The OMICs

Applications in Neuroscience

Edited by
Giovanni Coppola

OXFORD

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

常州大学图书馆
藏书章

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

© Oxford University Press 2014

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, without the prior permission in writing of Oxford University Press, or as expressly permitted by law, by license, or under terms agreed with the appropriate reproduction rights organization. Inquiries concerning reproduction outside the scope of the above should be sent to the Rights Department, Oxford University Press, at the address above.

You must not circulate this work in any other form
and you must impose this same condition on any acquirer.

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

THE OMICS

Tail suspension 2015-7-12

右
左 A1A2 → A4A5 → A3B1 → B2B3 → D1D2
→ D3D4 → G1G2 → G3G4 → J1J2 →
J3J4 → B4B5 → C1D5 → C2C3 →
C4C5 → G5H1 → H2H3 → J5K1 →
K2K3 → E1E2 → E3E4 → L5F1 → F2F3
→ H4H5 → I1I2 → K4K5 → L1L2
→ F4F5 → I3I4 → I5 3 → L4L5

TOYOBO

⑫ H4 - I2
⑬ K4 - L2
⑭ F4F5I3I4
⑮ I5L3-L5

① A1 - A4
② A5 - B3
③ D1 - D4
④ G1 - G4
⑤ J1 - J4
⑥ B4 B5 C1 D5
⑦ C2 - C5
⑨ G5 - H3
⑩ J5 - K3
⑧ E1 - E4
⑪ E5 - F3

TOYOBO

open field 2015-7-12

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

此为试读,需要完整PDF请访问: www.ertongbook.com

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

CONTENTS

PART FIVE: *THERAPEUTICS*

PART SIX: *OMICsOME: INTEGRATION OF OMICs DATA*

PART I

DNA

此为试读,需要完整PDF请访问: www.ertongbook.com

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

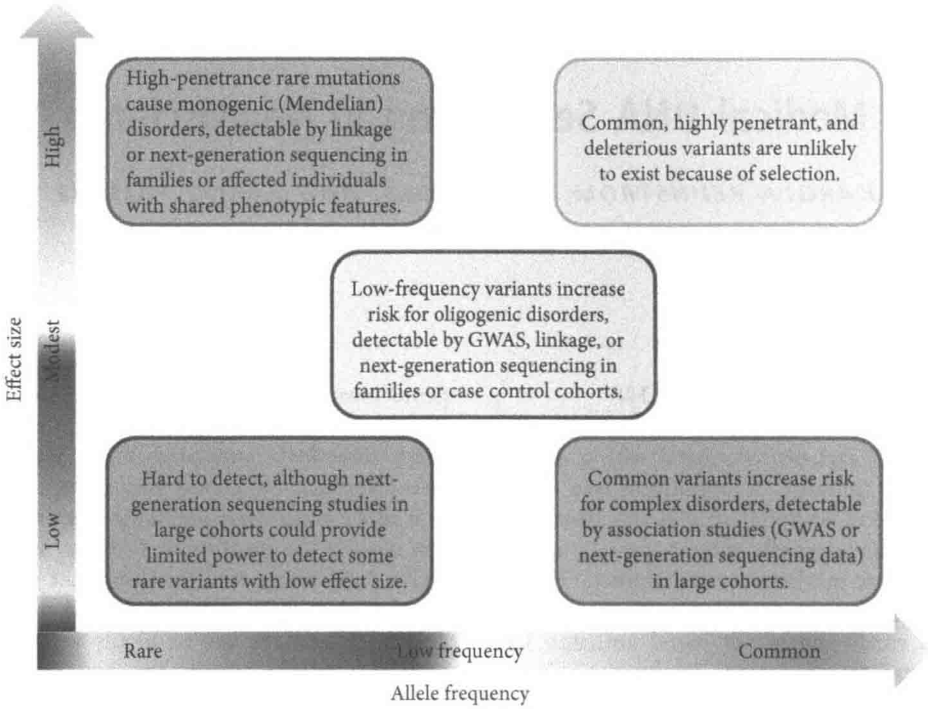
High-penetrance rare mutations cause monogenic (Mendelian) disorders, detectable by linkage or next-generation sequencing in families or affected individuals with shared phenotypic features.
Common, highly penetrant, and deleterious variants are unlikely to exist because of selection.
Low-frequency variants increase risk for oligogenic disorders, detectable by GWAS, linkage, or next-generation sequencing in families or case control cohorts.
Hard to detect, although next-generation sequencing studies in large cohorts could provide limited power to detect some rare variants with low effect size.
Common variants increase risk for complex disorders, detectable by association studies (GWAS or next-generation sequencing data) in large cohorts.
Effect size
High
Modest
Low
Rare
Low frequency
Common
Allele frequency

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.

共识编
序列

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.

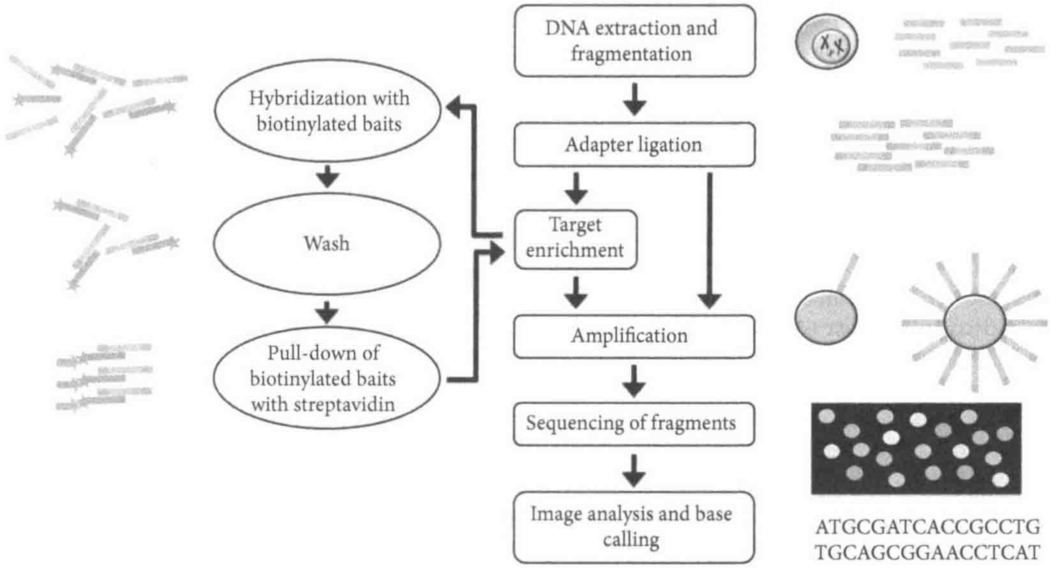
DNA extraction and fragmentation
Adapter ligation
Target enrichment
Amplification
Sequencing of fragments
Image analysis and base calling
Hybridization with biotinylated baits
Wash
Pull-down of biotinylated baits with streptavidin
ATGCGATCACCGCCTG
TGCAGCGGAACCTCAT

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

此为试读,需要完整PDF请访问: www.ertongbook.com