur W. Toga
The Institute for Neuroimaging and Informatics (INI) and Laboratory of Neuro Imaging (LONI)
Keck School of Medicine of USC
University of Southern California
Los Angeles, CA

Jonathan C. Trinidad
Department of Pharmaceutical Chemistry
University of California, San Francisco
San Francisco, CA and
Department of Chemistry
Indiana University
Bloomington, IN

Hua-an Tseng
Biomedical Engineering Department
Boston University
Boston, MA

John D. Van Horn
The Institute for Neuroimaging and Informatics (INI) and Laboratory of Neuro Imaging (LONI)
Keck School of Medicine of USC
University of Southern California
Los Angeles, CA

Saul A. Villeda
Department of Anatomy and the Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research
University of California, San Francisco
San Francisco, CA

Robert W. Williams
Department of Anatomy & Neurobiology
Center for Integrative and Translational Genomics
University of Tennessee Health Science Center
Memphis, TN

Tony Wyss-Coray
Department of Neurology and Neurological Sciences
Stanford University School of Medicine
Stanford, CA and
VA Palo Alto Health Care System
Palo Alto, CA

Yuan Yuan
Laboratory of Molecular Neuro-Oncology
The Rockefeller University
New York, NY

PART I

DNA

1

Medical DNA Sequencing in Neuroscience

KAROLA REHNSTRÖM, ARVID SULS, AND AARNO PALOTIE

INTRODUCTION

The aim of medical genetic studies is to identify genetic variants associated with a disorder or trait of interest. A hypothesis-free way to conduct gene mapping studies has been available ever since genetic variants, usually referred to as genetic markers, were identified. The first genetic markers used in gene mapping studies were a small number of blood antigens; later, microsatellites were used. The human reference genome and the Hap Map project identified millions of single nucleotide polymorphisms (SNPs) spread all across the genome, which provided a much denser map of genetic markers. Today high-throughput sequencing technology has made it possible to decode every base pair in the human genome, enabling the identification not only of sites, which are polymorphic in a population, but also of private mutations, which are present in only one individual. Despite the feasibility of producing enormous datasets for medical genetic studies, the path from generating the data to identifying the variants involved in the disease and further converting this to an understanding of biological mechanisms is still in its early stages.

THE HISTORY OF GENE MAPPING STUDIES

Traditionally, human genetic disorders have been divided into monogenic and complex types. This somewhat simplified division reflects the underlying genetic architecture. Monogenic (or Mendelian) disorders are caused by mutations in one gene. These mutations are highly penetrant and rare in the population (Figure 1.1). Depending on the mode of inheritance, loss of one or two copies is required for the disease to manifest. More than 3,000 such disorders are listed in the Online Mendelian Inheritance in Man (OMIM, www.ncbi.nlm.nih.gov/omim) database, and the causative genes have been identified in one third to half of these (Bamshad, Ng, et al. 2011). Although many disorders, particularly monogenic recessive disorders, are clearly caused by mutations in a single gene, there are likely other genes that can modify the phenotypic features. This could prove particularly true for dominant disorders, because they often display reduced penetrance and the phenotype can be highly variable, even within a family where the primary genetic lesion is shared by all affected individuals.

Genetic mapping of monogenic disorders has been successful. Linkage analysis and subsequent sequence analysis in a small number of families has often resulted in identification of the causative gene. An excellent example of the power of these approaches, and the power of genetic homogeneity in isolated populations, is successful mapping of genes for monogenic, often recessive disorders in population isolates such as the Finns or the Hutterites (Boycott, Parboosingh, et al. 2008; Norio 2003). Although linkage studies have identified genes for many monogenic disorders, there are still numerous disorders for which the causative gene or genes are not known. These include disorders where families are too small to provide a linkage signal or cases where genetic heterogeneity between families is very high and traditional methods have not been able to identify the disease genes.

Complex disorders are caused by a combined load of a large number of genetic variants, each of which confers a very small increase in risk (Figure 1.1). These variants are relatively common in the population. The genetic background of complex disorders has been extensively characterized during the last decade using genome-wide association studies (GWAS). In these studies, very large cohorts of samples are genotyped at loci known to be polymorphic in the population. Statistical tests

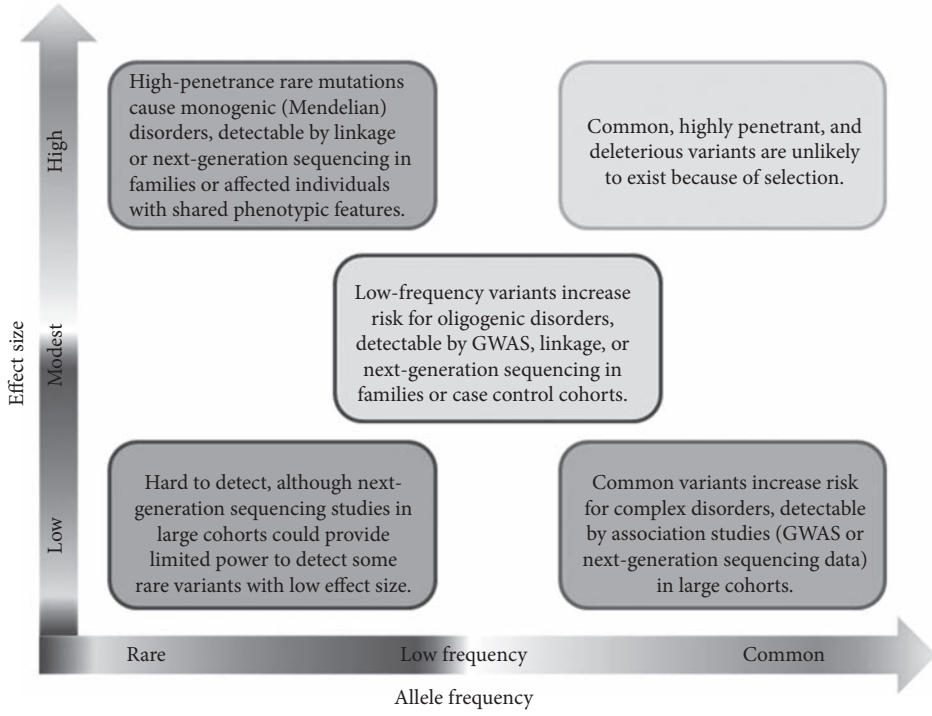

FIGURE 1.1: The genetic architecture of diseases and traits ranges from disorders caused by only one highly disruptive and fully penetrant variant to those caused by the additive effects of numerous genetic variants of very small effect, often in combination with environmental factors. Highly disruptive variants (i.e., variants with a large effect size) are rare in the population as they are subject to strong negative selection, whereas variants with lower effect sizes can become more common in the population as one variant alone is insufficient to cause the disorder. Currently available technologies and analysis methods for the identification of these variants have their limitations; choice of the most efficient approach for gene mapping studies depends on the genetic architecture of the trait.

are then performed to determine if a genetic marker is more common in cases than controls. The combination of large-scale SNP identification projects allowing for dense coverage of the whole genome combined with technological advances in high-throughput genotyping technology enabling the genotyping of tens of thousands of samples has resulted in identifying the association of thousands of SNP markers with hundreds of diseases and traits (http://www.genome.gov/gwastudies/). However, in most cases the GWAS loci explain only a small to moderate part of the heritability of the traits. For complex disorders, the environment is also likely to play a much larger role than for monogenic disorders and will probably prove to be the main susceptibility factor for some of them. In addition to common variation, rare variants with large effect sizes have also been found to play a role in several complex disorders. GWAS technologies have been poorly equipped to identify such risk variants, whereas large-scale sequencing studies are better equipped to identify them.

Many disorders cannot be distinguished as being either monogenic or complex, since there are numerous complex disorders that also have monogenic, very severe, and often early-onset forms. For example, meta-analyses of tens of thousands of individuals have revealed dozens of common susceptibility variants for both type 1 and type 2 diabetes (Bradfield, Qu, et al. 2011; Saxena, Elbers, et al. 2012). At the same time, rare mutations in *GCK* (Froguel, Vaxillaire, et al. 1992) and *HNF1A* (Yamagata, Furuta, et al. 1996) cause maturity-onset diabetes of the young (MODY), and mutations in *KCNJ11* (Gloyn, Pearson, et al. 2004) and *ABCC8* (Babenko, Polak, et al. 2006) cause neonatal diabetes, two monogenic forms of diabetes.

Similarly, GWAS analyses of blood lipid levels have revealed significant overlap between genes with common susceptibility variants and previously identified genes in familiar forms of dyslipidemias (Teslovich, Musunuru, et al. 2010). For many disorders where the molecular etiology is not known, it is not possible to differentiate between monogenic and complex forms of the disorder based on the phenotype alone; therefore several complementary gene mapping efforts are needed to further our understanding of the genetic architecture of genetic disorders and traits.

CURRENT STATUS

The development of genotyping and sequencing technologies along with a good partnership between academia and industry has been essential in changing the landscape on how human disease genomics research is done. During the past 10 years genotyping studies have moved from linkage panels based on 400 microsatellites to genotyping up to a million markers for GWAS and lately to sequencing the complete genome in each study sample. As summarized above, gene mapping technologies have successfully identified genes for monogenic as well as more complex disorders. However, there are many cases where neither approach has been successful. Traditional automated Sanger sequencing is very costly and laborious if large linkage intervals must be sequenced, and GWAS are limited in their power to identify susceptibility factors with a very low allele frequency.

Next-Generation Sequencing Technology

The initial draft of the human genome was produced using automated Sanger sequencing, a technology where modified fluorescent bases are incorporated into a strand of DNA using polymerase chain reaction (PCR) and then separated by gel electrophoresis (Lander, Linton, et al. 2001). However, the completion of the draft sequence took a large consortium of 20 collaborating research groups a decade and cost $3 billion. Clearly technological advances were required to enable large-scale DNA sequencing projects. The term *next-generation sequencing* (NGS) is used for the high throughput technologies that have been developed to complement and ultimately replace Sanger sequencing. These methods have been available from 2004 (Margulies, Egholm, et al. 2005) and have brought with them an immense drop in sequencing cost. Until 2007 the reduction in sequencing cost was well modeled by Moore's law (which describes a long-term trend in the computer hardware industry that involves the doubling of "compute power" every two years and is often used as a standard to assess whether technological development is being successful). Since the beginning of 2008 the drop in sequencing cost has been much faster than predicted by Moore's law, allowing for the generation of ever-growing datasets. (Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program Available at: www.genome.gov/sequencingcosts). NGS has been successfully applied to several areas of genetics and epigenetic research, including but not limited to medical genetic studies, population genetics, evolutionary studies, transcriptomics, and epigenomics.

Currently two main approaches are used to generate large-scale resequencing data for medical genetic studies: selective capture of specific genomic regions and whole-genome sequencing (WGS). Capture of selected genomic regions is suitable for projects where targeted genomic regions, such as loci identified in GWAS, or predefined sets of genes (such as synaptically expressed genes) are being targeted. The benefit of targeted sequencing is that because limited amounts of is being generated, data from several samples can be pooled together in one run on the sequencing instrument; thus a large number of samples can be included in the study. WGS generates a huge amount of data and requires much more sequencing capacity and storage space per sample. Furthermore, the additional data volume results in analytical and interpretational challenges. On the other hand, WGS data is totally hypothesis-free as it allows the assessment of all variation present in an individual's genome. An often used compromise between the two extremes is whole-exome sequencing (WES), a form of selective capture where all known protein coding regions (exons) are sequenced. The genetic variants causing monogenic disorders usually affect protein structure and function and are thereby located in exons (Kryukov, Pennacchio, et al. 2007; Stenson, Ball, et al. 2009). Therefore focusing sequencing efforts on the exome will likely reveal variants with large effect sizes that are acting by disrupting or altering protein function. However, the

basic assumption that all disorders are probably caused by coding variants is likely untrue. It is possible that the majority of identified variants are exonic because gene identification efforts have been concentrated on exons. In addition, prediction of the consequence of a coding variant on protein function is somewhat easier than prediction of the consequence of noncoding variants. WGS is likely to provide unbiased information about the true genetic architecture of traits.

Currently it is widely accepted that WES is well powered to detect variants involved in human disease. WES has so far identified genes for over 100 monogenic disorders (Rabbani, Mahdieh, et al. 2012). The same approach has also been applied to complex disorders, although with more modest success. In addition to the successes, the challenges of this approach have also become evident. Interpretation of the sequence data and identification of functional disease-causing mutations from the multitude of variants in each exome sample is not a trivial task. Developing the statistical framework guiding the interpretation of WES data is still in progress. Firm guidelines will help in the interpretation of the sequence data.

Sample Preparation and Targeted Sequence Capture

The NGS sequencing instruments will sequence every molecule of DNA in the template library loaded onto the instrument. If sequencing is to be limited to specific regions of interest, enrichment of these regions from the entire genome must be performed before the sample is sequenced. In traditional automated Sanger sequencing this was primarily achieved by PCR amplification of regions of interest, and PCR-based methods have also been used for NGS (Meuzelaar, Lancaster, et al. 2007; Varley and Mitra, 2008). Today, however, enrichment of regions of interest is primarily achieved by targeted hybrid capture methods.

Hybrid capture can be used to enrich for any regions of interest, such as a subset of genes (Figure 1.2). One of the most common applications, however, is to capture all protein coding regions of the genome. The protein coding exome comprises only 1.2% of the human genome (Dunham, Kundaje, et al. 2012). However, what today is called exome capture is actually an enrichment not only for protein coding regions but also other possible functional regions of the genome, such as micro RNAs (miRNAs) and noncoding exons. In practice, different manufacturers have slightly different content on their exome capture reagents. Comparisons of the most popular products available suggest that certain kits cover a slightly larger amount of protein coding and miRNA genes, but none of the kits cover all Consensus Coding Sequence (CCDS) exons (Asan, Xu, et al. 2011; Coffey, Kokocinski, et al. 2011; Sulonen, Ellonen, et al. 2011). Analogous to GWAS chips, the exome capture assays get updated as new annotation information becomes available to include as much of the coding sequence and other functional regions as possible. Usually the baits included in the exome capture assays are based on information from several different databases and annotation resources, such as genes from the CCDS project (Pruitt, Harrow, et al. 2009), RefSeq (Pruitt, Tatusova, et al. 2012), Gencode/Encode (Harrow, Frankish, et al. 2012) and miRbase (Kozomara and Griffiths-Jones 2011) or other miRNA databases.

It is highly likely that WES is a temporary compromise that is currently employed for convenience to limit data generation and ease the interpretation of results. It will be routinely replaced by WGS as prices drop, sequencing capacity increases, and better annotation workflows are available. Therefore, in the future, many of the problems and pitfalls associated with WES will be surpassed. Although the limited amount of data produced by WES can simplify interpretation of results, it will limit variant detection to a small part of the genome. Sample preparation using pull-down reagents also increases cost per base pair sequenced compared with WES. On the other hand, the small size of the target DNA allows for cost-efficient sequencing of samples at relatively high coverage (usually 30- to 60-fold coverage), increasing the power to detect rare variants compared with lower-coverage WGS. Despite the improvement of exome capture assays, the coverage of individual exomes is still highly variable even in high-coverage data. A fraction (up to 0.5%) of the target regions are not captured at all or at very low coverage, making the individual exon coverage highly variable (Asan, Xu, et al. 2011). WGS often produces a more even coverage of the genome, as no bias is introduced by hybrid capture. The uneven distribution of sequence depth in WES data makes the detection of copy number variants (CNVs) more challenging than for WGS data.

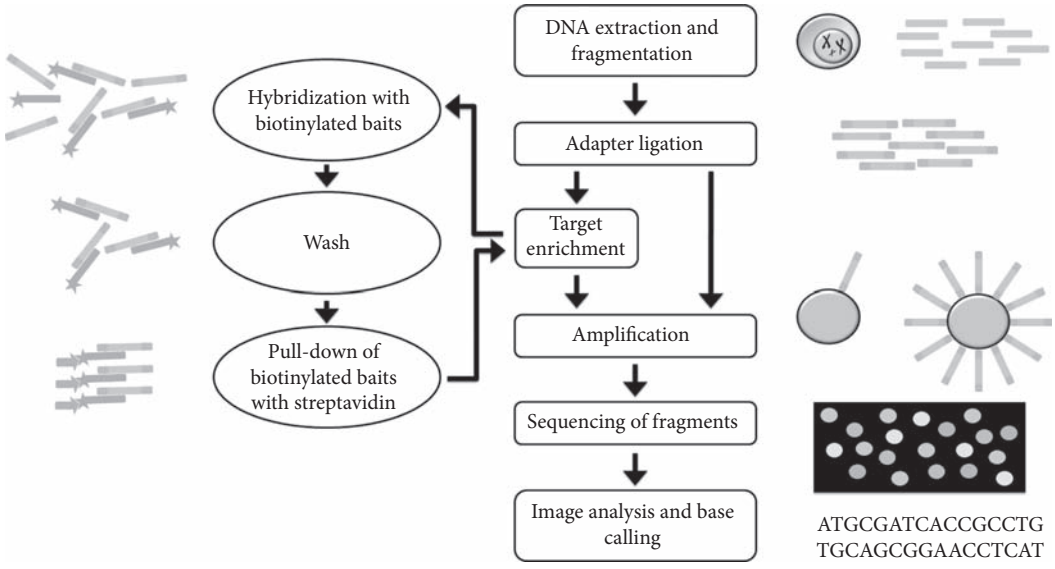

FIGURE 1.2: The main steps of next-generation sequencing: First DNA is extracted and fragmented and adapters that serve as PCR primers are added to the ends of the DNA fragments. If DNA from several samples is sequenced in the same lane of the sequencing instrument, oligonucleotides that serve as barcodes for each individual sample are also added to the fragments (not shown). If only a subset of the genome is to be sequenced, DNA or RNA baits are used to enrich for the desired genomic regions and a biotin-streptavidin–based pull-down reaction is used to obtain the desired DNA fragments. These are then amplified and sequenced and the images produced by the sequencing instrument are processed to extract the DNA sequence for each amplified DNA fragment.

The workflow for WES consists of three basic steps—template preparation, sequencing, and imaging—followed by bioinformatic analysis (Figures 1.2 and 1.3). To construct a template, a relatively large amount (several micrograms) of genomic DNA is randomly sheared to form fragments, and adaptors (short oligonucleotides) are added to the sequences. Enrichment of the exonic sequence is done by hybridizing the sheared DNA with biotinylated DNA or RNA baits, and the hybridized fragments are then captured by biotin-streptavidin–based pull-down. The exome library is then massively amplified by using the adapters as primers, and the amplified DNA molecules are sequenced. As current technologies allow for the sequencing of several samples in the same lanes of the sequencing instrument, barcoded indexing tags are introduced at the library preparation stage for identification, after sequencing, of sequences belonging to individual samples.

Sample preparation for WGS is simpler as it does not require any template selection. The sequencing library is created from sheared segments of DNA, which are attached to adapters to allow amplification of the DNA. Although most current technologies rely on amplification before sequencing, some technologies can sequence unamplified DNA (Treffer and Deckert 2010).

Amplification and Sequencing Technology

Before the actual sequencing takes place, most currently available sequencing technologies require that the DNA library be massively amplified to provide multiple copies of each DNA fragment. Various approaches are used by the different NGS technologies for the amplification and sequencing steps (Metzker 2010).

Amplification can occur by emulsion PCR (Dressman, Yan, et al. 2003) where single-stranded DNA is attached to beads and then amplified by PCR (used by Roche/454 and Applied Biosystems/SOLiD). The conditions are optimized so that only one template molecule attaches to each bead and is therefore a clonal copy of the original fragment after amplification. Beads can then be cross-linked to glass surfaces or deposited in microscopic wells for sequencing.

Amplification can also be performed in solid phase (Adessi, Matton, et al. 2000; Fedurco, Romieu, et al. 2006) (Illumina/HiSeq). The DNA

with the attached adapters is immobilized onto a two-dimensional surface with oligonucleotides that are complementary to the adapters. PCR is then performed, using primers designed to target the adapters of the DNA fragments until clusters of about a million copies of the original DNA molecule are formed.

After amplification, the actual sequencing reaction is performed, which involves the steps of base determination, imaging, and initial image processing to decode the order of bases in the DNA fragment (Anderson and Schrijver 2010; Mardis 2008; Metzker 2010). Sequencing can be performed either by synthesis or by ligation. Sequencing by synthesis can be further divided into cyclic reversible termination, single-nucleotide addition, and real-time sequencing.

Cyclic reversible termination involves the addition of either one or all four nucleotides, which will bind in a template-defined manner and are added by a mutant DNA polymerase that can incorporate the modified nucleotides. The nucleotides are capped to prevent additional extension reactions and have a fluorescent label. Following incorporation, the unincorporated nucleotides are washed away and imaging by lasers is performed to determine the identity of the nucleotide. Subsequently, the terminating group and fluorescent label are cleaved to allow for another round of template-directed extension. In this method, with the addition of all four bases, each cycle is used by the Illumina/HiSeq, whereas the Helicos BioSciences single molecule sequencing technology uses a cyclic reversible termination with only one base added to each cycle of the sequencing (Braslavsky, Hebert, et al. 2003).

Pyrosequencing (Ronaghi, Uhlen, et al. 1998), used by the Roche/454 (Margulies, Egholm, et al. 2005), is also a DNA polymerase–driven method that detects the bioluminescence generated by the release of inorganic pyrophosphate when the DNA sequence is being extended by a complementary nucleotide. The order and intensity of the bioluminescence is recorded by the charge-coupled device (CCD) camera in the instrument. The signal strength is proportional to the number of nucleotides; for example, homopolymer stretches generate a greater signal than single nucleotides.

Sequencing by ligation is also a cyclic method but uses a DNA ligase instead of a DNA polymerase (Tomkinson, Vijayakumar, et al. 2006). The process uses either one-base-encoded probes or two-base-encoded probes. A fluorescently labeled probe hybridizes to the target in a template-guided manner and a DNA ligase is added to join the probe with the primer. After nonincorporated probes are washed away, fluorescence detection will determine which nucleotide has been incorporated. Again, the fluorescent dye will then be removed and another set of probes will be added. The Life/SOLiD technology uses two-base-encoded probes, which yield a sequence every five base pairs because of three degenerate bases on each dinucleotide probe (Shendure, Porreca, et al. 2005; Valouev, Ichikawa, et al. 2008). After finishing the first round of ligation, the template is stripped and another primer is used, this time starting at (n-1) position relative to the first round. This way, after doing five rounds of elongation, the whole sequence will have been twice covered by template-specific interrogation bases.

Data from the sequencing run is stored in image files, which are processed to determine the base-pair composition of each fragment that has been sequenced. The manufacturers supply algorithms for base calling, but other base-calling algorithms have been developed that provide improvement over the manufacturer-developed methods at the cost of higher computational intensity (Kao, Stevens, et al. 2009; Kircher, Stenzel, et al. 2009; Quinlan, Stewart, et al. 2008; Wu, Irizarry, et al. 2010).

The different NGS platforms introduce different biases depending on the strengths and weaknesses of the technology used. For example, the 454 has increased error rates in homopolymer reads due to the wide variety in the observed fluorescence intensity for a homopolymer of a specific length. For Illumina data, the rate of error increases toward the end of the reads as the synthesis process becomes desynchronized between different copies of the DNA template in the clusters. The SOLiD technology suffers from errors due to biases in fluorescence intensities that appear in later cycles. All of these biases must be accounted for in image processing and subsequent analysis steps to produce a reliable dataset.

Bioinformatic Analyses

Multiple steps of bioinformatic analyses are required to transform the base call data obtained from the next-generation sequencers into variant lists that can be used in medical genetic studies (Figure 1.3). The first step is to align the sequence data to a known reference sequence

to determine the most likely location in the sequenced genome for each of the individual reads (Flicek and Birney 2009; Li and Homer 2010). If a reference genome is not available, in some cases alignment can be performed using the assembled genome of a closely related species. In some instances sequence data can also be assembled de novo (i.e., without using a reference). De novo assembly is more challenging and requires more computational resources. However, the increase in sequence read length as well as advances in algorithm development have made de novo assembly possible even for large genomes, and over 20 different de novo assemblers are available (Lin, Li, et al.; Zhang, Chen, et al. 2011).

Each NGS platform produces a per-base quality score by using noise estimates from image analysis. After assembly or alignment, quality scores are usually recalibrated to better reflect the true base-calling error rates. After initial alignment, realignment is often

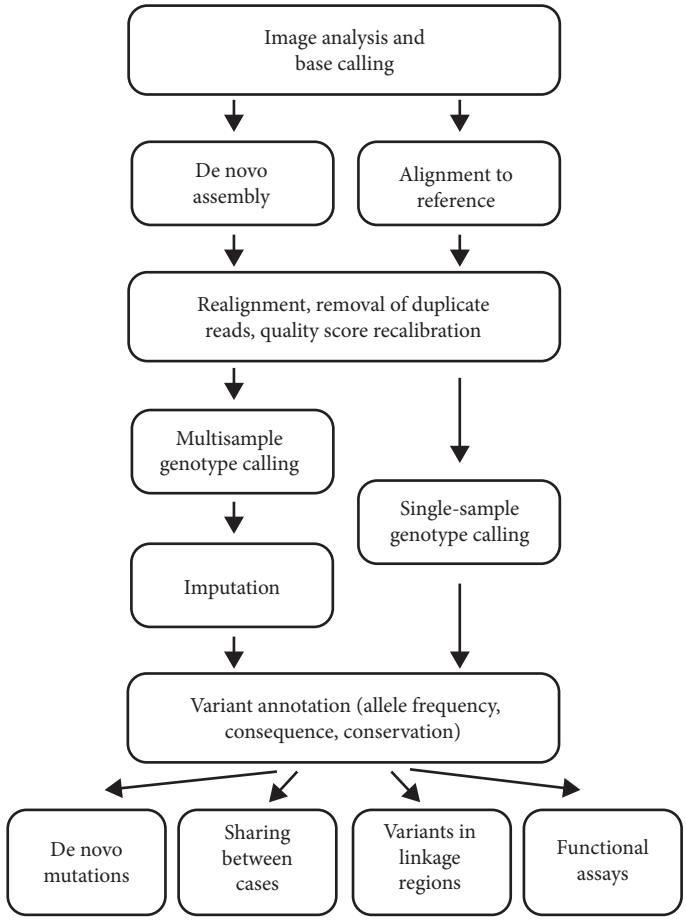

FIGURE 1.3: Basic workflow of bioinformatic analyses applied to the DNA sequence data obtained from the sequencing instrument. Raw DNA sequence reads must be assembled or aligned to a reference to determine their location in the genome before sites that differ from the reference (or between samples) can be identified. These variant sites are then annotated with information that will be useful in subsequent analysis, such as allele frequencies of the variants in control databases, predicted consequence on protein function, conservation of the site between species, or other information that could help to identify disease-associated variants. The analytical steps needed to identify the disease-associated variant depend on the study design and the genetic architecture of the trait. In some cases, variants that are not inherited from the parents (i.e., de novo in the affected patient) could be causative. In other cases, sharing of variants between multiple related or unrelated cases can help to identify the causative variants. Usually replication in large datasets as well as functional proof of the effect of the variant are needed to lend further support to the role of the identified variant in the trait of interest.

performed around known insertion/deletion polymorphisms (indels)—such as those identified in the 1,000 Genomes project (Abecasis, Auton, et al. 2012)—to decrease mapping errors and improve variant call accuracy.

Following alignment, a genotyping step is performed. This can be done either for one sample at a time or, as is more common, across multiple samples. Genotyping is split into two steps, SNP or variant calling followed by genotype calling. In the first phase, the aim is to determine in which positions there is at least one nonreference allele. Genotype calling is then performed only for sites where nonreference alleles are observed to determine the genotype for each sample at the site (Nielsen, Paul, et al. 2011).

Early SNP calling methods were simply based on comparing the number of reads with an alternative allele to those with the reference in a set of high-confidence bases and call SNPs based on fixed cutoffs. However, simple counting methods are not suitable for low-coverage data, as fixed cutoffs result in undercalling of heterozygous genotypes and simple filtering on quality score leads to loss of information regarding individual read qualities. Therefore current SNP callers use probabilistic methods (DePristo, Banks, et al. 2011; Le and Durbin 2010; Li, Handsaker, et al. 2009; Li, Yu, et al. 2009), which lead to genotype calls of higher accuracy. In addition, they provide a measure of the statistical uncertainty (in the form of a posterior probability) for each genotype and can incorporate information regarding allele frequencies and linkage disequilibrium (LD) patterns. For single-sample calls, priors may be chosen to assign equal probability to all genotypes, or information from dbSNP or other collections of known variant sites can be used to determine priors. For multiple-sample calls, the priors can be derived from jointly analyzing multiple individuals by using allele frequencies or genotype frequencies. Once allele frequencies are estimated, genotype probabilities can be calculated using the Hardy-Weinberg equilibrium assumption, and uncertainty in estimates of the allele frequency themselves can be incorporated by assigning a prior to the allele frequency itself. Imputation-based methods can also include information of the pattern of LD at nearby sites to improve genotype calls, which leads to a significant improvement in genotype-calling accuracy for common and moderate frequency SNPs (Nielsen, Paul, et al. 2011).

Alignments are most commonly stored in BAM files, which are binary versions of Sequence Alignment/Map (SAM) files (Li, Handsaker, et al. 2009). These files can efficiently store information from the large number of reads produced in NGS runs, and only the parts of the alignment which are of interest can be accessed without the need for reading in the whole alignment file for analysis. The called variants, such as the single nucleotide variants (SNVs) and indels are commonly stored in variant call format (VCF) files (Danecek, Auton, et al. 2011). In addition to the genotypes, VCF files contain information on call quality, read depth, and other necessary quality parameters of the variants. The VCF file also includes a large header containing metainformation about the analytical steps that were taken during the genotype calling as well as information about the fields that were added during annotation of the variants. VCF files can be compressed and are indexable, allowing for quick analysis of the variants, such as retrieval of variants from regions of interest.

The final step of data generation usually involves annotation of variants. The type of annotation depends on the needs of downstream analyses. Commonly added annotation includes the frequency of the variant in control databases. Another useful annotation is the predicted consequence of the variant on protein structure. Predicting such consequences is often problematic, although several methods such as PolyPhen (Adzhubei, Schmidt, et al. 2010) and SIFT (Ng and Henikoff 2003) are available. The annotation information can then be used in downstream analysis to aid in the identification of disease-causing variants. Obviously any errors in the annotations can have severe consequences in downstream analyses if variants are erroneously attributed to be conferring loss-of-function (LoF) effects or vice versa if a true LoF variant is not annotated as such. Annotation of WGS data is even more problematic than that of WES data, as very little is known about the functionality of noncoding variants.

Identification of Disease-Associated Variants

NGS and subsequent data processing steps produce a list of loci where the sequenced sample differs from a reference genome. The 1,000 Genomes Project reported 36.7 million autosomal SNPs and 1.38 million autosomal indels

in 1,092 low-coverage WGS samples from 14 populations. The average autosomal number of variant SNP sites per individual was around 3.6 million. WES data consisting of the autosomal GENCODE regions contained almost 500,000 SNPs and 1,800 indels in the same amount of samples. Individual exomes contained on average 24,000 variant SNP sites and 440 indels (Abecasis, Auton, et al. 2012). The large number of variants identified in every sample included in an NGS study presents a challenge for gene identification, and various analytical approaches must be employed to identify which individual variants are associated with a phenotype.

For most published studies to date, the assumption has been that disease-associated variants are highly penetrant and not found in dbSNP or other control datasets. This reduces the number of possible disease-causing variants to 1% to 2% of the original list. Ideally, if there are several cases sharing the same disease mutation, only a handful or even one variant will remain after filtering on control frequency and sharing between all samples. However, often the reality is that after filtering, no variants remain at all. Alternatively, filtering will not reduce the candidate variant list sufficiently, or the remaining variants will not overlap between cases. Many published studies have used the 1,000 Genomes Project, dbSNP, and NHLBI GO Exome Sequencing Project as controls. These datasets are useful because they are large; the 1,000 Genomes Project, particularly, includes individuals from a large number of populations. On the other hand, no phenotypic data are supplied for the 1,000 Genomes samples, and variant annotation in dbSNP is poor on phenotypic information. Therefore it is possible that individuals affected with the disorder being studied are included in these reference datasets. Also, particularly for recessive disorders, it is possible that carriers of disease variants are present in the general population. A specific problem with dbSNP is that it contains poorly validated variants. However, using a filter for variants with low frequency—such as 1% in the general population for recessive disorders and 0.1% for dominant disorders—could decrease the risk of missing true variants owing to disease allele carriers in the control data but still remain powerful (Bamshad, Ng, et al. 2011). As in GWAS studies, the controls should be from the same population as the cases to minimize the risk of false-positive variants due to population stratification.

It is tempting to assume that any LoF variant identified in an individual would be a strong candidate for being associated with the disorder. However, studies in healthy reference populations have shown that each person carries, on average, 100 LoF variants. Further, each person has on average 20 genes with two deleterious variants, resulting in complete inactivation of these genes (MacArthur, Balasubramanian, et al. 2012).

Large population-based studies have shown that over half of the variants identified in WGS or WES of large population samples are novel (Abecasis, Auton, et al. 2012; Tennessen, Bigham, et al. 2012); that is, they are not found in reference databases. Each individual carries hundreds of private or very rare variants. Again, assuming a correlation between the lack of a variant in control databases and association with a disease is not necessarily correct. The majority of protein coding variation is evolutionarily recent, rare, and enriched for deleterious alleles, so that analysis of WES in itself enriches for this type of variation. Extra care is needed to link this type of variation to phenotypes (Tennessen, Bigham, et al. 2012). Because control databases include more and more individuals, the probability of seeing multiple copies of very rare variants becomes higher and the risk of identifying very rare benign variants decreases.

Study Designs

The choice of study design is guided by the expected frequency and effect size of the underlying variant and the nature of the disease (prevalence, age of onset, etc.). In the case of monogenic traits, where individual variants have a very high impact on the trait, relatively small sample sizes can be sufficient to demonstrate disease causality of a variant. However, because sequencing studies identify a very large number of potential variants, the number of tests will inevitably be large. Thus knowing which variants/mutations is/are disease causing is not always trivial, even in the case of monogenic traits. For monogenic traits the generally accepted criteria developed for positional cloning studies provide a good reference base. In positional cloning studies the chromosomal location was typically first pinpointed by linkage, applying generally agreed significance

thresholds. If a variant in the linked region was not seen in a control population, the same or different variants in the same gene had to be replicated in several pedigrees with the same phenotype. Further, at least some functional data had to be presented to convince the field and the reviewers that this variation/mutation was associated with the phenotype. Similar rigor should be applied in WES-based variant identification.

Family-Based Studies

Family studies are the default in monogenic traits but have been expanded to more complex traits as well. The hypothesis is that an excess of disease susceptibility variants are clustered and more frequent in families with a specific disease than in the control population. Only a few family members need to be fully sequenced, whereas the remaining relatives can be more sparsely genotyped and the full genome variation imputed. The segregation of a disease-associated haplotype can then be followed in the full pedigree (Figure 1.4). Yet the optimal statistical family-based analysis in complex traits is not fully worked out. So far we are lacking publications that would provide a good understanding of the power and limitations of this approach. Unpublished work suggests that, with this strategy, one cannot hope to capture low-hanging fruits. It is likely too that family-based analyses will need large sample sizes to achieve statistically robust results.

De Novo Mutations

Spontaneous mutations that arise in parental germ cells are frequent causes of some diseases (Figure 1.5). These mutations are not observed in parents, only in the offspring. A classic example is achondroplasia (Bellus, Hefferon, et al. 1995; Shiang Thompson, et al. 1994) and, in CNS disorders, the Dravet syndrome (also known as severe infantile epileptic encephalopathy or SMEI) (Claes, Del-Favero, et al. 2001). Identification of de novo mutations (DNMs) using WES is especially advantageous. When both parents and the proband are sequenced at high coverage, identifying inherited variants is relatively easy, leaving a short list of DNMs. Yet because, on average, each individual carries about 0.8 to 1.3 DNMs in his or her exome, the causality of the DNM still needs verification. To be convincing, deleterious variant in the same gene must be identified in several individuals.

Case Control Studies

In case control studies, the sequences of sporadic cases and healthy or population controls are compared. The case control setting is the classic study design in complex traits. Currently, this study design aims to identify rare or low-frequency variants contributing to a complex trait. When WGS or WES becomes more cost-effective than chip genotyping, sequencing might be used also to identify common variants. As the typical effect sizes of variants associated with complex traits ranges between 1.1 and 1.5, typical sample sizes in GWASs range between a few thousand to tens of thousands of samples. In searching for low frequency and rare variants using sequencing, we can foresee a need for sample sizes that are even bigger than in GWASs. Because sequencing is still quite costly if applied in large sample sets, new, more focused low-cost genotyping chips are being developed. These chips (e.g., the exome chip) are based on low-frequency-variant catalogues developed in large sequencing studies, such as the 1,000 Genome Project. This makes it realistic to genotype large enough samples to enable statistically robust low-frequency association studies.

Population Isolates

It is hypothesized that population isolates provide a middle ground between family and case control studies. This is seen as an extension of the "megapedigree" concept. Because of bottleneck effects, genetic drift, and population expansion, some rare alleles are enriched in population isolates (Figure 1.6). The hypothesis is that some of these alleles, which are extremely rare (population frequency < 0.1%, as seen, for example, in most European populations), have been enriched to frequencies between 1% and 5% in a population isolate such as Finland. Further, selection has not had time to act in recently founded isolates, enabling a higher population frequency of harmful variants. Some of the enriched variants could possibly contribute to common diseases. Even though they could be neutral in an environment where the founder population was established more than thousand years ago, they could contribute to diseases in populations sharing the modern lifestyle and environment. An enrichment of low-frequency alleles in the study population should boost the power significantly compared with more mixed populations. Therefore the expectation is that smaller discovery sets will be needed to achieve significant

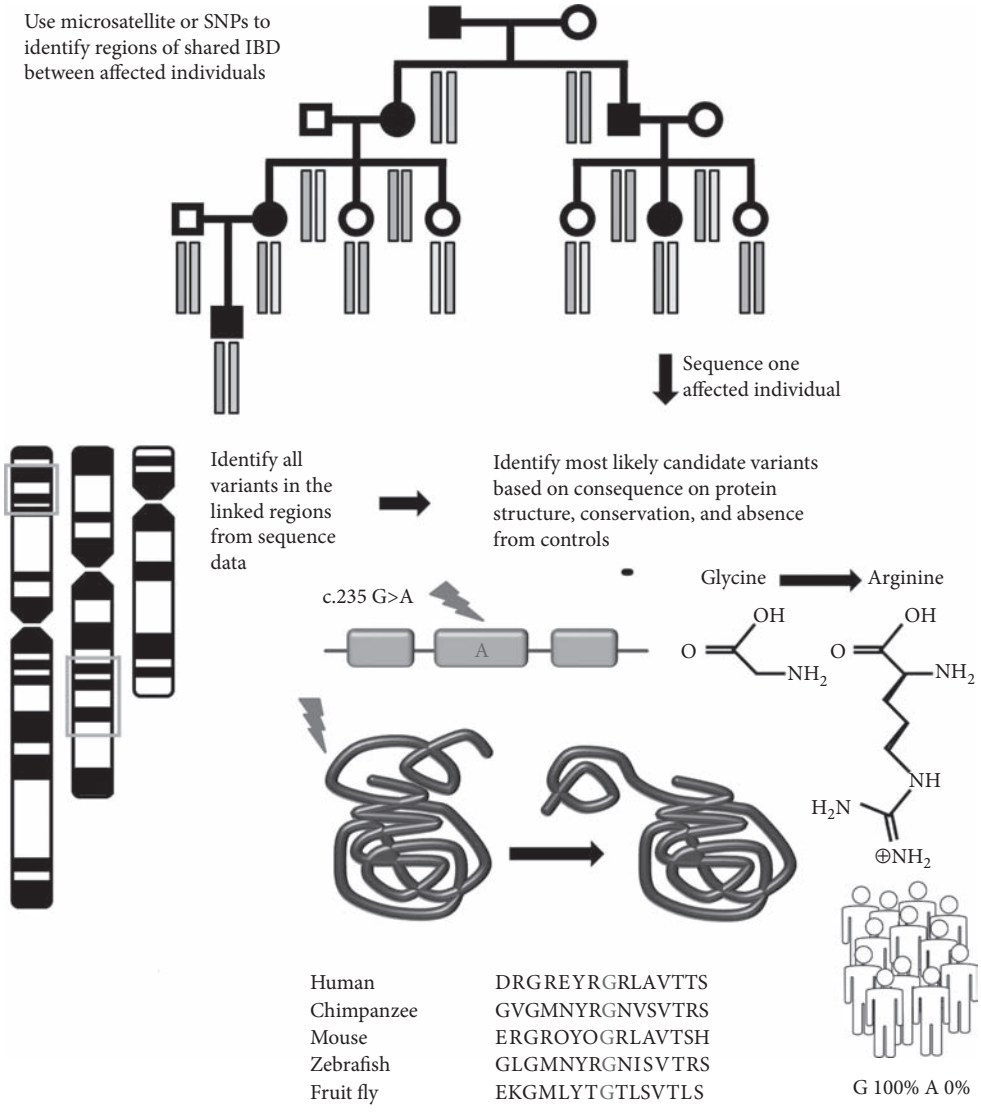

FIGURE 1.4: When large families with multiple affected individuals are available, sequencing is required for only a small subset of affected individuals. Microsatellites or SNPs can be used to identify regions shared identical-by-descent (IBD) by all cases; thus sequencing of one index case is enough to survey the full variation of these regions. Candidate variants are identified based on predefined criteria, such as effects on protein function, absence or low frequency of the variant in control databases, or evolutionary conservation. If a large number of meioses separates the individuals in the family, a small number of candidate regions and thus a small number of candidate variants remain after the analysis. Sanger sequencing may be needed to verify the cosegregation of the variant in the pedigree. Replication of the finding in other pedigrees is usually needed to separate benign but extremely rare variants from true disease-associated mutations.

association in population isolates. The success of this strategy has best been demonstrated in the Icelandic population (Holm, Gudbjartsson, et al. 2011; Jonsson, Atwal, et al. 2012; Sulem, Gudbjartsson, et al. 2011). Also, by sequencing a small subset of the study population and using SNP genotyping in the large majority, efficient imputation of whole-genome sequences can be enabled. Although imputation-based studies are possible in any population, a smaller number of individuals need to be sequenced to capture the majority of all genetic variation in isolated

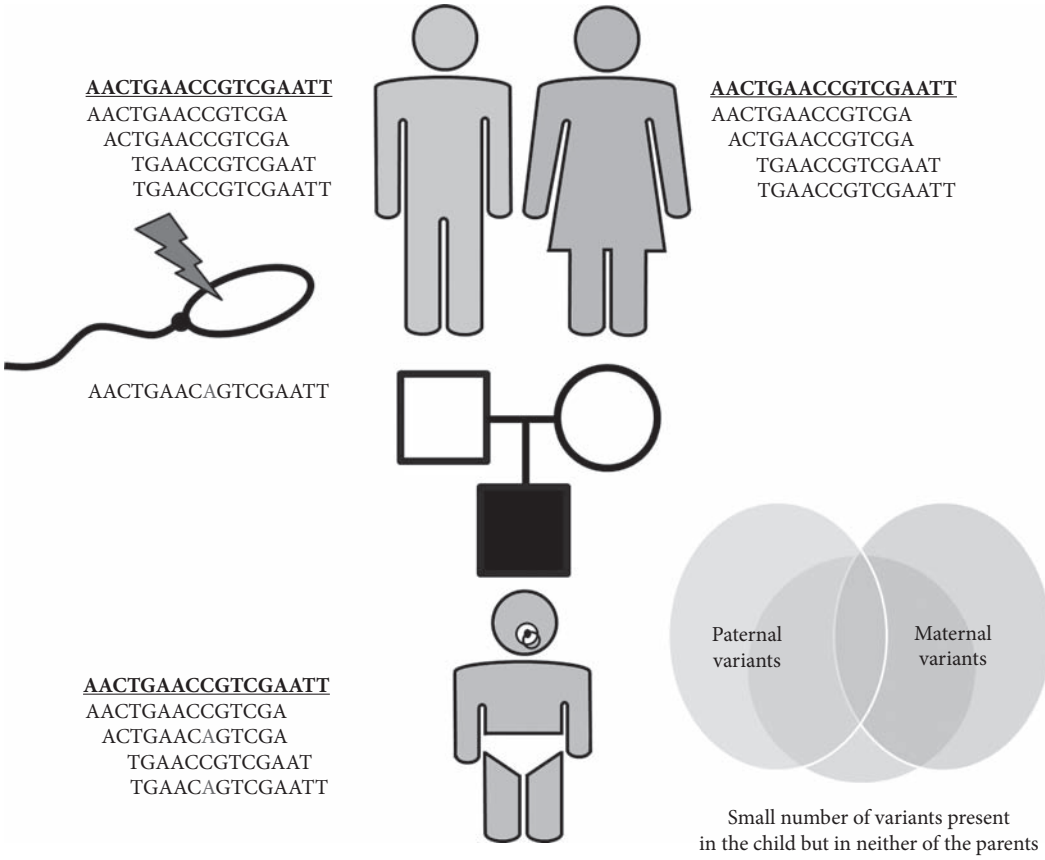

FIGURE 1.5: Disorders such as autism spectrum disorders, which are subject to strong negative selection in the population, have been shown in some cases to be caused by de novo mutations (i.e., mutations present in the affected individual but in neither of the parents). These can arise in either the paternal or maternal gamete or during early embryogenesis. De novo mutations are relatively easy to identify if both parents and the affected child are sequenced at high coverage. Only a handful of possible candidate variants remain if the data are of high quality and appropriate filtering is used in the analysis. Proving causality of individual variants can be hard, particularly if the disorder has a large mutational target, as a large number of families are needed to identify another family with a de novo mutation in the same gene.

populations, as the number of founder chromosomes is lower than in admixed populations.

Current Review of Results

So far, over 100 mutations in monogenic disorders have been identified by NGS, mainly by WES (Rabbani, Mahdieh, et al. 2012). There is an obvious publication bias toward successful studies, so the success rate of gene identification by NGS still remains unclear and will also be strongly dependent on the genetic architecture and availability of samples. One estimate suggests that WES identified the major disease gene in at least 50% of projects focused on rare but clinically well-defined monogenic diseases (Gilissen, Hoischen, et al. 2011). Experience has shown that for most disorders this is an overly positive prediction; realistically, much smaller yields are often to be expected.

Large-scale resequencing can be applied to several different study designs and can identify several different types of risk variants, as summarized above. Several of these approaches have been used to identify genes for intellectual disability (ID), and are described in more detail in the following paragraphs to provide an overview of the different analytical approaches that can be used depending on the expected genetic architecture of the trait being investigated.

A large number of monogenic traits with neurological and neurodevelopmental symptoms have been subjected to WGS and WES.

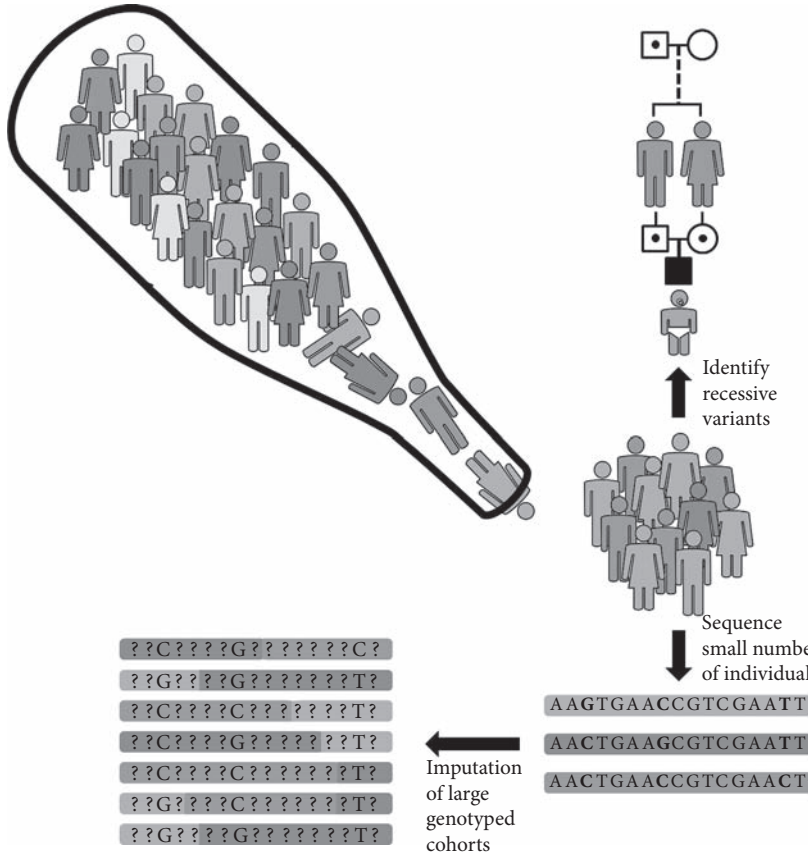

FIGURE 1.6: Population isolates can be powerful in the identification of disease-associated variants. The founding bottleneck has reduced the genetic diversity in the population and drift can enrich disease-associated variants. This is particularly true for recessive disorders, as negative selection is not acting on the asymptomatic disease carriers. Because of an enrichment of the disease allele in the population due to the bottleneck, it is more likely that two individuals are distantly related and carry the same recessive disease mutation, resulting in the risk of having affected offspring. The reduced genetic diversity also makes imputation studies particularly feasible. The founding bottleneck has reduced the number of founding chromosomes, so only a small number of individuals need to be sequenced to capture them. When all founding chromosomes have been sequenced, imputation is efficient and highly accurate. Large numbers of individuals can be genotyped using cheap SNP chips and then imputed using the reference panel generated from the sequenced individuals, enabling large case-control association studies.

This list is constantly growing. Thus we do not aim to provide a comprehensive list of these diseases but have rather selected a few examples of more complex and challenging phenotypes.

Intellectual Disability

ID often has a genetic basis, and positional cloning has shown that at least a subset of ID is caused by monogenic, fully penetrant mutations. ID can present together with other clinical symptoms such as metabolic or structural abnormalities. These syndromic forms of ID make it possible to identify patients with similar phenotypes, often revealing an underlying shared genetic etiology. However, ID can also present as the only observable phenotype, referred to as nonsyndromic ID (NSID). In these cases, it is impossible a priori to identify cases with a shared genetic etiology. Substantial genetic heterogeneity underlies NSID, since numerous genes have been identified. In fact, most identified genes account for only a very small fraction of cases, and over 100 genes have already been implicated in ID (Ropers 2010). Over 90 of these are located on the X chromosome. It is probable that this bias in

identification is largely due to ease of gene identification in large X-linked pedigrees, although unbiased exome and WGS studies will give a more unbiased estimate of the proportion of X-linked versus autosomal ID genes. It has been estimated that genes on the X chromosome account for 10% to 20% of male X-linked ID (Ropers and Hamel 2005). In addition to the high level of genetic heterogeneity, ID is known to be caused by many types of genetic abnormalities ranging from duplication or deletion of large chromosomal segments to small indels and SNVs. Further, many different inheritance patterns have been observed, ranging from autosomal dominant, autosomal recessive, and X-linked to DNMs.

Family Studies and Population Isolates

One of the most successful approaches to the identification of ID genes has been to use large families with a X-linked pattern of inheritance. Today there is a large collection of these families, which have been thoroughly studied (http://goldstudy.cimr.cam.ac.uk, http://www.euromrx.com). Traditionally, microsatellite markers have been used to identify regions of maximum linkage in these families, and Sanger sequencing of all genes in the linkage regions has resulted in the identification of numerous ID genes (Ropers and Hamel 2005). However, in many families, the linkage intervals have been too large to allow for the sequencing of all genes using traditional Sanger sequencing.

The first large-scale resequencing study of ID genes was published in 2009, where all known exons of genes on the X chromosome of 208 families with X-linked ID were sequenced (Tarpey, Smith, et al. 2009). Nine genes were deemed to be associated with ID. However, the authors also discuss extensively the difficulty of identifying true disease-associated variants. More than half of the gene truncating variants did not segregate with the disorder in the families or were found in controls; the authors therefore caution against concluding that truncating mutations in genes are sufficient on their own to be considered causative. Particularly, 8 of the 19 genes with truncating variants that did not segregate with the phenotype or were found in controls have only a single exon. This suggests that some of these genes might be retrotransposed copies without important function that therefore tolerate LoF mutations. The authors also note that although they screened most of the protein coding exons on the X chromosome, the likely genetic basis for ID was established in only 25% of families. Variants could be missed owing to low coverage, unannotated genes, the presence of copy number changes large enough to go undetected by the sequence data, nonexonic variants, and the presence of autosomal variants despite the appearance of X-linked inheritance in the families. Also, only LoF variants were considered, although it is highly likely that in some families the causative mutation is a missense or even noncoding variant.

To identify autosomal ID genes, studies have been performed in consanguineous families or founder populations. Traditionally, microsatellite markers or SNPs have been used to identify regions of homozygosity in affected relatives, and genes in these regions have been resequenced to identify disease-causing mutations. Today, WES allows for a shortcut directly to the causative variants. Still, the identification of homozygous regions either from SNP data or from the exome data themselves is useful to limit the amount of variation that could be considered to be pathogenic. A novel autosomal ID gene, *TECR*, was identified by linkage mapping followed by WES in a large consanguineous family with 5 of 13 children affected with ID (Caliskan, Chong, et al. 2011). Linkage analysis first identified a gene-rich region on 19p13 that cosegregated with the phenotype in the family; an SNP array was used to narrow the region to a 2-Mb homozygous segment with over 30 genes. WES was performed for both parents and filtering performed for novel disruptive variants heterozygous in the parents. Only a single variant fulfilling these criteria was identified, and follow-up genotyping in the rest of the family showed cosegregation with ID in the family. Further, homozygosity for the variant was not observed in over 1,000 individuals from the same population. In another study, sequencing was performed in 136 consanguineous families from Iran with autosomal recessive ID. Mutations were identified in 23 previously known ID genes as well as 50 novel candidate genes (Najmabadi, Hu, et al. 2011). In this study, the targeted genes were not the whole exome but genes from previously identified regions of homozygosity, thus significantly reducing the number of genes sequenced.

Another example of how WES has been used to identify autosomal recessive genes is in

the population isolate from the Ashkenazi Jews. Two individuals (the affected offspring and the mother) from a family with Joubert syndrome, an autosomal recessive ID syndrome, were exome sequenced (Edvardson, Shaag, et al. 2010). The search was concentrated on a linkage region that had previously been identified by homozygosity analysis of a larger pedigree from the same population. Seven variants homozygous in the child and heterozygous in the parent were identified, of which two remained after filtering on dbSNP; only one of the remaining variants (in *TMEM216*) was nonsynonymous. The added benefit of the WES data was to show that no other disruptive mutations existed in the previously identified region of homozygosity. The mutations segregated with the phenotype in the larger pedigree.

Analysis of non-consanguineous families must take into account that the mutation is likely to be compound heterozygous (i.e., a different mutation is inherited from each parent). This can in some cases prove to be more challenging, but as long as the inheritance model is known to be recessive, gene identification can often be successful. In a family with three affected offspring with hyperphosphatasia mental retardation syndrome (Marby syndrome), WES was performed for all three affected offspring (Krawitz, Schweiger, et al. 2010). First, all common variants and variants not found in all affected persons were excluded, leaving 14 candidate genes. A Hidden Markov Model was used to infer all loci where the offspring shared both alleles identical by descent (IBD = 2), reducing the number of candidate genes to two, *PIGV* and *SLC9A1*, located in a 13-Mb homozygous block. Mutation screening of *PIGV* revealed homozygous and compound heterozygous rare variants in other families with the same phenotype, identifying *PIGV* as the causative gene.

Exome Sequencing in Unrelated Cases Sharing Syndromic Forms of ID

WES has successfully been applied to several syndromic forms of ID. Here, patients with similar phenotypes are sequenced; after filtering for variants shared by all or the majority of affected individuals but which are not found in control datasets such as the 1,000 Genomes Project or dbSNP, only a small handful of variants remain. Schizel-Giedion syndrome is characterized by ID, distinctive facial features, multiple congenital abnormalities, and a high prevalence of tumors. Hoischen and colleagues (Hoischen, van Bon, et al. 2010) performed WES for four of these patients. After filtering known variants (dbSNP and variants observed in other WES projects from the same laboratory), only two genes were identified where all four affected individuals carried a mutation. One of these variants was of low quality, leading to the identification of *SETPB1* as the causative gene. Targeted resequencing of this gene in nine additional cases identified a variant in *SETPB1* in eight of these patients.

Although in general the identification of genes for syndromic forms of ID is easier as larger patient groups can be collected for study, this can sometimes prove challenging, since genetic heterogeneity can lead to false findings particularly for disorders with a dominant inheritance pattern. A WES study of Kabuki syndrome, characterized by ID and distinctive facial features (Ng, Bigham, et al. 2010), identified only one gene, *MUC16*, with novel variants in all 10 unrelated patients included in the study. However, this was deemed as an unlikely candidate because of its function and expression pattern. In addition, *MUC16* is one of the largest genes of the genome and would be expected to show numerous variants based on random chance. Because the only gene carrying mutations in all affected individuals was an unlikely candidate, the authors then focused on nonsynonymous variants present in most but not all of the cases. A truncating mutation in *MLL2* was identified in 7 of 10 patients. In two of the three remaining patients, a small indel, missed by the WES but identified from CGH arrays, was detected in *MLL2*, strongly implicating this gene in the etiology of Kabuki syndrome. Sanger sequencing of *MLL2* identified mutations in 26 of 43 additional patients and among the subset of samples (n = 12) with both parental DNAs available. All mutations were de novo.

De Novo Variants in ID

The identification of mutations underlying nonsyndromic ID without any family history has so far been challenging. However, with access to DNA from both parents and the affected child, WES enables relatively straightforward identification of DNMs present in children but not parents. These variants are in general more deleterious than variants segregating in

the population because they have not been subjected to evolutionary selection, making them excellent candidates for causing sporadic severe disorders (Eyre-Walker and Keightley 2007).

The first study of DNM rates in humans that was based on WGS suggested that on average 74 germline SNVs occur de novo in one individual's genome but also that there is huge variability in the DNM rate between trios. However, these conclusions were drawn from only two trios and should be interpreted with caution (Conrad, Keebler, et al. 2011). Later studies have tackled this issue only using WES, and consensus estimates in larger datasets suggest that, on average, 0.82 to 1.3 DNMs are observed per exome (Neale, Kou, et al. 2012; O'Roak, Vives, et al. 2012). However, these estimates have been derived from trios where the offspring are affected with autism spectrum disorders (ASDs), and although the consensus seems to be that the rate of DNMs is no higher in affected cases than in controls, this caveat should be kept in mind in interpreting these results. Because of the slightly different GC content of the exome compared with the whole genome, the DNM rate for exomes is expected to be higher than for whole genomes. One estimate placed the genome-wide DNM rate at 1.2×10^{-8} per base per generation, whereas the exome mutation rate has been estimated at 1.5×10^{-8} (Neale, Kou, et al. 2012). Although different studies estimate the exomic DNM rate to be different, the consensus ranges broadly in the same magnitude, around 1.2 to 2.2×10^{-8} per base per generation. Larger studies of population trios will undoubtedly narrow the confidence interval of these estimates. One very consistent finding from multiple studies is that paternal de novo SNVs are much more common than maternal ones, and there is a striking correlation between de novo SNV rate and paternal age (Neale, Kou, et al. 2012; O'Roak, Vives, et al. 2012), but there seems to be a large variation between trios (Conrad, Keebler, et al. 2011). Despite the overall consensus of the DNM rate of SNVs, the calling of small indels and CNVs still needs to be significantly improved to shed light on the rates of DNM in these classes of genetic variation.

Previous studies have implied that de novo CNVs are causal in about 15% of cases of ID (Cook and Scherer 2008; de Vries, Pfundt, et al. 2005). DNMs should be particularly common in disorders that have a relatively high prevalence in the population despite a strong reproductive bias due the fact that the early onset of the disorder will preclude transmission of the disease to subsequent generations. It is likely that DNMs will account for both very rare as well as more common phenotypes depending on the size of the mutational target. Diseases caused by DNMs in just one gene will be rare, but if the mutational target is large enough, DNMs can cause even common disorders, such as ID, ASD, and schizophrenia.

So far, only a small proof-of-concept WES study has been published assessing the role of DNMs in NSID, although bigger studies are ongoing (Vissers, de Ligt, et al. 2010). WES was performed in 10 trios with no family history of ID, no clear syndromic features, no evidence of Fragile X syndrome, and no de novo CNVs detected using CGH to enrich for families with de novo SNVs. The study identified six likely causative variants in six different genes, of which two had previously been implicated in ID. Further work is still required to validate the causative roles of these variants, but the study showed a proof of principle that WES is an appropriate tool to screen for DNMs. The same problem of establishing causality that affects inherited variants also applies to DNMs. Variants in the same gene in multiple cases need to be identified before any claims can be made about causality; particularly for disorders with large mutational targets, very large samples sizes are likely required to obtain sufficient power for replication. Often biological function is used to assess the significance of findings for DNM analyses, but for disorders with large mutational targets this can enrich for false-positive findings, such as assuming that all DNMs in brain-expressed genes are likely to be causative of ID. So far, mutational type seems to be the best predictor, with LoF variants the most likely to be causative, particularly if several LoF DNMs are observed in the same gene (Sanders, Murtha, et al. 2012).

Autism Spectrum Disorders and Schizophrenia

ASD and schizophrenia are among the most heritable neuropsychiatric disorders, but specific susceptibility genes remain elusive. Several monogenic forms of ASDs are known (Abrahams and Geschwind 2008), whereas no monogenic forms of schizophrenia have been reported. The role of CNVs as susceptibility

factors for ASDs and schizophrenia is well established. A substantial number of CNVs are de novo (Gilman, Iossifov, et al. 2011; Xu, Roos, et al. 2008). This has prompted several WES studies evaluating the role of DNMs, of which the first generation of studies has recently been published.

Based on four studies (Iossifov, Ronemus, et al. 2012; Neale, Kou, et al. 2012; O'Roak, Vives, et al. 2012; Sanders, Murtha, et al. 2012) encompassing over 900 trios or quads (trios with one unaffected sibling sequenced), the overall rate of DNMs in individuals with ASDs is no higher than that in controls. As the number of sequenced trios keeps increasing, the probability of hitting the same genes in several studies also increases. Simulation experiments taking into account the distribution of gene sizes and GC content across the genome suggest that focus should be on the severe LoF variants, since two or more nonsense and/or splice-site DNMs are highly unlikely to occur in the same gene. This conclusion remains robust to sample size and estimates of locus heterogeneity (Sanders, Murtha, et al. 2012), whereas if nonsynonymous sites are also included for sample sizes of 1,000 trios or more, at least four hits in one gene are needed to establish causality. These estimates vary strongly depending on the genetic model used and are not nearly as stable as the estimates for LoF variants. So far a total of five genes, *CHD8*, *DYRK1A*, *KATNAL2*, *SCN2A* and *POGZ*, have two LoF de novo hits in the 900 published ASD trios. Further, these studies have reported that the proteins encoded by genes with DNMs are more closely linked by protein-protein interaction networks than similarly sized sets of random genes. Especially intriguing is the result that genes with DNMs in a study of over 300 ASD trios found that many of the genes are linked with *FMRP*, a gene very robustly linked to ASDs and involved in synaptic plasticity.

To date, the studies of DNMs in schizophrenia are not as extensive as the data for ASDs, although several large-scale studies are under way. One study of 14 trios identified 15 DNMs (Girard, Gauthier, et al. 2011). These included four nonsense variants and eleven missense variants. Unsurprisingly for such a limited dataset, no gene was hit twice, and none of the genes had previously been implicated in schizophrenia etiology. The DNM rate was reported to be significantly higher than any of the DNM rates reported in population studies, which led the authors to conclude that there is a DNM burden in schizophrenia. However, the conclusions were drawn on a very limited number of trios; larger replication studies are needed to validate this observation. The second study involved 53 schizophrenia trios (Xu, Roos, et al. 2011) and identified a total of 40 DNMs in as many genes in 27 individuals. The authors also concluded that DNMs play an important role in schizophrenia and estimated that the mutational target is large, which would explain the high incidence of the disorder worldwide. The third study was larger and included WESs from 231 trios with schizophrenia (Xu, Ionita-Laza, et al. 2012). The authors reported an excess of both nonsynonymous and LoF variants in cases, but the control group consisted of only 34 trios. One nonsynonymous and one LoF DNMs was identified in four genes (*LAMA2*, *DPYD*, *TRRAP*, and *VPS39*). No gene with two LoF variants was identified. Interestingly five genes (*DGCR2*, *TOP3B*, *CIT*, *STAG1*, and *SMAP2*) were identified where a missense DNM and a de novo CNV were present in the same individual.

It seems possible that DNMs—in the form of CNVs, small indels, and SNVs—play a role in ASDs and schizophrenia. It seems likely that the risk is not conferred by an overall increase in mutation rate but by the severe interruption of genes involved in brain development and function. The next question that needs to be addressed is what proportion of these disorders can be explained by these highly penetrant variants. Current studies have found likely DNMs in only a small fraction of the studied individuals, but they are likely to suffer severely from lack of power and false negatives, since most studies have assumed missense variants to be benign. However, several ID genes with missense variants in conserved positions have been identified, so it seems likely that this will also be the case for ASDs and schizophrenia as well as other complex disorders. Further, it seems probable that many variants will be noncoding regulatory variants, which are beyond the scope of WES studies. More data are also needed to determine whether these variants are truly monogenic risk variants (i.e., fully penetrant and sufficient to develop the disease). Interestingly, simulations have been reported where models assuming a large number (such as 100) of rare, fully penetrant monogenic

genes are inconsistent with the observed data, whereas models where functional mutations in hundreds of genes that would increase the risk of the disease by 10- or 20-fold fit the observed data much better (Neale, Kou, et al. 2012). This could suggest that although DNMs play a role in ASDs (and possibly other complex diseases also), they are not necessarily sufficient for disease. There is also evidence suggesting that common variants confer susceptibility to ASDs (Klei, Sanders, et al. 2012), although all GWASs so far have failed to identify genome-wide hits, most likely owing to small sample sizes (Ma, Salyakina, et al. 2009; Wang, Zhang, et al. 2009; Weiss, Arking, et al. 2009). In schizophrenia, an unpublished large case-control GWAS consortium has identified dozens of robustly associated loci. This suggests that several different study designs are needed to identify all possible risk factors for these diseases. A recent WES study assessed the role of rare variation in 166 individuals with schizophrenia and subsequently genotyped 2,617 individuals with schizophrenia and 1,800 controls. The results suggested that schizophrenia susceptibility is unlikely to be significantly affected by low-frequency variants that are just outside the range of detectability using GWAS (Need, McEvoy, et al. 2012). The study did, however, detect several variants that were identified in a small number of cases and no controls. These variants could possibly play a role in disease etiology.

It will also be interesting to see if DNMs in the same genes cause disease all across the neuropsychiatric spectrum. It is generally accepted, that CNVs in the same genes can cause susceptibility for ASDs, schizophrenia, and ID (Mefford, Batshaw, et al. 2012). However, large sets of tens of thousands of cases have been genotyped on comparative genomic hybridization arrays, making these comparisons statistically powerful. It will take time to accumulate exome data from such large datasets to make comparison possible across disorders. Despite a significant overlap between rare variants, a recent study could not detect significant overlap of common variation between ASDs and schizophrenia (Vorstman, Anney, et al. 2012).

The contribution of DNMs to late-onset disorders, such as Alzheimer's disease (AD), will be harder to evaluate, since the analysis requires DNA from both parents and usually, for late-onsets disorders, the parents are no longer available for study. However, late-onset disorders might not be under such strict selective pressure as early-onset ones, allowing for inherited variants to play a larger role in disease etiology.

Monogenic Epilepsies

Epileptic seizures are a part of many syndromic developmental diseases but can also be the main or only symptom, thus being nonsyndromic. A small percentage of these genetic epilepsy syndromes, known as the rare epilepsy syndromes (RESs), are monogenic. By studying these Mendelian disorders, mainly via parametric linkage analysis and positional candidate gene sequencing in large multiplex families, the main concepts of the genetic architecture of epilepsies we have today were unraveled. Many of the known genes implicated in the development of Mendelian forms of epilepsies encode for subunits of ion channels, although it becomes more and more evident that risk variants are not limited to this class of genes.

Next to these familial RESs, the availability of genome-wide sequencing technologies has finally made it possible to study the interesting group of epileptic encephalopathies (EEs) genetically in a systematic manner. EEs are severe disorders with early onset, often within the first year of life. They present as distinct epilepsy syndromes often in combination with dysfunctions in the brain, such as ID and spasticity. These disorders severely interfere with reproductive fitness, and evolution strongly selects against the transmission of mutations. Most EE patients present as isolated cases owing to heterozygous de novo dominant mutations. The concept that de novo dominant mutations underlie EE was firmly proved by our observation that de novo LoF mutations in the *SCN1A* gene result in Dravet syndrome, the prototypical EE (Claes, Del-Favero, et al. 2001). To date, several distinct EEs are known to be caused by DNMs in genes, like *STXBP1*, *KCNQ2*, and many others (Saitsu, Kato, et al. 2008; Weckhuysen, Mandelstam, et al.). Many studies on sequencing EE patient-parent trios for the identification of novel genes harboring causal DNMs are in progress to gain more insight in the missing heritability of different EEs. On the other hand, the more common genetic epilepsy syndromes are usually considered to be complex genetic traits. Recently two large-scale studies (Klassen, Davis, et al. 2011; Heinzen, Depondt, et al. 2012) reported the

sequencing of over a hundred sporadic idiopathic epilepsy patients. Klassen, et al. focused on ion channel genes, whereas Heinzen and colleagues used WES. The Heinzen group tried to replicate almost 4,000 identified candidate epilepsy-susceptibility variants in 878 cases. Both studies failed to convincingly identify any disease associated variant. Both studies were small, but they suggest a similar picture as in many other complex traits; much larger study samples are needed to shed light on the potential contribution of low-frequency variants to epilepsy. Such studies are in progress, and we expect to have results from them in the next two years.

Alzheimer's Disease

AD is the most common form of dementia in the elderly. It is known that low-frequency and rare variants can contribute to the risk of AD, especially for early-onset forms of the disease (Goate, Chartier-Harlin, et al. 1991; Raux, Guyant-Marechal, et al. 2005; Sherrington, Rogaev, et al. 1995). Less is known about the genetics of late-onset AD, the more common form of the disorder.

A WES study of 14 individuals with early-onset AD revealed nonsense or missense mutations in 5 individuals in *SORL1* (Pottier, Hannequin, et al. 2012). The mutations were identified by using a simple filtering strategy where all variants were filtered against dbSNP and 1,000 genomes, HapMap, and an in-house database of 72 WES samples. After validation of variants in genes where multiple individuals were carrying a missense or LoF variant, *SORL1* was the gene with the largest number of variants. This gene binds *APP*, previously known to confer risk for AD. Analysis of 1,500 controls confirmed that the *SORL1* variants were not present in the control population. One of the sequenced individuals also had an affected mother, who also had the mutation. *SORL1* was sequenced in another 15 index cases, and two more mutations were identified. This study shows that genes can be identified even in sample sets with genetic heterogeneity if several individuals share a mutation in the same gene.

The most informative studies of late-onset AD have been reported in the Icelandic population. Iceland is a population isolate with a well-known genetic history (Helgason, Yngvadottir, et al. 2005). Much of the population has been genotyped using SNP chips, and Iceland has proved to be a treasure chest for GWASs. The extensive genetic information available combined with good genealogical records has also proven to be a useful resource for NGS studies. Some members of this population have undergone WGS, followed by imputation into essentially the entire Icelandic population. The genetic information combined with easily accessible phenotypic data has led to the identification of numerous susceptibility variants, both common and rare, for complex disorders. This approach identified a variant in *TREM2* associated with late-onset AD (Jonsson, Stefansson, et al. 2012). WGS of 2,261 Icelanders was used to identify over 34 million variants, 190,000 of them functional. These variants were imputed into 3,550 patients with AD and 1,236 controls who were over 85 years of age and had no symptoms of AD. When a case control study was performed, only one marker in addition to the already known *APOE* locus, a substitution of histidine for arginine in *TREM2*, reached genome-wide significance. This rare variant, with a population allele frequency of 0.63%, confers significant risk for AD with an odds ratio (OR) of 2.92. The finding was replicated in 2,000 cases from other populations with a combined OR of 2.83. Interestingly, compared with noncarriers, the variant also confers risk for worse cognitive function in individuals between 80 and 100 years of age but no diagnosis of AD.

Another study in the Icelandic population identified a low-frequency variant in the *APP* gene (population frequency < 0.5%) to be protective of AD (Jonsson, Atwal, et al. 2012). Variants in APP have previously been linked to early-onset monogenic forms of AD (Alonso Vilatela, Lopez-Lopez, et al. 2012). The variant was identified from WGS of 1,795 Icelanders and was then chip genotyped in 71,743 individuals and subsequently imputed into 296,496 relatives. These two reports demonstrate the power of imputation in a well-organized cohort in a population isolate. The latter also demonstrates that variants within the same gene can be either predisposing or protective.

Family-based studies have not been quite as successful in AD. A study of an individual with AD from a Turkish consanguineous family with a complex history of neurological and immunological disorders identified a nonsynonymous mutation in *NOTCH3*, which has previously been linked with cerebral autosomal dominant arteriopathy with subcortical infarcts and

leukoencephalopathy (CADASIL) (Guerreiro, Lohmann, et al. 2012). However, the mutation did not cosegregate with the neurological phenotype in the family, leaving the results of the study inconclusive.

FUTURE DIRECTIONS

Currently the true success in NGS studies has been achieved for monogenic diseases. However, the rapid reduction in sequencing costs makes larger studies possible, promising hope also for the identification of variants conferring risk for more complex disorders. It does not seem too implausible to predict that the course of sequencing studies will closely resemble that of GWASs a few years ago. In the beginning, the small GWASs identified risk variants only for disorders with relatively high risks. When more data was produced and pooled, robust associations were also identified for variants with very small effect sizes.

Increase in sample size is only one avenue of increasing power to detect genetic variants associated with traits and disorders. Sequencing technology keeps improving, and current data already allow for relatively robust genotype determination for SNVs. However, better data and genotyping methods are needed to reliably identify other types of variation from NGS data, such as indels and copy number variants. Improvements in data quality include increased read lengths as well as improvement of the sequencing chemistries. Particularly, the "third generation" of sequencing technologies offers great promise for single molecule sequencing. This would not only reduce the amount template but also significantly reduce the problems in variable read depth caused by the capture and amplification steps.

To be able to identify rare disease-associated variants, good-quality datasets with low false-positive and false-negative rates are needed. It seems very likely that data that is of very high quality from a technical viewpoint can soon be achieved. However, at present the bigger challenge is our limited knowledge of the functionality of the genome. Much work is still needed when it comes to the annotation of identified variants. Improved methods of predicting the consequences of variants on protein structure are needed. Increased understanding of the expression patterns (both spatial and temporal) of transcripts can help determine which variants could possibly be involved in the pathogenesis of genetic disorders. Improved predictions of the pathogenicity of missense and splice variants would decrease the need for downstream in vitro assays to determine the true functional consequences of variants. And current knowledge has only scratched the surface of consequence prediction of noncoding variants, although large consortia such as the ENCODE project are starting to shed light on those shady parts of the genome that we currently understand very little.

It is easy to get stuck on the technical aspects and problems of NGS and not see the forest for the trees. Current evidence seems quite convincing that rare variation will play a role in many neurological and neuropsychiatric phenotypes. Until the advent of NGS, this type of variation could be accessed only for very small targeted regions. Now large patient cohorts can be sequenced and it will be possible to assess the role of rare variation in these phenotypes. This also poses a challenge to improve phenotyping, as the subgrouping of patients based on endophenotypes could identify groups who share a genetic etiology, thus improving the probability of identifying disease-associated variants, as has been demonstrated by studies of syndromic ID. For many neuropsychiatric disorders, this is extremely challenging, as there are no biomarkers for diagnosis; that is, the definition of the phenotype relies entirely on observational data.

The analysis of rare variants will also pose challenges for the ever-increasing collaboration among researchers. If mutations are identified in one individual or family and genetic heterogeneity in the disorder is high, as has been seen for ID, large replication cohorts will be needed to identify other patients sharing the same genetic etiology. Good examples have already been set up, such as the DECIPHER online repository containing genotype and phenotype information of research subjects with developmental disorders (Firth, Richards, et al. 2009). Researchers all over the world can access the database and search for other patients with the same genetic variants. This has led to the identification of a number of new ID syndrome genes (Firth, Richards, et al. 2009).

NGS is already used in a clinical setting for gene identification in Mendelian diseases, but there are many unanswered questions (Anderson and Schrijver 2010). As in the use of NGS in a research setting, one of the main challenges remaining is the interpretation of results. Proponents of NGS in a clinical setting argue

that once NGS data have been generated, such data could be a useful resource for the individual all through his or her life. Sequence data could be useful for other purposes besides the identification of disease genes, such as personal pharmacogenomics. The many pros of using NGS in the clinical setting are weighed down by a large number of problems and unsolved questions. How can quality control of such large datasets be guaranteed to the same level as current genetic tests, which usually produce little data that can be visually inspected? How are results to be validated? How are incidental findings handled? What will be the impact on relatives who might not wish to know about possible genetic susceptibility factors for diseases? However, there is no reason to assume that NGS cannot be included as part of clinical testing in cases where this procedure provides added clinical utility compared with targeted tests, and the problems identified now should not be thought of as reasons to forever ban the clinical use of NGS; however, such problems must be solved before the widespread use of NGS in medical care can become a reality.

SUMMARY

Current technology allows for the interrogation of every base pair in an individual's genome. Many successful reports of gene identification using NGS have already been published, mostly for monogenic disorders. For several neurological and neuropsychiatric disorders, such as ID, autism and schizophrenia, the technology has been applied successfully to identify genes. Technological advances provide ever-improving data quality, but analytic approaches must keep up with the technological development to be able to make use of the data and convert the raw sequence to biological understanding. Currently NGS offers a promise that is starting to be realized in a research setting, whereas the routine use of NGS in a clinical setting still faces several challenges.

REFERENCES

Abecasis, G. R., A. Auton, et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature 491*(7422), 56–65.

Abrahams, B. S., & Geschwind, D. H. (2008). Advances in autism genetics: on the threshold of a new neurobiology. *Nat Rev Genet 9*(5), 341–355.

Adessi, C., Matton, G., et al. (2000). Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res 28*(20), E87.

Adzhubei, I. A., Schmidt, S., et al. (2010). A method and server for predicting damaging missense mutations. *Nat Methods 7*(4), 248–249.

Alonso Vilatela, M. E., Lopez-Lopez, M., et al. (2012). Genetics of Alzheimer's disease. *Arch Med Res 43*(8), 622–631.

Anderson, M. W., & Schrijver, I. (2010). Next generation DNA sequencing and the future of genomic medicine. *Genes 1*, 38–69.

Asan, Y. Xu, et al. (2011). Comprehensive comparison of three commercial human whole-exome capture platforms. *Genome Biol 12*(9), R95.

Babenko, A. P., Polak, M., et al. (2006). Activating mutations in the ABCC8 gene in neonatal diabetes mellitus. *N Engl J Med 355*(5), 456–466.

Bamshad, M. J., Ng, S. B., et al. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet 12*(11), 745–755.

Bellus, G. A., Hefferon, T. W., et al. (1995). Achondroplasia is defined by recurrent G380R mutations of FGFR3. *Am J Hum Genet 56*(2), 368–373.

Boycott, K. M., Parboosingh, J. S., et al. (2008). Clinical genetics and the Hutterite population: a review of Mendelian disorders. *Am J Med Genet A 146A*(8), 1088–1098.

Bradfield, J. P., Qu, et al. H. Q., (2011). A genome-wide meta-analysis of six type 1 diabetes cohorts identifies multiple associated loci. *PLoS Genet 7*(9), e1002293.

Braslavsky, I., Hebert, B., et al. (2003). Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A 100*(7), 3960–3964.

Caliskan, M., Chong, J. X., et al. (2011). Exome sequencing reveals a novel mutation for autosomal recessive non-syndromic mental retardation in the TECR gene on chromosome 19p13. *Hum Mol Genet 20*(7), 1285–1289.

Claes, L., Del-Favero, J., et al. (2001). De novo mutations in the sodium-channel gene SCN1A cause severe myoclonic epilepsy of infancy. *Am J Hum Genet 68*(6), 1327–1332.

Coffey, A. J., Kokocinski, F., et al. (2011). The GENCODE exome: sequencing the complete human exome. *Eur J Hum Genet 19*(7), 827–831.

Conrad, D. F., Keebler, J. E., et al. (2011). Variation in genome-wide mutation rates within and between human families. *Nat Genet 43*(7), 712–714.

Cook, E. H., Jr., & Scherer, S. W. (2008). Copy-number variations associated with neuropsychiatric conditions. *Nature 455*(7215), 919–923.

Danecek, P., Auton, A., et al. (2011). The variant call format and VCF tools. *Bioinformatics 27*(15), 2156–2158.

de Vries, B. B., Pfundt, R., et al. (2005). Diagnostic genome profiling in mental retardation. *Am J Hum Genet 77*(4), 606–616.

DePristo, M. A., Banks, E., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet 43*(5), 491–498.

Dressman, D., Yan, H., et al. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci U S A 100*(15), 8817–8822.

Dunham, I., Kundaje, A., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature 489*(7414), 57–74.

Edvardson, S., Shaag, A., et al. (2010). Joubert syndrome 2 (JBTS2) in Ashkenazi Jews is associated with a TMEM216 mutation. *Am J Hum Genet 86*(1), 93–97.

Eyre-Walker, A., & Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet 8*(8), 610–618.

Fedurco, M., Romieu, A., et al. (2006). BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res 34*(3), e22.

Firth, H. V., Richards, S. M., et al. (2009). DECIPHER: database of chromosomal imbalance and phenotype in humans using Ensembl resources. *Am J Hum Genet 84*(4), 524–533.

Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nat Methods 6*(11 Suppl), S6–S12.

Froguel, P., Vaxillaire, M., et al. (1992). Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus. *Nature 356*(6365), 162–164.

Gilissen, C., Hoischen, A., et al. (2011). Unlocking Mendelian disease using exome sequencing. *Genome Biol 12*(9), 228.

Gilman, S. R., Iossifov, I., et al. (2011). Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. *Neuron 70*(5), 898–907.

Girard, S. L., Gauthier, J., et al. (2011). Increased exonic de novo mutation rate in individuals with schizophrenia. *Nat Genet 43*(9), 860–863.

Gloyn, A. L., Pearson, E. R., et al. (2004). Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N Engl J Med 350*(18), 1838–1849.

Goate, A., Chartier-Harlin, M. C., et al. (1991). Segregation of a missense mutation in the amyloid precursor protein gene with familial Alzheimer's disease. *Nature 349*(6311), 704–706.

Guerreiro, R. J., Lohmann, E., et al. (2012). Exome sequencing reveals an unexpected genetic cause of disease: NOTCH3 mutation in a Turkish family with Alzheimer's disease. *Neurobiol Aging 33*(5), 1008 e17–e23.

Harrow, J., Frankish, A., et al. (2012). GENCODE: the reference human genome annotation for the ENCODE Project. *Genome Res 22*(9), 1760–1774.

Heinzen, E. L., Depondt, C., et al. (2012). Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. *Am J Hum Genet 91*(2), 293–302.

Helgason, A., Yngvadottir, B., et al. (2005). An Icelandic example of the impact of population structure on association studies. *Nat Genet 37*(1), 90–95.

Hoischen, A., van Bon, B. W., et al. (2010). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. *Nat Genet 42*(6), 483–485.

Holm, H., Gudbjartsson, D. F., et al. (2011). A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet 43*(4), 316–320.

Iossifov, I., Ronemus, M., et al. (2012). De novo gene disruptions in children on the autistic spectrum. *Neuron 74*(2), 285–299.

Jonsson, T., Atwal, J. K., et al. (2012). A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature 488*(7409), 96–99.

Jonsson, T., Stefansson, H., et al. (2012). Variant of TREM2 associated with the risk of Alzheimer's disease. *N Engl J Med 368*(2), 107–116.

Kao, W. C., Stevens, K., et al. (2009). BayesCall: a model-based base-calling algorithm for high-throughput short-read sequencing. *Genome Res 19*(10), 1884–1895.

Kircher, M., Stenzel, U., et al. (2009). Improved base calling for the Illumina Genome Analyzer using machine learning strategies. *Genome Biol 10*(8), R83.

Klassen, T., Davis, C., et al. (2011). Exome sequencing of ion channel genes reveals complex profiles confounding personal risk assessment in epilepsy. *Cell 145*(7), 1036–1048.

Klei, L., Sanders, S. J., et al. (2012). Common genetic variants, acting additively, are a major source of risk for autism. *Mol Autism 3*(1), 9.

Kozomara, A. & Griffiths-Jones, S. (2011). miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res 39*(Database issue), D152–D157.

Krawitz, P. M., Schweiger, M. R., et al. (2010). Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet 42*(10), 827–829.

Kryukov, G. V., Pennacchio, L. A., et al. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet 80*(4), 727–739.

Lander, E. S., Linton, L. M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature 409*(6822), 860–921.

Le, S. Q., & Durbin, R. (2010). SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res 21*(6), 952–960.

Li, H., Handsaker, B., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics 25*(16), 2078–2079.

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform 11*(5), 473–483.

Li, R., Yu, C., et al. (2009). SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics 25*(15), 1966–1967.

Lin, Y., Li, J., et al. (2011). Comparative studies of de novo assembly tools for next-generation sequencing technologies. *Bioinformatics 27*(15), 2031–2037.

Ma, D., Salyakina, D., et al. (2009). A genome-wide association study of autism reveals a common novel risk locus at 5p14.1. *Ann Hum Genet 73*(Pt 3), 263–273.

MacArthur, D. G., Balasubramanian, S., et al. (2012). A systematic survey of loss-of-function variants in human protein-coding genes. *Science 335*(6070), 823–828.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet 9*, 387–402.

Margulies, M., Egholm, M., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature 437*(7057), 376–380.

Mefford, H. C., Batshaw, M. L., et al. (2012). Genomics, intellectual disability, and autism. *N Engl J Med 366*(8), 733–743.

Metzker, M. L. (2010). Sequencing technologies—the next generation. *Nat Rev Genet 11*(1), 31–46.

Meuzelaar, L. S., Lancaster, O., et al. (2007). MegaPlex PCR: a strategy for multiplex amplification. *Nat Methods 4*(10), 835–837.

Najmabadi, H., Hu, H., et al. (2011). Deep sequencing reveals 50 novel genes for recessive cognitive disorders. *Nature 478*(7367), 57–63.

Neale, B. M., Kou, Y., et al. (2012). Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature 485*(7397), 242–245.

Need, A. C., McEvoy, J. P., et al. (2012). Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia. *Am J Hum Genet 91*(2), 303–312.

Ng, P. C., & Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res 31*(13), 3812–3814.

Ng, S. B., Bigham, A. W., et al. (2010). Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet 42*(9), 790–793.

Nielsen, R., Paul, J. S., et al. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet 12*(6), 443–451.

Norio, R. (2003). The Finnish Disease Heritage III: the individual diseases. *Hum Genet 112*(5-6), 470–526.

O'Roak, B. J., Vives, L., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. *Nature 485*(7397), 246–250.

Pottier, C., Hannequin, D., et al. (2012). High frequency of potentially pathogenic SORL1 mutations in autosomal dominant early-onset Alzheimer disease. *Mol Psychiatry 17*(9), 875–879.

Pruitt, K. D., Harrow, J., et al. (2009). The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res 19*(7), 1316–1323.

Pruitt, K. D., Tatusova, T., et al. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res 40*(Database issue), D130–D135.

Quinlan, A. R., Stewart, D. A., et al. (2008). Pyrobayes: an improved base caller for SNP discovery in pyrosequences. *Nat Methods 5*(2), 179–181.

Rabbani, B., Mahdieh, N., et al. (2012). Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. *J Hum Genet 57*(10), 621–632.

Raux, G., Guyant-Marechal, L., et al. (2005). Molecular diagnosis of autosomal dominant early onset Alzheimer's disease: an update. *J Med Genet 42*(10), 793–795.

Ronaghi, M., Uhlen, M., et al. (1998). A sequencing method based on real-time pyrophosphate. *Science 281*(5375), 363, 365.

Ropers, H. H. (2010). Genetics of early onset cognitive impairment. *Annu Rev Genomics Hum Genet 11*, 161–187.

Ropers, H. H., & Hamel, B. C. (2005). X-linked mental retardation. *Nat Rev Genet 6*(1), 46–57.

Saitsu, H., Kato, M., et al. (2008). De novo mutations in the gene encoding STXBP1 (MUNC18-1) cause early infantile epileptic encephalopathy. *Nat Genet 40*(6), 782–788.

Sanders, S. J., Murtha, M. T., et al. (2012). De novo mutations revealed by whole-exome sequencing are strongly associated with autism. *Nature 485*(7397), 237–241.

Saxena, R., Elbers, C. C., et al. (2012). Large-scale gene-centric meta-analysis across 39 studies

identifies type 2 diabetes loci. *Am J Hum Genet 90*(3), 410–425.

Shendure, J., Porreca, G. J., et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science 309*(5741), 1728–1732.

Sherrington, R., Rogaev, E. I., et al. (1995). Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature 375*(6534), 754–760.

Shiang, R., L.Thompson, M., et al. (1994). Mutations in the transmembrane domain of FGFR3 cause the most common genetic form of dwarfism, achondroplasia. *Cell 78*(2), 335–342.

Stenson, P. D., Ball, E. V., et al. (2009). The Human Gene Mutation Database: providing a comprehensive central mutation database for molecular diagnostics and personalized genomics. *Hum Genomics 4*(2), 69–72.

Sulem, P., Gudbjartsson, D. F., et al. (2011). Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat Genet 43*(11), 1127–1130.

Sulonen, A. M., Ellonen, P., et al. (2011). Comparison of solution-based exome capture methods for next generation sequencing. *Genome Biol 12*(9), R94.

Tarpey, P. S., Smith, R., et al. (2009). A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nat Genet 41*(5), 535–543.

Tennessen, J. A., Bigham, A. W., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science 337*(6090), 64–69.

Teslovich, T. M., Musunuru, K., et al. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature 466*(7307), 707–713.

Tomkinson, A. E., Vijayakumar, S., et al. (2006). DNA ligases: structure, reaction mechanism, and function. *Chem Rev 106*(2), 687–699.

Treffer, R., & Deckert, V., (2010). Recent advances in single-molecule sequencing. *Curr Opin Biotechnol 21*(1), 4–11.

Valouev, A., Ichikawa, J., et al. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res 18*(7), 1051–1063.

Varley, K. E., & Mitra, R. D. (2008). Nested Patch PCR enables highly multiplexed mutation discovery in candidate genes. *Genome Res 18*(11), 1844–1850.

Vissers, L. E., de Ligt, J., et al. (2010). A de novo paradigm for mental retardation. *Nat Genet 42*(12), 1109–1112.

Vorstman, J. A., Anney, R. J., et al. (2012). No evidence that common genetic risk variation is shared between schizophrenia and autism. *Am J Med Genet B Neuropsychiatr Genet*.

Wang, K., Zhang, H., et al. (2009). Common genetic variants on 5p14.1 associate with autism spectrum disorders. *Nature 459*(7246), 528–533.

Weckhuysen, S., Mandelstam, S., et al. (2012). KCNQ2 encephalopathy: emerging phenotype of a neonatal epileptic encephalopathy. *Ann Neurol 71*(1), 15–25.

Weiss, L. A., Arking, D. E., et al. (2009). A genome-wide linkage and association scan reveals novel loci for autism. *Nature 461*(7265), 802–808.

Wu, H., Irizarry, R. A., et al. (2010). Intensity normalization improves color calling in SOLiD sequencing. *Nat Methods 7*(5), 336–337.

Xu, B., Ionita-Laza, I., et al. (2012). De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. *Nat Genet 44*(12), 1365–1369.

Xu, B., Roos, J. L., et al. (2011). Exome sequencing supports a de novo mutational paradigm for schizophrenia. *Nat Genet 43*(9), 864–868.

Xu, B., Roos, J. L., et al. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet 40*(7), 880–885.

Yamagata, K., Furuta, H., et al. (1996). Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1). *Nature 384*(6608), 458–460.

Zhang, W., Chen, J., et al. (2011). A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS One 6*(3), e17915.

2

Epigenomics: An Overview

KEVIN HUANG AND GUOPING FAN

INTRODUCTION

Epigenetics is the study of mechanisms that can alter gene expression without changing the underlying DNA sequence. Under this broad term many mechanisms have been considered epigenetic, including DNA methylation, histone modifications, and noncoding RNA. Often these epigenetic mechanisms work in concert to influence both gene expression and each other. Epigenetic landscapes are extremely complex with a vast spectrum of variations that are used to fine-tune gene expression. In order to fully understand the regulatory domains in the genome, all epigenetic regulatory forces must be considered. Recent advances in high-throughput technology have afforded the opportunity to survey epigenetic features across entire genomes, bringing forth a vibrant field of "epigenomics"-based research. This chapter focuses on how epigenetic mechanisms shape the transcriptome, the tools we use to study these pathways on a genome scale, and the insights we have gained from these epigenomic-driven studies. Throughout the text, we highlight the impact and relevance of epigenomic studies on illuminating novelties in neurobiology.

HIGH-THROUGHPUT TECHNOLOGIES PAVING THE WAY FOR EPIGENOMICS

Microarrays

The advent of microarray technology provided a phenomenal method of measuring multiple events in a single experiment. Expanding on classical complementary hybridization-based detection methods, the microarray platform relies on hybridization of fluorescently labeled DNA to predefined probes that uniquely represents portions of the genome (Heller, 2002; Schulze & Downward, 2001; Young, 2000). DNA probes are usually evenly spaced and attached to a solid surface commonly referred to as a chip. Because of the limited size of the chip, only a finite number of probes can be placed onto a single chip. For experimental designs that attempt to exhaustively represent the genome by having probes found every few kilobases across the genome (so-called tiling arrays), multiple chips are required. Other experimental designs that focus on promoters alone may require fewer chips to fully represent all mammalian promoters. Microarray had clear advantages compared with previous approaches, in particular the ability to sample large portions of the genome in a cost-effective and less time-consuming manner.

Next-Generation Sequencing

In more recent years, high-throughput DNA sequencing has supplanted most microarray technologies for many reasons, including improved high throughput, sensitivity, and accuracy (Metzker, 2010; Shendure & Ji, 2008). However, it is worth mentioning that many laboratories are continuing to use microarray-based platforms primarily because of matured analytical tools (Allison et al., 2006; Gentleman et al., 2004; Li, 2008), and the costs are still lower for studies geared more for sample sizes in the hundreds and thousands.

Sequencing offers many advantages over the microarray platform, including base-pair resolution and unbiased surveying of the genome. In general, all library construction protocols share fundamental commonalities (Metzker, 2005, 2010; Shendure & Ji, 2008). Since the goal is to generate short reads, the majority of library construction protocols share common procedures such as DNA or RNA fragmentation to a desired size distribution, adapter ligation, and PCR amplification. The PCR step is

necessary because most library construction methods yield small amounts of DNA that may not be easily detected on the sequencer. On the other hand, the PCR step also remains one of the banes of library construction because PCR amplification introduces a variety of biases that confound quantitative analyses. Indeed, several groups are working on methods for circumventing the PCR step in library construction, which will simplify library construction and data analysis in the future.

DNA METHYLATION

Background

DNA methylation is one of the best-studied epigenetic mechanisms and involves the covalent attachment of a methyl group to the 5 carbon position of cytosine (Bird, 1986; Reik, 2007). In mammals, this action is catalyzed by a family of DNA methyltransferases (Dnmts), including Dnmt1, Dnmt3a, and Dnmt3b. Loss of any of these enzymes during embryogenesis is lethal, indicating an essential role for DNA methylation during development (Li et al., 1992; Okano et al., 1999). The prevailing hypothesis on the mechanism of action for DNA methylation involves repression via its presence on the proximal promoter (Miranda & Jones, 2007). It is thought that DNA methylation suppresses gene activity either by acting as part of a signaling pathway that recruits repressor complexes or by sterically hindering transcription factor binding (Huang & Fan, 2010; Moore et al., 2012). Nevertheless, with some exceptions, global mapping of gene promoters indicates a negative correlation between promoter methylation and gene activity (Suzuki & Bird, 2008). Epigenomic studies using mouse embryonic stem cells (ESCs) revealed that promoters can be subclassified based on their CpG content (Fouse et al., 2008; Meissner et al., 2008; Mikkelsen et al., 2007; Mohn et al., 2008). For example, proximal promoters with a high density of CpG dinucleotides tend to be hypomethylated, whereas promoters with a low density of CpGs are hypermethylated. However, the absence of DNA methylation does not necessarily predict gene activity; many gene promoters that lack DNA methylation can also be transcriptionally inactive (Fouse et al., 2008; Lister et al., 2009; Meissner et al., 2008; Mohn et al., 2008; Weber et al., 2007). Furthermore, DNA methylation patterns differ in various cell types. We now know that different cells have their own unique DNA methylation signature, and these characteristics are important for governing cell identity. For example, neural genes are repressed in nonneural tissues by promoter DNA methylation but are unmethylated in neural cells, indicating a direct role for DNA methylation (Meissner et al., 2008; Mohn et al., 2008). These types of studies have revealed an immense amount about the methylation status of gene promoters in regulating gene expression and cellular differentiation.

Gene Body Methylation

Global DNA methylation mapping has revealed many novel facets of DNA methylation beyond the classical model of gene regulation. For example, outside of gene promoters, DNA methylation appears to be highly enriched within the gene body (the transcribed portion of the gene). In many species, gene body methylation appears to have both repressive and enhancer roles (Zemach et al., 2010). Recent methylome studies across phyla found that gene body methylation is mostly enriched for genes with moderate expression. In other words, genes that are expressed either highly or lowly are depleted of gene body methylation. However, mammals do not seem to share this trait. Gene expression in both humans and mice does not correlate tightly with CG methylation in the gene body of protein coding genes (Feng et al., 2010b; Lister et al., 2009). Furthermore, there is still no conclusive evidence to indicate that gene body methylation plays a role in regulating gene expression.

On the other hand, gene body methylation for repetitive elements (such as retrotransposons) seems to be widely conserved across a diverse array of species (Feng et al., 2010b; Zemach et al., 2010). In many cases, heavy methylation across the repeat gene body results in stable silencing. Indeed, experiments that artificially remove DNA methylation within an organism result in a dramatic induction of repeat elements and may lead to cell death (Chen et al., 2007; Fan et al., 2005; Hutnick et al., 2009.). It is thought that DNA methylation has evolved as a defense mechanism to silence foreign DNA such as from viruses, which are capable of invading a host cell for viral replication (Zemach et al., 2010). So although foreign viral DNAs have successfully integrated into host genomes over time, the cell has used DNA methylation as a way of providing genomic