Lee-Jun C. Wong *Editor*

Next Generation Sequencing

Translation to Clinical Diagnostics

Springer

Next Generation Sequencing

Lee-Jun C. Wong
Editor

Next Generation Sequencing

Translation to Clinical Diagnostics

Springer

Editor
Lee-Jun C. Wong
Department of Molecular and Human Genetics
Baylor College of Medicine
Houston, TX, USA

ISBN 978-1-4614-7000-7 ISBN 978-1-4614-7001-4 (eBook)
DOI 10.1007/978-1-4614-7001-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013939348

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The recent development of high-throughput next-generation sequencing (NGS) technology has transformed the way DNA-based molecular diagnostic testing is performed in clinical laboratories. NGS allows parallel sequencing analyses of multiple genes effectively at any desirable depth of coverage. It is often difficult for clinicians to fully understand and utilize NGS-based tests effectively due to the sophisticated instrumentation, the enormous amount of data generated, the complex analytical tools involving sequence alignment, and the bioinformatics for variant annotation. While professional interpretation can address these issues, the purpose of this book is to help physicians to better understand the science behind and clinical utility of NGS to maximize benefit for their patients.

In Part I, DNA sequencing principles, the underlying chemistries, and the development of NGS technology are described. This is followed by an overview of the methods used in the traditional molecular diagnosis of human genetic diseases. Part II details the instrumentations, bioinformatics, and computational techniques involved in NGS. This section of the book includes a comparison of the different methods used for target gene enrichment and the various sequencing platforms, as well as the algorithms and bioinformatics used for sequence analysis, variant annotation, and interpretation of the final results. Part III describes specific applications of NGS to the molecular diagnosis of clinically defined diseases (e.g., congenital disorders of glycosylation), genetically heterogeneous disorders involving many different genes leading to the same or similar clinical phenotypes (e.g., retinitis pigmentosa), diseases associated with a particular region or a whole chromosome (e.g., X-linked intellectual disability), and dual genome mitochondrial respiratory chain disorders involving nuclear and mitochondrial DNA. The utility of NGS for noninvasive prenatal diagnosis is also described. Part IV discusses the College of American Pathologists (CAP) / Clinical Laboratory Improvement Amendments (CLIA) guidelines for establishing a clinically based test using novel technology and compliance issues for laboratories offering NGS-based testing.

I am indebted to the contributing authors who have made this book entitled *Next-Generation Sequencing: Translation to Clinical Diagnostics* possible. I particularly appreciate the patience of the authors who submitted their chapters on time.

Houston, TX, USA Lee-Jun C. Wong

Contents

Contributors

Monica J. Basehore Greenwood Genetic Center, Greenwood, SC, USA

Rui Chen Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Ephrem L. H. Chin Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Cristian Coarfa Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

William J. Craigen Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Hong Cui Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Michael J. Friez Greenwood Genetic Center, Greenwood, SC, USA

Amy S. Gargis Division of Laboratory Science and Standards, Centers for Disease Control and Prevention, Atlanta, GA, USA

Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA

Si Houn Hahn Seattle Children's Research Institute, Seattle, WA, USA

Department of Pediatrics, University of Washington School of Medicine, Seattle Children's Hospital, Seattle, WA, USA

Madhuri Hegde Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

Whitehead Biomedical Research Building, Emory University School of Medicine, Atlanta, GA, USA

Ammar Husami Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Department of Pediatrics, University of Cincinnati Medical School, Cincinnati, OH, USA

Melanie Jones Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

Lisa Kalman Division of Laboratory Science and Standards, Centers for Disease Control and Prevention, Atlanta, GA, USA

Megan L. Landsverk Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Yuk Ming Dennis Lo Li Ka Shing Institute of Health Sciences and Department of Chemical Pathology, Centre for Research into Circulating Fetal Nucleic Acids, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China

Ira M. Lubin Division of Laboratory Science and Standards, Centers for Disease Control and Prevention, Atlanta, GA, USA

Margherita Milone Department of Neurology, Neuromuscular Division, Mayo Clinic, Rochester, MN, USA

Susan Pacheco Department of Pediatrics, Division of Allergy/Immunology, University of Texas Health Science Center, Houston, TX, USA

Theru A. Sivakumaran Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Department of Pediatrics, University of Cincinnati Medical School, Cincinnati, OH, USA

Brad T. Tinkle Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Department of Pediatrics, University of Cincinnati Medical School, Cincinnati, OH, USA

Nancy Bo Yin Tsui Li Ka Shing Institute of Health Sciences and Department of Chemical Pathology, Centre for Research into Circulating Fetal Nucleic Acids, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China

C. Alexander Valencia Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Department of Pediatrics, University of Cincinnati Medical School, Cincinnati, OH, USA

Valeria Vasta Seattle Children's Research Institute, Seattle, WA, USA

Feng Wang Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Jing Wang Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Lisa D. White Departments of Molecular & Human Genetics and Molecular & Cellular Biology, Baylor College of Medicine, Houston, TX, USA

Lee-Jun C. Wong Medical Genetics Laboratories, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Fuli Yu Department of Molecular and Human Genetics, Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA

Kejian Zhang Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

Department of Pediatrics, University of Cincinnati Medical School, Cincinnati, OH, USA

Victor Wei Zhang Medical Genetics Laboratories, Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

Part I
Overview

Chapter 1
History of DNA Sequencing Technologies

Lisa D. White

Abstract DNA sequencing technologies have a relatively short history – with the first published report in 1973 – a mere 39 years. But what a history! In 1973, Walter Gilbert and Allan Maxam published the 24 base pair sequence of the *lac* operator using chemical sequencing techniques Gilbert and Maxam (Proc Nat Acad Sci USA 70(12):3581–3584, 1973). Through innovation and automation, we are now able in 2012 to sequence the entire human genome (three billion base pairs) in a little more than a week. This chapter will review this short but amazing story of brute force, automation, and innovation from its infancy to the present. Suffice it to say that advances and technological leaps are continuing apace and this chapter will be outdated by the time you are reading it.

1 The Early Days of Sequencing

The 1953, landmark publication suggesting the structure of DNA by Watson and Crick [1] began what has become one of the most successful modern endeavors carried out by humans. The Nobel Prize in Physiology or Medicine 1962 was awarded jointly to Francis Harry Compton Crick, James Dewey Watson, and Maurice Hugh Frederick Wilkins "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"[2].

After the 1962 Nobel recognition for the discovery of the structure of DNA, it was only a short 18 years later that the Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg "for his fundamental studies of the biochemistry of

L.D. White, Ph.D. (✉)
Departments of Molecular & Human Genetics and Molecular & Cellular Biology,
Baylor College of Medicine, Houston, TX 77030, USA
e-mail: lisaw@bcm.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_1,

nucleic acids, with particular regard to recombinant-DNA," the other half jointly to Walter Gilbert and Frederick Sanger "for their contributions concerning the determination of base sequences in nucleic acids" [3]. The first report of chemical sequencing in 1973 [4] showed our new ability to determine the nucleotide sequence of the *lac* operator (24 base pairs in length). This report was quickly followed by a paper by Sanger et al. [5] reporting an "easy" method for sequencing using synthetic primers and DNA polymerase. Then in 1977, two sequencing methods were described in detail [6, 7]. The Maxam-Gilbert method utilized chemical breakage, radioisotope end labeling, and gel electrophoresis to sequence DNA, while Sanger described sequencing by using radiolabeled ddNTPs and gel electrophoresis. The Maxam-Gilbert method eventually fell out of favor due the improvements in the Sanger method. Additionally the toxic chemicals used in the method were considered unsafe.

1.1 Maxam-Gilbert Method

The Maxam-Gilbert method [8] utilizes chemical degradation of a radioisotope end-labeled DNA fragment. The fragment is partially cleaved in five separate chemical reactions that are specific to either an individual nucleotide base or a type of nucleotide base. These reactions generate five separate populations of radiolabeled molecules that extend from a common point (the end-labeled terminus of the strand). These populations are then size separated using PAGE (polyacrylamide gel electrophoresis) and the radiolabeled molecules are visualized by autoradiography.

This method has remained relatively unchanged with the exception of additional cleavage reactions developed to supplement the original reactions (see review [9]). The method relies on the specificity of the cleavage reactions, which are processed in two stages. In stage one, specific nucleotide bases (or types of bases) undergo a chemical modification. In stage two, the modified base gets removed from the sugar and the phosphodiester bonds on the 5′ and 3′ side of the modified bases are cleaved (see Fig. 1.1). The reactions are carefully controlled so on average only one base per strand is modified. This method is optimal for DNA strands that are less than 250 bases from the radiolabeled end. This range of fragment sizes is less than that of the Sanger method.

Initially the Maxam-Gilbert method was more reproducible and accessible than Sanger's method because of the requirements for single-stranded templates, specific oligonucleotide primers, and access to high-quality preparations of the Klenow fragment of *E. coli* DNA polymerase I. Further developments in the Sanger method and in the field, however, resulted in more widespread use of Sanger and the Maxam-Gilbert method was eventually supplanted by Sanger for simple determination of DNA sequence.

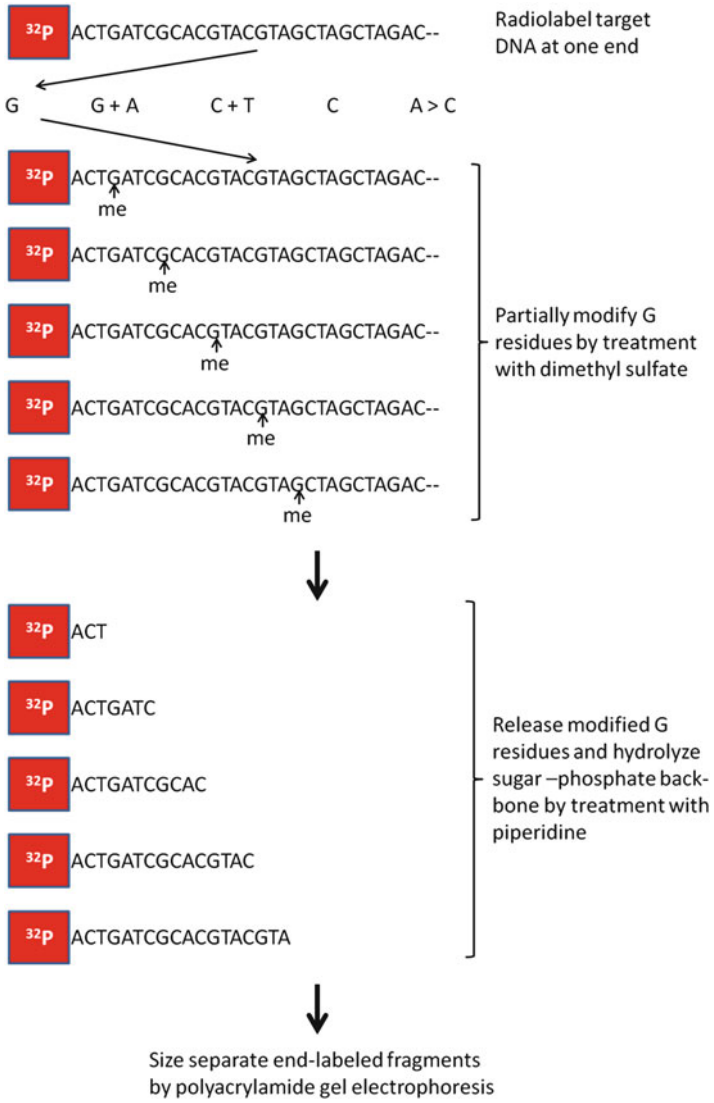

Fig. 1.1 Example of G nucleotide chemical cleavage in Maxam-Gilbert DNA sequencing

1.2 Sanger Method

The current Sanger sequencing-by-synthesis (enzymatic) method using chain-terminating dideoxynucleoside triphosphates (ddNTPs) [7] grew from the original ± sequencing technique first reported in 1975 [5]. Chain-terminating ddNTPs lack a

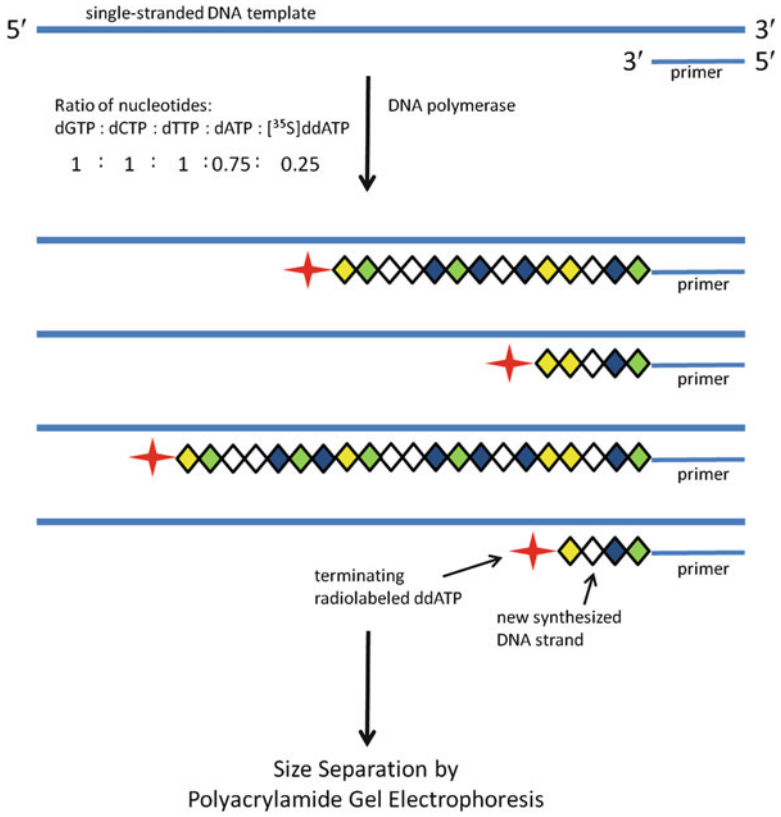

Fig. 1.2 Example of the "A" reaction (one of four to be performed on the same template) of the Sanger chain termination method. A small amount of radiolabeled ddATP is added to the reaction to produce a population of DNA chains terminating in an "A." After size separation and comparison to adjacent "G," "C," and "T" lanes, the DNA sequence can be read from smallest to largest on the autoradiograph

3′ hydroxyl residue and, while able to be incorporated into a growing DNA chain through their 5′ triphosphate group, cannot extend the chain further.

The reaction is relatively simple. In a series of four separate reactions (A, C, G, and T), a small competing amount of the ddNTP (ddATP, ddCTP, ddGTP, or ddTTP) is added to the reaction. This small amount of ddNTP mixed with conventional dNTPs competes to infrequently incorporate into the growing chain causing termination. An example of the "A" reaction is shown in Fig. 1.2. The product of the reaction is a population of oligonucleotide chains whose lengths are determined by the distance between the terminus of the primer used to initiate DNA synthesis and the site of premature termination.

The four reactions are loaded into four adjacent lanes on a polyacrylamide gel for size separation and the bands are visualized by autoradiography.

2 Automated DNA Sequencing

The discovery of thermostable DNA polymerases led to the development of the polymerase chain reaction (PCR) and subsequently improved methods for DNA sequencing, namely, thermal cycle sequencing [10]. This method has advantages over traditional chain-terminating sequence. First, it uses double-stranded DNA as the starting template instead of single stranded, and second, it can start with very small amounts of starting DNA.

The method is similar to conventional PCR but only a single primer is used along with a radiolabeled ddNTP in four separate reactions (A, C, G, T). The result is linear, not exponential, accumulation of a single-stranded product that can then be size separated and read by polyacrylamide gel electrophoresis and autoradiography.

Replacement of the radioisotopes (usually ^{35}S or ^{33}P) with fluorescently labeled molecules along with thermal cycle sequencing opened the door to automating DNA sequencing [11] because the detection systems for fluorescent labels can easily be adapted for automation (Fig. 1.3). Instruments originally utilized slab polyacrylamide gel electrophoresis for size separation and reading of the sequence. Fluorescently labeled molecules can be separated with capillary electrophoresis so that a population of fluorescent terminated fragments can be read. Additionally, this method allows all four reactions to be performed in a single tube. The result (shown in Fig. 1.3) is called a "trace" or an electropherogram. Capillary electrophoresis sequencing can involve a single capillary or up to 96 capillaries allowing increased throughput.

The chain termination method can generate an average of 500 bases of sequence per run. This constraint does not become an issue until larger regions of sequence are required. The human genome is approximately three billion bases (3 gigabases (gb)).

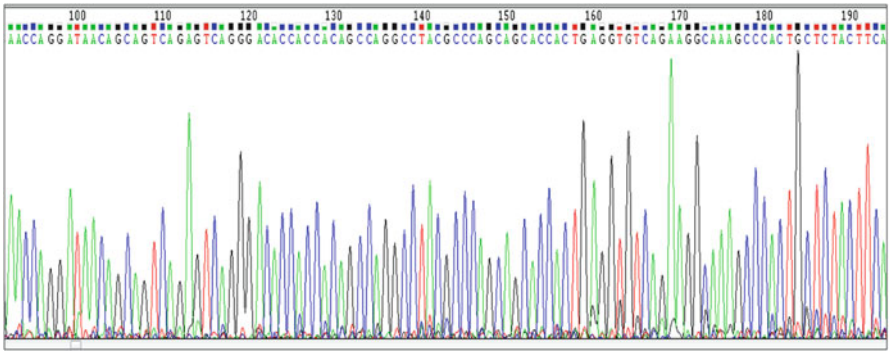

Fig. 1.3 Electropherogram "trace" of thermal cycle sequencing with each peak representing fluorescently labeled ddNTP that is color coded for A (*green*), C (*blue*), G (*black*), or T (*red*) in this image. The sample was sequenced on the ABI PRISM® 310 Genetic Analyzer from Applied Biosystems

That means that one run of a 96 capillary machine would generate only 48,000 bases or 48 kilobases (kb) or 0.000048 gb. So, a researcher would have to run the 96 capillary instrument 62,500 times to generate a 1X coverage of the human genome. At 10 runs a day, 7 days a week for one instrument (each run takes approximately 2 h), it would take more than 17 years to generate the 1X coverage referenced above.

Only through combining the advances in sample processing technology and detection methods along with robotics, automatic data collection, and a multitude of instruments can the sequencing of whole genomes in a reasonable amount of time finally be realized.

3 Human Genome Project

With the advent of automated high-throughput sequencing, the idea of analyzing the whole human genome was first proposed seriously by academics in the 1980s. The US Department of Energy (DOE) became involved in 1986 and established an early genome project in 1987. This was followed in 1988 by the US Congress funding both the National Institutes of Health (NIH) and the DOE to establish a joint project to explore the idea of a human genome project.

James Watson was appointed to lead the NIH contingent: the Office of Human Genome Research. This group evolved into National Center for Human Genome Research (NCHGR) in 1989.

In 1990, there was a joint publication entitled "Understanding Our Genetic Inheritance: The U.S. Human Genome Project, The First Five Years, FY 1991–1995" [12]. At this point the project completion of the human genome was slated for 15 years and the project was described as an international effort to develop genetic and physical maps and determine the DNA sequence of the human genome and the genomes of several model organisms. An updated plan was published in 1993 [13] by Francis Collins who was the Director of the NCHGR. This report stressed that continued development and improvements in technology would be required to keep the project on schedule. This project represented a truly international effort with an unprecedented amount of collaboration and extensive data, information, and resource sharing.

In 1998, whole genome shotgun sequencing was proposed to increase sequencing data acquisition [14]. This method entails randomly breaking up genomic DNA and then cloning the fragments into vectors before sequencing. This method obviates the need to increase computational power in order to make sense of the raw data, assemble contiguous blocks (contigs), and align the sequences.

The 2001 publication of the draft human genome [15] covered approximately 90 % of the genome with the remaining unsequenced bases predominantly located in tightly condensed heterochromatin. The focus then became on "finishing" the genome and announcement of the final human genome came in 2003 and coincided with the 50th anniversary of the Watson and Crick discovery of the DNA double helix [1]. The paper was published in 2004 [16] and the published sequence contained approximately 99.7 % of the euchromatic genome, was interrupted by only

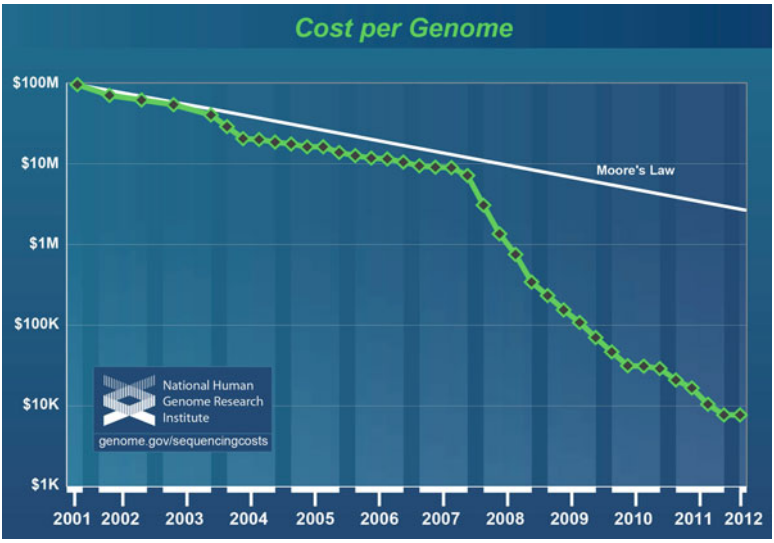

Fig. 1.4 Graph of sequencing costs per genome (estimated on a genome the relative size of the human genome). Moore's law (hypothetical computer industry that predicts the doubling of technology every 2 years) is illustrated on the graph. The abrupt drop in sequencing costs in 2008 represents the change from Sanger-based (dideoxy chain termination) sequencing to the "second-" or next-generation sequencing technologies (Data from NIH [17])

300 gaps, and contained only one nucleotide error per 100,000 bases. There were 28 megabases (Mb) of gaps consisting of euchromatic sequence, which were predominantly repetitive regions of the genome, and 200 Mb of heterochromatic sequence including the large centromeres and the short arms of acrocentric chromosomes.

The Human Genome Project took 13 years and was estimated to cost approximately $3 billion. In order to translate the data and make resequencing and personalized genomic analysis achievable, the cost per genome must be reduced to what is optimistically referred to as the $1,000 genome. Figure 1.4 is a current graph provided by the NIH on the costs of sequencing a genome. These figures do not reflect "nonproduction" costs which include the following:

- Quality assessment/control for sequencing projects
- Technology development to improve sequencing pipelines
- Development of bioinformatics/computational tools to improve sequencing pipelines or to improve downstream sequence analysis
- Management of individual sequencing projects
- Informatics equipment
- Data analysis downstream of initial data processing (e.g., sequence assembly, sequence alignments, identifying variants, and interpretation of results)

The next section touches on the changes in technology that have resulted in the amazing drop in cost per genome.

4 Next-Generation Sequencing

After the announcement of the completion of the first sequenced human genome, the NIH began a program that would quickly reduce the cost per genome. In 2004, the cost for a single genome was still greater than $10 million (Fig. 1.4). The goal was to promote the development of technologies that would increase the throughput of sequencing – translation of these technologies into the medical arena and therapeutics relies on reduction of the turn-around-time (patients can't very well wait for 13 years to get results). Additionally, the throughput goal is invariably linked to decreasing costs per genome.

In 2004, the first projects receiving grants included 11 groups chosen to begin technology "near term" development and attempt to reduce the cost to $100,000/genome. The second set of projects were awarded to seven groups that were tasked with longer-term development and reduction of the per genome cost to $1,000. Technologies involved in these endeavors ranged from sequencing-by-synthesis to nanopore DNA sequencing.

The following discussions will lightly cover the different technologies and commercial platforms that have been and/or are important in the development of next-generation sequencing. More detailed discussion will follow later in this book.

4.1 Technologies

The abrupt reduction of cost per genome shown in Figure 1.4 points to the significance and return on investment of the NIH effort to jump start technology development. In 2004, the cost of a genome was greater than $10 million and now the average cost per genome is $5,000–$8,000. Turnaround time (TAT) and cost continue to be the object of intense developmental effort.

How did this happen? Simultaneous improvements in chemistry, engineering, and instrumentation development combined to allow methods that have been dubbed "massively parallel sequencing." The following sections will touch on PCR-based next-gen sequencing, single-molecule next-gen sequencing, and the promising future.

4.1.1 PCR-Based Sequencing

The major players in next-generation or "massively parallel" sequencing came to the market with sequencing technologies that rely on PCR-based amplification to allow signal detection. 454 Life Sciences (Roche Diagnostics), SOLiD (Applied Biosystems), and Illumina sequencing rely on either emulsion PCR or bridge amplification (cluster) to generate enough molecules to make the sequencing process more easily imaged. Of course, PCR bias introduced into the system by known

issues with amplification such as preferential amplification or uneven amplification requires the users to understand that although each bead emulsion amplification or cluster amplification (although amplifying only a single molecule) can exhibit bias.

454 Life Sciences (Roche): Sequencing-by-Synthesis (Pyrosequencing)

Next-generation sequencing in a massively parallel fashion [18] was commercialized by 454 Life Sciences, originally named 454 Corporation, as a subsidiary of CuraGen Corporation. In 2007, Roche Diagnostics purchased the company and became the sole provider of genome sequencers utilizing the 454 pyrosequencing method [19].

The technology involves generation of a single-stranded template DNA library, emulsion-based clonal amplification of the library, data generation via sequencing-by-synthesis, data analysis using different bioinformatics tools. Purified DNA (or cDNA) undergoes fragmentation and then ligation with adapters that will aid in further purification, quantification, amplification, and sequencing of the DNA library.

The next step requires that the adapter library be attached to capture beads and subsequently emulsified to generate "microreactors" that will amplify a single fragment in a "one fragment = one bead" reaction. The emulsified beads then undergo emulsion PCR to generate millions of clonally amplified fragments on each bead. The emulsification is broken and the beads are loaded into a PicoTiterPlate where each well of the plate contains a single bead and thus millions of clonal amplification products from a single fragment. After placing the loaded PicoTiterPlate into the sequencing instrument, nucleotides are "flowed" across the plate in sequence (A, C, G, T). As each nucleotide complementary to the template is incorporated, a chemiluminescent light is emitted and recorded by the instrument camera (Fig. 1.5).

Applied Biosystems SOLiD: Sequencing-by-Hybridization/Ligation (Fluorescent Detection)

Massively parallel sequencing by hybridization-ligation from Applied Biosystems (SOLiD) is based on the chemistry of the polony method of sequencing published by Shendure et al. [20]. Sequencing library preparation begins with emulsion PCR-based amplification of a single molecule linked to a bead in a method similar to the 454 method described above. The beads are then deposited on a glass slide and sequencing occurs by multiple rounds of hybridization and ligation of di-base probes (Fig. 1.6) fluorescently labeled with four different fluorescent dyes. The method interrogates two nucleotides by using the four-dye schema and the sequence is determined by identifying the color resulting from successive reactions. This method allows distinction of a sequencing error and a true sequence polymorphism.

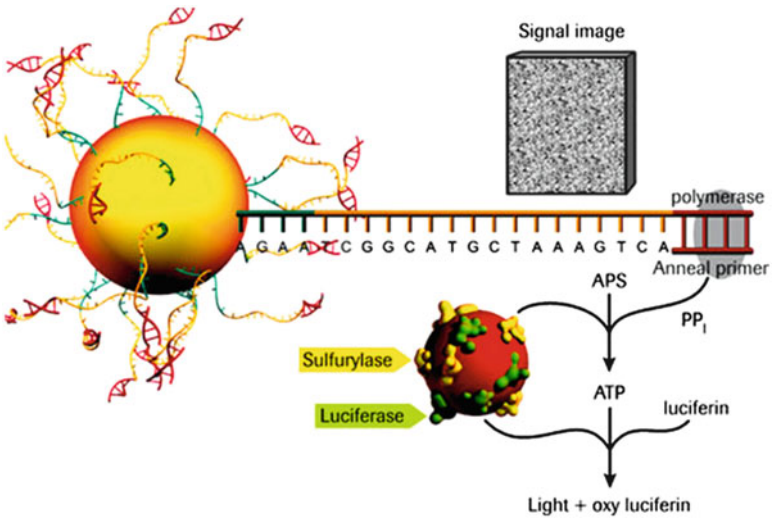

Fig. 1.5 454 Pyrosequencing reaction (Image used with permission: 454 Sequencing © Roche Diagnostics)

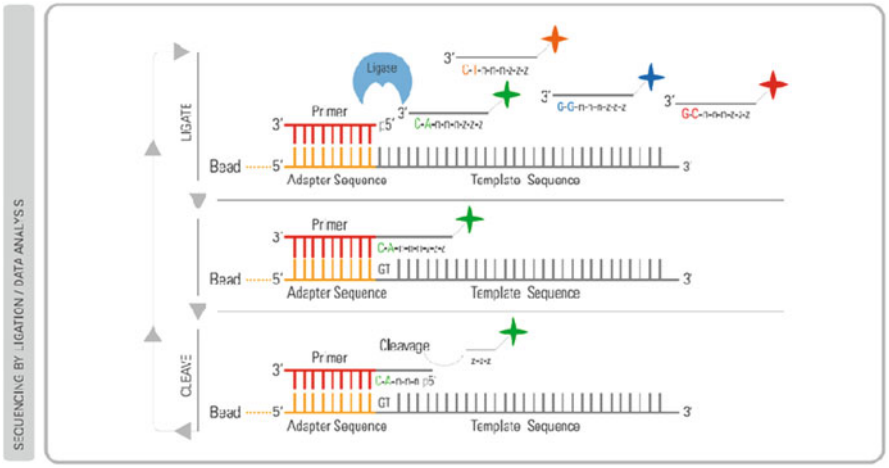

Fig. 1.6 Sequencing by hybridization-ligation (Image used with permission from Life Technologies)

Illumina (Solexa technology): Sequencing-by-Synthesis with Reversible Fluorescent Terminators

Illumina Inc. acquired Solexa Ltd. in 2006 to gain control of Solexa's proprietary next-generation genetic analysis system. The approach, unlike 454 and SOLiD, does not utilize emulsion PCR but instead uses bridge amplification of single-molecule DNA strands [21–23]. In this method, adapter sequences are ligated to

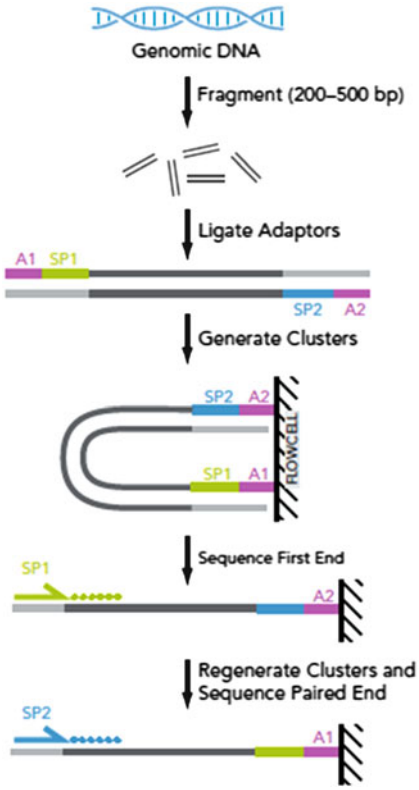

Fig. 1.7 A. Library preparation and bridge amplification of fragments on an Illumina flow cell. Paired End sequencing is illustrated (Image used with permission from Illumina Inc.)

fragmented DNA. Denatured single-strand DNA molecules are attached to a solid surface called a flow cell and the molecules are amplified into clusters by solid-state bridge amplification. Each cluster is composed of about 1,000 clonal copies of the molecule. The sequencing occurs in a massively parallel fashion using sequencing-by-synthesis and sequential addition of reversible fluorescent dye terminators (Fig. 1.7).

Life Technologies Ion Torrent/Proton: Sequencing-by-Synthesis (pH Detection)

Life Technologies Corp. acquired Ion Torrent Corp. in 2010 in an effort to enhance the DNA sequencing offerings of the company. Ion Technology offers what has been coined "PostLight™" technology: semiconductor sequencing and pH detection. Like 454 and SOLiD, emulsion PCR is used to generate sequencing libraries bound to beads. The beads are flowed across a semiconductor chip (Fig. 1.11) that consists of wells sized to accept a single bead and an ion sensor that will read pH. As nucleotides are incorporated into the template strand, a hydrogen ion (H^+) is released as a by-product. The ion's charge is detectable by the ion sensor in the chip and as nucleotides are added, the voltage change is detected and data is displayed as a peak of voltage (Fig. 1.8).

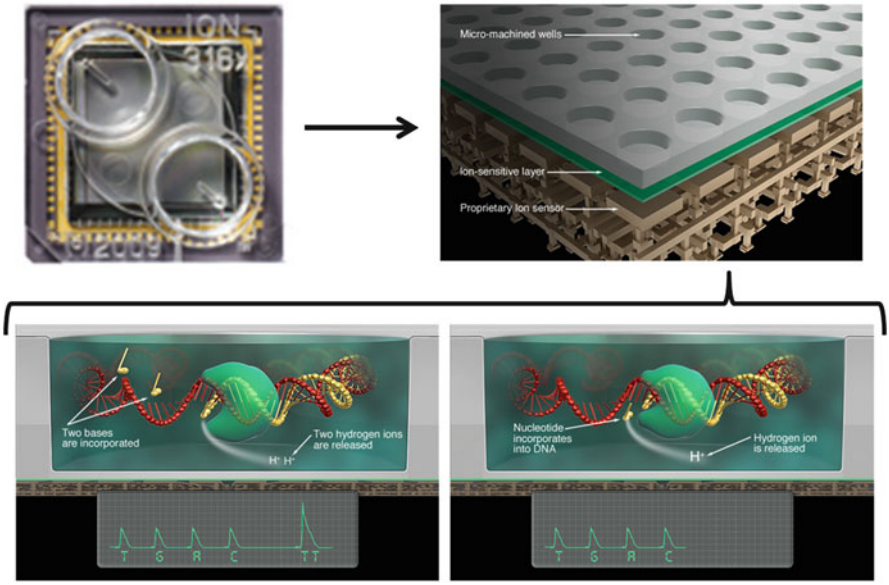

Fig. 1.8 Ion semiconductor sequencing. As nucleotides are flowed across the chip, incorporation results in the release of a hydrogen ion that is detected by the chip. If identical nucleotides are incorporated (homopolymer), each H^+ released will increase the signal detection. In this example two bases are incorporated and the readout shows concurrent increase in signal peak (Images used with permission Life Technologies (Ion))

Complete Genomics: Sequencing-by-Hybridization/Ligation (Fluorescent Detection)

Unlike the platforms described above, Complete Genomics Inc. does not produce a commercialized instrument for sale to consumers. Instead, the company provides a whole human genome sequencing service. The company has created a sequencing center that will accept genomic DNA and deliver sequencing data to the customer. Several diverse technologies were integrated to produce the Complete Genomics sequencing solution. The sequencing technology generates a total read length of 35 bases. The methodology follows.

Sequencing library construction begins with fragmentation of genomic DNA and the addition of four adapter sequences to the fragment. Fragments are circularized and a head-to-tail concatemer of approximately 200 fragments is generated with a rolling circle amplification method. The concatemers are formed into ball shapes (DNA nanoballs or DNBs) and the DNBs are flowed across an array with "sticky" spots that allow the binding of only a single DNB.

Sequencing occurs with a combination hybridization-ligation reaction using Complete Genomics' proprietary combinatorial Probe-Anchor Ligation (cPAL™). cPAL™ uses pools of probes labeled with four different dye molecules to perform sequencing that allows 10 bases to be read starting with each adapter sequencing (Fig. 1.9).

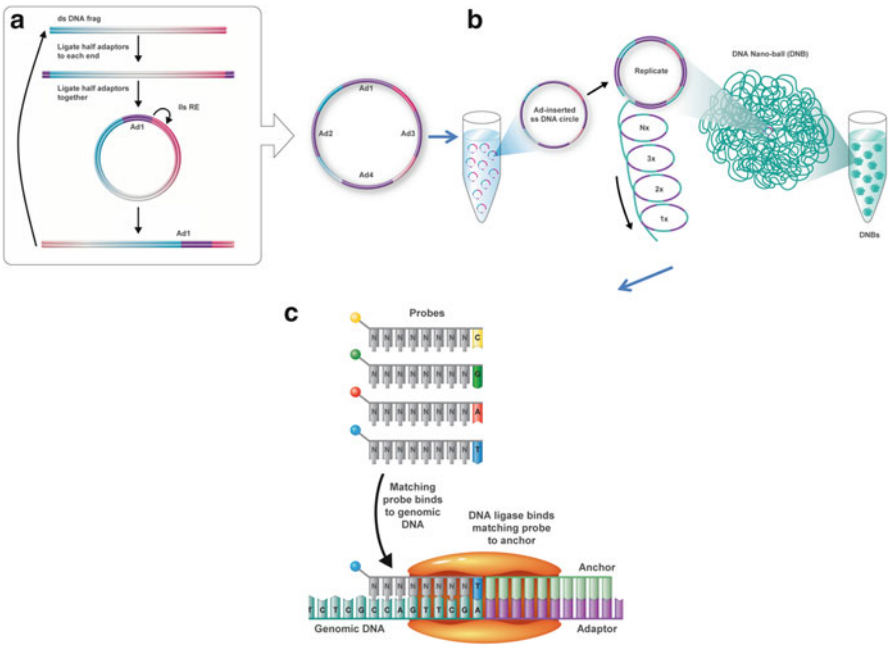

Fig. 1.9 (**a**) Multiple adaptor library construction process. (**b**) DNA nanoball (DNB) formation. (**c**) Combinatorial Probe-Anchor Ligation (cPAL™) (Images used with permission from Complete Genomics)

4.1.2 Single-Molecule Sequencing

Of course, the Holy Grail in the field is fast, inexpensive, accurate single-molecule (non-PCR-based) sequencing. Several commercial groups have marketed technologies for single-molecule sequencing including Helicos and Pacific Biosciences. Other possibilities include nanopore-based sequencing and we have hopes that this technology will achieve commercial success.

Pacific Biosciences

Pacific Bioscience markets SMRT® (Single Molecule Real Time) sequencing technology. This platform brings together different technologies to allow very long sequencing reads in mere minutes (Fig. 1.10). A SMRT cell (glass slide coated with 100 nm metallic film that has been fabricated to hold approximately 75,000 small holes only 10 s of nanometers in diameter – the ZMW or zero-mode waveguide). DNA polymerase is immobilized to the floor of the ZMW and the cell is flooded with phospholinked nucleotides. Because the ZMW is smaller than the wavelength of the light, only the bottom of the ZMW is illuminated allowing detection of the individual nucleotides as the DNA polymerase incorporates them into a growing DNA chain.

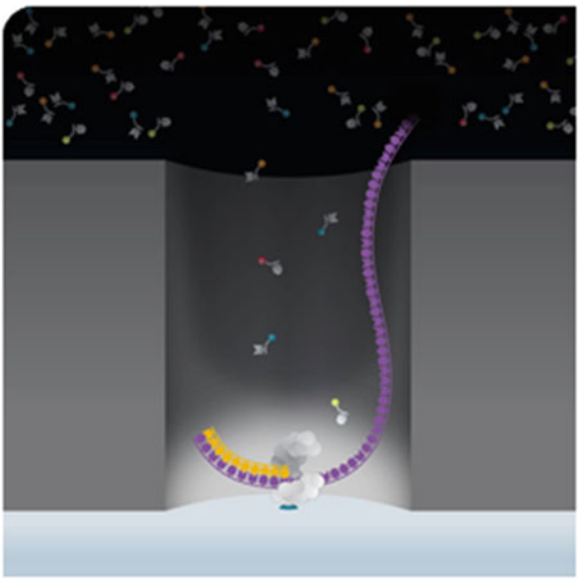

Fig. 1.10 With an active polymerase immobilized at the bottom of each ZMW, nucleotides diffuse into the ZMW chamber. In order to detect incorporation events and identify the base, each of the four nucleotides A, C, G, and T is labeled with a different fluorescent color. Since only the bottom 30 nm of the ZMW is illuminated, only those nucleotides near the bottom fluoresce (Image and figure legend used with permission from Pacific Biosciences)

5 Conclusion

Sequencing technologies have come a long way in a very short time and advances in material sciences and other innovative technology breakthroughs continue to occur that can offer faster, cheaper, more accurate sequencing. Sequencing is already in the clinic and consumer demand for better testing with faster turnaround will require that these advances be incorporated and implemented. The fast evolution of these technologies is unprecedented and exciting.

References

1. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171(4356):737–738
2. www.nobelprize.org (1962) The Nobel prize in physiology or medicine 1962. http://www.nobelprize.org/nobel_prizes/medicine/laureates/1962/index.html
3. www.nobelprize.org (1980) The Nobel prize in chemistry 1980. http://www.nobelprize.org/nobel_prizes/chemistry/laureates/1980
4. Gilbert W, Maxam A (1973) The nucleotide sequence of the lac operator. Proc Natl Acad Sci USA 70(12):3581–3584
5. Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol 94:441–448
6. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci USA 74(2):560–564

7. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(2):5463–5467
8. Maxam A, Gilbert W (1980) Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol 65:499
9. Ambrose BJB, Pless RC (1987) DNA sequencing: chemical methods. Methods Enzymol 152:522
10. Sears LE, Moran LS, Kissinger C, Creasy T, Perry-O'Keefe H, Roskey M, SUtherland E, Slatko BE (1992) CircumVent thermal cycle sequencing and alternative manual and automated DNA sequencing protocols using the highly thermostable VentR (exo-) DNA polymerase. Biotechniques 13(4):626–633
11. Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. Science 238(4825):336–341
12. Energy USDoHaHSaDo (1990) Understanding our genetic inheritance. The U.S. human genome project: the first five years. National Institute of Health, Washington, DC
13. Collins F, Galas D (1993) A new five-year plan for the U.S. human genome project. Science 262(5130):43–46
14. Venter JC, Adams MD, Sutton GG, Kerlavage AR, Smith HO, Hunkapiller M (1998) Shotgun sequencing of the human genome. Science 280(5369):1540–1542
15. Consortium IHGS (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
16. Consortium IHGS (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945
17. Wetterstrand KS (2013) DNA sequencing costs: data from the NHGRI genome sequencing program (GSP). www/genome.gov/sequencingcosts. Accessed January 2013
18. Leamon JD, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, deWinter AD, Berka J, Lohman KL (2003) A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. Electrophoresis 24(21):3769–3777
19. Diagnostics R (2007) Roche acquires 454 Life Sciences to strengthen presence in ultra-fast sequencing
20. Shendure J, Porreca PJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741):1728–1732
21. Bennett S (2004) Miscellaneous Solexa Ltd. Pharmacogenomics 5(4):433–438
22. Bennett ST, Barnes C, Cox A, Davies L, Brown C (2005) Toward the $1000 human genome. Pharmacogenomics 6(4):373–382
23. Bentley DR (2006) Whole-genome re-sequencing. Curr Opin Genet Dev 16(6):545–552

Chapter 2
Clinical Molecular Diagnostic Techniques: A Brief Review

Megan L. Landsverk and Lee-Jun C. Wong

Abstract The identification and characterization of the genetic basis of disease is often fundamental to diagnosis. Detection of pathogenic mutations in a DNA sample can lead to a diagnosis, possible prognosis, and prospective therapy treatments. Over the years, a variety of molecular biology techniques have been utilized in clinical diagnostic laboratories in the analysis of patient samples. The recent development of next-generation sequencing (NGS) techniques has revolutionized the field of clinical molecular diagnostics. In this chapter, we review the development of molecular diagnostic approaches and some of the most commonly used assays prior to the NGS era. Although PCR-based methods are the most commonly used assays in molecular diagnostics today, a number of caveats must be taken into consideration and are also discussed.

1 Introduction

Genetics and the study of the human genome have become an integral part of medicine and public health. Determining the full molecular characteristics of genetic disorders provides additional information in the diagnosis of a patient. In addition, identification of familial mutations leads to appropriate genetic counseling for families and possible prenatal diagnosis or preimplantation genetic diagnosis (PGD) for future pregnancies. The field of clinical molecular diagnostics has grown considerably in the last couple of decades, benefitting from advancements in

M.L. Landsverk (✉)
Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA
e-mail: meganl@bcm.edu

L.-J.C. Wong
Medical Genetics Laboratories, Department of Molecular and Human Genetics,
Baylor College of Medicine, One Baylor Plaza, NAB 2015, Houston, TX 77030, USA

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_2,

human genetics basic research and technologies. In the early years, research laboratories primarily developed the techniques used to analyze genetic mutations. Many of these assays were then implemented into the clinical molecular diagnostic repertoire. The earliest assays were generally targeted to common disorders such as hemoglobinopathies and cystic fibrosis. These early molecular diagnosis methodologies often involved indirect mutation detection through haplotype and linkage analyses, which are extremely laborious, required large amounts of patient DNA, generally required extensive knowledge of the genomic region in question, and did not always result in an easily interpretable result. Nevertheless, they provided a foundation for molecular diagnostics as we know it today, and some of these techniques are still currently in use.

The discovery of polymerase chain reaction (PCR) essentially revolutionized molecular diagnostics. First described by Mullis et al. in 1986 [1], PCR provided the ability to produce many copies of a target DNA region, allowing for faster analysis and direct mutation identification. Assays that were in use prior to the discovery of PCR were quickly modified to incorporate the use of PCR-amplified DNA rather than genomic DNA. These allele-specific detection assays rapidly developed into high-throughput systems to analyze patient samples on a larger scale. The implementation of PCR-based assays also provided laboratories with a means to analyze rare disorders in addition to common ones. Today, with resources such as the 1000 Genomes Project, a detailed catalogue of human genetic variation, diagnostic molecular laboratories have access to the sequence of all human genes and a continuously growing database of human variation. While next-generation sequencing (NGS) technologies are becoming more and more popular, automated Sanger sequence analysis appears to presently be the most common technique for analysis of many genetic disorders in clinical molecular diagnostic laboratories. However, the assay choice often depends on the gene or alleles of interest and the volume of patients to be screened. In general, most current molecular diagnostic assays are either targeted to specific alleles or analyze particular genes or groups of genes if there is no specific allele of interest. Here we describe some of the more common molecular techniques used in the analysis of both known and unknown mutations (Table 2.1) and discuss possible pitfalls of conventional PCR-based methodologies.

2 Targeted Analyses

Allele-specific mutation detection methods were the first assays implemented in clinical diagnostic laboratories. The initial techniques were developed in the early 1980s and some are still in regular use in clinical laboratories today. These assays are attractive in their ease of use and most are easily convertible to high-throughput applications. However, they can only be used to detect known mutations and polymorphisms and therefore need to be combined with additional assays if full comprehensive mutation detection is required.

Table 2.1 Select techniques used in a variety of clinical molecular diagnostic laboratories

Select techniques used in clinical diagnostic laboratories
Targeted mutation analysis
Southern/restriction fragment length polymorphism (RFLP)
Allele-specific oligonucleotide (ASO)
Allele refractory mutation system (ARMS)
Oligonucleotide ligation assay (OLA)
Pyrosequencing
Real-time PCR
Sanger sequence analysis (if mutation is know)
Detection of unknown mutations
Gradient gel electrophoresis (GGE)/denaturing (DGGE) and temperature (TGGE)
Single-strand conformation polymorphism (SSCP)
Heteroduplex analyses (HDA)
Denaturing high-performance liquid chromatography (DHPLC)
Protein truncation test (PTT)
Sanger sequence analysis
Detection of copy number variations
Southern blot
Multiplex ligation-dependent probe amplification (MLPA)
Array comparative genomic hybridization (aCGH)
Single-nucleotide polymorphism (SNP) arrays

2.1 Restriction Fragment Length Polymorphism

One of the first techniques utilized in clinical molecular diagnostics was the detection of genomic changes using Southern blotting and restriction fragment length polymorphism (RFLP). The Southern blot transfer hybridization assay was developed in 1975 [2]. Around that same time, cDNA synthesis and cloning provided the ability to determine the primary sequence of a number of genes [3]. Some of the first studies using cloned human cDNA were to identify the nucleotide sequences of the human alpha, beta, and gamma globin genes [4]. When combined with RFLP, the availability of the sequence of these genes provided a means to map normal and mutant genomic DNA. For example, genetic variations in a restriction enzyme site close to the beta-globin structural gene were identified only in people of African origin [5]. These polymorphic sites were then used in the diagnosis of sickle-cell anemia. These early studies set the stage for the use of RFLP and Southern blotting in diagnostic tests such as linkage analysis and prenatal diagnosis. Disorders such as the thalassemias, cystic fibrosis, and phenylketonuria were among the first to be

described [6, 7]. However, in the early days to detect mutations using RFLP was laborious. To identify a disease causing mutation in a gene meant first cloning the gene in question to create probes for Southern blot analysis. Genomic DNA was then digested with a variety of restriction enzymes and probed for products that were polymorphic in size. If carrier parents of an affected proband could be identified, the polymorphic fragment sizes could then be used for prenatal diagnosis. It was also possible to further analyze families that were non-informative by single enzyme digestion using a combination of restriction enzymes to determine their haplotype and possible carrier status. This technique was used to identify different forms of beta-thalassemia by taking advantage of the fact that specific mutations were generally found on a particular haplotype background [8]. This analysis method avoided the repeated isolation of the same mutation by selecting genes based on their associated haplotype.

Since RFLP was already a mainstay in molecular analysis when PCR was developed, PCR amplification of a region of DNA followed by RFLP quickly became a widely used approach. In this case, the mutation of interest was known and an enzyme that cuts at the mutation site was used. Patients that were carriers could be distinguished from those that were either homozygous wild type or homozygous mutant by the banding pattern of the PCR products on a gel. Some of the first applications of PCR-based RFLP analysis were in the characterization of sickle-cell anemia alleles [9].

2.2 Allele-Specific Oligonucleotide Hybridization

Allele-specific oligonucleotide (ASO) hybridization, or dot blot analysis, was also an early approach to detect specific mutations in particular disorders. This assay is based on the principle that when probing a region of DNA, even a single-base-pair change between a target region and the probe can destabilize the hybrid. In general, two synthetically created probes are designed to the region of interest, one complementary to the wild-type allele, one complementary to the mutant allele. The digested DNA is separated by gel electrophoresis and immobilized on a membrane. It is then hybridized with radioactively labeled probes. If both probes react, then the individual is heterozygous for the mutation of interest, if only one probe reacts then that individual is homozygous for either the wild-type or mutant allele. This technique was used in the early 1980s in the detection of the sickle-cell allele [10] and prenatal diagnosis of β-thalassemia [11].

ASO was also a technique that benefitted from the use of PCR. Instead of probing non-amplified genomic or cloned DNA, regions of interest could be PCR amplified first and then probed. The addition of PCR to ASO allowed for a more rapid detection of mutations [12] and became one of the more widely used techniques to study targeted alleles in the mid-1980s. Early on, the ASO probes were labeled with radioactivity; however, subsequent protocols used probes that were conjugated to biotin. This allowed for detection using streptavidin conjugated to horseradish

peroxidase (HRP) and colorimetric or chemiluminescent detection without the use of radioactivity. The original method of ASO is generally known as forward ASO, where patient DNA is immobilized on a membrane and hybridized with probes targeted to a specific allele. This technique is most useful when a large number of patient samples are to be screened for a small number of mutations. However, each oligonucleotide probe must be labeled separately and as the number of mutations in the screen increases, the assay becomes more complex. Reverse ASO was developed as a solution to this problem. In reverse ASO, also known as a reverse dot blot, the probes are immobilized onto the membrane and the PCR-amplified patient DNA sample is hybridized to the membrane. This allows for multiple mutations in a variety of genes to be assayed simultaneously in a single patient sample.

2.3 Amplification Refractory Mutation System

Amplification refractory mutation system (ARMS) is a PCR-based method developed in the late 1980s also for the analysis of known point mutations. ARMS is based on the fact that DNA amplification is inefficient if there is a mismatch between the template DNA and the 3′ terminal nucleotide of a PCR primer [13]. A primer with a 3′ terminal nucleotide that is complementary to the wild-type allele will not have efficient extension when a mutation is present and vice versa. Therefore, one can differentiate between two alleles by simple PCR amplification. The design and optimization of ARMS assays is primarily a function of the alleles of interest and the nucleotides surrounding them. Often, incorporating additional mismatched nucleotides near the target allele can enhance the reaction [13]. Multiple sets of primer pairs can be used simultaneous in a single tube allowing for the analysis of many mutations at one time. This particular technique has been used to identify patients that carry known mutations in many disorders such as cystic fibrosis and phenylketonuria and to determine heteroplasmy levels of mitochondrial mutations [14].

2.4 Oligonucleotide Ligation Assay

The oligonucleotide ligation assay (OLA) combines PCR with ligation in one reaction at a target allele site. After PCR amplification around the target region is performed, three oligonucleotides are added to the reaction. One, generally known as the reporter, is a common probe that is complementary to the target DNA sequence immediately 3′ to the allele of interest. The other two "capture" probes are complementary to the target DNA sequence immediately 5′ to the target allele and differ only in their final 3′ terminal nucleotide which is the target allele. Only if there is a perfect match between the capture probe and the target allele can ligation between the capture probe and the reporter probe occur. A number of different detection methods for OLA have been developed including detecting different lengths of

ligated products for the two target alleles and alternate labels on the capture probes such as fluorescence or biotin [15]. While optimization of the assay is often required, the detection methods of OLA allow for rapid and sensitive detection of alleles in a high-throughput capacity at a decreased expense. The OLA has been utilized in the detection of mutations in a number of metabolic disorders, cystic fibrosis, and pharmacogenetics [16–18].

2.5 Pyrosequencing

Pyrosequencing is a DNA sequencing technology based on real-time detection of DNA synthesis monitored by luminescence. First described in 1985 as an enzymatic method for continuous monitoring of DNA polymerase activity [19] and modified in subsequent years to optimize the reaction [20–22], the assay is based on a reaction in which each sequential nucleotide incorporated during DNA synthesis releases a pyrophosphate. ATP sulfurylase converts that pyrophosphate to ATP in the presence of adenosine 5′ phosphosulfate. That ATP then drives a luciferase-mediated conversion of luciferin to oxyluciferin that generates visible light that can be measured. Unincorporated dNTPs are degraded by apyrase. The nucleotides are added in a specific order such that there is an expected pattern for the wild-type or mutant allele. When compared to other sequencing techniques such as Sanger sequencing, pyrosequencing offers the advantage of short read length when analyzing genetic variants for applications such as SNP genotyping or detection of known mutations. PCR fragments can be small and the assay is relatively fast in which 96 samples can be processed in approximately 20 min. Pyrosequencing is used in clinical laboratory settings for a variety of tests including pharmacogenetic testing in the analysis of polymorphisms within genes such those involved in drug metabolism [23, 24]. A quick analysis of these polymorphisms provides information on whether a patient may be a poor or rapid metabolizer of a particular drug, allowing clinicians to make more informed choices as to patient dosage.

2.6 Real-Time PCR

All of the previously discussed assays require post-PCR manipulation. In the mid-1990s, a technique involving the analysis and quantification of DNA or RNA in real time was developed [25, 26]. This sensitive assay allows for accurate quantification of a PCR product during the exponential phase of PCR. The first reports of real-time PCR were performed using hydrolysis or TaqMan probes [25, 26]. These probes specifically hybridize to the region around a target allele internal to the primer binding sites. The TaqMan probe is generally labeled on each end with a fluorescent molecule, a reporter dye and a quencher. As long as the two are in close proximity, the quencher prevents the reporter from fluorescing. As the PCR cycle progresses,

the exonuclease activity of Taq polymerase degrades the probe and the fluorophore become separated from the quencher allowing fluorescence emission which can then be measured. The increase in fluorescence is measured at every cycle and directly correlates to the amount of PCR product formed [26, 27].

Another method utilized in real-time PCR is the use of fluorescent DNA intercalating dyes. The first use of this method measured the increase in ethidium bromide fluorescence in double-stranded DNA molecules and was referred to as kinetic PCR [25]. In later years, SYBR Green I was used since it incorporates into double-stranded DNA and is less toxic than ethidium bromide. As the amount of double-stranded DNA increases exponentially during the PCR reaction, the amount of dye incorporation and emission also increases and is measured.

There are multiple pros and cons in the use of either TaqMan probes or SYBR Green dye in real-time PCR. In a TaqMan assay, specific hybridization between probe and target is required for fluorescent signal, which greatly decreases background and false positives. In addition, probes can be labeled with different reporters so two distinct assays can be performed in one tube. However, individual probes must be constructed for every allele of interest, so it can be costly. Off-the-shelf kits containing probes for a variety of disorders are commercially available. An advantage to using SYBR Green is that since no special probes are required, the cost is much cheaper. However, SYBR Green will bind to any double-stranded DNA species including nonspecifically amplified products leading to an increase in background and false positives. The reproducibility and accuracy of real-time PCR assay is also highly dependent on factors such as normalization of samples and controls. Regardless, the ability to quickly measure factors such as DNA copy number in real time with little DNA manipulation makes this assay common in molecular diagnostic laboratories.

3 Detection of Unknown Mutations

All of the techniques described in the previous section require a prior knowledge of the mutation in question and the nucleotide sequences around it. Here we describe techniques that were developed to screen unknown changes in targeted genomic regions.

3.1 Gradient Gel Electrophoresis

Gradient gel electrophoresis (GGE), including temperature (TGGE) and denaturing (DGGE), is based on the principle that the electrophoretic mobility of double-stranded DNA fragments is altered by their partial denaturation. The technique was first used in characterizing human mutations in the mid-1980s when it was applied to detect β-thalassemia mutations [28]. At that time both RFLP-Southern blotting

and ASO were being used to analyze single-base-pair mutations leading to a disease state or polymorphisms linked to mutant alleles. Still, many base pair substitutions did not lead to altered restriction sites and using ASO probes required knowledge of the DNA sequence around the allele of interest. In addition, as more mutations were identified in disorders such as β-thalassemia, the number of probes required for ASO continued to increase. GGE allowed for the detection of allelic changes without the requirement of knowing the exact DNA sequence of the region in question and multiple nucleotide changes in a single region could be simultaneously screened. The initial GGE assays were performed using digested genomic DNA mixed with a synthesized oligonucleotide probe of the region of interest. In later years, amplified PCR fragments of the region of interest were used. These DNA fragments are denatured then re-annealed, followed by the analysis on denaturing gradient gels. Fragments move through the gel based on their melting temperatures (T_m). Since the T_m is dependent on the overall DNA sequence, even a single-nucleotide substitution can alter the dissociation and mobility. Heteroduplexes of wild-type and mutant DNA fragments generally migrate slower than homoduplexes in polyacrylamide gels under denaturing condition due to mismatching of alleles and can therefore be separated by gradients of linearly increasing denaturant such as urea (DGGE) or temperature (TGGE). However, some base substitutions will not lead to a shift in position for the heteroduplex. As duplex DNA moves through the gradient, dissociation occurs in discrete regions known as "melting domains" that are 50–300 base pairs in size. All of the nucleotides in a particular region dissociate in an all-or-nothing fashion in a given temperature interval. If the mutation is located in the highest temperature region, or if the entire fragment dissociates as a single domain, no shift is observed.

Both DGGE and TGGE require a gradient of either denaturant or temperature. Temporal temperature gradient gel electrophoresis (TTGE) was first introduced by Yoshino et al. as a modification of TGGE [29]. In TTGE, the temperature of a gel plate increases gradually and uniformly with time which allows for easier temperature modulation. This increases the sensitivity as the separation range expands. One of the first reports showing successful application of this method to clinical diagnosis was in the detection of mutations in mitochondrial DNA [30]. Subsequently, TTGE has been used as a method of detection for germline mutations in a variety of disorders including cystic fibrosis [31] and somatic mutations in cancer tissues [32].

3.2 Single-Strand Conformation Polymorphism and Heteroduplex Analyses

Single-strand conformation polymorphism (SSCP) and heteroduplex analyses (HDA) were developed shortly after the introduction of PCR amplification [33, 34]. SSCP is based on the theory that single-stranded short DNA fragments migrate in a non-denaturing gel as a function of their sequence as well as size. In SSCP, during electrophoresis the single-stranded fragments adopt a unique conformation

depending on their nucleotide sequence. Even a single-base-pair change can alter the conformation leading to a change in migration on a gel. Fluorescent SSCP (F-SSCP) using fluorescently labeled PCR products and an automated DNA sequencer was developed in the early 1990s [35]. Its advantages included nonradioactive labeling of PCR products, greater reproducibility, and lower overall cost. HDA is also based on the migration of PCR products through a non-denaturing gel in a similar fashion to GGE in which heteroduplexes are analyzed in relation to homoduplexes. These heteroduplexes are formed by mixing denatured, single-stranded wild-type and mutant DNA PCR products, followed by slowly reannealing them to room temperature to form duplexes. These duplexes migrate differently depending on whether they are heteroduplexes of wild-type and mutant PCR fragments or homoduplexes of wild-type or mutant PCR fragments. Therefore, mutations can quickly be detected through simple gel migration analysis. One of the first reported uses of HDA in molecular diagnostics was in the detection of the p.F508del three base pair deletion in cystic fibrosis [36]. These techniques have been used for years in the detection of mutations in a number of disorders including a variety of cancers, phenylketonuria, and retinoblastoma [37–39].

3.3 Denaturing High-Performance Liquid Chromatography

Denaturing high-performance liquid chromatography (DHPLC) was first reported in 1997 [40] and was designed to combine some of the best features of methods currently available at that time. Sensitive methods such as DGGE were very labor intensive and required much optimization and analysis by gel electrophoresis whereas less complex methods such as SSCP and HDA lacked sensitivity. DHPLC was introduced as a highly sensitive method that facilitates the analysis of a large number of samples in a high-throughput capacity. Briefly, in a manner similar to HDA, DNA fragments are denatured and allowed to reanneal and homo- or heteroduplexes are formed. However, instead of gel electrophoresis, the DNA duplexes are applied to a positively charged chromatography column. The PCR fragments then bind to the column at different strengths depending on whether they are homo- or heteroduplexes and will elute from the column at different times generating distinct chromatograph patterns.

While relatively fast with easy automation and high specificity since no labeling or purification of the PCR products is required, DHPLC does have drawbacks. Each allelic change in any given PCR fragment will have a characteristic heteroduplex elution pattern. Although this technique has been used routinely to analyze a large number of samples for known mutations, it does not distinguish among different mutations in the same fragment. Therefore, its utility in clinical diagnostic laboratories, similar to other mutation detection methods, is limited to the detection of mutations, not the final identification, which will require Sanger sequencing. In addition, the elution conditions must be optimized for each assay in order to get the correct degree of denaturation and separation.

3.4 The Protein Truncation Test

The protein truncation test (PTT), also known as the in vitro synthesized protein assay (IVSP), does not rely on analysis of changes at the genomic level. Instead, this method is based on the change in size of proteins resulting from in vitro transcription and translation of a gene target [41]. Briefly, an RNA template is reverse transcribed to generate a cDNA copy. That cDNA is then amplified with primers specifically designed to facilitate in vitro transcription and translation. The resulting proteins are then analyzed by SDS-PAGE electrophoresis. Proteins of lower mass than the expected full-length protein represent translation products derived from truncating frameshift or nonsense mutations. The PTT was initially developed in the early 1990s to detect early termination mutations in the dystrophin gene responsible for Duchenne and Becker muscular dystrophy [42]. At that time, analysis of genes with as many exons as the dystrophin gene (79 exons over 2.4 Mb) was extremely time consuming and laborious, and a substantial number of cases were a result of truncating mutations. The most frequent application of the PTT is in detection of premature truncation mutations in cancer-causing genes in which many truncating mutations have been identified such as APC and BRCA1 [43, 44]. However, the PTT has a number of limitations and is not commonly used in most clinical diagnostic laboratories today. It only detects mutations that lead to truncated proteins, and missense mutations are not detected. Also, the requirement for electrophoresis of translation products does not translate easily to high-throughput analyses. In addition, the dependence on RNA as an amplification source precludes its easy use in most clinical diagnostic laboratories which generally work with genomic DNA. Although it is possible to use genomic DNA as a source, exons must be then analyzed individually. Therefore, while effective in the determination of truncation mutations in specific target genes, the PTT is not a commonly used test to screen for mutations in most genes.

3.5 Sanger Sequencing

Although all the mutation scanning methods described in the above sections are relatively easy to perform and fairly sensitive, they often require an extensive amount of design and optimization. In addition, any mutations detected by these scanning methods need to be ultimately confirmed by Sanger sequencing. Therefore, today, the capillary electrophoresis-based Sanger sequencing has become the most widely used approach for DNA analysis in molecular diagnostic laboratories. For a more in-depth review of the history and development of sequencing technologies, see Chap. 1. First described by Sanger et al. in 1977 [45], it has become the “gold standard” for mutation analysis particularly for very rare disorders and genes that do not harbor common mutations. In general, Sanger sequencing-based clinical assays use amplified PCR products of a particular region of interest. This may be a single-PCR product if a specific target allele is of interest or all coding exons plus flanking intronic sequences

of a gene. Each amplicon must then be sequenced independently. For clinical diagnostic laboratories, 2X coverage of a sequence is generally required. This is most often accomplished by sequencing once each in the forward and reverse directions. However, sometimes that is not possible due to repetitive sequence around the region in question, and two separate forward or reverse sequences are used. Therefore, while faster, safer, and often cheaper than some of the other techniques, it is still somewhat laborious with a high operating cost. In recent years, "next-generation" or massively parallel sequencing technologies have been developed that will provide clinical diagnostic laboratories the ability to offer analysis of multiple genes or even the whole exome at a cost that is competitive with single-gene Sanger testing.

4 Detection of Copy Number Variations

Chromosome analysis is important in the diagnosis of conditions such as intellectual disability, developmental delay, and congenital anomalies. Routine chromosomal analysis has the capacity to detect both balanced and unbalanced structural rearrangements as well as deletions and duplications larger than ~5 Mb in size, in addition to whole chromosome aneuploidy. However, conventional karyotype analysis is unable to detect submicroscopic deletions and duplications that are a common cause of intellectual disability. While Southern blotting is able to detect copy number changes in a number of genes, as previously discussed, it is very labor intensive, requires large amounts of DNA, and is only able to analyze a single region at a time. In addition, copy number changes in small regions such as single exon deletions may not be detected by Southern blot. Real-time PCR can be used to detect copy number changes in small regions; however, its use in a multiplex assay is also limited by the number of fluorescent dyes available and quantification of the data can be problematic if multiple primer pairs are desired. The requirement for the ability to analyze copy number variations across multiple exons/genes simultaneously is essential for disorders like DMD, BRCA1-related breast cancer, and mental retardation. Therefore, a variety of techniques have been developed to increase the resolution of detection for chromosomal alterations, such as multiplex ligation-dependent probe amplification (MLPA), array comparative genomic hybridization (aCGH), and single-nucleotide polymorphism (SNP) arrays. The clinical implementation of these assays has revolutionized the ability of diagnostic laboratories to detect copy number variations down to the exonic level in a multitude of genes simultaneously, with SNP arrays providing the additional ability to detect large regions of homozygosity in the genome.

4.1 Southern Blotting

As previously noted in Sect. 2.1, Southern blotting and RFLP were commonly used to track mutations in particular disorders. However, after the discovery of PCR,

mutations and deletions that previously required Southern blot mapping were routinely analyzed using PCR-based techniques. Even though RFLP analysis became PCR based, Southern blotting techniques still provide additional information for some diseases such as fragile X, although prescreening or tandem PCR analyses are often performed. The fragile X mental retardation syndrome was one of the earliest disorders in which Southern blotting and RFLP were used in clinical diagnostics. Mapped in the late 1980s and early 1990s using linkage and RFLP [46–50], analysis of the number of CGG repeat expansions and their methylation status in the 5′ untranslated region of the FMR1 gene has become one of the most common assays performed in clinical diagnostic laboratories today. PCR-based methods can be used to amplify the region containing the repeats and the size of the PCR product is therefore indicative of the number of repeats. However, the efficiency of the reaction is somewhat inversely related to the number of repeats and the larger the size, the more difficult it is to PCR. In addition, no methylation information is provided by PCR. Southern blotting allows both the size of the repeat region and its methylation status to be assayed at the same time. During restriction enzyme digestion, methylation-sensitive restriction enzymes can be used to distinguish between methylated and unmethylated species. Even though it is laborious and requires a large amount of DNA, the Southern blot is still used today in many clinical molecular diagnostic laboratories in the analysis of many diseases in particular trinucleotide repeat expansion disorders.

4.2 Multiplex Ligation-Dependent Probe Amplification (MLPA)

For many clinical diagnostic laboratories, MLPA is an attractive assay to use for the detection of copy number variations. MLPA has the advantage of analyzing multiple regions of interest simultaneously with a low operating cost requiring only a PCR thermocycler and capillary electrophoresis equipment. Briefly, MLPA is essentially a combination of two techniques: amplified fragment length polymorphism (AFLP) in which up to 50 different multiple DNA fragments are amplified in a single reaction with a lone primer pair and multiplex amplifiable probe hybridization (MAPH) in which multiple target oligonucleotide probes are hybridized to specific nucleotide sequences [51]. These probes are then also amplified with a single primer pair. However, similar to Southern, MAPH requires the immobilization of samples to a membrane and multiple washing steps to remove unbound probes. MLPA allows for the amplification of multiple oligonucleotide probes in a single reaction without the immobilization of sample to a membrane and removal of excess probe is not necessary. Each MLPA probe set consists of two oligonucleotides that hybridize to adjacent sides of the target sequence. Only when both oligonucleotides are hybridized to the correct nucleotide sequence can they be ligated into a single

probe. Therefore, only ligated probes are amplified using M13 primers at the 5′ends of the oligonucleotides. Each probe set gives rise to a unique amplification product of a particular size that can then be separated using capillary electrophoresis.

As previously discussed, using MLPA in a clinical laboratory setting is cost-effective and fast, and the universal tags allow multiple amplicons to be produced in a single reaction. However, MLPA does have its disadvantages. Due to the limits of multiplexing, each gene in a kit generally only has a limited number of probes per exon, with a maximum probe number of about 50 per reaction. SNPs in probe regions can cause a decrease in the binding efficiency of oligonucleotides resulting in false positives. Also, if a deletion is detected and the break points of that deletion are desired, further studies requiring additional PCR reactions are necessary. Finally, extensive design "rules" sometimes make the development of an MLPA assay difficult. Therefore, the inclusion of MLPA in a clinical diagnostic laboratory can increase the detection rate of mutations in a number of genes; however, the caveats listed above must be taken into consideration.

4.3 Array Comparative Genomic Hybridization (aCGH)

First described in the early 1990s [52], aCGH is now widely used for the identification and characterization of chromosomal abnormalities in many different cell types. The principle of aCGH analysis is the detection of chromosomal deletions and duplications by comparing equal amounts of genomic DNA from a patient and a normal control. Briefly, patient and control DNA are each labeled with a different fluorescent dye, typically Cy5 (green) for the patient and Cy3 (red) for the control. Equal amounts of labeled patient and control DNA are then mixed together and co-hybridized to the array, which is a microscope slide onto which small DNA fragments (targets) of known chromosomal location have been affixed. Current oligonucleotide arrays use targets made from short oligomers of approximately 60 base pairs in length. If the oligo density in a particular region is high enough, even small single exon deletions and duplications may be detected [53, 54]. Some of the first clinical diagnostic arrays were constructed using bacterial artificial chromosome (BAC) clones as targets [55]. At that time, constructing microarrays for use in a clinical laboratory was complicated due to the fact that mapping information for some BAC clones was inaccurate, and cross hybridization to multiple regions of the genome often occurred. While these first studies highlighted multiple challenges in using aCGH technology such as equipment costs, proper mapping and FISH confirmation of BAC clones, and interpretation of data, most of those challenges have now been overcome. Today, some clinical molecular diagnostic laboratories continue to use arrays based on BAC clones; however, cDNA clones, PCR products, or synthesized oligonucleotides immobilized on glass slides are increasingly more common [56].

4.4 Single-Nucleotide Polymorphism Arrays

Single-nucleotide polymorphism (SNP) arrays were originally designed to genotype human DNA by simultaneously analyzing thousands of SNPs across the genome [57]. Since their inception, SNP arrays have been used for a variety of other applications including detection of copy number changes and absence of heterozygosity. Like aCGH, SNP arrays are also based on oligonucleotide probes immobilized to glass slides. However, unlike CGH arrays that use both patient and control samples for comparison, SNP arrays use only a single patient DNA. The patient DNA binds to the oligonucleotide probes differently depending on the target SNP allele. Therefore, the resolution of the array is limited by SNP distribution. One major advantage of SNP arrays is their ability to detect copy number neutral differences in cases of absence of heterozygosity (AOH) that may occur as a result of uniparental isodisomy (UPD) or consanguinity (two copies), or deletion (one copy) such as loss of heterozygosity (LOH) associated with tumors. They can also detect copy number variants, but do not have the exon-by-exon coverage that most CGH arrays have nowadays.

5 Pitfalls of Conventional PCR-Based Methods

In most clinical molecular diagnostic laboratories, a single set of primers is used for PCR amplification of regions of interest. As a result, an SNP present within a primer site may disrupt the binding of that primer, and allele dropout could unknowingly occur. If a mutation is located within that region of interest, it may be missed. If a heterozygous deletion encompasses the region of amplification, only one chromosome will be amplified which would also result in incorrect analysis. Failure of allele amplification for one chromosome will also cause heterozygous mutations to appear homozygous. These problems can be minimized by continuous reassessment of the presence of SNPs in primer sites using the constantly updated dbSNP database. In addition, the identification of an apparently homozygous point mutation in an affected proband with an autosomal recessive disease should be followed up by parental testing whenever possible. If testing of the parents does not confirm their carrier status, additional molecular analyses can be performed to identify the underlying molecular etiology [58]. In general, allele dropout due to SNPs at primer sites should always be ruled out first for any PCR-based analyses. Capture-based next-generation sequencing will not have the problem of allele dropout since they do not rely on PCR primers. However, some regions of the genome may have poor coverage due to high GC content or the interference of pseudogenes, which will still require Sanger sequence analysis, and all positive results obtained by next-generation sequencing should be confirmed by a secondary method, which is usually Sanger sequencing.

6 Conclusions

With the advent of next-generation sequence analysis, we are entering a new era for molecular diagnostics. However, PCR-based testing methodologies still currently predominate most clinical molecular diagnostic laboratories. The choice of detection method used in the analysis of gene mutations depends on a variety of factors and can range from laboratory to laboratory. Sample volume, the spectrum of mutations in a given gene of interest, and equipment investment required can all play a role in what type of assays a molecular diagnostic laboratory chooses to perform.

References

1. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51(Pt 1):263–273
2. Southern EM (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol 98(3):503–517
3. Rougeon F, Mach B (1976) Stepwise biosynthesis in vitro of globin genes from globin mRNA by DNA polymerase of avian myeloblastosis virus. Proc Natl Acad Sci USA 73(10):3418–3422
4. Wilson JT, Wilson LB, deRiel JK, Villa-komaroff L, Efstratiadis A, Forget BG, Weissman SM (1978) Insertion of synthetic copies of human globin genes into bacterial plasmids. Nucleic Acids Res 5(2):563–581
5. Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proc Natl Acad Sci USA 75(11):5631–5635
6. Woo SL, Lidsky AS, Guttler F, Chandra T, Robson KJ (1983) Cloned human phenylalanine hydroxylase gene allows prenatal diagnosis and carrier detection of classical phenylketonuria. Nature 306(5939):151–155
7. Farrall M, Law HY, Rodeck CH, Warren R, Stanier P, Super M, Lissens W, Scambler P, Watson E, Wainwright B et al (1986) First-trimester prenatal diagnosis of cystic fibrosis with linked DNA probes. Lancet 1(8495):1402–1405
8. Orkin SH, Kazazian HH Jr, Antonarakis SE, Goff SC, Boehm CD, Sexton JP, Waber PG, Giardina PJ (1982) Linkage of beta-thalassaemia mutations and beta-globin gene polymorphisms with DNA polymorphisms in human beta-globin gene cluster. Nature 296(5858):627–631
9. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science 230(4732):1350–1354
10. Conner BJ, Reyes AA, Morin C, Itakura K, Teplitz RL, Wallace RB (1983) Detection of sickle cell beta S-globin allele by hybridization with synthetic oligonucleotides. Proc Natl Acad Sci USA 80(1):278–282
11. Orkin SH, Markham AF, Kazazian HH Jr (1983) Direct detection of the common Mediterranean beta-thalassemia gene with synthetic DNA probes. An alternative approach for prenatal diagnosis. J Clin Invest 71(3):775–779
12. Saiki RK, Bugawan TL, Horn GT, Mullis KB, Erlich HA (1986) Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. Nature 324(6093):163–166. doi:10.1038/324163a0

13. Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, Markham AF (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). Nucleic Acids Res 17(7):2503–2516
14. Venegas V, Halberg MC (2012) Quantification of mtDNA mutation heteroplasmy (ARMS qPCR). Methods Mol Biol 837:313–326. doi:10.1007/978-1-61779-504-6_21
15. Jarvius J, Nilsson M, Landegren U (2003) Oligonucleotide ligation assay. Methods Mol Biol 212:215–228
16. Schwartz KM, Pike-Buchanan LL, Muralidharan K, Redman JB, Wilson JA, Jarvis M, Cura MG, Pratt VM (2009) Identification of cystic fibrosis variants by polymerase chain reaction/oligonucleotide ligation assay. J Mol Diagn 11(3):211–215. doi:S1525-1578(10)60230-9 [pii] 10.2353/jmoldx.2009.080106
17. Bathum L, Hansen TS, Horder M, Brosen K (1998) A dual label oligonucleotide ligation assay for detection of the CYP2C19*1, CYP2C19*2, and CYP2C19*3 alleles involving time-resolved fluorometry. Ther Drug Monit 20(1):1–6
18. Chakravarty A, Hansen TS, Horder M, Kristensen SR (1997) A fast and robust dual-label nonradioactive oligonucleotide ligation assay for detection of factor V Leiden. Thromb Haemost 78(4):1234–1236
19. Nyren P, Lundin A (1985) Enzymatic method for continuous monitoring of inorganic pyrophosphate synthesis. Anal Biochem 151(2):504–509
20. Nyren P (1987) Enzymatic method for continuous monitoring of DNA polymerase activity. Anal Biochem 167(2):235–238. doi:0003-2697(87)90158-8 [pii]
21. Hyman ED (1988) A new method of sequencing DNA. Anal Biochem 174(2):423–436
22. Ronaghi M, Pettersson B, Uhlen M, Nyren P (1998) PCR-introduced loop structure as primer in DNA sequencing. Biotechniques 25(5):876–878, 880–872, 884
23. Soderback E, Zackrisson AL, Lindblom B, Alderborn A (2005) Determination of CYP2D6 gene copy number by pyrosequencing. Clin Chem 51(3):522–531. doi:clinchem.2004.043182 [pii] 10.1373/clinchem.2004.043182
24. Rose CM, Marsh S, Ameyaw MM, McLeod HL (2003) Pharmacogenetic analysis of clinically relevant genetic polymorphisms. Methods Mol Med 85:225–237. doi:10.1385/1-59259-380-1:225
25. Higuchi R, Fockler C, Dollinger G, Watson R (1993) Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. Biotechnology (N Y) 11(9):1026–1030
26. Gibson UE, Heid CA, Williams PM (1996) A novel method for real time quantitative RT-PCR. Genome Res 6(10):995–1001
27. Heid CA, Stevens J, Livak KJ, Williams PM (1996) Real time quantitative PCR. Genome Res 6(10):986–994
28. Myers RM, Lumelsky N, Lerman LS, Maniatis T (1985) Detection of single base substitutions in total genomic DNA. Nature 313(6002):495–498
29. Yoshino K, Nishigaki K, Husimi Y (1991) Temperature sweep gel electrophoresis: a simple method to detect point mutations. Nucleic Acids Res 19(11):3153
30. Chen TJ, Boles RG, Wong LJ (1999) Detection of mitochondrial DNA mutations by temporal temperature gradient gel electrophoresis. Clin Chem 45(8 Pt 1):1162–1167
31. Alper OM, Wong LJ, Young S, Pearl M, Graham S, Sherwin J, Nussbaum E, Nielson D, Platzker A, Davies Z, Lieberthal A, Chin T, Shay G, Hardy K, Kharrazi M (2004) Identification of novel and rare mutations in California Hispanic and African American cystic fibrosis patients. Hum Mutat 24(4):353. doi:10.1002/humu.9281
32. Tan DJ, Bai RK, Wong LJ (2002) Comprehensive scanning of somatic mitochondrial DNA mutations in breast cancer. Cancer Res 62(4):972–976
33. Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T (1989) Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci USA 86(8):2766–2770
34. White MB, Carvalho M, Derse D, O'Brien SJ, Dean M (1992) Detecting single base substitutions as heteroduplex polymorphisms. Genomics 12(2):301–306. doi:0888-7543(92)90377-5 [pii]

35. Makino R, Yazyu H, Kishimoto Y, Sekiya T, Hayashi K (1992) F-SSCP: fluorescence-based polymerase chain reaction-single-strand conformation polymorphism (PCR-SSCP) analysis. PCR Methods Appl 2(1):10–13
36. Wang YH, Barker P, Griffith J (1992) Visualization of diagnostic heteroduplex DNAs from cystic fibrosis deletion heterozygotes provides an estimate of the kinking of DNA by bulged bases. J Biol Chem 267(7):4911–4915
37. Suzuki Y, Orita M, Shiraishi M, Hayashi K, Sekiya T (1990) Detection of ras gene mutations in human lung cancers by single-strand conformation polymorphism analysis of polymerase chain reaction products. Oncogene 5(7):1037–1043
38. Dockhorn-Dworniczak B, Dworniczak B, Brommelkamp L, Bulles J, Horst J, Bocker WW (1991) Non-isotopic detection of single strand conformation polymorphism (PCR-SSCP): a rapid and sensitive technique in diagnosis of phenylketonuria. Nucleic Acids Res 19(9):2500
39. Hogg A, Onadim Z, Baird PN, Cowell JK (1992) Detection of heterozygous mutations in the RB1 gene in retinoblastoma patients using single-strand conformation polymorphism analysis and polymerase chain reaction sequencing. Oncogene 7(7):1445–1451
40. Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res 7(10):996–1005
41. Den Dunnen JT, Van Ommen GJ (1999) The protein truncation test: a review. Hum Mutat 14(2):95–102. doi:10.1002/(SICI)1098-1004(1999)14:2<95::AID-HUMU1>3.0.CO;2-G [pii] 10.1002/(SICI)1098-1004(1999)14:2<95::AID-HUMU1>3.0.CO;2-G
42. Roest PA, Roberts RG, van der Tuijn AC, Heikoop JC, van Ommen GJ, den Dunnen JT (1993) Protein truncation test (PTT) to rapidly screen the DMD gene for translation terminating mutations. Neuromuscul Disord 3(5–6):391–394. doi:0960-8966(93)90083-V [pii]
43. Friedl W, Aretz S (2005) Familial adenomatous polyposis: experience from a study of 1164 unrelated german polyposis patients. Hered Cancer Clin Pract 3(3):95–114. doi:1897-4287-3-3-95 [pii] 10.1186/1897-4287-3-3-95
44. Hogervorst FB, Cornelis RS, Bout M, van Vliet M, Oosterwijk JC, Olmer R, Bakker B, Klijn JG, Vasen HF, Meijers-Heijboer H et al (1995) Rapid detection of BRCA1 mutations by the protein truncation test. Nat Genet 10(2):208–212. doi:10.1038/ng0695-208
45. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74(12):5463–5467
46. Filippi G, Rinaldi A, Archidiacono N, Rocchi M, Balazs I, Siniscalco M (1983) Brief report: linkage between G6PD and fragile-X syndrome. Am J Med Genet 15(1):113–119. doi:10.1002/ajmg.1320150115
47. Mulligan LM, Phillips MA, Forster-Gibson CJ, Beckett J, Partington MW, Simpson NE, Holden JJ, White BN (1985) Genetic mapping of DNA segments relative to the locus for the fragile-X syndrome at Xq27.3. Am J Hum Genet 37(3):463–472
48. Oberle I, Rousseau F, Heitz D, Kretz C, Devys D, Hanauer A, Boue J, Bertheas MF, Mandel JL (1991) Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. Science 252(5010):1097–1102
49. Richards RI, Holman K, Kozman H, Kremer E, Lynch M, Pritchard M, Yu S, Mulley J, Sutherland GR (1991) Fragile X syndrome: genetic localisation by linkage mapping of two microsatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. J Med Genet 28(12):818–823
50. Yu S, Pritchard M, Kremer E, Lynch M, Nancarrow J, Baker E, Holman K, Mulley J, Warren S, Schlessinger D et al (1991) Fragile X genotype characterized by an unstable region of DNA. Science 252(5009):1179–1181. doi:252/5009/1179 [pii] 10.1126/science.252.5009.1179
51. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30(12):e57
52. Kallioniemi A, Kallioniemi OP, Sudar D, Rutovitz D, Gray JW, Waldman F, Pinkel D (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 258(5083):818–821

53. Landsverk ML, Wang J, Schmitt ES, Pursley AN, Wong LJ (2011) Utilization of targeted array comparative genomic hybridization, MitoMet, in prenatal diagnosis of metabolic disorders. Mol Genet Metab 103(2):148–152. doi:S1096-7192(11)00064-3 [pii] 10.1016/j.ymgme.2011.03.003
54. Wang J, Zhan H, Li FY, Pursley AN, Schmitt ES, Wong LJ (2012) Targeted array CGH as a valuable molecular diagnostic approach: experience in the diagnosis of mitochondrial and metabolic disorders. Mol Genet Metab 106(2):221–230. doi:S1096-7192(12)00106-0 [pii] 10.1016/j.ymgme.2012.03.005
55. Bejjani BA, Saleki R, Ballif BC, Rorem EA, Sundin K, Theisen A, Kashork CD, Shaffer LG (2005) Use of targeted array-based CGH for the clinical diagnosis of chromosomal imbalance: is less more? Am J Med Genet A 134(3):259–267. doi:10.1002/ajmg.a.30621
56. Stankiewicz P, Beaudet al (2007) Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. Curr Opin Genet Dev 17(3):182–192. doi:S0959-437X(07)00074-3 [pii] 10.1016/j.gde.2007.04.009
57. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J, Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris MS, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson TJ, Lipshutz R, Chee M, Lander ES (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280(5366):1077–1082
58. Landsverk ML, Douglas GV, Tang S, Zhang VW, Wang GL, Wang J, Wong LJ (2012) Diagnostic approaches to apparent homozygosity. Genet Med. doi:10.1038/gim.2012.58 gim201258 [pii]

Part II
The Technologies and Bioinformatics

Chapter 3
Methods of Gene Enrichment and Massively Parallel Sequencing Technologies

Hong Cui

Abstract Thirty years after the invention of dideoxy sequencing (a.k.a. Sanger sequencing), the advent of massively parallel sequencing technologies became another biotechnical revolution that enables the acquisition of genetic information in gigabase scale within an acceptable period of time. As a consequence, causal mutations underlying clinically heterogeneous disorders are more efficiently detected, paving the way for deciphering the pathogenecities of complex diseases. With the huge potential impact in modern medicine and health care, progress has been rapid in further optimizing the technology in both the academic and industrial fields. Third-generation sequencing technologies, although still facing multiple challenges, have shed light on the direct analyses of DNA and RNA at single-molecule level. Currently, before whole genome sequencing becomes routine, an in-depth assessment of targeted genomic regions is more feasible and has been widely applied in both clinical applications and basic research. Various enrichment methods, either PCR-based or hybridization-based, have been developed and gradually improved during application. This chapter provides detailed information on various target gene enrichment methods as well as massively parallel sequencing platforms. Hopefully this could assist the project-based approach/platform selections tailored for individual needs.

1 Introduction

The massively parallel sequencing (MPS) era was marked in 2005, when the complete genome sequences of two bacteria were published by 454 Life Sciences Corporation [1]. Since then, there has been a fundamental shift from automated

H. Cui Ph.D (✉)
Department of Molecular and Human Genetics, Baylor College of Medicine,
Houston, TX 77030, USA
e-mail: hcui@bcm.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_3,

Sanger sequencing to high-throughput MPS, which is also referred to as next-generation sequencing (NGS). From tens of megabases (Mb) to hundreds of gigabases (Gb) of data generated per run, the capacity of MPS instruments has increased tremendously in recent years, enabling many studies that were not feasible before. With steadily reducing costs, genome-wide studies have been more widely carried out in both basic and translational research [2–4]. More recently, whole exome sequencing (WES) was implemented in clinical diagnostic laboratories for the purpose of molecular diagnosis of heterogeneous genetic diseases. Announced in early 2008 by an international consortium, the 1000 Genomes Project proposed to sequence the genome of at least 1,000 individuals with different ethnic backgrounds, aiming to document common SNPs and discover rare disease-causing variants [5]. The pilot project was designed to sequence three groups with various depths: (1) whole genome sequencing of 179 individuals at very low coverage (2X), (2) whole genome sequencing of two trios at high coverage (20X), and (3) targeted exon sequencing of 697 individuals at high coverage (20X). Released in 2010, data generated from the pilot studies achieved its main goal on enriching the public SNP database (dbSNP) for variant imputation [5]. However, false positive calls that were inevitably produced due to low-coverage sequencing errors have caused concerns in relying on these data. To date, deep sequencing of the whole genome of a large cohort is still not practical for most research-based projects due to the high cost associated with sequencing itself and the bioinformatics infrastructure required for downstream data analyses and storage. Consequently, targeted enrichment of a group of genomic regions has become an attractive alternative and has been widely adapted for various applications. From a practical point of view, MPS-based sequence analysis of a group of candidate genes responsible forparticular diseases is especially beneficial for clinical diagnostic purposes when cost and turnaround time are of particular interest [6, 7]. This chapter provides the technical details of different target enrichment methods and their overall performance, as well as the comparison of the most widely adapted massively parallel sequencing platforms.

2 Target Gene Enrichment Methods

The aim of the target gene enrichment is to selectively pull out or enrich a subset of genomic regions of interest prior to MPS. Multiple factors, including the target size, sample volume, and the specificity and sensitivity associated with each method, affect the choice of selection. Along with the quickly evolving MPS Technologies, a number of enrichment approaches have been developed over the past years. These methods differ greatly in not only the principle of enrichment (PCR-based versus hybridization-based) but also other aspects including ease of handling, throughput, and cost.

2.1 PCR-Based Enrichment

2.1.1 Multiplex PCR

The principle of conventional multiplex PCR is to include multiple primer pairs in a singleplex reaction to amplify a group of genomic loci under the same thermal cycling condition. The target size is limited to tens of amplicons while the design of the primers could be problematic as well in regarding to the high levels of nonspecific amplifications generated by the primer cross amplification between different pairs [8, 9]. To overcome these bottlenecks and make PCR-based enrichment more compatible with MPS, multiple new methods have been developed, among which the microfluidic chip-based Access Array (Fluidigm) and microdroplet PCR-based RainDance (RainDance) Technologies have been commercialized [10–12].

Microfluidic Technology (Fluidigm)

The Access Array from Fluidigm managed to increase both throughput and assay sensitivity by using a microfluidic chip [12]. With its current configuration, 48 samples can be tested simultaneously for 48 assays. Each reaction contains a unique primer pair and is carried out separately in 2,304 reaction chambers on a microfluidic chip. The reaction volume is only 35 nanoliter, which greatly reduces reagent costs. Comparing to the conventional PCR which needs ~10–100 ng genomic DNA as template for a single reaction, the required template amount is dramatically decreased to 50 ng for setting up 48 reactions using Access Array. Another advancement is the increased assay sensitivity due to compartmentation. Since each reaction is physically separated, single PCR with unique primer pair avoids potential cross-reactions. The workflow is further tailored to be more compatible with MPS platforms by synthesizing long primers with adaptor sequences added to the 5′ end; therefore, amplified products are ready to be used after thermal cycling without additional steps required for sequencing library construction. With a capacity of a maximum of 48 amplicons per sample, the Access Array system is tailored for small target enrichment. Jones et al. used this method to analyze a set of 24 genes involved in congenital disorders of glycosylation (CDG) [6]. With a target set containing 387 amplicons ranging from 201 to 617 bp in length, 12 samples were enriched by the Access Array. Five exons failed completely, presumably due to regions challenging for primer design. Using the same method, another group analyzed two genes, *CYP7B1* and *SPG7*, in 187 patients diagnosed with sporadic spastic paraplegia [13]. The authors claimed an 80 % success rate for enrichment. Sequence complexity and GC-rich regions are a major cause for failed amplifications. In addition, technical issues, such as introducing bubbles during assay setup, contributed to some failures as well but can be prevented with experiences.

For the sequencing to be carried out economically, all inlets for both samples and reactions should be occupied which is often not feasible. Lack of flexibility, difficulty in finding universal optimal condition for all 2,304 PCR reactions, and a fixed capacity are the major drawbacks of this system that need to be addressed in the future.

Microdroplet Technology (RainDance Technologies)

An alternative approach has been developed for PCR-based enrichment by RainDance Technologies using microdroplet PCR [11]. Similar to emulsion PCR [14, 15], the PCR reactions are performed independently in droplets encapsulated by oil layer. Each droplet contains a single primer pair along with template DNA and other reagents necessary for PCR amplification. An instrument, RDT 1000, manufactured by RainDance Technologies is required to generate reaction droplets by merging a primer droplet with a template-containing droplet at high speed. With as little as 250 ng genomic DNA, ten million picoliter-size droplets are produced within an hour. All merged droplets are collected in a single PCR tube for subsequent amplification in a standard thermal cycler. Upon completion, a droplet destabilizer is added to the reaction tube to break the emulsion and amplified products are harvested by collecting the aqueous phase. An extra post-amplification concatenation step is required before shearing during downstream library construction. Efforts were made to circumvent this step by adding adaptor sequence to the primers. Although this may help to simplify the library construction procedure, as a trade-off, it complicates the primer design process and introduces noninformative sequences from primer binding and linker sequences into the MPS reads, leading to reduced analytical specificity and more complicated data analysis.

Although constraints exist, improvements have been made as well. By producing droplets with consistent size, similar amount of reagent is provided in each isolated reaction, thus minimizing the difference of amplification efficiency between various primer pairs. Consequently, this approach renders more uniformed enrichment of targeted regions. Moreover, with 200–20,000 primer pairs covering up to 10 Mb per library, both the flexibility and throughput are improved relative to the Fluidigm Access Array, making it an alternate choice for enrichment of small- to medium-sized targets. Twelve genes associated with congenital muscular dystrophies (CMDs) were enriched by this approach and by in-solution-based capture (SureSelect, Agilent) for comparison [16]. While the same sensitivity was achieved by both methods, microdroplet-based PCR enrichment displayed higher specificity and reproducibility. Nevertheless, all PCR-based methods suffer from the same PCR-specific limitations, such as allele dropout, limit scalability, and difficulty of one-condition-fit-all.

2.1.2 Long-Range PCR (LR-PCR)

A haploid human genome contains three billion nucleotides, within which only 1 % are protein-coding sequences [17, 18]. Spread out across ~180,000 regions, the average size of an exon is calculated to be around 170 bp [19, 20]. LR-PCR is capable of amplifying long genomic regions that are tens of kilobases in size [21]. Although it is not suitable for enrichment of nuclear gene exons due to the presence of intronic and intergenic regions, LR-PCR has been shown to be an efficient approach for enriching the intronless mitochondrial genome [22–25]. More importantly, it avoids the unintended problems caused by the highly polymorphic feature of mtDNA and the abundant nuclear mtDNA homologues (NUMTs) [26]. By using one pair of tail-to-tail primers, the entire 16.6 kb mtDNA is amplified as a single amplicon. Coupled with MPS, this method can detect not only single-nucleotide changes and small indels but also large deletions with exact deletion junctions mapped. In addition, heteroplasmy levels at every nucleotide position are simultaneously quantified as well.

2.1.3 Quantitative PCR (qPCR)

As a derivative of the PCR-based target enrichment method, a qPCR-based novel approach was reported recently for targeting 16 Leber congenital amaurosis (LCA) disease genes [27]. Using a liquid-handling system, 375 amplicons were amplified in real time followed by pooling and concatenation before shearing into small fragments. According to the authors, the qPCR method efficiently solved the difficulties associated with normalization of the amplified product for downstream library construction. Another superiority of this method is the direct visualization of the amplification results, which enables instant discovery of failed amplicons.

2.2 Capture-Based Enrichment: By Hybridization with Oligonucleotide Probes

2.2.1 The Molecular Inversion Probe (MIP) Hybridization

In order to rectify the limitation of PCR-based enrichment, several hybridization-based methods have been developed. The molecular inversion probe (MIP) approach uses single-stranded oligonucleotides (oligos) to hybridize with target regions. Each oligo is 70 bases long consisting of common linkers flanked by target-specific sequences on both ends. During hybridization, an individual MIP forms an arch by annealing with genomic regions that are complementary to its

unique ends, leaving a gap in between spanning from 60 to 190 bases of the target region. DNA polymerase is subsequently used to fill in the gaps representing the target sequences, and the resulting nick is closed by ligation. Upon removing free oligos by exonuclease digestion, the target sequences within the circularized probes are amplified using common primers complementary to the linker sequences followed by sequencing the library construction [28].

Based on the circularization method used by MIP, several derivatives were developed, including Selector (Olink Genomics), gene-collector, and CIPer techniques [29–33]. Although over 90 % of the regions were covered well when the targets are restricted to several hundred exons [29, 32], only less than 20 % of the targets were captured in a study in which 55,000 oligos were used to enrich 10,000 exons in human genome [33]. Restrained by low specificity, these methods require additional optimizations before application to genome-wide studies.

Synthesizing large set of long oligos for hybridization is quite expensive in MIP-based strategies. Microarray-based in situ oligo synthesis dramatically reduces the cost and has been commercialized for production of large amount of oligonucleotide probes suitable for both in solution- and array-based enrichment at affordable prices.

2.2.2 Solid-Phase Hybridization

Array-based hybridization directly captures the target sequences without PCR amplification. Both Roche NimbleGen and Agilent provide commercialized products with similar performance. In general, fragmented DNA is hybridized with probes on a high-density DNA array. After incubation, excess DNA and nonspecifically bound DNA are removed by extensive washing. Captured DNA is then eluted and used for library construction. Two forms of arrays, 385 K and 2.1 M, are offered by NimbleGen, targeting 5 and 34 Mb regions, respectively. In comparison, the Agilent array contains relatively fewer probes (244 K) with a smaller target size. As the first method used for genome-wide studies, array-based hybridization renders a hundred- to thousand-fold enrichment covering over 90 % of targets [34–36]. Two major drawbacks associated with array-based hybridization are as follows: first, additional hardware and software need to be purchased for hybridization; second, as solid-phase enrichment, it is not compatible with liquid-handling automation systems.

2.2.3 Solution-Phase Hybridization

In the solution-based hybridization method, oligonucleotide probes are cleaved off from the microarray after synthesis. Depending on the sample volume, hybridization can be set up in single PCR tube or 96-well plate, rendering flexibility and choice of scaling up. Without additional dedicated instrument, hybridization is carried out in regular thermal cycler. Consequently, by overcoming the bottlenecks

affiliated with solid-phase hybridization, solution-phase hybridization has been more widely adapted [3, 4, 7, 37–39]. Both Roche NimbleGen and Agilent offer commercialized products with slight differences regarding the probe characters and experimental procedures. In SeqCap EZ Choice (NimbleGen), 2.1 million, biotinylated DNA probes are used to target custom regions ranging from 100 kb to 50 Mb. With the extensive tiling design, each target region is deeply covered with probes that are 50–105 bases in length. This built-in redundancy increases capture efficiency and uniformity. Hybridization is carried out at 47 °C for 65–72 h. Probe-target hybrids are captured by streptavidin magnetic beads followed by washing and elution. The Agilent SureSelect target enrichment system uses 120-mer biotinylated RNA probes to capture target regions. Long RNA probes are generated by extra in vitro transcription step after full-length synthesized DNA oligos are cleaved off from array. The longer probes and the more stable RNA-DNA hybrid increase capture efficiency yet require relatively higher temperature, 65 °C, during hybridization to reduce nonspecific interactions. Under such conditions, seamless sealing of the reaction tubes is of great importance to prevent excessive evaporation during the 24 h incubation which may lead to suboptimal or even failed hybridization. In addition, due to the unstable nature of RNA, extra care must be taken during storage and handling of RNA libraries.

So far, solution-phase hybridization has become the most prevailing method for enriching target regions in megabases. In combination with commercialized automation systems, which are tailored specifically for this application, handling a large volume of samples under high-throughput platform is also possible.

2.3 Additional Considerations for PCR- and Oligonucleotide Hybridization-Based Enrichment Methods

Target enrichment enables the study of a portion of genomic regions at an affordable price. With relatively fixed cost, probes at same amount are synthesized on an array in hybridization-based enrichment, targeting regions of interest up to 50 Mb. Conversely, it is not feasible to apply PCR-based methods for megabase target enrichment due to the parallely increasing cost of primer synthesis associated with the growing size of the target. With extra investment, the target size can be extended to 10 Mb by using RainDance microdroplet technology for PCR setup. Therefore, in various genome-wide studies targeting 35 Mb human exome, hybridization-based enrichment becomes the only choice.

An advantage of PCR-based enrichment is its ability to avoid pseudogene interference. In most cases, target regions can be selectively amplified using primers designed to regions that are not homologous to pseudogenes. Additionally, leaving a target-specific nucleotide at a primer's 3′ end increases its specificity. In some cases, where highly homologous regions in pseudogenes extend over kilobases, LR-PCR can be considered instead by designing primers at target-specific regions. A major concern associated with PCR-based strategies is the intrinsic polymorphic

feature of the human genome, which could cause allele dropout if there are SNPs residing in the primer binding sites. Although designing two or even more pairs of primers for the same target region can help to dramatically reduce the incidence of allele dropout, it also increases the cost as well as the experimental and analytical workload. Conversely, with long oligos and dense tiling design, the hybridization method is able to tolerate the existence of SNPs in any target regions. However, as a trade-off, oligonucleotide probes cannot differentiate real targets from pseudogenes and may skew allele coverage ratios when mutations are solely present in either one of them.

Poor uniformity can also be a concern in PCR-based methods due to unequally pooled amplicons or the "end sequencing effect." As for most of the MPS Technologies, templates are read from the same ends using common sequencing primers. Consequently, the closer a position is to the end of a read, the higher coverage it will get. For a target enriched by the PCR method, all amplicons share the same sequence, leading to an uneven coverage pattern with the highest at the ends and the lowest in the middle. To solve this problem, an extra concatenation step is incorporated in some PCR-based enrichment methods to join individual targets into long fragments before shearing. This helps to generate templates with random ends, at the price of an extended workflow. In comparison, hybridization-based enrichment uses tiled probes to anneal with randomly sheared targets, rendering better uniformity.

3 Overview of Massively Parallel Sequencing Technologies

3.1 Second-Generation Sequencing Technologies

3.1.1 Sequencing by Synthesis-Parallelized Pyrosequencing (Roche 454)

Evolved from the principle of pyrosequencing, the 454 Sequencer (Genome Sequencer-20, Roche 454) released in 2005 was the first commercially available next-generation sequencing instrument [1]. Today, the GS FLX+system typically produces 700 Mb data with read lengths of up to 1 kilobase within 23 h.

The workflow of the 454 GS sequencer can be divided into five steps (Fig. 3.1a). In the first step, adaptors containing universal primer binding sites are ligated to the target fragments. In order to facilitate the collection of fragment-enriched beads in the downstream pipeline, one of the adaptors is biotinylated. Next, denatured single-strand adaptor-ligated library fragments are mixed with primer-coated beads under a condition that favors equal bead to DNA molecule ratio, ensuring that a single DNA molecule is trapped with an individual bead in a microreactor. Immersed in oil, emulsion PCR (emPCR) is then carried out in water droplets containing single bead and other components required for PCR amplification, generating millions of

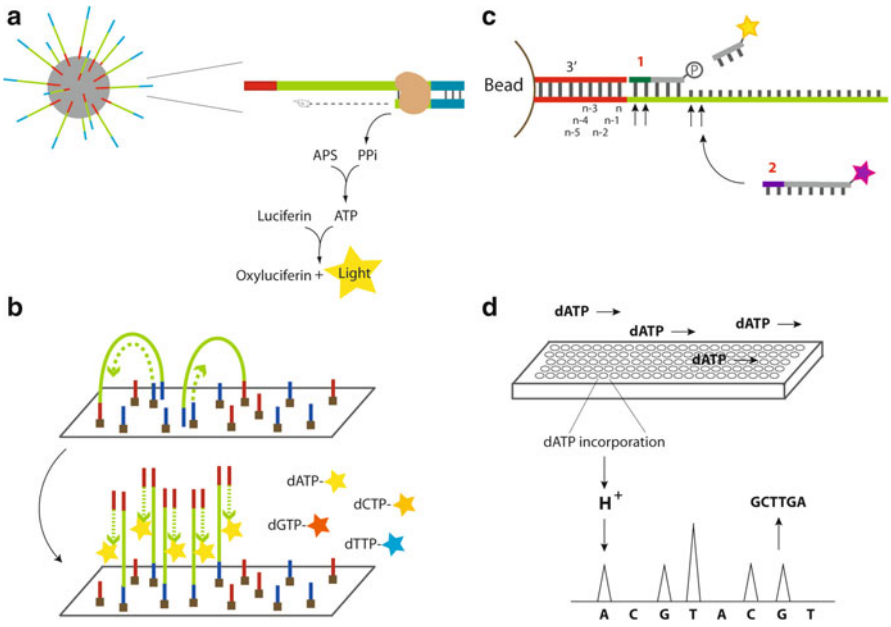

Fig. 3.1 Overview of the second-generation sequencing technologies. (**a**) Pyrosequencing (Roche 454) (**b**) Sequencing by synthesis (GAII/HiSeq, Illumina) (**c**) Sequencing by ligation (SOLiD, Life Technologies) (**d**) Sequencing by synthesis and semiconductor chip (Ion Torrent, Life Technologies)

copies of identical template on each bead. The emulsion is broken at the end of amplification, and beads carrying biotinylated templates are captured by streptavidin magnetic beads to enrich the population with successful amplification. Millions of template-carrying beads are subsequently seeded into PicoTiterPlate (PTP) wells where parallel pyrosequencing will take place. The size of each well allows for only one template-carrying bead together with additional smaller beads coupled with sulfurylase and luciferase needed in downstream enzymatic reactions to produce light. Sequencing starts by releasing a single type of deoxyribonucleoside triphosphate (dNTP) into the sequencing chamber in a predetermined order. As they flow across the PTP wells, a polymerase-mediated incorporation event occurs in certain wells where the released dNTP is complementary to the next nucleotide in the growing strand. Catalyzed by ATP sulfurylase, the released inorganic pyrophosphate (PPi) together with adenosine phosphosulfate (APS) forms adenosine triphosphate (ATP), which will be utilized by luciferase to catalyze luciferin and emit light [40]. The chemiluminescent light signals generated from each well are captured by a high-resolution charge-coupled-device (CCD) camera to record the incorporation event. Unincorporated dNTPs as well as residual PPi and ATP are washed away before DNA synthesis resumes. At the end of sequencing, the recorded signal intensity of each incorporation cycle within individual wells is analyzed by bioinformatics software to determine the sequence of millions of reads in parallel.

As the first generation of massively parallel sequencer, the Roche 454 instrument has been applied to sequencing of relatively small targets, such as various microbial genomes and small RNAs, producing around 500 Mb of data in a single run [41–45]. Although the intensity of the emitted light is proportional to the number of incorporated nucleotides, sequencing of homopolymer regions is still a big hurdle for the pyrosequencing-based 454 platform when the length of the homopolymer stretch is longer than six nucleotides [46]. Other second-generation sequencers are more widely used on human genome sequencing projects with increased throughput and lower cost per Mb.

3.1.2 Sequencing by Synthesis-Reversible Terminator (Illumina GA/HiSeq)

Introduced to the market a year later than the Roche 454, the Illumina GA system was able to produce much more data with dramatically reduced sequencing cost per base. Instead of emPCR, its sequencing template was generated by clonal bridge amplification on a solid surface (flow cell). A microfluidic cluster generation station (cBot) is used for clonal amplification. Oligos complementary to the Illumina adaptor sequences are covalently bound to the surfaces of the flow cell. By hybridizing with another oligo in the nearby region, a bridge is formed and the complementary strand is synthesized (Fig. 3.1b). Two strands are subsequently separated by denaturing and ready for additional cycles of bridge amplification. By the end of the process, each cluster will contain approximately 1,000 copies with a cluster density at 400 k/mm^2. With improved software and hardware systems, the cluster density is further increased to ~700 k/mm^2 for V3 flow cell, rendering 300 Gb data per flow cell per run with 100 bp, pair-end reads. The density of the cluster is positively correlated to the size of the sequence data. While a reduced data size is expected with loose clusters, clusters that are too dense may also fail to produce enough data since many overlapping signals generated by overly densed clusters will be filtered out. Therefore, a good estimation of the total library DNA concentration is of great importance for generating properly densed clusters.

For Illumina platforms, sequence analysis is based on DNA synthesis chemistry using fluorescent-labeled, block-reversible dNTPs. Three steps are repeated in each cycle of sequencing: nucleotide incorporation, imaging, and removing blocks. With different dyes attached, all four dNTPs are released together at the beginning of the cycle. DNA synthesis is temporarily paused following incorporation of a modified dNTP. Unincorporated dNTPs are washed away before imaging. Fluorescent signals are then scanned followed by cleavage of the nucleotide base-attached fluorophores and 3′-blocks, allowing the incorporation of the next nucleotide in the following cycle. Unlike the dideoxynucleotides used in traditional Sanger sequencing as chain terminators, the use of the reversible terminators allows the chain elongation to be paused temporarily after each nucleotide incorporation and continuation of the synthesis after the pause for imaging. This feature greatly improves the sequencing quality of homopolymeric regions which is a major drawback for methods using pyrosequencing chemistry.

3.1.3 Sequencing by Ligation (SOLiD, Applied Biosystems)

Introduced to the market in 2007, SOLiD (Sequencing by Oligonucleotide Ligation and Detection) platform employed a new sequencing chemistry: sequencing by ligation. Similar to the emulsion PCR used by Roche 454 system, adaptor-ligated template DNA is diluted and hybridized with a P1 adaptor on magnetic beads which are 1 μm in diameter. Emulsion PCR is carried out to clonally amplify the unique template on individual bead. Beads with successful amplification are enriched by density-gradient media followed by deposition on a glass surface. The SOLiD system uses a two-base color coding scheme in which each dye encodes four different two-base combinations of the A, T, C, and G nucleotides. Each probe is an octamer with the first two bases interrogated in every ligation cycle. The third to fifth bases are degenerate (n) whereas the last three bases are universal. Once the primer is annealed to the adaptor, all probes are released to compete for the best match between the template and the first two nucleotides of the probes. Connection of the primer's 5′ end with a probe's 3′end is accomplished by DNA ligase (Fig. 3.1c). After imaging, fluorescent dyes as well as the last three nucleotides at the 5′ end of each probe are cleaved, exposing the 5′ phosphate group for the next cycle of ligation. The resulting product is then denatured and removed, allowing additional four rounds of extension using primers staggered by a single base. Since each nucleotide is interrogated twice, the measurement errors therefore can be more accurately distinguished from the real polymorphisms. Without the need of any manual decoding, the SOLiD system automatically compiles and converts the color space data into standard base calls in the final step.

3.1.4 Sequencing-by-Synthesis Polymerase with Semiconductor Chip (Ion Torrent)

The founder of 454 Life Sciences, Dr. Jonathan M. Rothberg, founded another next-generation sequencing company, Ion Torrent (Life Technologies), in 2007. Designed for sequencing relative small targets, the Ion Personal Genome Machine (IPGM) is a bench-top high-throughput sequencer generating 50 Mb to near 1 Gb of data within a single run depending on the choice of chips. Running time is dramatically reduced to 1–2 h compared to days for the Illumina GA/HiSeq or SOLiD system. This improvement in running time is attributed to the novel signal detection system employed by IPG [47]. Instead of scanning signals by a CCD camera, a semiconductor chip which acts as a pH meter can sense minor changes in pH caused by a nucleotide incorporation event. Templates for IPG are enriched by emPCR in the same way as those for Roche 454. Template-carrying Ion Sphere™ particles together with DNA polymerase and sequencing primers are injected into a semiconductive chip which is then loaded on a PGM machine for sequencing. Unmodified dNTPs are released in order and diffuse to each well (Fig. 3.1d). Once a nucleotide is incorporated into the growing strand, a hydrogen ion is released, leading to a pH change due to its positive charge. Changes in pH are converted to voltage alterations and the electrical signals will be further processed and translated into sequence reads.

The number of released ions is proportional to that of the incorporation events. Once two identical nucleotides are incorporated, two hydrogen ions will be released and the reading of the digital signal will also double. Whereas if the released nucleotide is not complementary to the template, no hydrogen ion will be generated and the pH in solution will not change, indicating there is no incorporation event. The whole process is presented in a video at "http://ioncommunity.lifetechnologies.com/videos/1016." With relatively low capital investment and short running time, the PGM has been quickly adapted into many areas including microbial sequencing and targeted resequencing (e.g., Ion AmpliSeq™ Cancer Panel) [48–51].

3.2 A Glance at the Third-Generation Sequencing Technologies

As discussed in the previous sections, second-generation sequencing technologies utilize either emPCR (Roche 454, SOLiD, and Ion Torrent) or bridge amplification (Illumina) to clonally amplify sequencing templates in order to strengthen the signals for detection. However, amplification-based errors and biases are inevitable. The major characteristic of third-generation sequencing technologies is to sequence at the single-molecule level, avoiding the requirement for pre-amplification of sequencing templates. Using this strategy, the sample preparation time is also shortened. Another advantage is that the required amount of initial DNA is greatly reduced, making it a better choice for sequencing samples with limited quantity. Three major third-generation sequencers, HeliScope™ Single Molecule Sequencer (Helicos Biosciences), Single Molecule Real Time (SMRT™) (Pacific Biosciences), and nanopore sequencer (Oxford Nanopore Technologies), are discussed below.

3.2.1 HeliScope™ Single Molecule Sequencer (Helicos Biosciences)

The HeliScope™ Single Molecule Sequencer utilizes the True Single Molecule Sequencing (tSMS™) technology to conduct sequencing at the single-molecule level [52]. With poly A tails added to 3′ ends, fragmented single-stranded templates hybridize with poly Ts that are covalently bound on the flow cell surface. Fluorescently labeled dNTPs are added sequentially. After washing away excess dNTPs, fluorescent signals are generated from incorporated nucleotides by excitation with laser. Thousands of pictures are taken in order to record both the incorporation events and their corresponding positions. Fluorophores are subsequently cleaved from the incorporated nucleotides, followed by the start of next sequencing cycle. Once all cycles are completed, signals recorded from each cycle are converted into sequence reads of individual template molecule for analysis. HeliScope™ Single Molecule Sequencer generates 21–35 Gb of data per run with read lengths ranging from 25 to 55 bases. The claimed error rates are 0.2 %, 1.5 %, and 3.0 % for single-nucleotide substitutions, insertions, and deletions, respectively.

3.2.2 Single Molecule Real Time (SMRT™) (Pacific Biosciences)

The Single Molecule Real Time (SMRT™) technology from Pacific Biosciences advances the single molecule sequencing by monitoring the sequencing progress in real time and generating long reads [52]. Similar to the flow cell used by Illumina and the semiconductive chip in the Ion Torrent technology, sequencing templates are deposited into "wells" on an SMRT cell which contains thousands of zero-mode waveguides (ZMWs). The ZMWs with nanometer-scale diameter provide spaces for parallel sequencing to take place. To record DNA synthesis in real time, the focus of the camera needs to be kept constant, avoiding the requirement for minor adjustment through the whole sequencing procedure. Therefore, the physical position of each incorporation event should be predetermined and fixed during synthesis. To achieve this, the DNA polymerase, instead of the template, is immobilized inside the ZMWs. Templates can be either linear or circular. Linear template generates single pass, long reads (up to 10 kb), while circular template generates multiple pass, short reads (250 bp). Templates are held by the DNA polymerase. A fluorophore with a unique color is linked to each dNTP through the triphosphate chain, omitting the extra cleavage required for the dNTPs with dyes attached to the bases. By constantly providing the fluorophore-labeled dNTPs at high concentration in the reaction solution, the performance of the polymerase, including the enzyme's accuracy and processivity, is greatly improved. The size of the ZMWs is too small to hold free fluorophore-labeled dNTPs. The duration for each stay within the detection distance is determined at tens of milliseconds which is short enough to cause only low-level background fluorescence. Unmatched dNTPs quickly diffuse out of the ZMWs while the complementary dNTP is "held" by the polymerase during incorporation. The resulting pause leads to extended fluorescent signals, and the type of the incorporated dNTP can be identified by the unique fluorescent emission spectrum. Once a phosphodiester bond is created, the diphosphate together with the attached fluorophore is cleaved off by DNA polymerase at the end of the incorporation event. Released fluorophores quickly diffuse away from ZMWs, restoring the fluorescent signals to background level. A cartoon illustrating the essence of the SMRT technology is available at "http://www.youtube.com/watch?v=v8p4ph2MAvI." Using this strategy, strand synthesis is performed continuously and signals are captured in real time. Multiple sequencing steps, including focusing, signal scanning, washing, and releasing of fresh dNTPs, are removed in this novel technology, greatly shorten the sequencing time. By mimicking the natural synthesis procedures of the DNA polymerase to maximally extend the DNA, the speed of strand elongation is much higher than those in any other technologies, enabling the generation of long reads that are kilobase in length (Table 3.1).

Another attractive feature of the SMRT technology is its ability to directly sequence and detect methylated DNA without the need for conventional bisulfite conversion. DNA methylation, histone modifications, and chromatin remodeling are three types of epigenetic modifications that regulate various biological processes, e.g., gene expression, genomic imprinting, and X chromosome inactivation [53–55]. Defects in these processes have been linked to various human diseases

Table 3.1 Comparison of massively parallel sequencing platforms

	Second-generation MPS Technologies				Third-generation MPS Technologies		
Platforms	454 GS	HiSeq2000	SOLiD	Ion torrent	HeliScope™ single molecule sequencer	SMRT	GridION
Producer	Roche	Illumina	ABI, Life Technologies	Life Technologies	Helicos Biosciences	Pacific Biosciences	Oxford Nanopore Technologies
Template preparation	Emulsion PCR-based clonal amplification on beads	Bridge amplification on solid surface	Emulsion PCR-based clonal amplification on beads	Emulsion PCR-based clonal amplification on beads	Clonal amplification is not required, but need poly A addition to 3′ end of the fragmented templates	Clonal amplification is not required.	Clonal amplification is not required
Sequencing mechanism	Pyrosequencing, one type of dNTP at a time	Sequencing by synthesis with reversible dye terminators	Sequencing by ligation	Sequencing by synthesis with semiconductor technology	Sequencing by synthesis with reversible dye terminators	Sequencing by synthesis with nature cleavage of the phospho-dye group	No chemistry involved. Use parameters measured during current change to interrogate the nucleotide
Signal source	Chemiluminescent lights	Fluorescence	Fluorescence	Hydrogen ion	Fluorescence	Fluorescence	Molecule-based characteristic current disruption sensed by digital device
Read length (bases)	400	50–100, paired end	75+35, paired end	200	25 to 55, 35 in average	1,000 in average	10,000
Run time	10–20 h	4–11 days	Up to 8 days	2.5–4.5 h	8 days	<1 h	Variable, depends on yield
Throughput per run	400 Mb	Maximum 300 Gb per flow cell	50–70 Gb per FlowChip	10 Mb (314 chip), 100 Mb (316 chip), 1 Gb (318 chip)	21–35 Gb	36 Mb per hour	40–100 Mb

Error rate per incorporation event (%)	1	0.1	5	1	≤1	15	≥4
Final error rate (%)	1	0.1	≤0.1	1	≤1	1	4
Cost per Mb ($)	10	0.07	0.13	1 (318 chip)	0.45–0.6	>7	1
Instrument cost ($)	<500 k	700 k	530 K	50 K	~1 million	695 K	N/A
Advantages	Better alignment at repetitive regions due to long reads	Low error rate; lowest cost per Mb; good performance in homopolymer tract	Very low error rate	Low capital investment; short run time	Sequencing at single-molecule level	Sequencing at single-molecule level; direct sequencing of methylated DNA; long reads; short run time	Sequencing at single-molecule level; capable of direct analyzing of DNA, RNA, and protein; ultra-long reads; predicted low capital investment
Limitations	Not cost-effective; low throughput; high error rate in homopolymer track	Short reads; long run time	Short reads; long run time	Low throughput; high error rate in homopolymer track	Short read; high capital investment; long run time	High capital investment; high cost per Mb; high error rate	High error rate; not available in market yet

[56–58]. Mapping of genome-wise methylation profiles has been carried out in different research areas using various methods to enrich methylated regions followed by massively parallel sequencing, such as methylated DNA immunoprecipitation sequencing (MeDIP-seq), methyl-CpG binding domain sequencing (MBD-seq), and bisulfite sequencing, all of which require additional treatments during library constructions [59–62]. In SMRT technology, the detection of an incorporation event is characterized by two parameters: pulse width and interpulse duration. Pulse width measures the duration between binding of a nucleotide with template and the cleavage of the fluorophore-associated phosphate group, while interpulse duration represents the interval of two successive pulses. A proof of concept study proposed that methylated DNA can be readily detected by SMRT technology since the kinetics of DNA polymerase was predicted to be sensitive to DNA modifications and therefore will generate distinct profiles for pulse width and interpulse duration comparing to those of unmodified templates [63]. Indeed, the results demonstrated that different methylation patterns, including methyladenine (mA), methylcytosine (mC), and hydroxymethylcytosine (hmC), were readily detected and distinguished from unmethylated controls based on their unique pulse measurements.

Along with all the advancements offered by the SMRT technology, there are challenges as well that need to be addressed. A high error is a major concern for the SMRT platform. Among a read with 158 bases, only 131 bases aligned correctly to the reference, corresponding to 82.9 % sensitivity [63]. Among these, indels account for 74 % of the errors generated. Several solutions, such as increasing pulse width and signal intensities, have been proposed to improve the accuracy. Another effective solution is to sequence circularized template molecule multiple times (>=15). This strategy dramatically increased the sensitivity to 99 % since most of the errors are generated stochastically and the probability of having two errors occurring at the same locus is greatly reduced with increased coverage. Although more improvements are pending to fine-tune the detection system and increase accuracy, SMRT technology advances MPS by providing single-molecule-based, high-throughput information at both genomic and epigenomic levels.

3.2.3 GridION: A Nanopore Sequencer (Oxford Nanopore Technologies)

Another new MPS technology, Nanopore sequencing, dissects the sequence of template molecule by recording the current change as a single-strand DNA or RNA passes through a 2 nm pore embedded in a semiconductor chip [64]. Without the addition of adaptors, double-strand DNA is used directly as template for sequencing. During sequencing, the ionic flow is disrupted as the molecule is trapped inside the channel. The unique base-specific parameters measured for each disruption are used to determine the sequence read. Instead of capturing light signals, this new approach simply records electrical signals, thereby dramatically speed up the sequencing process at reduced cost. Being a potential powerful candidate of the $1,000 genome competition [65], two nanopore technology-based

instruments are aimed for commercialization in 2012 by Oxford Nanopore technologies, moving a step further of this third-generation MPS platform from theory to application.

3.3 Comparison of the Second- and Third-Generation MPS Platforms

As a variety of MPS platforms coexist in the market, a cross comparison between them may facilitate the platform selection based on the applications and reported performances. While the second-generation MPS technologies rely on the detection of cluster generated signals, third-generation MPS technologies capture signals emitted by single molecule, enabling direct sequencing of DNA and RNA molecules at high speed. Furthermore, by increasing the read length from tens of bases to several kilobases, third-generation MPS-based de novo assembly and insertion/deletion (indel) detections will be dramatically enhanced. However, the prevailing use of the second-generation technologies in the biomedical and diagnostic fields will continue due to their stable systems and much lower error rates. Table 3.1 shows a comparison of platform performance using commonly adapted metrics.

4 Conclusions

With the arrival of the MPS era, single-gene studies have gradually evolved into panel and genome-wide studies. This transition challenges bioinformaticians to improve the algorithms for resequencing-based reads alignment and de novo assembly of unknown sequences, computer scientists to maintain and update public databases, as well as biologists to interpret filtered results and follow with functional studies. As feedback, unsolved biological puzzles will guide engineers to continue to enhance the performance of current MPS platforms or invent fundamentally novel technologies aimed for new applications. Clearly, the MPS race will continue [66] as does the advancement of our understandings of the human genome and unsolved genetic diseases.

Acknowledgments I sincerely thank Drs. Jing Wang, Megan L. Landsverk, Victor W. Zhang, and Lee-Jun Wong for reviewing this manuscript and their invaluable comments.

References

1. Margulies M, Egholm M, Altman WE et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380
2. Bilguvar K, Ozturk AK, Louvi A et al (2010) Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. Nature 467(7312):207–210

3. Campeau PM, Lu JT, Sule G et al (2012) Whole-exome sequencing identifies mutations in the nucleoside transporter gene SLC29A3 in dysosteosclerosis, a form of osteopetrosis. Hum Mol Genet 21(22):4904–4909
4. O'Roak BJ, Deriziotis P, Lee C et al (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 43(6):585–589
5. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073
6. Jones MA, Bhide S, Chin E et al (2011) Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. Genet Med 13(11):921–932
7. Wang J, Cui H, Lee N-C et al (2012) Clinical application of massively parallel sequencing in the molecular diagnosis of glycogen storage diseases of genetically heterogeneous origin. Genet Med 15(2):106–14
8. Gowrisankar S, Lerner-Ellis JP, Cox S et al (2010) Evaluation of second-generation sequencing of 19 dilated cardiomyopathy genes for clinical applications. J Mol Diagn 12(6):818–827
9. Baetens M, Van Laer L, De Leeneer K et al (2011) Applying massive parallel sequencing to molecular diagnosis of Marfan and Loeys-Dietz syndromes. Hum Mutat 32(9):1053–1062
10. Meuzelaar LS, Lancaster O, Pasche JP, Kopal G, Brookes AJ (2007) MegaPlex PCR: a strategy for multiplex amplification. Nat Methods 4(10):835–837
11. Tewhey R, Warner JB, Nakano M et al (2009) Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat Biotechnol 27(11):1025–1031
12. Kirkness EF (2009) Targeted sequencing with microfluidics. Nat Biotechnol 27(11):998–999
13. Schlipf NA, Schüle R, Klimpe S et al (2011) Amplicon-based high-throughput pooled sequencing identifies mutations in CYP7B1 and SPG7 in sporadic spastic paraplegia patients. Clin Genet 80(2):148–160
14. Ghadessy FJ, Ong JL, Holliger P (2001) Directed evolution of polymerase function by 0 self-replication. Proc Natl Acad Sci 98(8):4552–4557
15. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD (2006) Amplification of complex gene libraries by emulsion PCR. Nat Methods 3(7):545–550
16. Valencia CA, Rhodenizer D, Bhide S et al (2012) Assessment of target enrichment platforms using massively parallel sequencing for the mutation detection for congenital muscular dystrophy. J Mol Diagn 14(3):233–246
17. Wheeler DA, Srinivasan M, Egholm M et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872–876
18. Levy S, Sutton G, Ng PC et al (2007) The diploid genome sequence of an individual human. PLoS Biol 5(10):e254
19. Zhang MQ (1998) Statistical features of human exons and their flanking regions. Hum Mol Genet 7(5):919–932
20. Sakharkar MK, Chow VT, Kangueane P (2004) Distributions of exons and introns in the human genome. In Silico Biol 4(4):387–393
21. Barnes WM (1994) PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. Proc Natl Acad Sci USA 91(6):2216–2220
22. Tang S, Huang T (2010) Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. Biotechniques 48(4):287–296
23. He Y, Wu J, Dressman DC et al (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 464(7288):610–614
24. Zhang W, Cui H, Wong LJ (2012) Comprehensive 1-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem 58(9):1322–1331
25. Li M, Schönberg A, Schaefer M, Schroeder R, Nasidze I, Stoneking M (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet 87(2):237–249
26. Cui H, Li F, Chen D, et al (2013) Comprehensive next-generation sequence analyses of the entire mitochondrial genome reveal new insights into the molecular diagnosis of mitochondrial DNA disorders. Genet Med. 2013 Jan 3. [Epub ahead of print], PMID 23288206

27. Coppieters F, De Wilde B, Lefever S et al (2012) Massively parallel sequencing for early molecular diagnosis in Leber congenital amaurosis. Genet Med 14(6):576–585
28. Turner EH, Lee C, Ng SB, Nickerson DA, Shendure J (2009) Massively parallel exon capture and library-free resequencing across 16 genomes. Nat Methods 6(5):315–316
29. Dahl F, Stenberg J, Fredriksson S et al (2007) Multigene amplification and massively parallel sequencing for cancer mutation discovery. Proc Natl Acad Sci 104(22):9387–9392
30. Dahl F, Gullberg M, Stenberg J, Landegren U, Nilsson M (2005) Multiplex amplification enabled by selective circularization of large sets of genomic DNA fragments. Nucleic Acids Res 33(8):e71
31. Stenberg J, Dahl F, Landegren U, Nilsson M (2005) PieceMaker: selection of DNA fragments for selector-guided multiplex amplification. Nucleic Acids Res 33(8):e72
32. Fredriksson S, Banér J, Dahl F et al (2007) Multiplex amplification of all coding sequences within 10 cancer genes by Gene-Collector. Nucleic Acids Res 35(7):e47
33. Porreca GJ, Zhang K, Li JB et al (2007) Multiplex amplification of large sets of human exons. Nat Methods 4(11):931–936
34. Hodges E, Xuan Z, Balija V et al (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39(12):1522–1527
35. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4(11):907–909
36. Albert TJ, Molla MN, Muzny DM et al (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4(11):903–905
37. Gnirke A, Melnikov A, Maguire J et al (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27(2):182–189
38. Cancer Genome Atlas Research Network (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474(7353):609–615
39. Boileau C, Guo D-C, Hanna N et al (2012) TGFB2 mutations cause familial thoracic aortic aneurysms and dissections associated with mild systemic features of Marfan syndrome. Nat Genet 44(8):916–921
40. Glycogen storage diseases (GSDs) are a group of inherited genetic defects of glycogen synthesis or catabolism. GSDs are categorized into 14 subtypes, based on the specific enzyme deficiency and disease phenotype. Common symptoms include hypoglycemia, hepatomegaly, developmental delay and muscle cramps. Based on major clinical presentation, GSDs can be divided into two sub-forms: muscle and liver. This comprehensive panel includes genes involved in both the muscle and liver forms of GSDs
41. Metabolic myopathies are genetic disorders of energy metabolism due to defects in the pathways of carbohydrate and fatty acid catabolism, and the subsequent energy production through mitochondrial oxidative phosphorylation. Mutations in genes involved in three major pathways of energy metabolism, including glycogenolysis, fatty acid oxidation, and mitochondrial oxidative phosphorylation are the main causes of metabolic myopathies. These groups of diseases are clinically heterogeneous with variable penetrance, severity and age of onset. The predominant clinical symptoms associated with metabolic myopathy include chronic muscle weakness, myoglobinuria, and/or acute and recurrent episodes of irreversible muscle dysfunction related to exercise intolerance. Patients with metabolic myopathy are usually diagnosed based on their clinical features, abnormal metabolites, and enzymatic deficiency. However, the biochemical and molecular analytical procedures are time-consuming, costly, and often not confirmatory. Definitive diagnosis is made through the identification of deleterious mutations in the causative gene. Early diagnosis of these conditions is important for prompt clinical management and improved outcome of the patients
42. Oh JD, Kling-Backhed H, Giannakis M et al (2006) The complete genome sequence of a chronic atrophic gastritis Helicobacter pylori strain: evolution during disease progression. Proc Natl Acad Sci USA 103(26):9999–10004
43. Smith MG, Gianoulis TA, Pukatzki S et al (2007) New insights into Acinetobacter baumannii pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. Genes Dev 21(5):601–614

44. Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature 442(7099):199–202
45. Tarasov V, Jung P, Verdoodt B et al (2007) Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. Cell Cycle 6(13):1586–1593
46. Wicker T, Schlagenhauf E, Graner A, Close T, Keller B, Stein N (2006) 454 sequencing put to the test using the complex genome of barley. BMC Genomics 7(1):275
47. Rothberg JM, Hinz W, Rearick TM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352
48. Elliott AM, Radecki J, Moghis B, Li X, Kammesheidt A (2012) Rapid detection of the ACMG/ACOG-recommended 23 CFTR disease-causing mutations using ion torrent semiconductor sequencing. J Biomol Tech 23(1):24–30
49. Vogel U, Szczepanowski R, Claus H, Junemann S, Prior K, Harmsen D (2012) Ion torrent personal genome machine sequencing for genomic typing of Neisseria meningitidis for rapid determination of multiple layers of typing information. J Clin Microbiol 50(6):1889–1894
50. Whiteley AS, Jenkins S, Waite I et al (2012) Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. J Microbiol Methods 91(1):80–88
51. Junemann S, Prior K, Szczepanowski R et al (2012) Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. PLoS One 7(8):e41606
52. Eid J, Fehr A, Gray J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323(5910):133–138
53. Meissner A (2010) Epigenetic modifications in pluripotent and differentiated cells. Nat Biotechnol 28(10):1079–1088
54. Esteller M (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet 8(4):286–298
55. Delaval K, Feil R (2004) Epigenetic regulation of mammalian genomic imprinting. Curr Opin Genet Dev 14(2):188–195
56. Dobrovic A, Simpfendorfer D (1997) Methylation of the BRCA1 gene in sporadic breast cancer. Cancer Res 57(16):3347–3350
57. Plath K, Mlynarczyk-Evans S, Nusinow DA, Panning B (2002) Xist RNA and the mechanism of X chromosome inactivation. Annu Rev Genet 36:233–278
58. Horsthemke B, Wagstaff J (2008) Mechanisms of imprinting of the Prader–Willi/Angelman region. Am J Med Genet A 146A(16):2041–2052
59. Taiwo O, Wilson GA, Morris T et al (2012) Methylome analysis using MeDIP-seq with low DNA concentrations. Nat Protoc 7(4):617–636
60. Harris RA, Wang T, Coarfa C et al (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. Nat Biotechnol 28(10):1097–1105
61. Serre D, Lee BH, Ting AH (2010) MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 38(2):391–399
62. Ball MP, Li JB, Gao Y et al (2009) Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol 27(4):361–368
63. Flusberg BA, Webster DR, Lee JH et al (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7(6):461–465
64. Branton D, Deamer DW, Marziali A et al (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26(10):1146–1153
65. Pennisi E (2012) 23 July, 2012. New start date and first contestant for genomics X PRIZE. ScienceInsider
66. von Bubnoff A (2008) Next-generation sequencing: the race is on. Cell 132(5):721–723

Chapter 4
Sequence Alignment, Analysis, and Bioinformatic Pipelines

Fuli Yu and Cristian Coarfa

Abstract As DNA sequencing becomes more affordable and data 'tsunami' is practically here, it is clear that the inforamtics and analysis are the rate-limiting step currently for scientific discoveries as well as medical actions in the genomics enterprise. In this chapter, soon after the brief historical overview on the genomics field, we describe details on alignment and variation analysis algorithms and software packages. Remaining challenges and potential directions are discussed at the end of the chapter.

1 Historical Overview

1.1 From HGP and HapMap to the 1000 Genomes Project

The Human Genome Project (HGP) was a daunting scientific pursuit that lasted more than a decade (1990–early 2000) [1, 2] in the history of modern biomedical research. The completion of the HGP marked the introduction of "genomics" into the daily practice of almost every discipline in the biomedical research enterprise [3]. Since the publication of the human reference genome sequences, the biomedical research community has been enabled to leap forward in an unprecedented way in many fields of study. For example, it allows evolutionary biologists to compare the human genome to genomes from other species such as chimpanzees and rhesus

F. Yu (✉)
Department of Molecular and Human Genetics, Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston 77030, TX, USA
e-mail: fyu@bcm.edu

C. Coarfa
Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston 77030, TX, USA

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_4,

[4, 5] and gain tremendous insights into evolutionary history at the nucleotide level, where the "tree of life" is no longer based on phenotypic similarities rather on genotypes – the core "engine" for species evolution. The availability of the physical map along with the genetic map allows investigators to quickly narrow down a reasonable list of genes of interest in a large interval by browsing the reference genome sequences, so the time and efforts in disease gene mapping have been vastly eased. More importantly, the HGP for the first time provided a reference sequences for individual resequencing data to compare against and therefore global survey of the various kinds of genetic variations that are commonly present among individuals became feasible. This is central to genetic studies.

These genetic polymorphisms are considered the key factors that make each of us phenotypically unique and have a major impact on how we are differentially predisposed to hereditary diseases and how we respond to environmental insults (such as bacteria, viruses, and chemicals), drugs, and other treatments [6]. Thus, genetic variation is of great value for biomedical research and molecular diagnostics.

There are different types of genetic variations. Single-nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genomic sequence is altered and are by far the most abundant form of polymorphisms. Their high density and the ease of the scoring methods make SNPs ideally suitable as markers for genetic mapping. The International Haplotype Map Project (HapMap) has characterized most of the common SNPs in the genome [7–10]. Small-size insertions-deletions (INDELs) are the second most frequent polymorphism type and are more polymorphic than SNPs with potentially more drastic protein functional consequences when they alter the codon sequences (i.e., frameshift INDELs) [11, 12]. Even in case where the codon sequences are not disrupted, INDELs may manifest themselves as a triplet repeat (CGG or CAG) in diseases such as Huntington's and Fragile X in either the protein coding sequences or outside of the protein coding sequences [13]. Triplet repeat exceeding certain threshold (i.e., expansion) results in toxicity due to abnormal level of aggregation due to polyQ tract [14]. A fraction of the INDELs are manifested as variable number of tandem repeats (VNTR), also known as short tandem repeats (STRs). They consist of variable length sequence motifs that are repeated in tandem with variable copy numbers (i.e., microsatellites and minisatellites). They have tremendous size heterogeneity and can have hundreds of alleles per locus. Copy number variations (CNVs) and structural variations (SVs), including genomic region duplications, deletions, inversions, and translocations, are being uncovered more frequently than expected [15] in normal populations. They may affect the susceptibility to disease and differential drug responses by altering the expression dosages of the encompassed genes. A number of large-scale efforts have been carried out in the past decade to characterize the genomic landscape of CNV/SVs [16]. This type of genomic aberration is especially prevalent in cancer genomes and rare inherited disorders.

Because of the informativeness of SNPs in various aspects of genetics, the International HapMap Consortium was formed with six participating countries, including Canada, China, Japan, Nigeria, the United Kingdom, and the United States in 2002 [8]. This enterprise initially aimed "to determine the common

patterns of DNA sequence variation in the human genome" by characterizing the common SNPs, their frequencies, and their correlation patterns in four different populations [17]. By the end of February 2005, the group has reached completion of its original goal (Phase I) of genotyping about one million SNPs across the entire human genome [17]. The group expanded their efforts to include more SNPs in Phase II [10] and more individuals from eleven diverse ethnicities in "HapMap 3" [9].

Tremendous progress has been made by the HapMap Project [7, 8] on cataloging common SNPs. Methodologies are being developed to utilize a much smaller subset of markers as proxies (tagging SNPs) to represent the underlying haplotype structures of almost all the common variants to scan for significant marker association to a phenotypic trait or a drug response [18]. This laid the foundation for the genotyping array platforms (e.g., Affymetrix 6.0 SNP array and Illumina BeadChip) that have been widely applied to genome-wide association studies (GWAS) in many common diseases [19].

The impact of the HapMap Project has been profound in various biomedical fields including complex disease gene mapping, population genetics, and evolutionary genetics. More than 1,617 associations were uncovered for 249 common traits; the correlation patterns among SNPs were characterized in great detail (haplotypes) [9, 10, 17]; great progress was achieved in understanding the demographic history and natural selection process [20, 21] in the last 100,000 years of human history; genomic features that demonstrated significant impact on biological mechanisms, such as recombination, were identified [22–24]; and computational algorithms were developed by imputing more markers for association mapping using HapMap as the reference [25]. A large number of software tools have been developed and widely applied in the GWAS era for analyzing SNP array data and visualizing the haplotype patterns [26].

As the GWAS era unfolded, it became evident there was a so-called missing heritability issue, that is, the identified associations can only account for a small fraction (<10 %) of the genetics for almost all the common traits [27]. A working hypothesis is that the combinatorial effects of a few genetic variants segregating at lower frequencies (1–5 %) can explain a large fraction of the heritability of common traits (so-called rare variants-common trait hypothesis). To this end, the International 1000 Genomes Project was launched in 2007 (http://www.1000genomes.org/sites/1000genomes.org/files/docs/1000Genomes-MeetingReport.pdf). The goal of the 1000 Genomes Project is to "find most genetic variants that have frequencies of at least 1 % in the populations studied" by surveying ~2,500 individuals from more than 20 ethnicities (http://www.1000genomes.org/about). The project design includes three sub-projects: (a) low-coverage sequencing at 4–6X read depth across the entire human genome, (b) targeted capture sequencing of coding sequence, and (c) high-coverage sequencing on a few samples [28]. So far, ~40 million genetic variants including 38 million SNPs, 1.4 million INDELs, and 14,000 CNVs/SVs have been discovered (http://www.1000genomes.org/phase1-analysis-results-directory). The realization of the 1000 Genomes Project depends on the technological advances of the "next-generation sequencing (NGS)" platforms.

1.2 NGS Platforms

By far, DNA sequencing/genotyping technologies have evolved into one of the most robust and automated biomedical assays that can be routinely applied on a large-scale fashion. The deployments of the NGS platforms (e.g., Roche 454, Life Technologies SOLiD and Ion Torrent, and Illumina's HiSeq and MiSeq) in not only the large genome centers like the BCM-Human Genome Sequencing Center but also the smaller laboratories, such as clinical laboratories, allow a wide spectrum of scientific questions to be explored. It strongly reflects the vision and determination of scientific leaders, such as James Watson and Francis Collins, that aspired for rapid developments in DNA technologies more than 20 years ago at the dawn of the HGP. There have been a number of very informative review articles on the technologies, prices, and comparisons among these platforms [29] that can be tremendously helpful for readers.

The core innovations of NGS platforms are massively parallel chemical reactions, ultrahigh-resolution optics, and computational methods to analyze very short reads. These revolutionary technological advances have drastically reduced the sequencing cost and shortened the turnaround time to merely a few days (instead of tens of years). According to the price chart by the NIH-NHGRI, we are in the $10 k/genome arena in 2012 (http://www.genome.gov/images/content/cost_per_genome.jpg). It will keep decreasing by the introduction of a number of 3rd generation platforms. A possible new platform is Oxford Nanopore although skepticism remains. As the field rapidly changes, we expect the landscape and the key players both to change quite a bit in the next 2–3 years.

1.3 Paradigm Shift Enabled by NGS

Human genetics studies that aim to map causal mutations for a certain phenotypic trait (both common and rare) have been considered to be lengthy for decades. The positional cloning approach can take years to find candidate intervals that are associated with the trait. To narrow down to a particular gene mutation can be an expensive pursuit. The prominent example is the positioning cloning of cystic fibrosis gene that took ~4 years [30]. In contrast to the older mapping approaches, now with whole-exome capture sequencing (WECS) technologies [31], getting almost all the functional variants from 30 to 60 Mb target regions can be a project for a few days that costs under $1,000.

WECS has become highly feasible to systematically interrogate genes that are inherited in a Mendelian manner. The NIH has initiated scientific programs "Mendelian Disorders Genome Centers Program" (http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-10-016.html) and "Clinical Sequencing Exploratory Research Projects" (http://grants.nih.gov/grants/guide/rfa-files/rfa-hg-12-009.html) to systematically discover the causal mutations for rare inherited diseases and

cancers and to examine the utilities and impact of DNA sequencing in a clinical setting. We are witnessing a paradigm shift in the clinical genetic testing field where WECS is replacing panel-by-panel tests. They are offered by a handful of institutions in the United States, including the BCM clinical laboratories. Rapid delivery (in terms of days) of sequencing results has been achieved by smart implementation of informatics tools [32].

2 Informatics in NGS Analysis

2.1 Pipeline Overview

The NGS platforms can be exploited to either interrogate the whole genome or the coding regions by using capture protocols to enrich targeted regions of the genome. Regardless of the sequencing strategies, the major data processing and analytical components include the following steps (Fig. 4.1), and a very exhaustive list of tools can be found in SEQanswers (http://seqanswers.com/wiki/Software/list):

1. Primary Data Analysis Using the Manufactures' Software. A FASTQ [33] file is produced that consists of both the nucleotides in a contiguous sequence (ranging from tens of bases to hundreds of bases) and their base quality scores, reflecting how accurate the nucleotide calls are. The quality score is in a so-called phred scale, which is the negative logarithmic scale of the error rate for calling a particular base multiplied by ten. So in the case of an A associated with a score Q20, it means the probability of the A being erroneous is 0.01.
2. Mapping/Alignment. This is an important step that maps the sequencing reads to the human reference genome and compares a subject's sequence to the reference genome. The mapping/alignment step has been undergoing intensive research and development, and a number of algorithms have become widely utilized, such as BWA and BFAST [34–38].

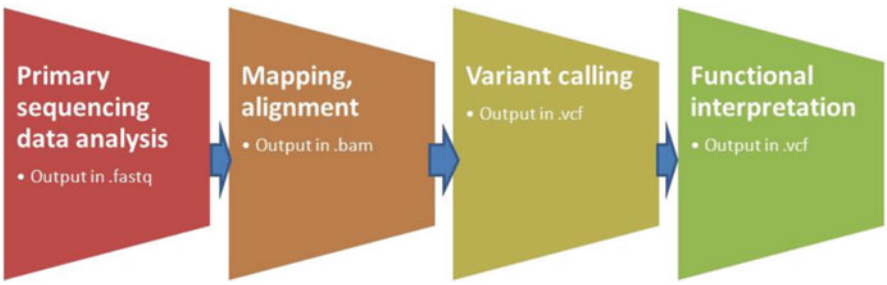

Fig. 4.1 Overview of a general bioinformatics pipeline for NGS data processing. The cartoon depicts the typical workflow and general output file formats, which are generally shared across different sequencer platforms

3. Variant Calling. The accuracy of detecting genetic variants including SNPs, INDELs, CNVs, and SVs is going to determine the signal to noise ratio in gene mapping for both association studies of common and complex diseases and Mendelian diseases. The International 1000 Genomes Project has motivated many groups, to spearhead the rapid informatics development and to release open source software packages to the NGS community. Here, we will summarize in more detail two particular software suites (Atlas2 and SNPTools) that our group developed in the past 3 years while we led the variant analysis in the 1000 Genomes Project.
4. Functional Interpretation of Variants. The understanding of the biological association of the genetic variations can be a rate-limiting step. However, several developments in genetic databases (such the 1000 Genomes, HGMD for germline mutations, and COSMIC for cancer) and computational algorithms (PolyPhen-2 and SIFT) made it possible to shorten the candidate list of loci quite substantially.

2.2 Nuts and Bolts

2.2.1 Alignment

The development of read alignment algorithms and software is a prolific research field, with new methods developed and published at a fast pace. We will summarize in detail the general read alignment algorithms and tools. We will also briefly discuss the alignment of bisulfite-treated reads, as recent advances enabled querying at the same time both nucleotide variation and DNA methylation status.

While string matching is a venerable problem in the fields of information theory and computer science, the development of read mapping algorithms has been motivated and shaped by the advances in sequencing technologies. Whereas optimal algorithms exist, they tend to be impractical for the size of mammalian genomes, on the order of several billion base pairs. The technical challenges are multiple, namely, accuracy of mapping, ensuring that the resources used (memory, storage) fit within commodity workstations, and that the speed of analysis can match the ever increasing sequence throughput. An overview of alignment software and common applications is presented in Table 4.1.

Technology Evolution Dictates Alignment Methods

Prior to NGS, Sanger reads had sequencing lengths up to 1,000 bases, with low-quality bases at both ends. Alignment methods had to be sensitive to both gaps and mismatches. Due to the low throughput of Sanger sequencing, speed of the alignment was a secondary concern to accuracy. The large increase in the number of sequencing reads brought about by NGS demands faster run time and more

Table 4.1 Overview of various NGS applications and preferred alignment software packages

Application	Aligner software packages used
Whole-genome/whole-exome resequencing, Illumina	Bowtie, BWA, Bowtie2, SOAP2
Whole-genome/whole-exome resequencing, SOLiD	BFAST, SHRIMP
Whole-genome/whole-exome copy number variation	mrsFAST, RazerS, Hobbes, Novoalign
454 resequencing	SSAHA2, Pash 3.0, BWA-SW
Whole-genome DNA bisulfite-sequencing analysis	Bismark, BSMAP, Pash, LAST, BS-seeker, BRAT

optimized computational performances. However, the higher throughput combined with shorter read length enabled aligners to initially improve the alignment speed by using one or a combination of heuristics, such as performing ungapped alignment or restricting the number of acceptable differences between the reads and reference genome. These heuristics have had a generally negative impact on the ability to map reads onto the large fraction of the human genome that is semi-repetitive and to map reads that carry sequence variants not present in the reference sequence, either due to naturally occurring genomic variants [40]. The length of sequencing reads has continually increased; this opened opportunities to map more efficiently onto the large fraction of genomic DNA that contains repetitive elements and segmental duplications. As a result, newer aligners place again an emphasis on the ability to map across mismatches and gaps.

Optimal Alignment by Gold Standard Aligner

The gold standard Smith-Waterman alignment (S-W) algorithm [41], which performs base pair-level comparisons, is guaranteed to produce optimal alignment results. The S-W algorithm assigns a gain for base pair match and penalties for base pair mismatch and for gaps. Considering a sequence of length M and a reference of length N, the algorithm explores an $M \times N$ matrix, such that S(i,j) corresponds to the highest scoring local alignment ending in the i-th position in the sequence and the j-th position in the reference. The key idea is of trying the scoring all possible alignment extensions, such as a base pair match, a base pair mismatch, or a gap in either the sequence or the reference. Both the running time and memory footprint have an O(MN) complexity, making them impractical even if run on the fastest processors.

Seed-and-Extend Paradigm

The "seed-and-extend" paradigm for fast read mapping emerged during the early Sanger sequencing era and has been implemented in comparison tools such as FASTA [42], BLAST [34], BLAT [43], SSAHA [44], and Mosaik http://bioinformatics.bc.edu/marthlab/Mosaik). These "seed-and-extend" tools perform filtering of potential similarities using k-mer level matches, called "seeds," and limit base pair-level comparisons to the areas around the seeds, thus reducing the total

number of base pair-level comparisons; an important contribution was the rigorous statistical characterization of general scoring schemes (cite Karlin and Altschul 1990, PNAS) and the discovery of significant alignments above those that can be found by chance (cite the BLAST paper, J Mol Biol 1990 Oct 5;215(3):403–410. Basic local alignment search tool. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ.). To reduce the cost of extending spurious mappings, additional filters, such as minimum number of seeds or minimum seed occurrence in the reference, were added.

Seed-and-Extend Alignment with Gapped Patterns

Tools such as BLAST and BLAT initially relied on using contiguous seeds to guide the subsequent extend step. Shorter seeds yield more sensitive matches but also lead to longer alignment time due to spurious matches; longer seeds generate more accurate mappings but could miss mappings with high dissimilarity. An innovation was the use of gapped seeds [45–47], which achieve higher sensitivity by allowing mismatches, and higher specificity by having a larger length; seed patterns and spaced seeds for various sequence similarities have been determined via extensive simulations. For short reads generated by the early versions of NGS systems, such as Illumina, gapped seeds were used to ensure the detection of mappings with a small number of mismatches. Eland and SOAP [36] use six seeds to find reads with two mismatches; MAQ [48] finds reads with two mismatches in the first 28 bp. More generally, RMAP used seeds of size $(k + 1)$ nucleotides to detect alignments with at most k mismatches [49, 50].

Zoom inferred optimal size and location of the spaced seeds, depending on the desired alignment sensitivity [51]. Pash uses spaced seeds, indexes the query reads, and then relies on the index of the seed in the reads to run a k-mer-level approximation of the Smith-Waterman algorithm prior to the expensive extend step; it trades accuracy for execution speed by adapting the seed sampling frequency based on read length; it has been used successfully in the detection of interspecies conserved synteny blocks among the human genome and clinically relevant model organisms such as mouse and rat [35, 52, 53].

Burrows-Wheeler Transform Methods

A major breakthrough in the development of aligners for high-throughput sequencing reads was the use of the FM-index, based on the Burrows-Wheeler Transform (BWT), which enables detection of exact matches in a time period proportional with string length, regardless of the target size. Backtracking was then used to enable the mapping of reads with mismatches. The first generation of BWT-based tools such as BWA [37], SOAP2 [54], and Bowtie [55] provided ungapped alignments extremely efficiently; however, performance was degraded when processing reads containing small INDELs. Eventually, FM-index based mappers accounted for

gaps. BWA-SW employs similar principles as BWA and works for long reads, up to 100,000 base pairs, with sensitivity similar to mappers such as SSAHA2 [44] and Pash 3.0 [35, 52, 53]. The latest entry in the group is Bowtie2 [38], which uses the FM-index to seed alignments and implements local alignment using vectorized CPU instructions, similar to tools such as SHRIMP [56]. The FM-index based aligners have provided the best combination of speed and accuracy over the past 3 years, and for this reason they are also some of the most widely used, in premier applications such as 1000 Genomes [28].

q-Gram Filtering Methods

A number of spaced seeds methods further improve their performance by employing the q-gram filter. The key observation is that for a sequence of length w to map with at most m mismatches onto a reference, when using q-size seeds (called q-grams), the sequence and the reference need to share at least $w-(m+1)*q$ q-grams. By imposing m based on the read length or application needs, mappers then reduce the numbers of extensions performed. SHRIMP further speeds up the alignment step by using vectorization and thus taking advantage of low-level parallelism present in modern x86 Intel CPUs in the form of SSE2 instructions. mrsFast [57] aims to correctly identify all mapping locations of an input read, to enable accurate detection of structural polymorphisms in repeat rich regions; it further improves performance by utilizing a cache-oblivious strategy; it first determines a list of reads l1 that map to a set target regions l2, decomposes them until they fit in the CPU cache, and finally performs the alignment step. RazerS employs gapped q-grams; it precomputes optimal q-gram size and spacing, using dynamic programming, for a range of read lengths and intended mapping sensitivity [58]; suitable read mapping locations in q-gram space across the genome are determined efficiently with sliding diagonal parallelograms using the SWIFT algorithm [59]. Hobbes implements q-gram filtering and ensures that candidate seed enumeration is performed for the genome hash entries with the fewest number of entries [60]. It speeds up search by encoding in a binary format (with 1 bit used to represent 4 base pairs) the neighborhood of a seed and checking for mismatches using assembly level instructions; it improves cache performance by prefetching genome hash entries.

DNA Methylation Profiling via Bisulfite Sequencing

Bisulfite sequencing is an accurate method of determining base-level DNA methylation status. Sample DNA is treated with bisulfite, after which NGS is employed. Methylated cytosine bases are preserved as Cs in the reads, and unmethylated ones are converted to Us. A pioneering project employed whole-genome bisulfite sequencing to reconstruct two human methylomes, sequencing a total of 4.8 billion reads, or 376 Illumina lanes [61]. Alignment methods have to account for Us mapping to either Ts or Cs in the reference and on reads originating from either strand.

A common method employs mapping of the reads on the forward and reverse genome strand, both with all Cs maintained and with all Cs converted to Ts, performed using commodity mappers such as Bowtie. A subsequent processing step deconvolutes the mappings of each read and determines the most likely mapping locations. This method was employed in the first human methylome effort [61], and by a number of software packages such as Bismark [62], BS-seeker [63], and methylcoder [64].

A number of bisulfite-sequencing mapping tools have been purposely tailored to perform efficient and accurate mapping. BSMAP [65] performs bisulfite sequence mapping by hashing the reference genome and by hashing multiple seeds around the CpG sites, according to the possible methylation status. Pash 3.0 hashes all k-mers that may arise from bisulfite treatment of reads, after which regular mapping occurs; both the forward and the reverse complement strand of each chromosome are used as a reference for mapping; the resulting alignments can contain gaps. mrsFAST [57] employs a cache-oblivious strategy to perform ungapped mapping of bisulfite reads. RMAP-BS [49] performs gapped mapping of bisulfite-treated reads. The LAST [66] alignment software builds on the gapped seeds strategy employed by BLAST to achieve good sensitivity and practical mapping times for bisulfite-seq data. Other mappers tailored to bisulfite-seq data are BRAT [67, 68] and Novoalign (www.novocraft.com).

2.2.2 Variant Calling from NGS Data

It is essential for NGS projects to accurately detect genetic variants. The higher error rates of the NGS platforms compared to Sanger sequencing technologies [29] and the complexity of mapping shorter reads to the reference genome present challenges for variant calling, including the identifications of SNPs and short-range insertions-deletions (INDELs). One of the major challenges in SNP discovery from NGS data is to distinguish true individual variations from errors introduced by either sequencing artifacts or inaccurate alignments, by thoroughly examining error modes. There are a number of key software packages that are widely used by the NGS analysis community including Samtools [69], SOAPsnp [70], GATK [71], and FreeBayes (http://bioinformatics.bc.edu/marthlab/FreeBayes). In the following sections, we will review two software packages and their underlying algorithms that have been developed in our group – one is called Atlas2 [72, 73], another SNPTools [74]. Atlas2 is more relevant to the whole-exome capture sequencing data analysis, where SNPTools is suitable for population sequencing data analysis.

Atlas2: A Software Suite for SNP/INDEL Calling in Exome Capture Sequencing

Whole-exome capture sequencing (WECS) is a cost-effective approach to identify the mutations of highest biomedical importance, generating hypotheses for

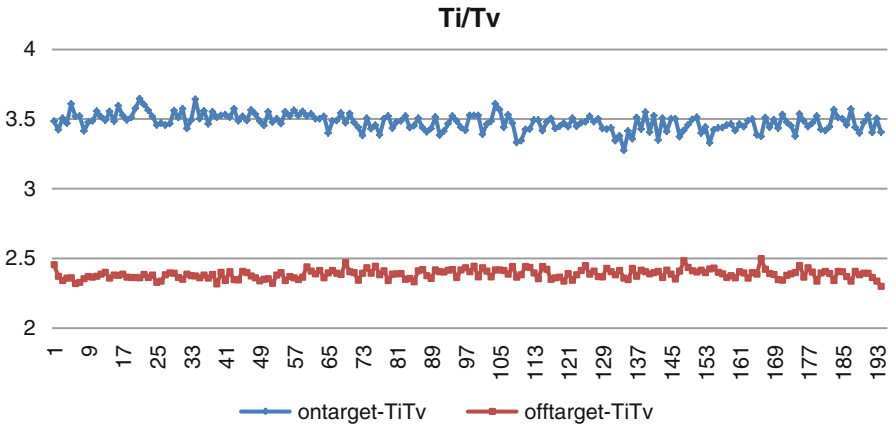

Fig. 4.2 The Ti/Tv ratio of discovered SNPs from WECS data using Atlas2. We sampled 193 exome data from the 1000 Genomes Project. The Ti/Tv of on-target SNPs (i.e., coding) is ~3.5, much higher than the genome average in mostly off-target regions where Ti/Tv is ~2.4

downstream follow-up. WECS data introduces a set of biases and error patterns distinct from whole-genome sequencing, such as heterogeneous depth coverage, and reference bias due to capture [75]. The coding regions of the human genome also modify the expectations of a number of quality metrics that are routinely used in variant calling. For example, the transition/transversion ratio (Ts/Tv) in the coding sequences is expected to be ~3.5, higher than the noncoding average of ~2.4 (Fig. 4.2). This higher ratio within the coding sequencing is resulted from the selection pressure on transversions, which likely causes codon changes. Few bioinformatics tools were primarily developed for WECS analysis, with requirement for extensive manual adjustments – such as implementing several additional ad-hoc filters pertinent to coding regions. We have developed the Atlas2 software suite to specifically accommodate the WECS data analysis by taking into consideration of features that are particular pertinent to exome data. Atlas2 detects and accounts for systematic sequencing errors caused by context-related variables in a logistic regression model. Variables such as proportion of reads with variants (SNPs or INDELs) are specifically tuned for WECS data. It then estimates the posterior error probability for each identified substitution through a Bayesian method that integrates prior knowledge of the error probability of the given substitution and SNP rate among humans and the results from the logistic regression model. Based on the estimated posterior error probability, one can better separate true SNPs from false positives. Parameters for Atlas2 are being extensively tuned using SNP array data, PCR validated sites, HapMap, and ENCODE resequencing data [8–10, 17]. Highly accurate and sensitive results were achieved in WECS by using the Atlas2 suite [28, 39].

SNPTools: An Integrative Suite for Variant Analysis in Large Cohort Studies

As a continuation of the HapMap and GWAS, studies with thousands of samples are a major focus in human genetics studies. This kind of large-scale cohort studies (e.g., CHARGE-S, http://web.chargeconsortium.com/) or population sequencing projects (e.g., the 1000 Genomes Project, http://www.1000genomes.org/) paves the way to understanding the genotype-phenotype associations and properties regarding human populations such as their demographics [76]. There is great need to produce integrative SNP calls in NGS data set from thousands of samples with phenotypic information. Until high-coverage sequencing becomes affordable for large cohort interrogation, many large cohorts studies aggregate NGS data with low-coverage data across thousands of subjects [77] to characterize the variants. Linkage disequilibrium (LD) between variants allows for imputation to refine the individual genotypes to improve specificity and sensitivity [78] and to phase the variants into haplotypes that are key to phenotypic associations.

Motivated by the 1000 Genomes Project [28], which are sequencing ~2,500 samples from ~30 ethnicities, we devised a framework that consists of four major steps: (1) A probabilistic model to re-weight each base so the base quality and mapping quality are both rescaled into a single quality score. (2) SNP site detection based on a variance ratio indexing framework that realizes high sensitivity and specificity. (3) Initial genotyping using a "BAM-specific binomial mixture model (BBMM)" to provide raw genotype likelihoods for three possible genotypes. The BBMM can effectively overcome the high level of heterogeneity in the population sequencing data and produce high-quality results. And (4) a novel imputation-based approach for genotype refinement to achieve high accuracy. This novel approach can effectively reduce computational burden. This pipeline can also integrate data from multiple sources (e.g., exome, low-coverage whole-genome sequencing, and SNP genotyping data) via an innovative "data integration" step (Fig. 4.3) [74].

Both Atlas2 and SNPTools produce high-quality SNP calls using either a single-sample schema or a multiple-sample schema. The selection of which approach to apply depends on the overall number of samples, data coverage, and computational resources available to the users. In a clinical sequencing project that consists of tens of WECS, Atlas2 returns high-quality results. In a cohort sequencing project (e.g., CHARGE-S) with thousands of samples, SNPTools is likely to be more suitable when high-performance computing resources are readily accessible.

2.2.3 Functional Interpretation of Genetic Variation

The ultimate challenge for geneticists is to interpret the variants and to prioritize a short list of potential mutations from hundreds of thousands and even millions of candidate loci. The NGS platforms and data processing pipelines for mapping, alignment, and variant calling can carry out the heavy-lifting to produce an impressive list of variants. Often, researchers are overwhelmed by large number of variants. This is increasingly becoming the most significant bottleneck both for studies mapping

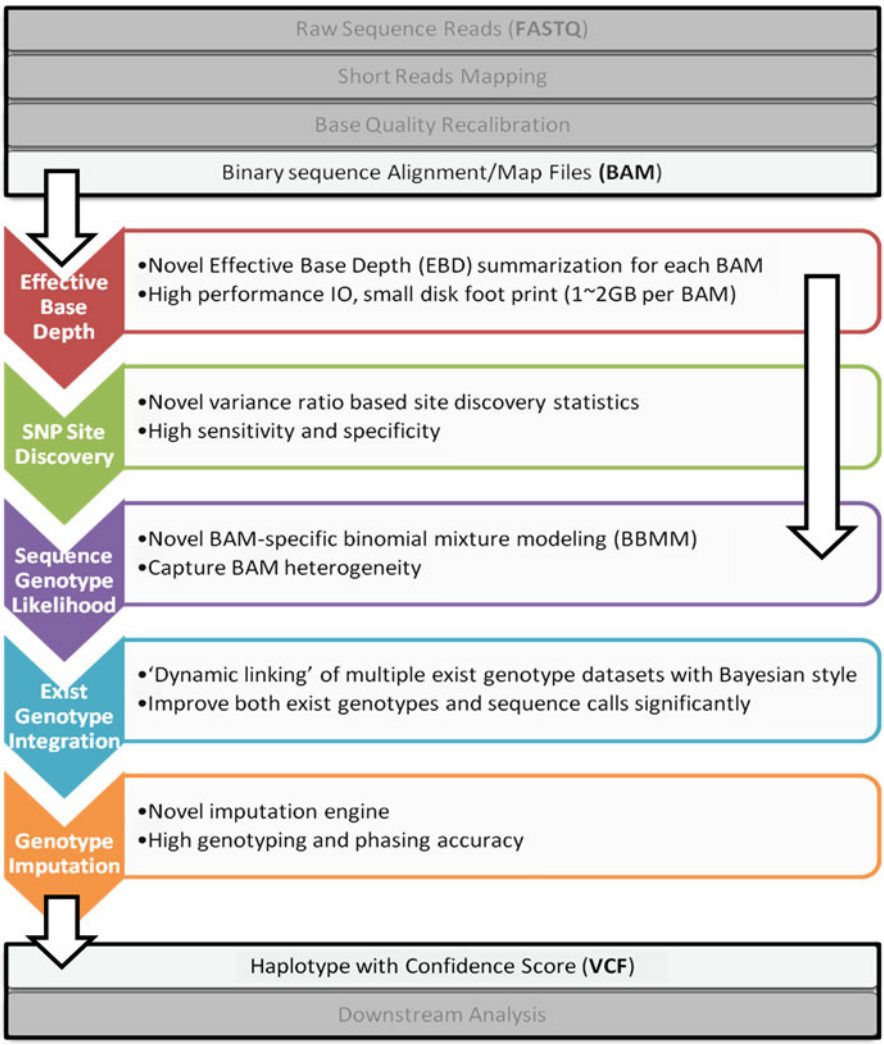

Fig. 4.3 Pipeline overview of our SNPTools software for large-scale population NGS projects. The SNPTools processes BAMs to produce "effective base depth (EBD)" for each base that reflects both the base quality and mapping quality. The candidate SNPs are discovered from considering all the BAMs. It next models the genotype clusters using a BAM-specific mixture modeling (BBMM) process to estimate genotype likelihoods. Final genotypes are determined by using a novel imputation algorithm implemented in the package

disease causes and for clinical genetic diagnosis leveraging the power of exome NGS technologies [79].

There are mainly three approaches currently for the functional interpretation of one's variant list. (1) Functional annotation based on existing knowledge known

(e.g., disease mutations, and functional databases). Such databases include the HGMD for germline mutations (http://www.hgmd.cf.ac.uk/ac/index.php) and Cosmic for cancer somatic mutations (http://www.sanger.ac.uk/genetics/CGP/cosmic/). Functional information from project such as ENCODE presents ambitious efforts that experimentally catalogue expression patterns of the genome (http://www.genome.gov/10005107). (2) In silico predictions made by PolyPhen-2 and SIFT. There are a number of software packages [80–85] available for scoring the functional impact of polymorphisms, using evolutionary conservation in the DNA sequences, and/or protein structural and biochemical properties in the coding region. This line of research is under very active development now. (3) Thresholding by the population frequency of the variant alleles. The rationale is that most of the deleterious mutations tend to be rare due to negative selection over time, with minor allele frequency <0.5–1 %. It is particularly useful for studies with Mendelian disease models. Many databases can serve this purpose – common practice includes using dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/), HapMap (http://hapmap.ncbi.nlm.nih.gov/), 1000 Genomes (http://www.1000genomes.org/), and NIEHS exome variant project (http://evs.gs.washington.edu/niehsExome/) for variant frequency filtration.

The users will utilize all three categories of information for functional interpretation of the genetic variants from the NGS data. One good example is the Cassandra annotation suite developed by Dr. Matthew Bainbridge at the BCM-Human Genome Sequencing Center (Matthew Bainbridge personal communication). The variants are assessed for their effects on both a conservative (RefSeq) and inclusive (UCSC) gene model set. This assessment includes whether a variant changes an amino acid residue, occurs proximally to an intron-exon boundary, creates or removes a stop-codon, or occurs intronically, within noncoding RNA, or intergenically. Nonsynonymous variants are further annotated based on their computationally predicted deleteriousness by different algorithms [83–85]. Variants are additionally annotated for their frequency and presence in, or proximity to, other known variants in multiple variant collections (dbSNP, 1000 genomes, NEIHS exome variant project, HGMD). Lastly, variants are annotated based on functional data for the gene in which they occur. These data include known function of the gene, previous association of the gene with disease, posttranslational modifications of the gene, and the expression profile of the gene across human tissues.

3 Discussion

3.1 Remaining Challenges in Informatics

With the increased sequencing throughput and broadened applications, there are significant data processing informatics challenges in terms of IT infrastructure, scientific algorithm development, and efficient implementation.

3.1.1 Storage

The immediate impact of the continually increased sequencing throughput is a growing strain on the existing storage infrastructure. Advanced data compression and retrieval techniques such as CRAM [86] are under active development, a more efficient BAM file format. Whereas local file servers are being overrun, commercial avenues such as cloud storage are gaining more interest (Amazon, Google).

3.1.2 Computation Time

In general, at every processing step there is a need for more computationally efficient methods with the most effective approaches leading to new algorithms and data structures. There are also engineering aspects such as utilizing languages that compile to efficient code such as C/C++ for time intensive computations. New algorithms should be designed to exploit the entire gamut of parallelism available: low-level parallelization using vectorization capabilities present in both CPUs and GPUs (made available by most vendors, CUDA), efficiently scaling to the ubiquitous multiple cores by using multithreaded applications and libraries, and finally dividing large computing jobs both at the levels of compute clusters but also cloud computing. The bioinformatics workbench of the near future is most likely powered by cloud computing and distributed databases.

3.2 Cloud Computing for Genomics

The deployment of genomic analysis software as a service within a cloud computing framework offers a unique solution for these problems. The concept behind cloud computing is to outsource computation to third-party servers or clusters at a remote location. This software as a service model removes the upfront investment requirement and any delays associated with building local computing infrastructure, enabling flexibility and scalability.

As a pilot, our group has integrated our genomic variant analysis pipeline – Atlas2 Suite – onto Amazon cloud [87]. We also performed a case study using this pipeline as a proof of concept to demonstrate the potential of personal genome analysis on the cloud [87]. With many MiSeq and HiSeq sequencers now streaming data to the cloud server in real time, we expect to see the genomics fields gravitate to cloud usage in the near future. It effectively streamlines the process and reduces the burden for many end users that do not have sufficient access to computational resources or expertise. Our pilot study demonstrated that the cost is reasonable for processing a few personal exomes or genomes. When scaling up to thousands of genomes, costs become an issue for current price. Data transfer into the cloud and the data

security are concerns too. We remain optimistic for the cloud usage in genomic sciences as we view the issues as technical ones.

3.3 Concluding Thoughts: Grand View of NGS Informatics

DNA technologies have become essential to many branches of research not restricted to SNP identification and gene mapping. They can be applied to understanding the more dynamic parts of the biological machinery: RNA-Seq, ChIP-Seq, epigenomic sequencing, and human microbiota sequencing are all being actively pursued. They hold promise to be part of the routine clinical diagnostic toolkits in the future. As we are zeroing in on a very affordable price tag for whole-genome sequencing, the prospect of having national level projects for genetics and environmental research is brightening. It has great potential to thoroughly characterize the risk factors arising from genotypes, environments, and the interaction between them. It will also advance medical practice for intervention and prevention, reducing healthcare costs in a long run [88].

NGS informatics is certainly a critical component to realize this overarching goal. As one of the most active research areas, many software packages have been developed. Open questions remain. Some are scientific ones: methods for INDEL and CNV/SV detections are under constant improvement; annotation of functional impacts of polymorphisms needs to integrate with biological knowledge; and robust statistical methods is required to improve signal to noise ratio in mutation detection for both Mendelian diseases or complex disorders. Others are engineering ones: scientific software needs to be designed for scalability in order to match with increase in the NGS throughput.

Acknowledgments We thank R. Alan Harris for critical comments on the earlier version of this chapter.

References

1. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431:931–945
2. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921
3. Green ED, Guyer MS (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470:204–213
4. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69–87
5. Gibbs RA et al (2007) Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234
6. Daly AK (2010) Genome-wide association studies in pharmacogenomics. Nat Rev Genet 11:241–246

7. Consortium TIH (2004) Integrating ethics and science in the International HapMap Project. Nat Rev Genet 5:467–475
8. Consortium TIH (2003) The International HapMap Project. Nature 426:789–796
9. Altshuler DM et al (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467:52–58
10. Frazer KA et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861
11. Mills RE et al (2006) An initial map of insertion and deletion (INDEL) variation in the human genome. Genome Res 16:1182–1190
12. Mills RE et al (2011) Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res 21:830–839
13. Orr HT, Zoghbi HY (2007) Trinucleotide repeat disorders. Annu Rev Neurosci 30:575–621
14. Gatchel JR, Zoghbi HY (2005) Diseases of unstable repeat expansion: mechanisms and common principles. Nat Rev Genet 6:743–755
15. Kidd JM et al (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453:56–64
16. Berger MF et al (2011) The genomic complexity of primary human prostate cancer. Nature 470:214–220
17. Consortium TH (2005) A haplotype map of the human genome. Nature 437:1299–1320
18. Lin M, Aquilante C, Johnson JA, Wu R (2005) Sequencing drug response with HapMap. Pharmacogenomics J 5:149–156
19. Cozen W et al (2012) A genome-wide meta-analysis of nodular sclerosing Hodgkin lymphoma identifies risk loci at 6p21.32. Blood 119:469–475
20. Sabeti PC et al (2006) Positive natural selection in the human lineage. Science 312:1614–1620
21. Sabeti PC et al (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449:913–918
22. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310:321–324
23. Myers S et al (2006) The distribution and causes of meiotic recombination in the human genome. Biochem Soc Trans 34:526–530
24. Myers S, Freeman C, Auton A, Donnelly P, McVean G (2008) A common sequence motif associated with recombination hot spots and genome instability in humans. Nat Genet 40:1124–1129
25. Li Y, Willer C, Sanna S, Abecasis G (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406
26. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265
27. Manolio TA et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753
28. 1000 Genomes Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073
29. Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11:31–46
30. Collins FS (1992) Cystic fibrosis: molecular biology and therapeutic implications. Science 256:774–779
31. Albert TJ et al (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4:903–905
32. Saunders CJ et al (2012) Rapid whole-genome sequencing for genetic disease diagnosis in neonatal intensive care units. Sci Transl Med 4:154ra135
33. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771
34. Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

35. Coarfa C et al (2010) Pash 3.0: a versatile software package for read mapping and integrative analysis of genomic and epigenomic variation using massively parallel DNA sequencing. BMC Bioinformatics 11:572
36. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714
37. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760
38. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
39. Marth GT et al (2011) The functional spectrum of low-frequency coding variation. Genome Biol 12:R84
40. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4:e7767
41. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197
42. Pearson WR (1994) Using the FASTA program to search protein and DNA sequence databases. Methods Mol Biol 24:307–331
43. Kent WJ (2002) BLAT–the BLAST-like alignment tool. Genome Res 12:656–664
44. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729
45. Ma B, Tromp J, Li M (2002) PatternHunter: faster and more sensitive homology search. Bioinformatics 18:440–445
46. Li M, Ma B, Kisman D, Tromp J (2003) PatternHunter II: highly sensitive and fast homology search. Genome Inform 14:164–175
47. Li M, Ma B, Kisman D, Tromp J (2004) Patternhunter II: highly sensitive and fast homology search. J Bioinform Comput Biol 2:417–439
48. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858
49. Smith AD et al (2009) Updates to the RMAP short-read mapping software. Bioinformatics 25:2841–2842
50. Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128
51. Lin H, Zhang Z, Zhang MQ, Ma B, Li M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics 24:2431–2437
52. Coarfa C, Milosavljevic A (2008) Pash 2.0: scaleable sequence anchoring for next-generation sequencing technologies. Pac Symp Biocomput:102–113
53. Kalafus KJ, Jackson AR, Milosavljevic A (2004) Pash: efficient genome-scale sequence anchoring by Positional Hashing. Genome Res 14:672–678
54. Li R et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967
55. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25
56. Rumble SM et al (2009) SHRiMP: accurate mapping of short color-space reads. PLoS Comput Biol 5:e1000386
57. Hach F et al (2010) mrsFAST: a cache-oblivious algorithm for short-read mapping. Nat Methods 7:576–577
58. Weese D, Emde AK, Rausch T, Doring A, Reinert K (2009) RazerS–fast read mapping with sensitivity control. Genome Res 19:1646–1654
59. Rasmussen KR, Stoye J, Myers EW (2006) Efficient q-gram filters for finding all epsilon-matches over a given length. J Comput Biol 13:296–308
60. Ahmadi A et al (2012) Hobbes: optimized gram-based methods for efficient read alignment. Nucleic Acids Res 40:e41
61. Lister R et al (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. Nature 462:315–322

62. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27:1571–1572
63. Chen PY, Cokus SJ, Pellegrini M (2010) BS Seeker: precise mapping for bisulfite sequencing. BMC Bioinformatics 11:203
64. Pedersen B, Hsieh TF, Ibarra C, Fischer RL (2011) MethylCoder: software pipeline for bisulfite-treated sequences. Bioinformatics 27:2435–2436
65. Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinformatics 10:232
66. Frith MC, Mori R, Asai K (2012) A mostly traditional approach improves alignment of bisulfite-converted DNA. Nucleic Acids Res 40:e100
67. Harris EY, Ponts N, Le Roch KG, Lonardi S (2012) BRAT-BW: efficient and accurate mapping of bisulfite-treated reads. Bioinformatics 28:1795–1796
68. Harris EY, Ponts N, Levchuk A, Roch KL, Lonardi S (2010) BRAT: bisulfite-treated reads analysis tool. Bioinformatics 26:572–573
69. Li H et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
70. Li R et al (2009) SNP detection for massively parallel whole-genome resequencing. Genome Res 19(6):1124–1132
71. DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498
72. Shen Y et al (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res 20:273–280
73. Challis D et al (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13:8
74. Wang Y, Lu JT, Jin Y, Gibbs R, Yu F. Integrative imputation-based framework for variant analysis in population genomics studies. jtlu@bcm.edu (In revision)
75. Bainbridge MN et al (2010) Whole exome capture in solution with 3 Gbp of data. Genome Biol 11:R62
76. Gravel S et al (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA 108:11983–11988
77. Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR (2011) Low-coverage sequencing: implications for design of complex trait association studies. Genome Res 21:940–951
78. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12:443–451
79. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12:628–640
80. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814
81. Chun S, Fay JC (2009) Identification of deleterious mutations within three human genomes. Genome Res 19:1553–1561
82. Liu X, Jian X, Boerwinkle E (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32:894–899
83. Cooper GM et al (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901–913
84. Davydov EV et al (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6:e1001025
85. Adzhubei IA et al (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249
86. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res 21:734–740
87. Evani US et al (2012) Enabling Atlas2 personal genome analysis on the cloud. In Genomic Signal Processing and Statistics (GENSIPS), 2011 I.E. international workshop. San Antonio
88. Boerwinkle E (2012) Translational genomics is not a spectator sport: a call to action. Genet Epidemiol 36:85–87

Chapter 5
Protein Structural Based Analysis for Interpretation of Missense Variants at the Genomics Era: Using MNGIE Disease as an Example

Victor Wei Zhang

Abstract The analysis of a group of genes or the whole exome by massively parallel sequencing provides an efficient way to molecular diagnosis of genetic disorders. Meanwhile, tremendous amount of variant data produced brings great challenges for interpretation, particularly for novel variants without functional evidences. There is an urgent need for a nonexperimentally based method to understand their genotype and phenotype correlation. Moreover, missense variants contribute to about 50 % disease-causing changes. This chapter will demonstrate the value of in silico stereochemical analysis in the interpretation of disease-causing missense mutations using thymidine phosphorylase as an example. The integrated approaches including disease clinical phenotype, biochemical genetic analysis, computational prediction, and structural based modeling can be helpful in understanding the pathogenic mechanism of these missense variants. The protein structural based analysis is a valuable tool to visualize small but significant structural changes, thus, can be incorporated into routine variant interpretation pipeline.

1 Introduction

The development of massively parallel sequencing (MPS, also known as next generation sequencing, "NGS") technology has revolutionized the laboratory practice for clinical molecular diagnosis. The simultaneous analysis of a group of genes involved in a common pathway and/or similar phenotype, or the whole exome

V.W. Zhang (✉)
Medical Genetics Laboratories, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, NAB 2015, Houston, TX 77030, USA
e-mail: wzhang2@bcm.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_5,

sequencing (WES), has provided higher diagnostic yields with reduced cost and time. Meanwhile, the tremendous amount of sequence data produced brings great challenges for variant interpretation. The types of disease-causing mutations can be missense, nonsense, splice site mutation, and small insertions/deletions. The interpretations of nonsense, splice site mutation, and small insertions/deletions resulting in reading frameshift and truncated proteins are relatively straightforward according to ACMG (American College of Medical Genetics) guideline. However, in the absence of functional assays, the interpretation of the pathogenicity of novel missense variants is always difficult. Thus, in the absence of functional studies, the missense novel variants have always been classified as variants of unknown clinical significance (VUS) according to ACMG guideline. Publicly available mutation databases provide information regarding reported disease-causing mutations, but recently, publications indicated that about 25 % of reported variants are likely to be misclassified as mutations [1]. The recent performance evaluation of nine commonly used computational algorithms for missense variant classification revealed that these predicted results are not consistent among themselves [2, 3]. In addition, commonly used computational algorithms for the prediction of pathogenicity of missense variants do not have the analytic sensitivity and specificity to be reliably used in the clinical setting. Due to the large number of genomic variants generated from massively parallel sequencing studies, there is an urgent need for a nonexperimentally based method for the evaluation of the structural/functional consequences of a missense change in order to assist the interpretation of the variants in the context of patient's clinical symptoms, disease progression, and biochemical genetic evaluation.

In this chapter, I will demonstrate the value of the in silico stereochemical analysis in the interpretation of disease-causing missense mutations based on the available X-ray crystallographic structure, using thymidine phosphorylase (TP) as an example.

2 Thymidine Phosphorylase (TP) and Mitochondrial Neurogastrointestinal Encephalopathy (MNGIE) Syndrome

2.1 Biochemical Function of Human Thymidine Phosphorylase (TP)

Human thymidine phosphorylase (EC 2.4.2.4) plays an important role in cellular thymidine metabolism and pyrimidine homeostasis. It functions as a homodimer in cytosol that catalyzes the reversible phosphorylysis of thymidine or deoxyuridine by the breakage of the glycosidic bond to form thymine or uracil and 2-deoxyribose-1-phosphate. The expression of TP appears to be ubiquitous in a

variety of human tissues, with high expression levels in the digestive system and brain, but not in muscle and renal system [4]. Defects in TP (OMIM #603041) impair the nucleotide salvage pathway resulting in elevated plasma thymidine concentration and imbalanced deoxynucleotide pools, the common causes of the mitochondrial DNA (mtDNA) depletion syndromes [3].

2.2 Clinical Presentation and Pathogenesis of MNGE Disease

Human thymidine phosphorylase (TP) is encoded by the *TYMP* gene. Mutations in *TYMP* cause pan-ethnic autosomal recessive disease called mitochondrial neurogastrointestinal encephalomyopathy (MNGIE), affecting multiple organ systems. High levels of *TYMP* expression in tissues of digestive system and both central and peripheral nerve systems correlate with the observed early and severe gastrointestinal symptoms and leukoencephalopathy as well as peripheral neuropathy in affected patients [5]. Unlike the broad phenotypic spectrum caused by DNA polymerase gamma (POLG) deficiency, clinical features of TP deficiency are rather homogenous, characterized by early symptoms of gastrointestinal dysmotility and cachexia, that lead the patients to seek medical evaluation. These characteristic clinical features can be distinguished from the wide clinical spectrum of other mitochondrial disorders. The patients with severe enzyme deficiency due to null mutations often have the recognizable symptoms in early childhood, with mean age of onset at the second decade of life [6]. These patients usually do not survive beyond 30–40 years old. The high mortality is largely due to the malnutrition status, the complication of GI system, or infection. Patients with partial enzyme deficiencies have disease onset in the fifth or sixth decade of life [7, 8] with symptoms of ptosis, ophthalmoparesis, and peripheral neuropathy. The diffuse leukoencephalopathy by MRI presented among MNGIE patients is the important hallmark in differential diagnosis for patients with similar clinical presentations [6, 9–11].

2.3 Mutation Spectrum of TP Responsible for MNGE Disease

More than 70 disease-causing mutations have been identified in patients with human TP deficiency. The spectrum of mutations in *TYMP* includes missense, nonsense, splice site mutations, and small insertions/deletions. No gross genomic structure abnormalities have been reported in MNGIE patients. The disease association of nonsense, splice site mutation, and small insertions/deletions is relatively straightforward, because these mutations usually result in truncated proteins without activities. However, to interpret the missense change in the context of disease pathogenesis is always challenging. As shown in Fig. 5.1a and Table 5.1, the majority of reported mutations in the *TYMP* gene of MNGIE patients are missense changes (Table 5.2).

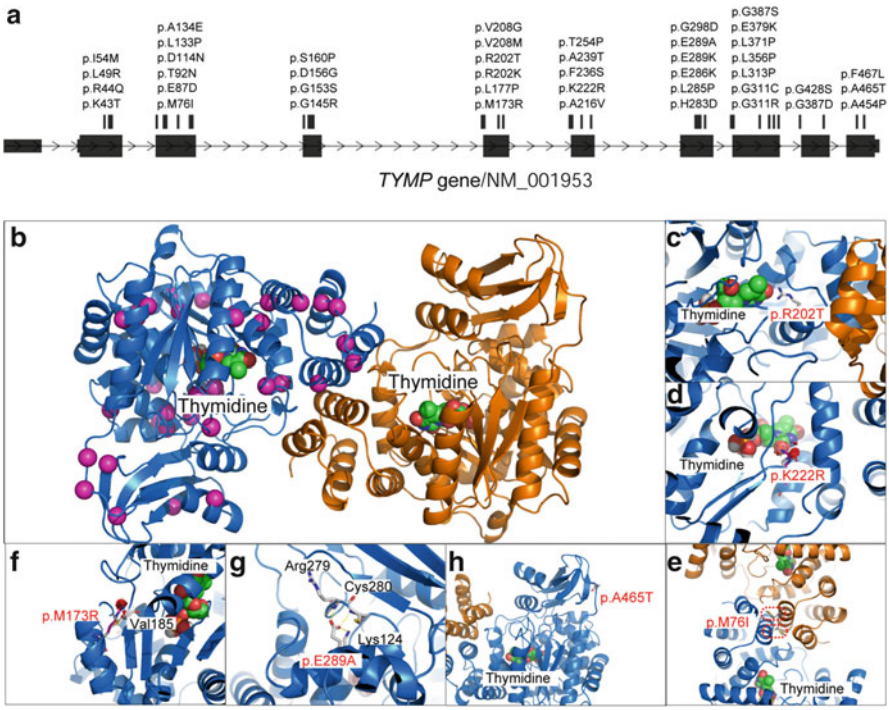

Fig. 5.1 Missense TYMP mutations distribution and their structural basis. (**a**) Distribution of reported missense mutation in the exons of human *TYMP* gene. (**b**) Ribbon representation of human TYMP protein complex structure with two monomeric subunits colored in *blue* and *orange*. Thymidine is represented in *spheres*. Missense mutations are mapped in the complex structure in *purple spheres*. (**c**) p.R202T. (**d**) p.K222R. (**e**) p.M76I. (**f**) p.M173R. (**g**) p.E289A. (**h**) p.A465T

Table 5.1 Summary of reported mutations of MNGIE disease

Mutation category	Number of mutations
Missense	44
Nonsense	4
Splicing	12
Frameshift	19
Total	79

Table 5.2 Integrated approaches to understand missense variant/mutation of MNGIE disease

#	Codon change	AA change	Predictions				Structural analysis	Class	Clinical severity	Enzyme activity	Ref.
			Conservation[a]	MutPred[b]	Polyphen2[c]	SNPs&GO[d]					
1	c.128A>C	p.K43T	6	0.88	1.00	8	Stabilize a helix in the dimer interface	B	44 years	0	[9]
2	c.131G>A	p.R44Q	4	0.83	1.00	4	Direct interact with b. Ala206 and b.Asp209 to form dimer interface	B	22/29 years, GI/ ophthalmo-paresis	0	[24]
3	c.146T>G	p.L49R	8	0.86	1.00	9	Stabilize a helix in the dimer interface	B	N/A	<10 %	[6]
4	c.162C>G	p.I54M	8	0.90	1.00	5	Stabilize protein core for helices in the dimer interface	B	24 years, diarrhea, ptosis	N/A	[29]
5	c.228G>A	p.M76I	6	0.94	0.98	8	Interact with residues of b subunit to form interface	B	60 years, *late onset*, no GI symptom	<1 %	[31]
6	c.261G>C	p.E87D	9	0.96	0.97	8	Stabilize a helix in the dimer interface by interacting act with a. lys43	B	12 years, cachexia, pseudo-obstruction	0	[25]
7	c.275C>A	p.T92N	8	0.88	1.00	7	Stabilize protein core helices	C	12 years, cachexia, GI dysmotility	N/A	[32]
8	c.340G>A	p.D114N	9	0.98	1.00	8	Located in the beta sheet of the protein core and close to active site	A,C	27 years, pseudo-obstruction, ophthal-moplegia	N/A	[25]

(continued)

Table 5.2 (continued)

#	Codon change	AA change	Predictions				Structural analysis	Class	Clinical severity	Enzyme activity	Ref.
			Conservation[a]	MutPred[b]	Polyphen2[c]	SNPs&GO[d]					
9	c.398T>C	p.L133P	7	0.91	1.00	9	Located in a helix and the protein core	C	23 years, diarrhea, pseudo-obstruction, ptosis	~2.5 %	[33]
10	c.401C>A	p.A134E	9	0.86	1.00	9	Close to surface, create clashes with a.Cys182	C	N/A	<10 %	[6]
11	c.433G>A	p.G145R	9	0.95	1.00	9	Adjacent to active site	A	N/A	<5 %	[23]
12	c.457G>A	p.G153S	9	0.98	1.00	9	Adjacent to active site	A	N/A	<5 %	[23]
13	c.467A>G	p.D156G	9	0.99	1.00	9	Close interaction with a.Arg408 and a. Arg146	C	24 years	N/A	[9]
14	c.478T>C	p.S160P	6	0.94	1.00	9	Close to surface and close packing of protein	C	Abnormal muscle biopsy	0	[28]
15	c.518T>G	p.M173R	7	0.88	0.99	7	Hydrophobic interacting with a.Val185 for packing	C	Abnormal muscle biopsy	0	[28]
16	c.530T>C	p.L177P	6	0.94	1.00	9	In the middle of helix	C	24 years	0	[9]
17	c.605G>A	p.R202K	9	0.89	1.00	7	Close to active site, may interact with thymidine	A	N/A	N/A	[34]
18	c.605G>C	p.R202T	9	0.94	1.00	9	Close to active site, may interact with thymidine	A	55 years, *late onset*	15 %	[8]
19	c.623T>G	p.V208G	8	0.77	1.00	9	Adjacent to active site, have hydrophobic interaction with thymidine	A	N/A	<10 %	[6]

20	c.622G>A	p.V208M	8	0.92	1.00	8	Adjacent to active site, have hydrophobic interaction with thymidine	A	61 years, *late onset*	16 %	[8]
21	c.647C>T	p.A216V	9	0.87	1.00	7	In the helix interacting with thymidine and tightly packed	A,C	N/A	<10 %	[6]
22	c.665A>G	p.K222R	9	0.93	1.00	8	Adjacent to active site, interaction with thymidine	A	N/A	0	[23]
23	c.707T>C	p.F236S	7	0.85	0.99	4	Forms the hydrophobic core with beta sheet and a helix	C,D	27 years, peripheral neuropathy, pseudo-obstruction	Normal for hetero-zygote	[11]
24	c.715G>A	p.A239T	7	0.77	0.99	4	Forms the hydrophobic interaction with adjacent residues	C	N/A	<10 %	[6]
25	c.760A>C	p.T254P	3	0.87	0.28	4	In the middle of helix	C,D	6 years	50 %, elevated thymi-dine	[9]
26	c.847C>G	p.H283D	1	0.92	0.15	5	Solvent exposure, likely to be tolerated	D	60 years *late onset*	??	[31]
27	c.854T>C	p.L285P	5	0.90	1.00	9	In the middle of helix spanning from His283 to Gly296, forming hydrophobic interac-tion with adjacent residues	C	57 years, *late onset*	9%	[8]

(continued)

Table 5.2 (continued)

#	Codon change	AA change	Predictions				Structural analysis	Class	Clinical severity	Enzyme activity	Ref.
			Conservation[a]	MutPred[b]	Polyphen2[c]	SNPs&GO[d]					
28	c.856G>A	p.E286K	7	0.94	1.00	9	Forming extensive hydrogen network with adjacent residues	C	10 years, recurrent diarrhea	N/A	[25]
29	c.866A>C	p.E289A	8	0.94	1.00	8	Solvent exposure, tightly packed with positively charged residues	C	N/A	0	[23]
30	c.865G>A	p.E289K	8	0.94	1.00	9	Solvent exposure, tightly packed with positively charged residues	C	N/A	N/A	[28]
31	c.893G>A	p.G298D	4	0.44	1.00	5	Solvent exposure, located in connecting loop	C,D	32 years, weight loss, ptosis, ophthalmo-paresis	0	[35]
32	c.931G>C	p.G311R	1	0.89	1.00	1	Tightly packed with adjacent helix	C	61 years, *late onset*	16 %	[8]
33	c.931G>T	p.G311C	1	0.72	1.00	5	Tightly packed with adjacent helix	C	N/A	<10 %	[6]
34	c.938T>C	p.L313P	3	0.95	1.00	9	Forming the protein hydrophobic core	C	N/A	0	[9]
35	c.1067T>C	p.L356P	4	0.53	0.99	6	In the middle of helix spanning from Asp353 to Gly363, forming hydrophobic interac-tions with adjacent residues	C	41 years	0	[6]
36	c.1112T>C	p.L371P	1	0.73	0.99	1, Neutral	At the end of a helix	C,D	19 years, diarrhea, ptosis, etc.	N/A	[29]

37	c.1135G>A	p.E379K	1	0.74	0.08	6, Neutral	Exposed to solvent, likely interacting with positive charges of p.Arg442	C	67 years, *late onset*, no peripheral neuropathy	18 %	[7]
38	c.1160G>A	p.G387D	3	0.89	1.00	9	Locate at the transition position from loop to beta sheet	C	19 years, anorexia nervosa	0	[25]
39	c.1159G>A	p.G387S	3	0.78	1.00	9	Locate at the transition position from loop to beta sheet	C	N/A	<10 %	[6]
40	c.1282G>A	p.G428S	2	0.94	1.00	9	Locate at the transition position from loop to beta sheet	C	33 years	0	[9]
41	c.1360G>C	p.A454P	2	0.57	0.92	4	In the middle of the helix from Trp437 to Pro450	C	N/A	<10 %	[6]
42	c.1393G>A	p.A465T	1	0.71	0.14	2, Neutral	Solvent exposed loop, no interacting residues	D	25 years, diarrhea, ptosis	N/A/ normal	[29, 30]
43	c.1401C>A	p.F467L	2	0.19	0.01	1, Neutral	Tightly packed with the hydrophobic atoms of charged residue	C	N/A	<10 %	[6]
44	c.1412C>T	p.S471L	2	0.88	0.46	6, Neutral	No interacting residues	D	GI, cachexia, peripheral neuropathy	N/A	[28]

N/A not available, *GI* gastrointestinal symptoms

[a] http://consurf.tau.ac.il (value represent conservation scores, 1 is variable and 9 is conserved)

[b] http://mutpred.mutdb.org (value represent the probability of deleterious mutation)

[c] http://genetics.bwh.harvard.edu/pph2 (value represent prediction confidence, 0 is benign and 1 is damaging)

[d] http://snps-and-go.biocomp.unibo.it/snps-and-go/ (value represent prediction reliability, 0 is unreliable and to 10 most reliable)

3 Current Approaches of Interpreting Missense Variants

3.1 Computational Prediction Without Structural Information

Amino acid substitutions, in general, account for the majority of the disease-causing mutations. It is also the most difficult to correlate missense variants with disease pathogenesis. Experimentally determined protein structure is highly accurate and informative in the inference of its function. However, to obtain the protein X-ray crystal structure is time-consuming and low throughput and may not even be successful. Therefore, due to the technical limitation or intrinsic properties of a protein, experimental structure is not always available. Differentiation of a benign polymorphism from a true mutation can be done by searching locus-specific mutation database, which can provide valuable information on the interpretation of known mutations. Nevertheless, these variants very often are novel or only have been reported in a limited number of cases without sufficient supporting evidences. To overcome these difficulties, many nonexperimental computational algorithms have been developed based upon a variety of theoretical models to assist with the interpretation of the enormous amount of variant data generated from NGS analyses, for example, PolyPhen and SIFT [12–15]. In general, these algorithms incorporate a wide spectrum of information, including sequence conservation, physical properties of amino acids, secondary structure information, and known locus-specific mutation database to construct the theoretic model. However, the recent performance evaluation of nine commonly used computational algorithms for missense variant classification revealed different results among these prediction methods themselves, and the best method may provide only 80 % accuracy [3]. The conclusion of that report is consistent with the results from another performance evaluation of 23 computational algorithms for the classification of cystathionine beta-synthase (CBS) missense mutations for homocystinuria (http://cagi.genomeinterpretation.org/content/CBS).

3.2 Structure-Based Homology Modeling

The comparison of the linear sequence of a protein to its orthologous protein across a variety of species has been extensively used to infer its biological function, which often applied to certain well-conserved critical residues. The detailed and thorough analysis of the 3D protein macromolecular structure can yield mechanistic insight into the stereochemical configuration at atomic resolution and expand our knowledge regarding the pathogenesis of the disease. In order to study the effect of an amino acid substitution on enzymatic function in terms of catalysis, the structure of enzyme-substrate complex must be available for the investigation of the atom-to-atom interactions between the substrate and the catalytic amino acid residues, and the effects on such interactions upon a missense alteration.

When the structure of a protein of interest cannot be experimentally determined, it may be constructed by homology modeling if the structure of a homologous protein in other species is available. Based on the homology model, the researchers can generate reasonable working hypothesis by in silico mutagenesis and design experiment to delineate the biological significance of missense substitutions to categorize the mode of action of these variations. The application of structural biology to clinical investigation has reviewed [16, 17] elsewhere.

The most commonly used methods for the investigation of macromolecular structural change are X-ray protein crystallography and nuclear magnetic resonance (NMR). A protein may adapt different structures in various biological forms. Multiple approaches can be used to obtain the atomic resolution of a macromolecular protein in order to examine individual atom and chemical bond. The protein structure can be determined by itself in the native state, in complex with substrate(s) or an inhibitor, or in a state with a mutated amino acid residue in the protein. While the importance of establishing structure-function relationships has been widely recognized, some of the relevant biological macromolecular structure, either small or large in size, cannot be crystallized or has intrinsic disordered protein for NMR study. Thus, the protein structures at atomic resolution are not always available at the time of investigation.

Using the experimentally determined structure for the protein of interest or a protein model generated from homologous structure template is generally considered to be more reliable for interpreting molecular outcome of missense variants than the theoretical computational prediction. By examining the atomic resolution of a protein structure, it can facilitate to infer the chemical interactions among atoms, the packing of a protein core, or the requirement for proper secondary structure formation [2].

4 Structural Basis of Analysis of TP and Its Missense Variants/Mutations

4.1 Catalytic Mechanism Deduced from TP Structure

The native structure of TP was determined at 3.0 Å resolution with the similar overall fold to other bacterial pyrimidine phosphorylases, despite low sequence similarity [18]. The monomer of TP is composed of two domains, the alpha helical amino sub-domain and alpha/beta carboxyl sub-domain. TP forms a dimer with coiled coil at the interface from amino terminal helices (Fig. 5.1b). The active site in each monomer is embedded between these two sub-domains. Based on the investigations of bacterial thymidine phosphorylase (bTP), the native bTP can adapt an open conformational change without substrate. Binding of substrates introduces new chemical interactions to stabilize the complex structure by inducing a relatively large domain movement to form the closed active conformation [19, 20]. However, this

conformational change has not been observed yet in the native structure of TP as compared to TP complex with a nucleotide analog, which is thought to be likely due to the intermolecular interactions which help to form the crystal packing lattice [18]. Multiple TP protein-substrate structures including complex with nucleotide analog, product thymine, or a chemical inhibitor are available [18, 21, 22]. Comparison of multiple crystal structures indicates that they can adapt very similarly closed active conformation, even in the absence of phosphate. The active sites generated by closing two sub-domains with different substrates are nearly identical.

4.2 Construction of Complete TP Substrate Complex

Although multiple TP protein-substrate structures are available, the structure of TP with two relevant substrates, phosphate and thymidine, has not been solved yet. To classify novel amino acid substitution in an incomplete structure is difficult because the missense variant can modulate the protein function in a variety of ways. The subtle but significant alternation in local stereochemical interactions can be easily overlooked. Based on the structures of both TP and bTP, construction of an atomic-resolution model will yield a highly accurate structure of TP-phosphate-thymidine complex. First, multiple structural alignments indicate that there is little difference among the TP structures, either native or other complexes. Second, the active site residues can be superimposed with high confidence. Third, different protein-substrate structures can complement with each other to create a more complete picture of the protein interactions. Fourth, the electron density from the original structure was examined when necessary for the configuration of the residues of interest. This rebuilt protein complex structure would be the desirable one to be used for the investigation of missense changes of the protein in the context of protein/substrate complex as a biological unit. Thus, the TP-phosphate-thymidine complex were built based on the three available TP structures, with both phosphate and thymidine molecules modeled. The configuration of the active site was also compared to that of the bacterial active site [19].

5 Evaluation of Disease-Causing Missense Mutation

5.1 Protein Structure Assisted Interpretation of Missense Variants

All reported missense mutations causing TP deficiency have been mapped onto the rebuilt TP-thymidine-phosphate complex structure to probe the structural and chemical changes (Fig. 5.1b). These structural perturbations can be examined and deduced at atomic resolution in the context of the local secondary or tertiary conformation.

The structure-function relationship was further delineated to correlate the genotype with the corresponding clinical phenotype. Based on the structural analysis, these reported missense mutations can be grouped into four classes: (A) those affecting catalytic site, (B) those interfering with dimerization interface, (C) those changing secondary structure and/or protein stability, and (D) those with unclear significance or likely to be benign.

5.1.1 Missense Mutations Affecting Catalytic Site (Class A)

In general, the conservation of residues around the catalytic site of the protein is often unusually high. Computational algorithms that incorporate such information in their model usually provide consistent prediction results among themselves and correlate well with biochemical findings. Probing structural perturbations around the active site for these missense mutations can provide critical information regarding pharmacological chaperone for treatment.

Mutations p.G145R and p.G153S are close to the catalytic site, substrate binding pocket, and the amino acid residues involved in catalysis of TP are highly conserved during evolution. Substitution of glycine with either arginine or serine would introduce a relatively large side chain into the active site configuration, either affecting the domain movement upon substrate binding or rendering the catalysis less effective. Both mutations in homozygous state have been reported in patients with severe enzyme deficiency (<5 % residual activity) [23]. Mutations p.R202T and p.V208M are also in the active site and interact with thymidine as shown in the TP complex model (Fig. 5.1c). All four computational algorithms made concordant deleterious prediction. However, both mutations have been reported in compound heterozygote state in a late onset MINGE patient with 15 % residual enzyme activity [8]. Upon close examination of the configuration of the complex model, it reveals that there are some spaces to allow them to adapt to a different conformation. Both changes can be considered as mild mutations, which do not completely abolish the enzymatic activity, despite that the location of these residues is at the active site.

Mutation p.K222R is also located at the active site. It interacts with the thymidine and participates in the formation of the key hydrogen bond with the phosphate [23]. While both lysine and arginine are positively charged residues, the bulky side chain of arginine is predicted to impair the substrate binding (Fig. 5.1d). The patient who carried the p.K222R mutation in compound heterozygote with a frameshift mutation had no TP enzyme activity [23], further confirmed the significant functional effect of this change.

5.1.2 Missense Mutations Interfering the Dimerization Interface (Class B)

TP protein forms a homodimer to carry out the conversion of thymidine to thymine. Some mutations in TP perturb neither the active site nor the stability but the

quaternary structure of the protein. The functional effect of simply quaternary structural perturbation is difficult to predict by computational approaches. The residues on the surface of the monomers can be visually inspected to assess their impact on quaternary structure. Moreover, the substitutions at the dimerization interface of homodimeric protein can have additive effect on the structure due to the symmetrical nature. The amino terminus of TP forms a coiled-coil structure with four tightly packed helices against the counterpart of the other subunit. Mutations p.K43T, p.R44Q, p.L49R, p.M76I, and p.E87D are located at the dimer interface (Fig. 5.1e). These substitutions would cause the perturbation of either the coiled-coil packing or the interaction with the other subunit [6, 9, 24, 25]. Unlike class A mutations, this class B mutation may exert less severe effect on the protein function, which may mildly reflect to the relatively late onset of the these patients.

5.1.3 Missense Mutations Changing the Secondary Structure or Protein Stability (Class C)

These secondary structural perturbations can be examined at atomic resolution in the context of the local secondary or tertiary configuration. As it has been shown previously for other pathogenic missense mutations, it has been estimated that the majority of amino acid substitutions lead to protein instability or reduced protein solubility [26, 27]. The methionine at position 173 has hydrophobic interactions with the adjacent valine residue at position 185 (Fig. 5.1f). Introduction of a highly positively charged arginine to this position would likely disrupt the local hydrophobic interaction with p.Val185 and destabilize the protein. Reduced stability can lead to abolished enzyme activity, which is consistent with the reported results of homozygous p.M173R [28].

Mutation p.E289A is located at a solvent exposure site, and the negatively charged side chain forms hydrogen bonding with nearby charged atoms from p.Lys124 and p. Arg279 and from p.Cys280 (Fig. 5.1g). The polar to nonpolar change of p.E289A is expected to disrupt the tightly coupled chemical interactions, consequently affecting the local structure. The enzyme activity of patients bearing homozygous p.E289A mutation or compound heterozygote with other deleterious mutations was almost totally diminished [23, 25]. Mutation p.L356P is located in the middle of the helix at the N-terminal that is packed closely with adjacent residues through hydrophobic interactions. Substitution of this residue with proline creates unfavorable phi-psi angle, which disrupts local helical structure. Three computational algorithms (conservation, MutPred, SNPs&GO) give relatively low pathogenic scores regarding this change, except Polyphen-2, which predicted it to be deleterious. The patient who was compound heterozygous for the p.L356P mutation and another deleterious mutation had severely reduced TP enzymatic activity [6].

Before the discovery of mutations responsible for late onset MNGIE patients, the genotype of *TYMP* mutations is found to be poorly correlated with the clinical phenotype of MNGIE patient [8–10]. The late onset MNGIE patients usually have a

diagnosis at around age 60, while the age of onset of typical MNGIE is at second decade, and patients usually die at the fourth decade. The patients with late onset disease usually do not have a full spectrum of clinical presentations and much less severity. The enzyme activity of these patients is in the range of 10–15 % of normal values.

The p.L285P change is located at a middle of a helix spanning from His283 to Gly296 around the active site, which provides hydrophobic interactions with adjacent residues, p.Val281, p.Ala421, and p.His441 at the substrate binding pocket. The substitution of the hydrophobic leucine residue with a helix destabilizing secondary amino acid proline residue very likely disrupts the helical structure. Both thymidine and phosphate substrates are likely to promote the stability of the mutant protein. The TP activity of the patient bearing the compound heterozygote p.L285P and p.G153S mutations was 9 % of control. Since the activity of the p.G153S mutant protein is <5 %, the residual activity of this patient is probably due to the presence of the mild mutation, p.L285P. Mutation p.E379K is located at the protein surface and exposed to solvent. It has ionic interaction with p.Arg442. The substitution of glutamic acid with a highly positively charged lysine is expected to result in unbalanced charge-charge interaction; such ionic interaction may be neutralized by solvent molecules on the protein surface. Thus, the impact of this change is likely to be mild. The patient bearing this mutation and a splice site mutation has only mild elevation of thymidine with 18 % residual TP activity in white blood cells.

5.1.4 Missense Change Likely to Be Benign Variants (Class D)

The normal biological function of protein can be abolished in multiple ways. During the evolution, the residues required for maintaining the structural integrity and proper function are likely to be conserved. However, benign polymorphic missense variants can also occur in the normal protein at different population frequencies. It also further complicates the interpretation by the fact that some rare variants in certain ethnic background are unlikely to cause disease. Variants p.A465T and p.S471L have been observed at high frequencies in the general population, with minor allele frequencies of 0.0543 and 0.1089, respectively. A homozygous p.A465T change has been previously reported in a patient with suspected clinical diagnosis of MNGIE [29]. Upon the examination of the TP complex structure, alanine at this position is exposed to solvent and does not have interacting residues at nearby region (Fig. 5.1h). Moreover, the p.A465T has been reported to have normal enzyme activity [30]. Thus, the p.A465T change may not be the genetic defect responsible for that patient's clinical symptom.

As for the p.S471L change, the high minor allele frequency indicates that it is a common variant occurring in multiple ethnic backgrounds. While a patient with p.S471L in homozygote state was reported to meet the clinical diagnosis [28], there is no biochemical evidence to support TP deficiency. The structural analysis of p.S471L also does not yield evidence for any obvious effect. Collectively, both p.A465T and p.S471L changes are likely to be benign polymorphisms.

6 Integrated Approaches to Understand Missense Variant/Mutation of MNGIE Disease

Molecular diagnosis of MNGIE disease is relatively straightforward if the clinical information and biochemical studies are indicative. However, it becomes challenging to interpret the pathogenicity of novel missense variants if only limited clinical and biochemical information is available. Computational and structural methods presented in this chapter can be used cautiously to help understand the likely pathogenic effects of these variants. The integrated approaches, which include clinical evaluation, biochemical genetic analyses, computational prediction, and structural analysis, can be helpful in understanding the molecular mechanism of disease pathogenesis. A single amino acid substitution in the protein may be subtle, but the structural change may be significant enough to lead to deleterious effects. The protein structural based analysis is a valuable tool to visualize such small structural changes and can be incorporated into routine variant interpretation pipeline.

References

1. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore SF (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3(65):65ra64. doi:10.1126/scitranslmed.3001756
2. Thusberg J, Vihinen M (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 30(5):703–714
3. Thusberg J, Olatubosun A, Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32(4):358–368
4. Matsukawa K, Moriyama A, Kawai Y, Asai K, Kato T (1996) Tissue distribution of human gliostatin/platelet-derived endothelial cell growth factor (PD-ECGF) and its drug-induced expression. Biochim Biophys Acta 1314(1–2):71–82
5. Shoffner JM (1993) Mitochondrial neurogastrointestinal encephalopathy disease. Gene Reviews, 1993
6. Garone C, Tadesse S, Hirano M (2011) Clinical and genetic spectrum of mitochondrial neurogastrointestinal encephalomyopathy. Brain 134(Pt 11):3326–3332
7. Massa R, Tessa A, Margollicci M, Micheli V, Romigi A, Tozzi G, Terracciano C, Piemonte F, Bernardi G, Santorelli FM (2009) Late-onset MNGIE without peripheral neuropathy due to incomplete loss of thymidine phosphorylase activity. Neuromuscul Disord 19(12):837–840
8. Marti R, Verschuuren JJ, Buchman A, Hirano I, Tadesse S, van Kuilenburg AB, van Gennip AH, Poorthuis BJ, Hirano M (2005) Late-onset MNGIE due to partial loss of thymidine phosphorylase activity. Ann Neurol 58(4):649–652
9. Hirano M, Nishigaki Y, Marti R (2004) Mitochondrial neurogastrointestinal encephalomyopathy (MNGIE): a disease of two genomes. Neurologist 10(1):8–17
10. Nishino I, Spinazzola A, Hirano M (2001) MNGIE: from nuclear DNA to mitochondrial DNA. Neuromuscul Disord 11(1):7–10
11. Said G, Lacroix C, Plante-Bordeneuve V, Messing B, Slama A, Crenn P, Nivelon-Chevallier A, Bedenne L, Soichot P, Manceau E, Rigaud D, Guiochon-Mantel A, Matuchansky C (2005) Clinicopathological aspects of the neuropathy of neurogastrointestinal encephalomyopathy (MNGIE) in four patients including two with a Charcot-Marie-Tooth presentation. J Neurol 252(6):655–662

12. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814
13. Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33(Web Server issue):W480–W482
14. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249
15. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25(21):2744–2750
16. Minor DL Jr (2007) The neurobiologist's guide to structural biology: a primer on why macromolecular structure matters and how to evaluate structural data. Neuron 54(4):511–533
17. Machius M (2003) Structural biology: a high-tech tool for biomedical research. Curr Opin Nephrol Hypertens 12(4):431–438
18. Mitsiki E, Papageorgiou AC, Iyer S, Thiyagarajan N, Prior SH, Sleep D, Finnis C, Acharya KR (2009) Structures of native human thymidine phosphorylase and in complex with 5-iodouracil. Biochem Biophys Res Commun 386(4):666–670
19. Pugmire MJ, Ealick SE (1998) The crystal structure of pyrimidine nucleoside phosphorylase in a closed conformation. Structure 6(11):1467–1479
20. Pugmire MJ, Cook WJ, Jasanoff A, Walter MR, Ealick SE (1998) Structural and theoretical studies suggest domain movement produces an active conformation of thymidine phosphorylase. J Mol Biol 281(2):285–299
21. El Omari K, Bronckaers A, Liekens S, Perez-Perez MJ, Balzarini J, Stammers DK (2006) Structural basis for non-competitive product inhibition in human thymidine phosphorylase: implications for drug design. Biochem J 399(2):199–204
22. Norman RA, Barry ST, Bate M, Breed J, Colls JG, Ernill RJ, Luke RW, Minshull CA, McAlister MS, McCall EJ, McMiken HH, Paterson DS, Timms D, Tucker JA, Pauptit RA (2004) Crystal structure of human thymidine phosphorylase in complex with a small molecule inhibitor. Structure 12(1):75–84
23. Nishino I, Spinazzola A, Hirano M (1999) Thymidine phosphorylase gene mutations in MNGIE, a human mitochondrial disorder. Science 283(5402):689–692
24. Gamez J, Ferreiro C, Accarino ML, Guarner L, Tadesse S, Marti RA, Andreu AL, Raguer N, Cervera C, Hirano M (2002) Phenotypic variability in a Spanish family with MNGIE. Neurology 59(3):455–457
25. Slama A, Lacroix C, Plante-Bordeneuve V, Lombes A, Conti M, Reimund JM, Auxenfants E, Crenn P, Laforet P, Joannard A, Seguy D, Pillant H, Joly P, Haut S, Messing B, Said G, Legrand A, Guiochon-Mantel A (2005) Thymidine phosphorylase gene mutations in patients with mitochondrial neurogastrointestinal encephalomyopathy syndrome. Mol Genet Metab 84(4):326–331
26. Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353(2):459–473
27. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17(4):263–270
28. Nishino I, Spinazzola A, Papadimitriou A, Hammans S, Steiner I, Hahn CD, Connolly AM, Verloes A, Guimaraes J, Maillard I, Hamano H, Donati MA, Semrad CE, Russell JA, Andreu AL, Hadjigeorgiou GM, Vu TH, Tadesse S, Nygaard TG, Nonaka I, Hirano I, Bonilla E, Rowland LP, DiMauro S, Hirano M (2000) Mitochondrial neurogastrointestinal encephalomyopathy: an autosomal recessive disorder due to thymidine phosphorylase mutations. Ann Neurol 47(6):792–800
29. Kocaefe YC, Erdem S, Ozguc M, Tan E (2003) Four novel thymidine phosphorylase gene mutations in mitochondrial neurogastrointestinal encephalomyopathy syndrome (MNGIE) patients. Eur J Hum Genet 11(1):102–104
30. Vissing J, Ravn K, Danielsen ER, Duno M, Wibrand F, Wevers RA, Schwartz M (2002) Multiple mtDNA deletions with features of MNGIE. Neurology 59(6):926–929

31. Martin MA, Blazquez A, Marti R, Bautista J, Lara MC, Cabello A, Campos Y, Belda O, Andreu AL, Arenas J (2004) Lack of gastrointestinal symptoms in a 60-year-old patient with MNGIE. Neurology 63(8):1536–1537
32. Schupbach WM, Vadday KM, Schaller A, Brekenfeld C, Kappeler L, Benoist JF, Xuan-Huong CN, Burgunder JM, Seibold F, Gallati S, Mattle HP (2007) Mitochondrial neurogastrointestinal encephalomyopathy in three siblings: clinical, genetic and neuroradiological features. J Neurol 254(2):146–153
33. Monroy N, Macias Kauffer LR, Mutchinick OM (2008) Mitochondrial neurogastrointestinal encephalomyopathy (MNGIE) in two Mexican brothers harboring a novel mutation in the ECGF1 gene. Eur J Med Genet 51(3):245–250
34. Poulton J, Hirano M, Spinazzola A, Arenas Hernandez M, Jardel C, Lombes A, Czermin B, Horvath R, Taanman JW, Rotig A, Zeviani M, Fratter C (2009) Collated mutations in mitochondrial DNA (mtDNA) depletion syndrome (excluding the mitochondrial gamma polymerase, POLG1). Biochim Biophys Acta 1792(12):1109–1112
35. Nalini A, Gayathri N (2011) Mitochondrial neurogastrointestinal encephalopathy in an Indian family with possible manifesting carriers of heterozygous TYMP mutation. J Neurol Sci 309(1–2):131–135

Chapter 6
Algorithms and Guidelines for Interpretation of DNA Variants

Jing Wang and Megan Landsverk

Abstract With the increasing amount of molecular genetic testing offered for clinical diagnosis in recent years, there is a rapid growth in the detection of novel or unclassified variants of unknown clinical significance. To determine whether a sequence change is a disease-causing pathogenic mutation or a non-causative variant becomes increasingly important in translational medicine. Interpretation of the clinical significance of an unclassified variant in the mitochondrial genome is even more complicated due to the highly polymorphic feature of the mitochondrial DNA and the unique characteristics of heteroplasmy. The degree of mutant mitochondrial DNA heteroplasmy varies among different tissues; in general, it correlates with the disease severity in affected tissues. In this chapter, we provide updated procedures of evaluating unclassified variants in both the nuclear and mitochondrial genomes by using various databases, computational tools, and structural analysis methods to assist in clinical interpretation.

1 Introduction

Molecular testing has been widely used not only for the diagnosis of genetic disorders but also for infectious disease and cancer prognoses. With the rapid growth of high throughput sequencing technology, in particular the "next-generation" massively parallel sequencing, simultaneously sequencing of multiple genes for certain groups of disorders, or sequencing of the whole exome is now widely offered as routine clinical testing. This has inevitably resulted in a tremendous increase in the detection of novel or unclassified variants of unknown clinical significance. There are several important steps in the clinical management of sequencing results. One is to

J. Wang (✉) • M. Landsverk
Department of Molecular and Human Genetics, Baylor College of Medicine,
One Baylor Plaza, NAB 2015, Houston, TX 77030, USA
e-mail: jwang7@bcm.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_6,

generate a list of accurate mutations/variants detected; the second is to interpret the detected sequencing changes for their potential functional consequence; and the third is to determine whether a sequence change is a disease-causing pathogenic mutation or just a non-disease-associated neutral polymorphism. Any novel or unclassified variants need to be evaluated correctly. The interpretation of genomic changes will have a direct clinical impact on the management of patient care. Thus, accurate analyses of these novel or unclassified genetic variants become increasingly important in translational medicine. Although research laboratories may have resources such as functional studies to further investigate the functional consequence of individual variants, it is not practical for clinical diagnostic laboratories to perform such functional studies for individual variants under a limited time frame and budget.

Since the mitochondrial genome is an integral part of an individual's inherited material, many diseases are due to the defects in the dual genome cross talk. Mitochondrial disorders, which can be caused by mutations in the nuclear or mitochondrial genome, have gained increasing attention in recent years. Interpretation of the sequencing results is challenging because of the dual genome consideration and the extreme clinical and genetic heterogeneity of the diseases. Unlike the disomic nature of most nuclear genes, mtDNA copy number varies from hundreds to several thousands among different tissues. Sequence variants in mtDNA can be present in each mtDNA molecule, known as homoplasmy, or may occur in a subpopulation of mtDNA molecules, known as heteroplasmy. In addition, mtDNA is highly polymorphic; mtDNA variants have been reported to occur in almost every nucleotide position of the 16,569 bp mitochondrial genome [1, 2]. Consequently, rare and novel variants in mtDNA must be assessed for their pathogenic potential. Interpretation of these variants thus becomes a great challenge and requires broad genetics knowledge, clinical experience, and molecular laboratory trainings. The complex molecular results need to be communicated to the referring individual properly to help physicians make accurate diagnoses for patient care. The molecular testing result is indispensable information in making clinical correlation for diagnosis, prognosis, treatment, prenatal assessment, and evaluation of family members.

It is essential that diagnostic laboratories have established standards for interpretation of the clinical significance of unclassified variants detected in routine genetic testing. In 2008, American College of Medical Genetics (ACMG) published a guideline for the interpretation of sequence variants using ACMG Standards and Guidelines [3]. In this chapter, we provide updated procedures of the classification and interpretation of a sequence variant using various available databases, computational tools, and structural analytical methods.

2 Classification of Sequencing Variants

In order to properly interpret sequencing results, the spectrum of sequence variants needs to be carefully classified. Based on ACMG's recommendations for standards for interpretation and reporting of sequence variations [3] and the practical

experience in our clinical laboratory, sequence variants can be classified into six categories:

1. *Reported pathogenic mutation*: Sequence variation has been reported in multiple unrelated patients/families with clinical correlation and/or supporting functional studies and is recognized cause of the disorder.
2. *Novel pathogenic mutation*: Sequence variation is not previously reported but is expected to cause disorder. Generally, truncation mutation (novel nonsense, frameshift), missense mutations affecting the translation initiation codon, or splicing mutations that alter the invariant AG/GT boundaries belong to this category.
3. *Novel variant may or may not be disease causative*: Sequence variation is *not* previously reported and has uncertain pathogenicity in genes known to cause inherited Mendelian diseases. The types of variation include the following:
   (i) Missense variants or small in-frame deletions/duplications for which the effect on protein structure/function cannot be inferred.
   (ii) Exonic or intronic variants that may potentially affect pre-mRNA splicing, but no direct evidence is available. For example, synonymous changes close to the first or last nucleotide of exon.
   (iii) Putative splice site variants outside the invariant boundaries.
   (iv) Variants in regulatory sequences.
4. *Novel variant likely to be benign*: Sequence variation is *not* previously reported and is probably a benign change. Synonymous changes and deep intronic changes beyond 20 bp of exon/intron boundaries are in this category.
5. *Known benign variant*: Sequence variation is *previously reported* and is a recognized neutral variant, for example, a variant reported in dbSNP with a frequency that highly suggests it is benign: the average heterozygosity >=2 % or MAF (minor allele frequency) >= 1 % in a total chromosome count of >= 100.
6. *Reported variant with unclear pathogenicity*: Variant reported as a "mutation" based on observation in a single patient/family or in a population association study, with no functional studies addressing pathogenicity, and/or reported in dbSNP as rare variant with average heterozygosity <2 % or MAF <1 % and total chromosome count <100. Please note that this category does not completely correlate to category 6 in ACMG guideline [3], which are variants found in association studies. In our classification criteria, the ACMG category 6 variants are classified into either category 5 or 6 based upon allele frequency of each individual variant.

3 Interpretation of Sequence Variations

Proper classification of sequence variations into different categories facilitates the interpretation of the clinical significance of a variant. A schematic algorithm for sequencing variants classification is showing in Fig. 6.1. Alleles in categories 1 and 2 are considered to be bona fide deleterious and are disease causative.

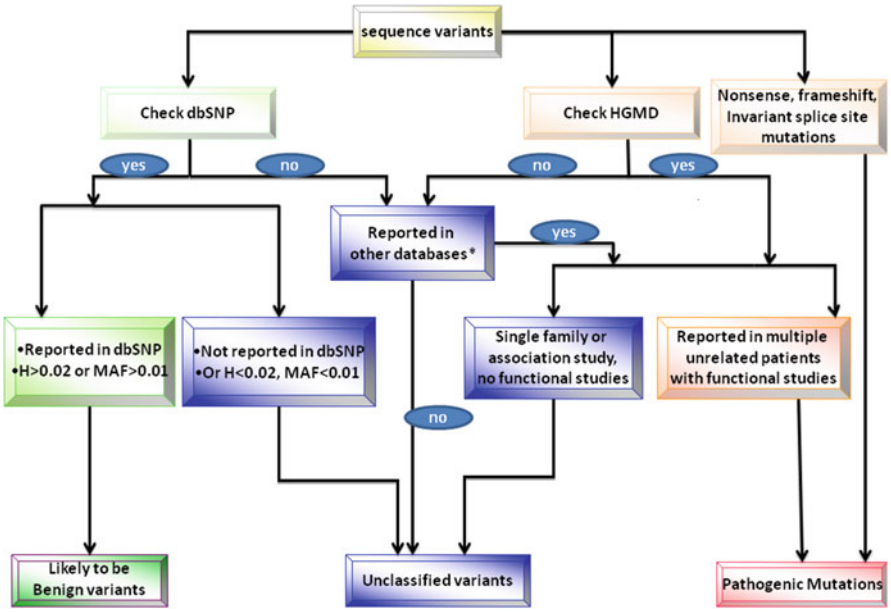

Fig. 6.1 Sequencing variants classification algorithm. *H* average heterozygosity, *MAF* minor allele frequency (*see Table 6.1 and Sect. 3.1 for list of databases)

The mutations in these two categories must be included in a patient's report with a discussion as to pathogenicity and proper reference to publications in which they are reported. Parental studies may be recommended in order to confirm the findings. For example, if the mutation is homozygous, parental testing is necessary in order to confirm that each is heterozygous for the finding, as opposed to being in one allele with failure to detect the other parental allele for technical or genetic reasons. If two heterozygous mutations are detected, parental studies facilitate the determination of phase (*cis* or *trans*) of the two mutations in the proband. Familial targeted analysis should also be provided for at risk family members.

Variations in category 5 are benign polymorphisms that are common in a healthy population. These benign variants may or may not be included in a patient's report but should be available upon request.

Sequence variations in categories 3, 4, and 6 are considered to be variants of unknown clinical significance. The assessment of the clinical relevance of unclassified variants is not straightforward and can be extremely challenging. Several resources such as variant databases (Table 6.1) and in silico prediction algorithms (Table 6.2) can be used to provide evidence for or against the pathogenicity of an unclassified variant to aid in genetic counseling. The report of unclassified variants should be issued to appropriately trained health-care professionals. It is essential to discuss the results with a clinical geneticist and to request additional specimens from the patient and family members for additional testing in order to facilitate the interpretation of the unclassified variants. However, please note that testing an unclassified variant for predictive context, for example, a prenatal diagnosis, is generally not recommended.

Table 6.1 Selected online databases and resources

Genome	Databases	Websites
nDNA variants	Human Genome Mutation Database (HGMD)	http://www.hgmd.cf.ac.uk
	Database of single nucleotide polymorphisms (dbSNP)	http://www.ncbi.nlm.nih.gov/projects/SNP
	Online Mendelian Inheritance in Man (OMIM)	http://omim.org/
	Locus-Specific Mutation Databases (LSDBs)	http://grenada.lumc.nl/LSDB_list/lsdbs
	Leiden Open Variation Database (LOVD)	http://www.lovd.nl/3.0/home
	MutDB	http://www.mutdb.org/
mtDNA variants	MITOMAP	http://www.mitomap.org/MITOMAP
	mtDB	http://www.mtdb.igp.uu.se/
	Mamit-tRNA database	(http://mamit-trna.u-strasbg.fr/human.asp)

Table 6.2 Selected in silico prediction tools for unclassified variants

Prediction category	Name	Website
Missense variants	PolyPhen	http://coot.embl.de/PolyPhen/
	SIFT	http://blocks.fhcrc.org/sift/SIFT.html
	Align GVGD	http://agvgd.iarc.fr/agvgd_input.php/
	Panther	http://www.pantherdb.org/tools/csnpScoreForm.jsp
	PhD-SNP	http://gpcr.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi
	PMut	http://mmb2.pcb.ub.es:8080/PMut/
Splicing prediction	NetGene2	http://www.cbs.dtu.dk/services/NetGene2
	BDGP	http://www.fruitfly.org/seq_tools/splice.html
	ESE Finder	http://rulai.cshl.edu/tools/ESE2

3.1 Interpretation of Deletions/Insertions

The size of deletion/insertions detected by sequencing analysis is relatively small (usually less than 100 bp) compared to those detected by array comparative genomic hybridization (aCGH), fluorescence in situ hybridization (FISH), and cytogenetic studies. If the observed deletion/insertion has not been previously reported but results in the alteration of the protein reading frame of a gene leading to a frameshift or nonsense stop codon, it is classified as a deleterious mutation according to ACMG guidelines [3]. If the deletion/insertion is novel, but it results in an in-frame insertion or deletion, depending on the size and location, it may or may not affect the function of the protein. A small in-frame deletion/insertion is classified as an unclassified variant. It is important to check if the deletion/insertion involves an important functional domain, which may have some impact on protein function.

3.2 Evolutionary Conservation

Analysis of the conservation of an amino acid at a specific position of the protein is the first step to infer structural/functional importance of an amino acid residue. The protein sequence containing the amino acid of interest should be used as a query for analysis using the NCBI protein BLAST website (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Orthologous proteins from various species throughout evolution need to be selected and retrieved. The different biological species should be selected from phylogenetic trees that represent the similarities and differences in their physical and genetic characteristics. As a minimal standard, the alignments should include the full length sequence of eight orthologous genes, in which at least five are from mammalian species. We recommend selecting data for the following species, if available: *Homo Sapiens*, *Bos Taurus* (cow), *Canis familiaris* (dog), *Mus musculus* (mouse)/*Rattus norvegicus* (rat), *Gallus gallus* (chicken), *Xenopus laevis* (frog), *Danio rerio* (fish), *Drosophila melanogaster* (fruit fly), *Strongylocentrotus droebachiensis* (sea urchin), *Caenorhabditis elegans* (worm), and *Saccharomyces cerevisiae* (yeast). ClustalW2, a multiple sequence alignment tool (www.ebi.ac.uk/Tools/msa/clustalw2/), is recommended for multiple sequence alignment.

3.3 Online Databases and Resources

There are many online databases that can be used to further determine if an allele has been previously reported with a disease association:

1. The Human Genome Mutation Database (HGMD, http://www.hgmd.cf.ac.uk) generally collects only the first report of a variant and includes some associated phenotype but not its recurrence in the population. It enables access to mutation queries and links to the published literatures as well as public resources like dbSNP and OMIM. Advanced search functionalities include the ability to find mutations based on the type of nucleotide or amino acid change or their location in a specific motif, splice site, or regulatory region. HGMD can assist in verifying if an observed mutation has been previously reported and in understanding the mutation spectrum of a given gene.
2. The Locus-Specific Mutation Databases (LSDBs) list sequence variants in a specific gene(s) causing a Mendelian disorder or a change in the phenotype, curated by an expert in that gene. The variants may just be observed in one or more tested individuals and may not be published in literature [4].
3. Leiden Open Variation Database (LOVD) is a web-based open source database developed in the Leiden University Medical Center to collect and display variants in the DNA sequence [5]. The focus of a LOVD is usually the combination between a gene and a genetic (heritable) disease. All sequence variants found in individuals are collected in the database, together with information about whether

they could be causally connected to the disease (i.e., a disease-causing variant or mutation) or not (i.e., a non-disease-causing variant).

4. Database of single nucleotide polymorphisms (dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP) is a free NCBI archive for genetic variation. It includes single-base nucleotide substitutions, small-scale insertions/deletions, retroposable element insertions, microsatellite repeat variations, and non-polymorphic variants.
5. PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) is a public recourse for biomedical literatures, life science journals, and online books. It also provides access to additional relevant websites and links to the other NCBI molecular biology resources. PubMed is developed and maintained by the National Center for Biotechnology Information (NCBI), at the US National Library of Medicine (NLM), located at the National Institutes of Health (NIH).
6. Online Mendelian Inheritance in Man (OMIM, McKusick-Nathans Institute of Genetic Medicine, http://omim.org/) is a comprehensive, authoritative compendium of human genes and genetic disorders. OMIM generally collects the first variants described and later some with unique characteristics. The OMIM database focuses on relationship between phenotype and genotype.
7. Google is a powerful internet search engine useful for finding a specific variant by standard and alternate nomenclatures [6]. One can search by gene name and all other possible ways to denote a variant to retrieve published literature. Google scholar is also very helpful in finding related literatures, but it is less comprehensive than PubMed.
8. Gene-specific databases: Some institutions maintain gene-specific databases that are open to the public, such as POLG (http://tools.niehs.nih.gov/polg/index.cfm?do=polg.home), PAH (http://www.pahdb.mcgill.ca/), OPA1 (http://lbbma.univ-angers.fr/lbbma.php?id=9), and ALDOB (http://www.bu.edu/aldolase/HFI/hfidb/hfidb.html). These disease-specific databases are good resources to check if a mutation/variant has been seen previously.

If the variant has been previously reported, the publication should be carefully reviewed to see if the published data provides sufficient evidence to address the pathogenicity of the findings. If the variant has never been reported in any of the above mentioned databases and does not belong to category 2, it is considered a variant of unknown significance and additional analyses are warranted.

3.4 Utilization of In Silico Prediction Algorithms for Pathogenicity Predictions of Unclassified Variants

A variety of in silico prediction tools are available to access the possible pathogenicity of a missense variant or splicing effect. Here, we list the prediction tools that are commonly used in clinical diagnostic laboratories (Table 6.2).

3.4.1 In Silico Prediction Tools for Missense Variants

1. The SIFT algorithm (Sorting Intolerant From Tolerant) (http://sift.jcvi.org) is mainly based on sequence homology. It can be used to predict the likely effect, tolerated or not tolerated, of a non-synonymous substitution on protein function [7].
2. The PolyPhen algorithm (polymorphism phenotyping) is a structure-sequence-based amino acid substitution prediction method. The current version is PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2). It utilizes the data available in UniProtKB/UniRef100 and is based on conservation, protein folding, and crystal structure [8]. This analysis classifies variants as likely benign, possibly damaging, or probably damaging.
3. Align GVGD is a web-based program that combines the biophysical characteristics of amino acids and protein multiple sequence alignments to predict where missense substitutions in genes of interest fall in a spectrum from enriched deleterious to enriched neutral [9, 10].

Caveats do exist for the use of predictive software; while limited studies evaluating the reliability of these programs have been reported [9, 11], no software has been clinically validated. Bioinformatics prediction tools may be valuable as screening tools for identifying alleles of high pathogenic potential in molecular and disease association studies. However, since the error rates are still high, current algorithms do not supplant the need for in vitro or in vivo functional studies [12].

3.4.2 In Silico Prediction Tools for Possible Splice Effect

The precise recognition of intron-exon junctions (splice sites) and the correct pairing of the 5′ splice site with its cognate 3′ splice site are critical for splice site selection. The splice donor site contains an almost invariant sequence GU at the 5′ end of the intron. The splice acceptor site terminates the intron with an almost invariant AG sequence at the 3′ end of intron. If a nucleotide substitution changes one of these invariant splice site sequences, it is considered to be a deleterious mutation according to ACMG guidelines [3]. Eukaryotic genomes contain large numbers of splice sites, known as cryptic splice sites (css), which are generally held to be disadvantageous sites that are dormant or used only at low levels unless activated by mutation of nearby authentic or advantageous splice sites. It appears that all types of genomic nucleotide variations can be deleterious by affecting normal pre-mRNA splicing via disruption/creation of splice site consensus sequences. It is important to use different splicing prediction algorithm to evaluate intronic or synonymous variants for possible effect on splicing.

1. NetGene2 (http://www.cbs.dtu.dk/services/NetGene2) is a neural network-based prediction of splice sites for human *C. elegans* and *A. thaliana* DNA [13]. Input the mutant sequence spanning the exon-intron boundary. The query sequence must be more than 200 and less than 80.000 nucleotides long.
2. The splice site predictor at Berkeley Drosophila Genome Project (BDGP) is based on neural network recognition of donor/acceptor splice site in a DNA

sequence [14]. This algorithm is applied to both human and *D. melanogaster* genes (http://www.fruitfly.org/seq_tools/splice.html). Input both wild-type and mutant sequences spanning the exon-intron boundary. Multiple sequences can be included to search at one time. The prediction also produces a confidence score based on the algorithm.

3. There are other splice algorithms available to predict the possible splicing effect: SpliceSiteFinder-like, MaxEntScan, NNSPLICE, GeneSplicer, and Known constitutive signals [15].
4. ESE Finder 2.0 is an algorithm for the prediction of how a sequence change in an exon interferes exonic splicing enhancer (ESE) [16, 17] (http://rulai.cshl.edu/tools/ESE2). Exonic enhancers are binding sites for specific serine-/arginine-rich (SR) proteins. SR proteins bound to ESEs can promote exon definition by directly recruiting the splicing machinery through their SR domain and/or by antagonizing the action of nearby silencer elements.

In silico predictions are useful for variants that decrease the strength of wild-type splice sites or create a cryptic splice site. Importantly, in silico predictions are not sufficient to classify variants as neutral or deleterious: they should be used as part of the decision-making process to detect potential candidates for splicing anomalies, prompting molecular geneticists to carry out transcript analyses in a limited and pertinent number of cases which could be managed in routine settings.

Table 6.2 listed commonly used in silico prediction tools for pathogenicity prediction of unclassified variants.

3.5 Additional Studies in Analysis of Unclassified Variants

3.5.1 Parental Testing

Testing the parents of an affect proband is an easy test to perform and can aid in the interpretation of molecular findings. When an apparently homozygous mutation or variant is detected in an autosomal recessive gene in the proband, parental testing is important in order to confirm each is heterozygous for the same finding. If the heterozygous variant was only detected in one, but not in other parent, further investigations should be followed up to determine if there was an allele drop out in the proband due to a rare SNP on primer sites, an intragenic deletion in the other allele causing apparent loss of heterozygosity (LOH) or uniparental disomy in the detected region. When two heterozygous mutations or variants are detected in an autosomal recessive gene in the proband, testing parents can help determine the phase or configuration of the two heterozygous changes. If a novel or rare missense unclassified variant is detected to be in a *cis* configuration with a known deleterious mutation, in general, this variant is unlikely to cause to clinical symptoms in the proband. On the other hand, if a novel or rare variant is in a *trans* configuration with a deleterious mutation and is located in highly conserved functional domain, then these evidences imply the pathogenic potential of this variant.

3.5.2 Co-segregation Analyses

The analysis of co-segregation of a variant with the disease in families is a powerful tool for the classification of unclassified variants [18]. It is relatively easy to perform and requires only gender, genotype, and age of onset as well as clinical and family history. Although in most cases, co-segregation analysis itself is insufficient to prove the pathogenicity of a given unclassified variant, absence of co-segregation provides strong evidence against pathogenicity of this variant.

3.5.3 Control Studies

Testing ethnicity matched healthy controls can determine whether an unclassified variant is segregating in the normal population and help to interpret the possible pathogenicity of an unclassified variant. Several factors need to be considered when using a control study: (1) The number of properly matched healthy controls that need to be screened. In general, in order to have 95 % chance to observe a variant with 1 % of allele frequency, at least 298 chromosomes have to be screened. (2) Some populations may have a high carrier frequency for certain pathogenic variants for recessive disorders due to founder effects. (3) In later onset disorders, some individuals in a control group may become affected later in life.

3.5.4 RNA Studies

RNA analysis is the essential test to confirm a splicing mutation. However, due to the limited availability of RNA specimens, it is not feasible to perform RNA studies for most unclassified variants. We recommend considering RNA studies for variants that are predicted to affect splicing by at least two independent in silico splicing prediction algorithms or patient samples in which only a heterozygous mutation was detected in an autosomal recessive disease gene with a highly suggestive clinical presentation. However, care must be taken in using an appropriate cell or tissue type since gene expression in the form of transcription may be tissue specific.

3.5.5 Protein Functional Studies

A reliable functional protein assay is the best way to evaluate the pathogenicity of a variant. Commonly used functional studies include measuring enzyme activity, in vitro or in vivo analysis of protein expression levels, immunochemistry for cellular protein localization, and tissue-specific expression. Due to the complexity of most of these assays, it is usually not feasible to provide protein functional studies as a routine assay in clinical diagnostic laboratories.

4 Other Considerations for the Interpretation of mtDNA Variants

Unlike the disomic nature of most nuclear genes, mtDNA copy number varies from hundreds to several thousands per cell among different tissues. Sequence variants in mtDNA can be present in homoplasmic or heteroplasmic states. The degree of mtDNA mutation heteroplasmy and its tissue distribution can affect an individual's clinical presentation. Therefore, the interpretation of the possible pathogenicity of rare and novel mtDNA variants can be further complicated by the level of heteroplasmy, tissue threshold of a particular mutation, and variable penetrance. While not always reliable, publically available mtDNA databases and algorithms that examine protein structure/function/evolution can be utilized to gauge the pathogenic potential of novel and/or rare mtDNA variants.

4.1 Publicly Available Databases for Mitochondrial Sequence Analyses

While not always reliable, publically available mtDNA databases and algorithms that examine protein structure/function/evolution can be utilized to gauge the pathogenic potential of novel and/or rare mtDNA variants.

MITOMAP (http://www.mitomap.org/MITOMAP) and the Human Mitochondrial Genome Database (mtDB) (http://www.mtdb.igp.uu.se/) provide useful data for the interpretation of mtDNA variants. MITOMAP compiles human mtDNA variations from both published and unpublished sources. All mtDNA variants are classified into two categories in MITOMAP: polymorphisms and mutations. The polymorphism category includes benign polymorphisms, somatic alterations, and collections of unpublished variants (e.g., mtDNA variants from PhyloTree). The mutation category contains confirmed mtDNA mutations and variants reported to have disease association. Relevant publications are listed for identified variants.

The Human Mitochondrial Genome Database (mtDB) is another resource with extensive documentation of human mitochondrial variants. It contains mtDNA variants from over 2,700 individuals who were healthy at the time of ascertainment and their frequency in the subject cohort, which is very helpful in interpreting of the nature of an mtDNA variant. However, it should be recognized that some variants may appear to be rare due to ethnic underrepresentation in the database. The allele frequencies are obtained from the mtDB database and our private database (comprised of mitochondrial whole genome sequences for over 3,000 unrelated individuals and partial mtDNA sequences of 420 matrilineal relatives as of June 2012). We use an allele frequency of $\leq$0.2 % as the definition of rare mtDNA variants. Considerable caution is needed in using data derived from diagnostic laboratories in that almost all samples are obtained from unhealthy individuals. Similarly, there are errors/inconsistencies in the medical literature concerning the classification of rare population-specific variants and pathologic alterations.

If the mtDNA variant occurs in the mitochondrial tRNA genes, the Mamit-tRNA database (http://mamit-trna.u-strasbg.fr/human.asp) should be checked. The Mamit-tRNA database contains mammalian mitochondrial tRNAs with an emphasis on the structural characteristics of these tRNAs. It contains extensive documentation on point mutations in mitochondrial tRNA genes related to human mitochondrial disorders including 2D cloverleaf representation of tRNA and a list of mutations, the associated disease phenotypes, and corresponding references [19].

4.2 Classification of mtDNA Variants

In general, mtDNA variants can be classified into three categories [20]:

1. *Benign variant (category 5)*: If a variant has been reported in MITOMAP as a polymorphism, has no report of disease association in the population or family studies, and has been reported in mtDB at a frequency greater than 0.2 %, then this variant is considered to be a benign variant.
2. *Unclassified variant*: Any variants that meet at least one of the criteria below:
   (a) A novel variant *(category 3)*
   (b) A rare variant that has been reported in MITOMAP as a polymorphism, but not in mtDB, or reported in mtDB at a frequency ≤0.2 %
   (c) A rare variant reported in the literature or MITOMAP as a "mutation" based on a single-family study or a single report with no functional studies addressing pathogenicity *(category 7)*
3. *Mutation (categories 1 and 2)*: mtDNA variants that have been listed in MITOMAP as "confirmed mutations" and have been reported in multiple unrelated patients/families with clinical correlation and/or supporting functional studies. Nonsense and frameshift mutations in the protein-coding genes of mtDNA are classified as categories 1 and 2 deleterious mutations.

4.3 Follow-Up Studies for mtDNA Variants

4.3.1 Test Matrilineal Relatives

Due to the uncertain biological/clinical significance of unclassified variants, targeted sequence analysis of the patient's mother and other matrilineal relatives is typically recommended. When a variant is homoplasmic in asymptomatic matrilineal adult relatives and is not co-segregating with disease phenotype, then that variant, by itself, is unlikely to be the primary cause of the clinical symptoms. If a variant is absent or at a low level of heteroplasmy in asymptomatic matrilineal relatives or is co-segregating with a disease phenotype, then this variant may be pathogenic. Additional studies, including mitochondrial functional studies and Western blot analysis, may be needed to further clarify the clinical significance of this variant.

4.3.2 Heteroplasmy Quantification and Verification

While the interpretation of a heteroplasmic change is inherently problematic, if the variant is absent or at lower heteroplasmy in asymptomatic matrilineal relatives, then there is a higher index of suspicion of pathogenicity. The degree of heteroplasmy of the variant in various tissues that correlates with the clinical features may also facilitate interpretation. Quantification of the level of heteroplasmy of variant/mutation in tissue samples from invasive (e.g., muscle, skin) or noninvasive (e.g., hair bulbs, urine sediment, and buccal mucosa cells) sources should be considered.

4.4 Tissue Specificity and Nuclear Modifier Genes

Both mtDNA replication and maintenance are controlled by nuclear genes. In addition to primary mtDNA mutations, mitochondrial diseases can also be caused by defects of nuclear gene encoded respiratory chain complex proteins, by nuclear genes mutations resulting in alteration of mtDNA, and by synergistic effect of an mtDNA mutation with a nuclear modifier gene. The examples of synergistic action of an mtDNA mutation with a nuclear modifier gene are m.11778G>A, 14484 T>C, and 3460G>A in the ND genes associated with Leber's hereditary optic neuropathy (LHON) and m.1555A>G in the 12S rRNA associated with sensorineural hearing loss (SNHL) [21]. The characteristic features of presence nuclear modifier gene for an mtDNA mutation are maternal inheritance, homoplasmic level of mtDNA mutation, tissue-specific presentation, variable penetrance, and large variability in clinical phenotype observed in different pedigrees [22].

For example, an individual can only develop LHON when a primary mtDNA mutation is present. However, due to reduced penetrance, only approximately 50 % of males and over 15 % of females who harbor a primary LHON mutation develop blindness, which certainly indicates that nuclear genetic factors are important in the expression of the disease [23]. Unknown environmental and genetic factors remain to be discovered that interact with primary mtDNA mutation to develop the disease. Therefore, when interpreting mtDNA mutation, the potential influence of nuclear genes should be considered.

4.5 Large mtDNA Deletions

Large mtDNA deletions usually cause one of the three conditions: Kearns-Sayre syndrome (KSS), Pearson Syndrome, or progressive external ophthalmoplegia (PEO). While the large mtDNA is a bona fide mutation, further test different tissues from the proband and matrilineal relatives can help to distinguish whether a large mtDNA deletion is a germline or somatic mutation, is inherited or de novo. This is important for genetic counseling. Most large mtDNA deletions arise de

novo. If the proband's mother does not carry the large mtDNA deletion, the risk to the proband's siblings is usually extremely low. While exceptions do occur, the offspring of a female patient is usually not at risk of inheriting the large mtDNA deletion [24].

4.6 Distinguishing Primary and Secondary mtDNA Mutations

Mitochondrial disorders have been called defects of intergenomic signaling or nuclear-mitochondrial communication disorders. In addition to sporadic or maternally inherited disorders due to mutations of mtDNA, nuclear-mitochondrial intergenomic signaling disorders are transmitted as Mendelian traits, either autosomal dominant or recessive, and are associated with the accumulation of molecular abnormalities of mtDNA. The consequences of these disorders are characterized by qualitative or quantitative alterations of mtDNA.

Most single mtDNA deletions occur in sporadic cases since they are de novo generated in the oocyte or during embryonic development. Maternal transmission of single large mtDNA deletion appears to be extremely rare. On the other hand, multiple deletions of mtDNA have been reported in familial forms of progressive external ophthalmoplegia (PEO) with both autosomal dominant and recessive inheritance. The accumulation of mtDNA multiple deletions is generally restricted to skeletal muscle and is usually secondary to defects in genes that are critical for the maintenance of mitochondrial integrity. Mutations in the *ANT1*, *Twinkle*, *POLG*, *POLG2*, *RRM2B*, *OPA1*, and *MFN2* genes have been associated with PEO syndrome and causing mtDNA multiple deletions.

The current massively parallel sequencing technology allows us to detect not only mtDNA point mutations and small indels but also single large deletion and multiple deletions [25]. When mtDNA multiple deletions are identified, further testing of nuclear genes that are responsible for the maintenance of mitochondrial integrity should be considered.

5 Summary

The next-generation sequencing technology has become increasingly important for molecular diagnosis, either to identify pathogenic mutations in known disease-causing genes or discover new disease-causing genes. Interpretation and reporting of sequencing results for clinical diagnosis are limited to qualified professionals based on ACMG recommendation [3]. In this chapter, we propose a classification scheme and interpretation guideline for variants detected in both nuclear and mitochondrial genome.

References

1. Ingman M, Gyllensten U (2006) mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34:D749–D751
2. Ruiz-Pesini E, Lott MT, Procaccio V, Poole JC, Brandon MC, Mishmar D, Yi C, Kreuziger J, Baldi P, Wallace DC (2007) An enhanced MITOMAP with a global mtDNA mutational phylogeny. Nucleic Acids Res 35:D823–D828
3. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE (2008) ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. Genet Med 10:294–300
4. Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehvaslaiho H, Liang P, Marsh S, Nebert DW, Povey S, Rossetti S, Scriver CR, Summar M, Tolan DR, Verma IC, Vihinen M, den Dunnen JT (2008) Recommendations for locus-specific databases and their curation. Hum Mutat 29:2–5
5. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT (2011) LOVD v. 2.0: the next generation in gene variant databases. Hum Mutat 32:557–563
6. Bandelt HJ, Salas A, Taylor RW, Yao YG (2009) Exaggerated status of "novel" and "pathogenic" mtDNA sequence variants due to inadequate database searches. Hum Mutat 30:191–196
7. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249
9. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res 34:1317–1325
10. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43:295–305
11. Tchernitchko D, Goossens M, Wajcman H (2004) In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clin Chem 50:1974–1978
12. Hon LS, Zhang Y, Kaminker JS, Zhang Z (2009) Computational prediction of the functional effects of amino acid substitutions in signal peptides using a model-based approach. Hum Mutat 30:99–106
13. Brunak S, Engelbrecht J, Knudsen S (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. J Mol Biol 220:49–65
14. Reese MG, Eeckman FH, Kulp D, Haussler D (1997) Improved splice site detection in Genie. J Comput Biol 4:311–323
15. Houdayer C (2011) In silico prediction of splice-affecting nucleotide variants. Methods Mol Biol 760:269–281
16. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR (2003) ESEfinder: a web resource to identify exonic splicing enhancers. Nucleic Acids Res 31:3568–3571
17. Cartegni L, Chew SL, Krainer AR (2002) Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3:285–298
18. Mohammadi L, Vreeswijk MP, Oldenburg R, van den Ouweland A, Oosterwijk JC, van der Hout AH, Hoogerbrugge N, Ligtenberg M, Ausems MG, van der Luijt RB, Dommering CJ, Gille JJ, Verhoef S, Hogervorst FB, van Os TA, Gomez Garcia E, Blok MJ, Wijnen JT, Helmer Q, Devilee P, van Asperen CJ, van Houwelingen HC (2009) A simple method for co-segregation analysis to evaluate the pathogenicity of unclassified variants; BRCA1 and BRCA2 as an example. BMC Cancer 9:211

19. Helm M, Brule H, Friede D, Giege R, Putz D, Florentz C (2000) Search for characteristic structural features of mammalian mitochondrial tRNAs. RNA 6:1356–1379
20. Wang J, Schmitt ES, Landsverk ML, Zhang VW, Li FY, Graham BH, Craigen WJ, Wong LJ (2012) An integrated approach for classifying mitochondrial DNA variants: one clinical diagnostic laboratory's experience. Genet Med 14:620–626
21. Guan MX, Fischel-Ghodsian N, Attardi G (1996) Biochemical evidence for nuclear gene involvement in phenotype of non-syndromic deafness associated with mitochondrial 12S rRNA mutation. Hum Mol Genet 5:963–971
22. Davidson MM, Walker WF, Hernandez-Rosa E, Nesti C (2009) Evidence for nuclear modifier gene in mitochondrial cardiomyopathy. J Mol Cell Cardiol 46:936–942
23. Man PY, Griffiths PG, Brown DT, Howell N, Turnbull DM, Chinnery PF (2003) The epidemiology of Leber hereditary optic neuropathy in the North East of England. Am J Hum Genet 72:333–339
24. DiMauro S, Hirano M (2011) Mitochondrial DNA deletion syndromes, GeneReviews™ [Internet]
25. Zhang W, Cui H, Wong LJ (2012) Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem 58:1322–1331

Part III
Application to Clinical Diagnosis

Chapter 7
NGS-Based Clinical Diagnosis of Genetically Heterogeneous Disorders

C.A. Valencia, T.A. Sivakumaran, B.T. Tinkle, A. Husami, and K. Zhang

Abstract Next-generation sequencing (NGS) has transformed genomic research by decreasing the cost of sequencing and increasing the throughput. NGS platforms have evolved to provide an accurate and comprehensive means for the detection of molecular mutations, and the recent focus is on using NGS technology in clinical diagnosis. NGS analysis has three major components: enrichment, sequencing, and analysis. In the last several years, enrichment technologies based on hybridization or amplification principles have emerged. Similarly, sequencing platforms have continued to improve by increasing the sequencing output and decreasing the sequencing time and cost. Various enrichment and sequencing platform combinations have been utilized for the diagnosis of genetically heterogeneous disorders, and it is the topic of discussion for this chapter. We describe the employment of NGS approaches to the diagnosis of genetically heterogeneous disorders and mention advantages and challenges of these technologies in a clinical laboratory setting.

Abbreviations

AF	Atrial fibrillation
APEX	Arrayed primer extension
ARPKD	Autosomal recessive polycystic kidney disease

C.A. Valencia, Ph.D. (✉) • T.A. Sivakumaran, Ph.D. • B.T. Tinkle, M.D., Ph.D.
A. Husami, B.Sc. • K. Zhang, M.D., M.B.A.
Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

Department of Pediatrics, University of Cincinnati Medical School, Cincinnati, OH 45229, USA
e-mail: Alexander.Valencia@cchmc.org; Siva.Theru_Arumugam@cchmc.org; Bradley.Tinkle@cchmc.org; Ammar.Husami@cchmc.org; Kejian.Zhang@cchmc.org

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_7,

ARVC	Arrhythmogenic right ventricular cardiomyopathy
BBS	Bardet–Biedl syndrome
CHD	Congenital heart disease
CHF	Congenital hepatic fibrosis
CMD	Congenital muscular dystrophy
CNS	Central nervous system
CSD	Conduction system disease
DCM	Dilated cardiomyopathy
ERG	Electroretinogram
HCM	Hypertrophic cardiomyopathy
HLH	Hemophagocytic lymphohistiocytosis
JBTS	Joubert syndrome-related disorders
LCA	Leber congenital amaurosis
LGMD1B	Limb-girdle muscular dystrophy type 1B
LVNC	Left ventricular noncompaction
MDC1A	Merosin-deficient congenital muscular dystrophy
MDC1C	Congenital muscular dystrophy type 1C
MEB	Muscle-eye-brain disease
MKS	Meckel–Gruber syndrome
NGS	Next-generation sequencing
NPHP	Nephronophthisis
NPHP-AC	Nephronophthisis-associated ciliopathies
PCD	Primary ciliary dyskinesia
PIDD	Primary immunodeficiency disorders
RP	Retinitis pigmentosa
RSS	Rigid spine syndrome
SCD	Sudden cardiac death
SVT	Supraventricular tachycardia
VT	Ventricular tachycardia
WGA	Whole genome amplification
WWS	Walker–Warburg syndrome

1 Introduction

The completion of the Human Genome Project in 2003 was regarded as the dawn of an era of genomic medicine, in which information from genomes would guide clinical decision making and contribute to the ultimate delivery of personalized medicine. It was expected that accelerated detection of disease-related mutations would improve genetic diagnosis and prognosis [1, 2]. The delivery of personalized genomic medicine requires not only access to the complete human genome but also availability of appropriate genetic tests for individual patients. Molecular genetic testing is important for clinical care, enabling assignment of risk, genetic

counseling, and prognosis, and will be essential for enrolling patients in future gene therapy trials [3–5]. However, the genetic heterogeneity of many Mendelian disorders is a major obstacle to obtaining molecular diagnoses in clinical practice [2]. Apart from the genetic heterogeneity, there are other obstacles that currently limit molecular diagnosis such as the lack of clearly defined genotype–phenotype correlations [6]. This makes it difficult to direct testing to specific candidate genes. The development of massively parallel or "next-generation" sequencing (NGS) techniques, which generate millions of DNA sequence reads in parallel during a single experimental run, offers a potential solution [2]. This is accomplished in two steps: an enrichment step and massive parallel sequencing using one of the commercially available NGS platforms. The aim of this chapter is to provide a broad, not comprehensive, review of the introduction of NGS, including enrichment technologies and sequencing platforms, into clinical practice by specifically focusing on how these technologies have helped with the genetic heterogeneity, cost, and poorly defined genotype–phenotype correlation challenges that physicians face.

2 Genetically Heterogeneous Disorders

2.1 Retinitis Pigmentosa

Retinitis pigmentosa (RP) is a genetically heterogeneous disorder, and the most common inherited retinal degeneration disorder, with prevalence of 1 in 4,000 [6]. RP is classified as nonsyndromic, or "simple" (not affecting other organs or tissues), syndromic (affecting other systems such as hearing), or systemic (affecting multiple tissues). Nonsyndromic RP can be inherited in an autosomal dominant, autosomal recessive, or X-linked manner [7]. In addition, rare digenic forms with heterozygous mutations in both *ROM1* and *RDS* also occur. Fifty-two genes are known to be associated with non-syndromic RP, involving all modes of inheritance, demonstrating its genetic heterogeneous nature [8]. A further 59 genes are known to underlie other subtypes of syndromic and non-syndromic retinal diseases (RetNet: http://www.sph.uth.tmc.edu/Retnet/). The genetic heterogeneity along with the lack of well-defined genotype–phenotype correlations makes molecular testing of a specific candidate gene challenging. For example, with a few exceptions, there are no ophthalmologic characteristics specifically associated with the genetic subtypes of RP, impeding the prioritization of genes for analysis by Sanger sequencing. Due to the unknown mode of inheritance of a large number of cases, mutations in any of the 52 known RP genes may be causative. Traditional genetic screening for RP is laborious, although technological advances have had some impact [9, 10]. The most widely applied diagnostic test for heterogeneous diseases, arrayed primer extension (APEX) chip technology, is only able to detect known mutations [11]. In addition,

these chips are designed to separately test for the presence of known mutations in autosomal dominant or recessive RP genes, resulting in a diagnostic yield for autosomal recessive RP of only ~10 % [11]. Altogether, the yield of diagnostic testing has remained disappointingly low for RP patients, despite the many important disease gene discoveries in the last two decades [8]. However, to date, applications of the NGS sequencing approach as a molecular diagnostic tool have been limited because of the costs and perceived technical and data-handling challenges.

2.2 Leber Congenital Amaurosis

Leber congenital amaurosis (LCA), a genetically heterogeneous disorder, is an early and severe autosomal recessive retinal dystrophy, causing progressive profound visual deficiency or blindness from birth. LCA has a worldwide incidence of about 1 in 30,000 and is the most frequent cause of childhood blindness [12]. LCA becomes evident in the first year of life. Visual deficiency is poor with nystagmus, sluggish or near-absent pupillary responses, photophobia, high hyperopia, and keratoconus [13]. Visual acuity is rarely better than 20/400. A characteristic finding is Franceschetti's oculo-digital sign, comprising eye poking, pressing, and rubbing. While the retina may initially appear normal, a pigmentary retinopathy reminiscent of RP is frequently observed later in childhood. The electroretinogram (ERG) is characteristically "nondetectable" or severely subnormal [13]. Genetic heterogeneity is a key obstacle in diagnosing and identifying LCA patients eligible for gene-specific treatment. Its mode of inheritance is mostly autosomal recessive; however, several autosomal dominant cases have been reported [14]. Approximately 70 % of LCA cases can be explained by mutations found in 16 disease genes, leaving the remaining 30 % cases unexplained [15]. Cases with an acuity of 20/50 or better have been reported, usually with *CRB1*, *LRAT*, or *RPE65* mutations [16]. Patients who exhibit mild improvements in visual function temporarily and then decline may have mutations in *CRB1*, *LRAT*, and *RPE65* genes. Generally, patients with a progressive course have mutations in *AILP1* and *RPGRIP1*, whereas those with severe but stable vision loss have mutations in *CEP290* and *GUCY2D* genes [16]. Moreover, some of these genes are also involved in syndromic diseases. Significantly, early molecular diagnosis offers the prospect of specific and adequate medical follow-up. Technologically, the use of an LCA microarray that evaluates 641 known mutations in 13 genes has been the most commonly used primary genetic test [17]. Unfortunately, it fails to detect new mutations, is costly for routine testing, and has a variable, population-dependent detection rate. Current genetic tests only offer a subset of causative genes [18]. There is a pressing necessity for a comprehensive approach that can to identify all mutations in all currently known LCA genes. NGS technologies offer an immediate solution for the molecular diagnosis of LCA as a genetically heterogeneous disorder.

2.3 Ciliopathies

Cilia are evolutionarily conserved hair-like structures with key roles in cell locomotion, fluid movement, and sexual reproduction [19]. Abnormal ciliary axonemal structure and function have been linked to the growing class of genetic disorders collectively known as ciliopathies [19]. The prototypical ciliopathy, primary ciliary dyskinesia (PCD), was the first human disorder linked to ciliary dysfunction [20–22]. The importance of cilia in other human diseases is just beginning to be elucidated, and defects have been associated with a growing number of pediatric conditions, including obesity, renal disease, hepatic fibrosis, skeletal dysplasias, endocrinopathies, neurodevelopmental defects, central nervous system (CNS) anomalies, laterality defects, and congenital heart disease (CHD) [21–24].

2.3.1 Motor Ciliopathies

PCD, a genetically heterogeneous disorder, is an autosomal recessive condition; however, rare cases of autosomal dominant and X-linked inheritance have been reported [25]. The incidence of PCD ranges from 1 in 15,000 to 30,000 live births [19]. Most patients with PCD have persistent hypoxemia or even acute respiratory failure during the immediate newborn period as the upper respiratory is almost always universally involved. Inadequate mucus clearance from the respiratory tract commonly manifests as chronic sinusitis, and some patients develop nasal polyposis. Middle ear disease is described in most with varying degrees of conductive hearing loss. Impaired mucociliary clearance of the lower respiratory tract leads to recurrent episodes of pneumonia or bronchitis. Chronic lung infection and inflammation result in persistent atelectasis and bronchiectasis in many patients, even young children, typically involving the right (or left) middle lobe and is most often progressive [26, 27]. Left-right laterality is also a cilia-dependent mechanism, and half those with PCD have situs inversus totalis (SI) with complete reversal of the thoracic and abdominal organs. Male patients with PCD are typically infertile as a result of impaired spermatozoa motility caused by defective sperm flagella, although male infertility is not a universal finding in this disease [20]. Male patients with PCD can have some spermatozoa motility, suggesting sperm tails retain some function or could actually be under different genetic control than cilia. Fertility issues in women have been reported, likely due to ciliary dysfunction in fallopian tubes [20]. Investigations of the genetic basis of PCD have focused on dynein arm proteins and have revealed the heterogeneous nature of the disease. To date, mutations have been identified in 11 genes. *DNAI1* was the first gene to be linked to PCD on the basis of the candidate approach. Mutations in *DNAI1* have been found in patients with outer dynein arm defects and functional ciliary abnormalities and have been estimated to occur in 10 % of patients with PCD [28]. Another gene, *DNAH5*, has also been identified as a causative gene with homozygosity mapping. A recent study has indicated that 53 % of patients with PCD with known outer dynein arm defects had

mutations in *DNAH5* clustered in five exons, making this a promising target for genetic screening [29]. Most recently, another dynein gene, *DNAH11*, has been linked to PCD with normal ultrastructure. The other PCD genes, including *DNAI2*, *KTU*, *TXNDC3*, *LRRC50*, *RSPH9*, *RSPH4A*, *CCDC40*, and *CCDC39*, have only been reported in a small number of patients, and their relative prevalence has not been defined [19]. A pressing necessity for comprehensive genetic testing that includes all causative genes for motor ciliopathies exists.

2.3.2 Sensory Ciliopathies

In some tissues, primary cilia serve as chemoreceptors and mechanoreceptors of the extracellular environment [22, 30, 31]. Autosomal dominant polycystic kidney disease, a common cause of chronic renal failure in adults, was one of the first diseases linked to sensory cilia dysfunction, caused by mutations in *PKD1* and *PKD2* genes, which encode polycystin 1 or polycystin 2, respectively [31–33]. In contrast, autosomal recessive polycystic kidney disease (ARPKD) is the most common childhood-onset ciliopathy, characterized by dilated renal collecting ducts resulting in progressive cystic degeneration of the kidneys and congenital hepatic fibrosis (CHF). ARPKD is caused by mutations in *PKHD1* that encodes polyductin, a protein involved in differentiation of cells lining the collecting ducts [34]. Another ciliopathy, nephronophthisis, is an autosomal recessive cystic renal disease of childhood caused by mutations in nine different genes encoding nephrocystins (*NPHP1-8* and *ALMS1*) that localize to cilia, basal bodies, centrosomes, adherens junctions, and focal adhesions [35]. Collectively, the various forms of nephronophthisis are the most common cause of end-stage renal failure in children. Genetically heterogeneous primary ciliary defects have been shown to cause several, overlapping syndromes, as a result of ciliary/centrosomal defects in various cell types such as retinal photoreceptors or renal tubular epithelial cells, and have been implicated in the pathogenesis of nephronophthisis-associated ciliopathies (NPHP-AC), including nephronophthisis (NPHP), Senior–Loken syndrome (SLSN), Joubert syndrome (JBTS), Meckel–Gruber syndrome (MKS), and Bardet–Biedl syndrome (BBS) [31, 36]. For instance, Bardet–Biedl syndrome (BBS) is a rare, genetically heterogeneous, autosomal recessive disorder with varied phenotypes, including RP, polycystic kidneys, truncal obesity, polydactyly, hypogonadism, intellectual disabilities, diabetes mellitus, and CHD. Approximately one-third of patients with BBS will have anosmia, caused by defective nonmotile sensory "9 + 2" cilia, present on olfactory neurons. There are at least 12 different BBS genes, found only in ciliated cells and localized to the basal body and ciliary axoneme, and expressed proteins are involved in microtubule anchoring and coordination of the cell cycle [37]. The most severe manifestation of the NPHP-AC clinical spectrum is seen in fetuses with MKS, a perinatally lethal ciliopathy, characterized by central nervous system malformations (typically occipital encephalocele), bilateral postaxial hexadactyly, hepatobiliary ductal plate malformation, and multicystic dysplastic kidneys [36]. It is caused by abnormalities of various proteins located in the cilia-basal body-centrosome complex, including *MKS1* and *MKS3*. In patients with JBTS, midbrain/hindbrain

malformations and cerebellar vermis hypoplasia/aplasia result in numerous neurological features including developmental delay, intellectual disability, muscle hypotonia, ataxia, oculomotor apraxia, nystagmus, and irregular breathing patterns in neonates [38]. Other less common features include retinal dystrophy, fibrocystic renal disease, CHF, occipital encephalocele, and polydactyly. Some genotype–phenotype correlations have emerged, and multiple causative genes are responsible for a syndrome, or mutations in a gene are responsible for several syndromes, illustrating locus and allelic genetic heterogeneity and shared pathologies of sensory ciliopathies. Mutations in 18 different recessive genes have been identified as the molecular cause in NPHP-AC [36]. Twelve genes have been implicated in NPHP and/or SLSN (*NPHP1*, *INVS*, *NPHP3*, *NPHP4*, *IQCB1*, *CEP290*, *GLIS2*, *RPGRIP1L*, *NEK8*, *TMEM67*, *TTC21B*, and *XPNPEP3*) [39–50]. Ten are known to cause JBTS (*AHI1*, *TMEM216*, *INPP5E*, *NPHP1*, *CEP290*, *RPGRIP1L*, *TMEM67*, *ARL13B*, *CC2D2A*, and *TTC21B*). Mutations in five genes (MKS1, *TMEM67*, *CEP290*, *RPGRIP1L*, *CC2D2A*, *TMEM216*) have been shown to cause MKS [51–55]. Moreover, multiple allelism within the NPHP-AC phenotypic spectrum has been recurrently reported for many of these genes, especially *CEP290*, *RPGRIP1L*, *TMEM67*, *CC2D2A*, *TTC21B*, and *TMEM216*. For example, hypomorphic missense mutations in the gene *TMEM67* (*MKS3*/*NPHP11*) are implicated in NPHP with liver fibrosis and JBTS type 6, whereas truncating mutations in *TMEM67*/*MKS3* have been reported in MKS cases with severe developmental and dysplastic phenotypes [49, 53]. The presence of multiple allelism and broad heterogeneity together with extensive phenotypic clinical overlap in patients with NPHP-AC requires extensive mutational analysis efforts in order to identify the underlying molecular etiology. The challenge of analyzing increasing numbers of candidate genes associated with disease in large cohorts of patients can now be met by applying NGS technologies.

2.4 Congenital Muscular Dystrophies

The congenital muscular dystrophies (CMDs) comprise a genetically and phenotypically heterogeneous group of disorders with preferentially autosomal recessive inheritance. CMDs manifest clinically at birth or early infancy and are characterized by congenital hypotonia, delayed motor development, progressive muscle weakness, respiratory insufficiency, bulbar dysfunction, arthrogryposis, and often involvement of other organ systems such as the brain and eyes [56]. Hypertrophy of the tongue and limb muscles, scoliosis, and contractures may develop with age [57]. Weakness is static or slowly progressive. Muscle biopsy shows typical dystrophic changes (degeneration and regeneration of muscle fibers and proliferation of fatty and connective tissue). Electromyography (EMG) is myopathic. Cerebral magnetic resonance imaging may show abnormalities of neuronal migration and white matter signal. Over the past decade, molecular understanding of the CMDs has greatly expanded [58–61]. Mutations in 12 different genes, comprising a total of 293 exons, have been shown to cause different forms of CMD (Table 7.1) [62–69]. Approximately one-third of all CMDs are caused by mutations in the *LAMA2* gene, which encodes

the α2 chain of laminin (Table 7.1). Mutations in *COL6A1*, *COL6A2,* and *COL6A3*, which encode the three chains of collagen type VI, give rise to Ullrich congenital muscular dystrophy and Bethlem myopathy. Mutations in the selenoprotein N gene (*SEPN1*) give rise to rigid spine syndrome, multiminicore disease, and a desmin-related myopathy with Mallory body-like inclusions. The other genes that are associated with CMD all code for molecules that affect cell surface receptors for the extracellular matrix molecule laminin. These genes include *ITGA7*—the gene that encodes integrin α7 (the predominant integrin α chain in skeletal muscle)—and six genes (fukutin [*FKTN*], fukutin-related protein [*FKRP*], protein O-linked mannose β1,2-N-acetylglucosaminyltransferase[*POMGnT1*],protein-O-mannosyltransferases 1 and 2 [*POMT1/2*], and like-glycosyltransferase [*LARGE*]), the products of which affect the glycosylation of α-dystroglycan [62–69]. The α-dystroglycan is an important membrane protein that, similar to integrins, binds to the extracellular matrix. Diseases resulting from mutations that affect α-dystroglycan glycosylation have been grouped together under the heading dystroglycanopathies. Mutations in these genes give rise to characteristic alterations in the glycosylation of the α-dystroglycan protein, which can alter its function, although the precise molecular mechanisms are poorly understood in some cases. The dystroglycanopathy genes now account for half of the genes that have been implicated in CMD (Table 7.1)[56]. Mutations in these genes have allowed for the better definition of molecular subgroups and their associated clinical phenotypes (Table 7.1). However, it is frequently not possible to maintain a one-to-one relationship between a given gene and a defined phenotype. The most striking example of this broadening genotype–phenotype relationship is the clinical spectrum associated with mutations in the FKRP (Fukutin-related protein) gene, ranging from Walker–Warburg syndrome to late adult-onset limb-girdle muscular dystrophy (LGMD). Two major themes have emerged concerning the molecular and clinical aspects of CMD. On the molecular side, it is striking that the majority of the genetic defects discovered either affect the posttranslational processing of α-dystroglycan or more directly involve molecules of the extracellular matrix itself, notably laminin-2 (the heavy chain of laminin-2/merosin), and the three alpha chains making up collagen type VI. On the clinical side, important themes include the potential additional involvement of the eyes and brain in the disorders of α-dystroglycan glycosylation and the combined involvement of muscle, tendon, and skin in the disorders of collagen VI. Thus, CMD shows considerable clinical as well as molecular heterogeneity, yet it seems that the majority of defined conditions involve a disturbed connection of muscle to its extracellular matrix [62].

2.5 Cardiac Diseases

Like many single-organ disorders, genetic conditions constitute an increasingly recognized group of cardiovascular disorders. Often, in the past, such conditions were listed as “idiopathic,” but as advances in our understanding of the human genome, genetic causation has been discovered, but often, multiple genes are

Table 7.1 Genetic heterogeneity of congenital muscular dystrophies shown by the multiple genes causing a similar phenotype

Gene	Reference	Location	Enzyme/protein	Disorder	Number of reported mutations	Transcript size (bp)	Number of coding exons	Number of amino acids
LAMA2	NM_000426.3	6q22	Laminin α2 chain of merosin	Merosin-deficient congenital muscular dystrophy (CMD1A)	127	9,708	65	3,122
FKRP	NM_024301.3	19q13	Fukutin-related protein	Fukutin-related proteinopathy (MDC1C), Muscle-eye-brain disease (MEB), Walker–Warburg syndrome (WWS), LGMD2I, FCMD	79	1,488	1	495
LARGE	NM_004737.3	22q12	LARGE-like glycosyltransferase	LARGE-related congenital muscular dystrophy (MDC1D)	9	2,268	14	756
FKTN	NM_001079802.1	9q31	Fukutin	Fukuyama congenital muscular dystrophy (FCMD), Walker–Warburg syndrome (WWS)	39	1,383	9	461
POMT1	NM_007171.3	9q34	Protein-O-mannosyltransferase 1	Walker–Warburg syndrome (WWS), LGMD2K	55	2,241	19	747
POMT2	NM_013382.4	14q24	Protein-O-mannosyltransferase 2	Muscle-eye-brain disease (MEB), Walker–Warburg syndrome (WWS)	35	2,250	21	750
POMGNT1	NM_017739.2	1p34	O-linked mannose β 1,2-N-acetyl glucosaminyltransferase	Muscle-eye-brain disease (MEB)	50	1,980	21	660

(continued)

Table 7.1 (continued)

Gene	Reference	Location	Enzyme/protein	Disorder	Number of reported mutations	Transcript size (bp)	Number of coding exons	Number of amino acids
SEPN1	NM_020451.2	1p36	Selenoprotein N	Rigid spine muscular dystrophy	43	1,770	13	590
COL6A1	NM_001848.2	21q22	α1 type VI collagen	Ulrich congenital muscular dystrophy and Bethlem myopathy	38	3,084	35	1,028
COL6A2	NM_001849.3	21q22	A 2 type VI collagen	Ulrich congenital muscular dystrophy and Bethlem myopathy	66	3,057	27	1,019
COL6A3	NM_004369.2	2q37	A 3 type VI collagen	Ulrich congenital muscular dystrophy and Bethlem myopathy	31	9,531	43	3,177
ITGA7	NM_002206.1	12q13	Integrin α7 precursor	Merosin-positive congenital muscular dystrophy	4	3,411	25	1,137

implicated in any single disorder and sometimes in multiple disorders. With the advent of technologies that can query multiple genes or larger segments of the genome, genetic testing of these disorders becomes feasible. Familial cardiovascular conditions often necessitate intensive monitoring of the proband and first-degree relatives. Recognition of those at risk can be beneficial in monitoring, activity-modifications, and interventions. A negative test would prevent the periodic cardiac evaluations, the anxiety of such testing and the overarching diagnosis, as well as unnecessary activity-restrictions. For example, relatives of patients with long QT syndrome (LQT) with normal cardiac studies remained at an increased risk of sudden cardiac death [70]. Indeed, several professional groups have recommended genetic testing in conditions such as dilated cardiomyopathy (DCM), hypertrophic cardiomyopathy (HCM), arrhythmogenic right ventricular cardiomyopathy (ARVC), LQT, and Brugada syndrome [71, 72]. NGS offers more comprehensive genetic analysis for genetically heterogeneous conditions such as HCM, DCM, and LQT at reduced costs. In addition, NGS can offer benefits in testing large single genes such as *TTN* (363 exons encoding 34,000 amino acids) which is a known genetic etiology for cardiomyopathy [73]. DCM is the most common form of cardiomyopathy accounting for nearly one-third of all cases with up to 30 % of all DCM cases attributable to Mendelian genetic causes [73]. DCM is characterized by ventricular dilatation and dysfunction. In contrast, HCM represents the most common cause of genetic cardiovascular disease. HCM is the thickening of the left ventricle that can lead to outflow obstruction and heart failure as well as rhythmic disturbances and stroke. ARVC is an autosomal (both dominant and recessive) disorder of the cardiac desmosome and involves the fibrofatty replacement of the myocardium initially in the right ventricle predisposing those affected to tachyarrhythmia and sudden death. LQT is the prolongation of the QT interval as seen on an electrocardiogram (ECG) and can result in arrhythmia leading to syncopal spells and even sudden death. As a matter of fact, during molecular autopsy, 35 % of sudden unexplained death and 10 % of sudden infant deaths may stem from genetic mutations associated with long QT or catecholaminergic polymorphic ventricular tachycardia [74]. Brugada syndrome is a channelopathy that can lead to rhythm disturbance and therefore susceptibility to syncope, tachyarrhythmia, and sudden cardiac death. Detection rates of the various disorders will vary depending on the disorder as well as the gene(s) included in each panel. DCM has been attributed to at least 28 genes whose products include cytoskeletal, contractile, and other proteins. Currently, there are at least 18 genetic causes of HCM accounting for roughly 50 % of familial HCM with 80 % isolated to *MYH7* and *MYBPC3* [75]. ARVC has at least 12 loci identified with nine available for genetic testing. Detection rates for a subset of these genes may be as high as 70 %. More than 13 genes are known to cause LQT with an overall detection rate greater than 75 %. At least eight genes can lead to Brugada syndrome, and all are clinically available with an overall detection rate of 20–38 % [76].

2.6 Hearing Loss

Hearing loss is the most common sensory disorder in human. Hearing loss results from obstructions in the transmission of the sound anywhere between the outer ear and auditory cortex in the brain. In a normal condition, the sound energy that is collected by the outer ear is amplified by the middle ear for transmission to the cochlea, which then converts this energy into electrical signals that is ultimately transmitted to the brain through the auditory nerves. Based on the defective anatomical structure involved, hearing loss can be classified as conductive, sensorineural, or mixed. Conductive hearing loss is a defect in conducting sound waves through outer and middle ear due to abnormalities of outer and/or the ossicles of the middle ear. Sensorineural hearing loss is due to a defect located anywhere from cochlea to the auditory cortex. Mixed hearing loss is a combination of both conductive and sensorineural hearing loss. Depending on the age at onset, hearing loss can be classified as prelingual, present before speech development, or postlingual, present after speech development. Severity of the hearing loss can be mild to profound, affecting anywhere from low to high frequencies (Box 7.1). One in 500 newborns is affected with bilateral permanent sensorineural hearing loss 40 dB or greater, and this number is increased to 3.5 per 1,000 during adolescence [77]. Approximately two-thirds of hearing loss is due to genetic factors, and in the remaining one-third of cases, it is caused by environmental factors [78, 79]. The environmental factors that cause hearing loss include both prenatal and postnatal infections, use of ototoxic drugs, and exposure to excessive noise. The majority of the inherited form of hearing loss is monogenic, and it can be syndromic or nonsyndromic. In the syndromic forms, hearing loss is accompanied by other physical manifestations, and it accounts for about 30 % of the inherited hearing loss [80]. Over 400 syndromes have been reported with hearing loss, and some of the common forms of syndromic hearing loss including Usher, Pendred, Jervell and Lange-Nielsen, Waardenburg, branchio-oto-renal, and Stickler syndromes, among the many [78, 81]. The nonsyndromic forms of hearing loss, with no other physical findings, accounts for about 70 % inherited hearing loss. They are categorized into four different groups according to their mode of inheritance: (1) autosomal recessive, (2) autosomal dominant, (3) X-linked, and (4) maternal inheritance due to mutations in mitochondrial genes. The autosomal recessive hearing loss is the most common type occurring in about 80 % of patients, followed by autosomal dominant in about 20 %. The X-linked and mitochondrial hearing loss are less common and accounting for only about 1 % of the patients [82–84]. Nonsyndromic hearing loss is extremely heterogeneous, and so far, over 150 loci responsible for such phenotype have been mapped (http://hereditaryhearingloss.org). These loci are designated as DFN followed by mode of transmission; DFNA refers to loci for autosomal dominant forms, DFNB refers to loci for autosomal recessive, and DFN to X-linked forms. The numbers following the designation are chronological order of locus identification (DFNB1 refers to first autosomal recessive locus). To date, 39 autosomal recessive, 25 autosomal dominant, 3 X-linked, and 2 mitochondrial genes have been identified. Many of

Box 7.1

Hearing is measured in decibels (dB). Severity of hearing loss is classified as:

- Mild (26–40 dB)
- Moderate (41–55 dB)
- Moderately severe (56–70 dB)
- Severe (71–90 dB)
- Profound (90 dB)

Based on frequency, hearing loss is classified as:

- Low (<500 Hz)
- Middle (501–2,000 Hz)
- High (>2,000 Hz)

these genes cause more than one form of hearing loss (Table 7.2); *SLC26A4*, *CDH23*, *MYO7A*, *DFNB31*, *USH1C*, and others cause both syndromic and nonsyndromic forms. *TMC1*, *GJB2*, *GJB6*, *MYO7A*, and others cause both autosomal dominant and autosomal recessive forms of hearing loss. Mutations in the *GJB2*, encoding connexin 26, that causes DFNB1 is the most common cause of hearing loss and account for about 50 % of the cases with autosomal recessive hearing loss in many populations [80, 81]. The remaining cases are attributable to the mutations in other genes, and among others *SLC26A4*, *MYO7A*, *OTOF*, *CDH23*, and *TMC1* are more prevalent [78]. Mutations in the rest of the genes are very rare (Table 7.2); many of them have been found to cause hearing loss in one or two consanguineous families [80, 85]. Similarly, none of the genes causing autosomal dominant hearing loss is a common cause of hearing loss, except *WFS1*, *KCNQ4*, *GJB2*, and *COCH* [78]. Elucidation of genetic basis of hearing loss is crucial for the clinical management of patients and their family. In addition, determination of genetic etiology in a large cohort of patients will provide better understanding of genotype–phenotype correlations, which could help developing specific therapeutic interventions. For syndromic hearing loss, candidate genes for molecular diagnosis are selected based on associated symptoms; whereas this approach is not viable for non-syndromic hearing loss as the phenotype caused by most of the genes is indistinguishable. Therefore, sequential screening of all hearing loss genes is critical to identify the genetic cause. Currently, genetic testing for hearing loss is conducted using different diagnostic algorithms in several institutes worldwide (Fig. 7.1). Mutation screening of coding and flanking intronic regions of the candidate genes using an automated Sanger sequencing is the most common approach in vast majority of these laboratories. However, the extreme genetic heterogeneity of non-syndromic hearing loss makes this strategy unfavorable in terms of cost and time. NGS technology offers the advantage of sequencing analysis of multiple genes in parallel [86]. Currently, only a few laboratories in the United States use this technology for mutation screening of hearing loss genes.

Table 7.2 Large number of genes responsible for hereditary hearing loss

Gene	Deafness locus	Reference	Location	Number of reported mutations	Transcript size (bp)	Number of coding exons	Number of amino acids
ACTG1	DFNA20/26	NM_001614.3	17q25	10	2,004	6	375
CCDC50	DFNA44	NM_178335.2	3q28	1	8,949	12	482
CDH23	DFNB12, USH1D	NM_022124.5	10q22.1	150	11,134	69	3,354
CLDN14	DFNB29	NM_144492.2	21q22.3	6	1,958	3	239
COCH	DFNA9	NM_004086.2	14q11.2-q13	13	2,558	12	550
COL11A2	DFNB53, DFNA13	NM_080680.2	6p21.3	28	6,425	66	1,736
DFNA5	DFNA5	NM_004403.2	7p15	5	2,521	10	496
DFNB31	DFNB31, USH2D	NM_015404.3	9q32	12	4,079	12	907
DIAPH1	DFNA1	NM_005219.4	5q31	1	5,804	28	1,272
ESPN	DFNB36	NM_031475.2	1p26.31	3	3,531	13	854
ESRRB	DFNB35	NM_004452.3	14q24.3	6	3,029	10	508
EYA4	DFNA10	NM_004100.4	6q23	8	5,697	20	639
GIPC3	DFNB15/72/95	NM_133261.2	19p13.3	10	4,317	6	312
GJB2	DFNB1A, DFNA3A	NM_004004.5	13q11-q12	>295	2,347	2	226
GJB3	DFNB91, DFNA2B	NM_024009.2	1p34	15	2,220	2	270
GJB6	DFNB1B, DFNA3B	NM_006783.4	13q12	14	2,110	3	261
GPSM2	DFNB82	NM_013296.4	1p13.3	2	3,039	15	684
GRHL2	DFNA28	NM_024915.3	8q22.3	1	5,231	16	625
GRXCR1	DFNB25	NM_001080476.2	4p13	4	1,003	4	290
HGF	DFNB39	NM_000601.4	7q21.1	3	2,820	18	728
ILDR1	DFNB42	NM_001199799.1	3q13.33	11	2,908	8	546
KCNQ4	DFNA2A	NM_004700.3	1p34	17	4,116	14	695
LHFPL5	DFNB66/67	NM_182548.3	6p21.31	6	2,147	4	219
LOXHD1	DFNB77	NM_144612.6	18q21.1	2	6,854	40	2,211
LRTOMT	DFNB63	NM_001145308.2	11q13.4	5	3,630	7	291
MARVELD2	DFNB49	NM_001038603.2	5q13.2	5	2,297	7	558

MIR96	DFNA50	NR_029512.1	7q32.2	3	78		
MSRB3	DFNB74	NM_198080.3	12q14.3	1	4,307	6	192
MYH14	DFNA4	NM_024729.3	19q13.33	7	6,807	39	1,995
MYH9	DFNA17	NM_002473.4	22q13.1	1	7,505	41	1,960
MYO15A	DFNB3	NM_016239.3	17p11.2	44	11,876	66	3,530
MYO1A	DFNA48	NM_005379.2	12q13-q14	7	3,621	28	1,043
MYO3A	DFNB30	NM_017433.4	10p11.1	3	5,798	35	1,616
MYO6	DFNA22, DFNB37	NM_004999.3	6q13	8	8,662	35	1,285
MYO7A	DFNB2, DFNA11, USH1B	NM_000260.3	11q13.5	253	7,465	49	2,215
OTOA	DFNB22	NM_144672.3	16p12.2	3	3,624	28	1,139
OTOF	DFNB9, AUNB1	NM_194248.2	2p23.1	80	7,171	47	1,997
PCDH15	DFNB23, USH1F	NM_033056.3	10q21.1	46	7,021	33	1,955
PJVK	DFNB59	NM_001042702.3	2q31.2	10	1,534	7	352
POU3F4	DFNX2 (DFN3)	NM_000307.3	Xq21.1	44	1,507	1	361
POU4F3	DFNA15	NM_002700.2	5q31	4	1,182	2	338
PRPS1	DFNX1 (DFN2)	NM_002764.3	Xq21.32-q24	4	2,156	7	318
PTPRQ	DFNB84	NM_001145026.1	12q21.2	3	8,066	45	2,299
RDX	DFNB24	NM_002906.3	11q23	4	4,498	14	583
SLC17A8	DFNA25	NM_139319.2	12q23.3	1	3,983	12	589
SLC26A5	DFNB61	NM_198999.2	7q22.1	2	2,697	20	744
DIABLO	DFNA64	NM_019887.4	12q24.31	1	2,265	7	239
SMPX	DFNX4 (DFN6)	NM_014332.2	Xp22.1	4	951	5	88
STRC	DFNB16	NM_153700.2	15q15.3	12	5,515	29	1,775
TECTA	DFNB21, DFNA8/12	NM_005422.2	11q22-q24	47	6,468	23	2,155
TJP2	DFNA51	NM_004817.3	9q13-q21	1	4,725	23	1,190
TMC1	DFNB7/11, DFNA36	NM_138691.2	9q32	38	3,201	24	760

(continued)

Table 7.2 (continued)

Gene	Deafness locus	Reference	Location	Number of reported mutations	Transcript size (bp)	Number of coding exons	Number of amino acids
TMIE	DFNB6	NM_147196.2	3p21	9	1,861	4	156
TMPRSS3	DFNB8/10	NM_024022.2	21q22.3	22	2,463	13	454
TPRN	DFNB79	NM_001128228.2	9q34.3	5	2,641	4	711
TRIOBP	DFNB28	NM_001039141.2	22q13.1	9	10,159	24	2,365
USH1C	DFNB18, USH1C	NM_153676.3	11p14.3	25	3,246	27	899
WFS1	DFNA6/14/38	NM_006005.3	4p16.1	238	3,640	8	890
SLC26A4	DFNB4, PDS	NM_000441.1	7q31	316	4,930	21	780

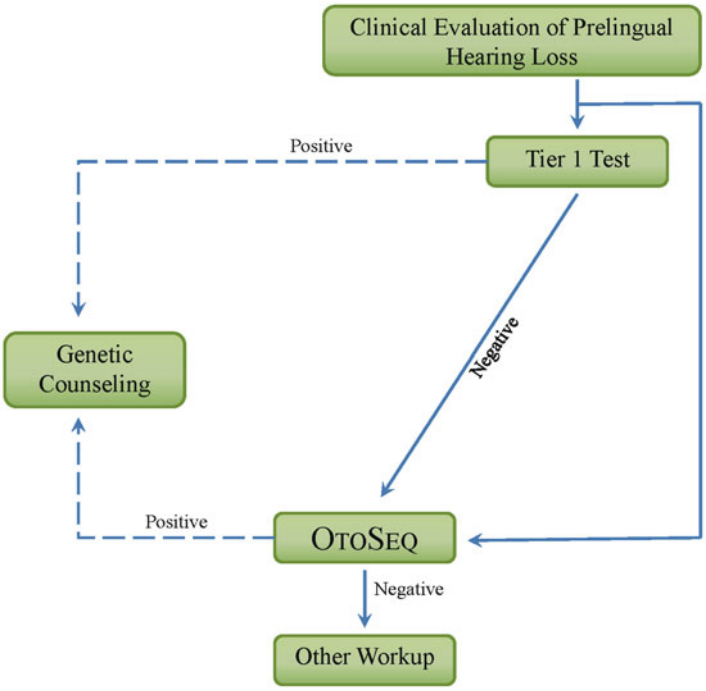

Fig. 7.1 Hearing loss diagnostic algorithm

2.7 Primary Immunodeficiencies

The primary immunodeficiency diseases (PIDD) are a large group of genetically heterogeneous disorders and estimated to affect 1 in 10,000 births [87, 88]. The hallmark of PIDD is the increased susceptibility to infections, which typically presents clinically with recurrent, severe, or unusual infections but are also associated with malignancies and autoimmune disorders [89]. Advances in basic research in immunology and DNA sequencing of the entire human genome have led to the discovery of precise molecular basis for more than 150 disorders of host immune defense in eight major categories [90–92]. These categories include combined T- and B cell immunodeficiencies; predominantly antibody deficiencies; other well-defined immunodeficiency syndromes; diseases of immune dysregulation; congenital defects of phagocyte number, function, or both; defects in innate immunity; autoinflammatory disorders; and complement deficiencies [90, 91]. For example, hemophagocytic lymphohistiocytosis (HLH), belonging to the category of diseases of immune dysregulation, is a rare immunodeficiency characterized by having prolonged fever, hepatosplenomegaly, hyperferritinemia, cytopenia, and hemophagocytosis [93, 94]. HLH has two forms, primary (inherited with genetic defects in *PRF1*, *MUNC13-4*, *STX11*, or *STXBP2*) and secondary (acquired or reactive), which are often difficult to distinguish from one another [94, 95]. Molecular genetic

testing is the only definite diagnostic tool to differentiate secondary HLH from the familial form, and it is critical in patient's management. Primary HLH is often lethal without an appropriate and timely chemotherapy followed by bone marrow or stem cell transplantation [96]. Therefore, it is necessary to envision an immunodeficiency NGS panel to aid in the diagnosis of these patients. A timely identification of the genetic defects is a critical factor for PIDD diagnosis, treatment, and prognosis. However, reaching a rapid diagnosis can be a challenge due to the genetic and phenotypic heterogeneity seen in PIDD patients. Specifically, locus heterogeneity is common in PIDD, and the example of severe combined immunodeficiency (SCID), with more than 20 causative genes, namely, *ADA*, *CD3D*, *CD3E*, *CD45*, *CORO1A*, *DCLRE1C*, *FOXN1*, *IL2RG*, *IL7R*, *JAK3*, *LIG4*, *NHEJ1*, *ORAI1*, *PNP*, *RAG1*, *RAG2*, *RMRP*, *STAT5B*, *STIM1*, and *ZAP70*, demonstrates the problem [97]. The issue is aggravated by the presence of allelic heterogeneity in some PIDDs, defined as the phenomenon in which different mutations at the same locus causes a similar phenotype. To date, only a few research and clinical laboratories offer individual genetic tests for PIDDs. Due to the more than 150 causative genes, it is technically challenging and financially prohibiting to use Sanger sequencing [88, 93]. Now, NGS offers the option to reach a diagnosis on a timely and cost-effective manner. Due to these advantages, several large-scale NGS genetic testing panels for PIDDs have been developed by several academic and commercial institutions (K. Zhang, personal communication).

3 NGS Technical Approaches: Enrichment, Sequencing, and Major Findings

Genetically heterogeneous disorders have been analyzed by several enrichment and sequencing platforms (Table 7.3). In this section, we discuss the enrichment methods in more detail in conjunction with the sequencing platforms used to analyze these large panels of genes belonging to aforementioned (Sect. 2) disorders.

3.1 Enrichment Method Comparisons

Solid capture (microarray capture), solution capture (SureSelect), and droplet-based PCR (RainDance) are the most common enrichment technologies that have been employed for the diagnosis of genetically heterogeneous disorders, and their advantages and limitations must be briefly considered here, namely, capture or amplification of homologous regions, enrichment of various interval sizes, allele dropout, and target sequences complexity (Table 7.3) [103]. PCR can be optimized to amplify target regions, whereas hybridization approaches may carry homologous regions together with real one, thus reducing specificity and lowering target percentages [100].

Table 7.3 Recent genetic heterogeneous disorders analyzed by next-generation sequencing. The size of the sequencing panels is shown by the number of genes, exons, and target interval that were analyzed. In addition, the enrichment and sequencing platform for each disorder is summarized below

Disorder	# of genes	# of exons	Target interval (bp)	Enrichment platform	NGS platform	Comments	Reference
Retinitis pigmentosa	45	681	359,000	Microarray capture	Illumina Genome Analyzer II		[6]
	46	504	249,267	WGA*+PCR	454 GS-FLX and Illumina Genome Analyzer IIx		[98]
	593	N/A	5,000,000	Microarray capture	454 FLX, 454 Titanium (Roche) and/or Illumina/ Solexa		[99]
Leber congenital amaurosis	16	252	152,000	Quantitative PCR (qPCR) amplicon ligation	Genome Analyzer IIx		[12]
Sensory ciliopathies	18	376	52,770	WGA+PCR	Illumina Genome Analyzer II	Panel includes genes for nephronophthisis-associated ciliopathy	[36]
Congenital muscular dystrophies	12	321	65,000	RainDance, solution capture	SOLiD 3		[100]
Cardiomyopathies	47	1,092	273,000	Microarray capture	SOLiD 3	Panel for AF, ARVC, CSD, LVNC, SCD, SVT, VT, DCM, HCM	[101]
	16	502	35,399	PCR	454 GS-FLX and Illumina Genome Analyzer	Panel for HCM	[102]

(continued)

Table 7.3 (continued)

Disorder	# of genes	# of exons	Target interval (bp)	Enrichment platform	NGS platform	Comments	Reference
Hearing loss	54	1,124	421,741	Microarray capture solution capture	454 GS-FLX and Illumina Genome Analyzer II		[86]
	24	731	117,041	RainDance	HiSeq2000		T. A. Sivakumaran, PC
Immunodeficiency disorders	395	3,439	559,937	Microarray capture	454 GS-FLX	Panel for PIDD	[97]
	124	1,569	301,417	RainDance	HiSeq2000	Panel for PIDD	K. Zhang, PC

WGA whole genome amplification by DNA polymerase strand displacement amplification, *HCM* hypertrophic cardiomyopathy, *DCM* dilated cardiomyopathy, *AF* indicates atrial fibrillation, *ARVC* arrhythmogenic right ventricular cardiomyopathy, *CSD* conduction system disease, *LVNC* left ventricular noncompaction, *SCD* sudden cardiac death, *SVT* supraventricular tachycardia, *VT* ventricular tachycardia, *N/A* not available, *PC* personal communication, *PIDD* primary immunodeficiency disorders

In the case of gene and pseudogene targets, RainDance and Sanger can readily address this issue by correctly choosing locations where the primers are to hybridize on the genomic DNA template. In contrast, the limitation of hybridization-based methods is not being able to distinguish between the gene and pseudogene targets. In this case, if an NGS panel for a genetically heterogeneous disorder has a pseudogene, it may be more appropriate to employ RainDance as an enrichment method, if the target region size is less than 1 Mb. In terms of interval size, solid and solution-hybridization captures can enrich the whole exome, whereas RainDance, by being a PCR-based method, is limited to the amplification of up to 1 Mb [100]. Hybridization-based capture methods are more flexible at addressing the genetic heterogeneity issue by virtue of accommodating larger panels. Allelic dropout due to SNPs in the PCR primer binding sites is a limitation inherent to all PCR-based assays, including Sanger sequencing [100]. RainDance uses a library of primers to amplify the target regions and is therefore also susceptible to allele dropout if specific SNPs are in the primer binding sites. To minimize the absence of amplification of specific exons due to allele dropout, primers are designed in regions where SNPs have not been reported by using the Single Nucleotide Polymorphism and the 1000 Genomes databases. However, even by applying these bioinformatic tools, allele dropout may occur. By contrast, hybridization-based technology that relies on 120-bp overlapping probes to capture the region of interest is less susceptible to allele dropout. The target sequences complexity has a strong effect on both the efficiency of DNA amplification and capture for individual exons. In many instances, first exons and GC-rich regions are problematic in both hybridization and droplet PCR-enriched samples [100]. For example, high GC content may explain the low coverage in the first exons of genes in RainDance samples, given that the mean GC content of the first coding exon of all CMD genes is 64 %. In contrast to RainDance, hybridization-based capture is also sensitive to sample base composition, and sequences at the extremes of high GC/AT content can be lost through poor annealing and secondary structure, respectively [104]. Depending on the NGS panel for a given heterogeneous disorder, full exon capture or amplification will depend on the GC percentage.

3.2 Microarray Capture or Solid-Phase Capture

Microarray captures have been a popular enrichment method for heterogeneous disorders that include RP, cardiomyopathies, hearing loss, and PIDDs (Table 7.3) [6, 86, 97, 99, 101]. A molecular diagnostic screen for patients with RP was recently developed [6]. A custom NimbleGen sequence microarray capture, containing 385,000 unique probes, was designed to target the coding regions and 100 bp flanking regions of all known RP genes and used to enrich a total of 359 kb of genomic sequence comprising 681 exons from 45 genes from DNA samples of five patients. Amplified enriched DNA was subjected to massively parallel sequencing on a Genome Analyzer II (Table 7.3). In this study, known homozygous *PDE6B* and

compound heterozygous *CRB1* mutations were detected in two patients in addition to a novel homozygous missense mutation (c.2957A>T; p.N986I) in the *CNGB1* gene. This homozygous mutation was predicted to have a deleterious effect and was absent in 720 normal control chromosomes. In contrast, a second microarray capture study, together with 454 FLX, 454 Titanium, and/or Illumina/Solexa NGS sequencing, was conducted using a comprehensive approach by including 593 genes covering a 5–10 Mb region with the intent of sequencing known RP genes and discovering novel ones from 21 affected families [99]. The detection of known and novel RP mutations in these studies establishes high-throughput DNA sequencing with DNA pooling as an effective diagnostic tool for heterogeneous genetic diseases like RP. Although NGS panels are clinically available for cardiovascular conditions, there are only a small number of peer-reviewed publications regarding the technology used and clinical efficacy. An NGS panel, including 47 genes and 273 kb region, using two microarrays with 15,000 probes was utilized for the clinical diagnosis of cardiomyopathies including both DCM and HCM (Table 7.3) [101]. The level of enrichment was 2169X. The NGS analysis was performed on ten patients with primary cardiomyopathies (5 HCM and 5 DCM). Disease-causing mutations, two microdeletions, and four point mutations were detected in six patients. Additionally, several novel nonsynonymous variants, predicted to be harmful, that are potentially disease mutations or modifiers for DCM or HCM were identified. Thus, this approach allows high-throughput mutation screening in cardiomyopathies using microarray-based target enrichment followed by SOLiD NGS with high sensitivity and specificity in order to accurately detect sequence variants in multiple disease-related loci. Using current methods, it is expensive and time consuming to diagnose nonsyndromic hearing loss because of the extreme genetic heterogeneity. Nine patients diagnosed with hearing loss were utilized to assess the target enrichment, using microarray (NimbleGen) or solution-based capture (SureSelect), and massively parallel sequencing by 454 or Illumina technologies to interrogate all exons of 54 genes implicated in nonsyndromic hearing loss [86]. Samples included one negative control, three positive controls (one biological replicate), and six unknowns (10 samples total), in which 605 single nucleotide polymorphisms (SNPs) were genotyped by Sanger sequencing to measure sensitivity and specificity for the solution capture-Illumina and microarray capture-454 methods at saturating sequence coverage (Table 7.3). In addition to the identification of five pathogenic mutations in six idiopathic hearing loss patients, as expected, causative mutations were found in the positive samples but not in the negative one [86]. This study demonstrates that targeted capture plus massively parallel sequencing has a sensitivity and specificity that will allow clinicians to improve patient care of genetically heterogeneous disorders by providing prognostic information and genetic counseling. The heterogeneity of primary immunodeficiencies, in which components of immunological pathways are either missing or dysregulated, has been explored by NGS [97]. The samples from two patients, and their parents, suspected to have an underlying immunodeficiency were analyzed by capturing 395 genes, known or predicted to be associated with primary immunodeficiencies, with a 385 K probe capture and GS-FLX Titanium 454 sequencing (Table 7.3) [97].

Indication of immunodeficiency included hepatosplenomegaly, recurrent infections, and an elevated IgM level in patient one and a SCID phenotype in patient two. Trio sequence analysis revealed *ATM* and *ARTEMIS* mutations in patient one and two, respectively. NGS expands our capacities to sequence large targeted DNA regions in a less laborious and time-consuming way and permits the identification of underlying gene mutations in genetically heterogeneous disorders like primary immunodeficiencies.

3.3 Solution-Based Capture or SureSelect

Solution-based capture has been utilized to study CMDs and hearing loss [86, 100]. Aside from genetic heterogeneity, CMDs are a diagnostic challenge because of the existence of their phenotypic variability, difficulties with muscle immunohistochemical stains that do not aid in gene candidate identification, and a general lack of clinician awareness [100]. The identification of mutations in all exons from 12 genes known to cause CMDs were assessed using two different enrichment technologies, namely, solution-based capture and microdroplet-based PCR, in conjunction with SOLiD three sequencing in 12 samples, including five positive, one normal, and six unknown samples. Genotyping data showed that both enrichment technologies produced suitable calls for use in clinical laboratories. In addition to the mutations identified in the unknown samples, the expected variants and mutations were identified in the positive controls [100]. This study demonstrates the successful application of targeted sequencing in conjunction with NGS to screen for mutations in hundreds of exons in a genetically heterogeneous human disorder.

3.4 Microdroplet-Based PCR or RainDance

Microdroplet-based PCR has been used to enrich the genes known to cause CMDs, hearing loss, and primary immunodeficiencies (Table 7.3) [100]. Many genetic causes of syndromic and nonsyndromic hearing loss exist, establishing this as a heterogeneous disorder. Diagnosing hearing loss is important in the clinical management, and NGS offers an opportunity to overcome the limitation of Sanger sequencing, such as the high investment of cost and time. In a study of eight samples, microdroplet-based PCR and Illumina HiSeq2000 sequencing was used to identify variants in 24 hearing loss genes, comprising a total of 117 kb target sequence (Table 7.3) (T. A. Sivakumaran, personal communication). A total of 1148 sequence variants were detected in eight samples in 24 genes. Significantly, results showed greater than 99.99 % concordance between NGS and Sanger sequencing in the four genes thoroughly evaluated, resulting in the analytical sensitivity and specificity of 100 % and 99.99 %, respectively. Due to high sensitivity and specificity, targeted enrichment with NGS sequencing is a feasible technology for identification

of variants in the multiple genetic causes of hearing loss. Primary immunodeficiencies, with more than 150 diseases divided into eight major categories, is a large group of genetically heterogeneous disorders. DNA from 21 PIDD patients, including 17 positive controls and four unknown samples, were subjected to microdroplet-based PCR enrichment of 124 genes, with a target region of 301 kb including 20 bp of exon/intron boundaries and up to 20 bases of 5′and 3′ UTR regions, and Illumina HiSeq2000 sequencing (Table 7.3) (K. Zhang, personal communication). Successful capture of over 98 % of the target bases yielded a 100 % concordance between Sanger and NGS identified mutations and SNPs. The sensitivity and specificity of NGS, using RainDance and HiSeq2000, is over 99.00 % and 99.99 %, respectively, for the detection of nucleotide base changes, small deletions, and insertions (<10 bases) in the genes currently known to be associated with PIDD. Additionally, this panel significantly addresses the problem of heterogeneity in these disorders.

3.5 Whole Genome Amplification in Conjunction with PCR

Whole genome amplification plus PCR (WGA+PCR) has been the choice of enrichment for RP and a sensory ciliopathy, namely, NPHP-AC (Table 7.3). RP is a challenging application of NGS because it has multiple patterns of inheritance, with mutations in many genes for each inheritance pattern and numerous, distinct, disease-causing mutations at each locus; further, many RP genes have not been yet identified [98]. NGS was used to determine whether it offers rapid and efficient detection of disease-causing mutations in pairs of affected individuals from 21 families with autosomal dominant RP, selected from a cohort of families without mutations in "common" RP genes [98]. After enrichment of 46 genes by WGA+PCR and NGS sequencing via the 454 GS-FLX Titanium and Illumina Genome Analyzer IIx, platforms identified an excess of 9,000 variants in 21 families. Most variants, over 8,000, were classified as benign on the basis of their presence in controls with clearly defined disease-causing mutations. After the completion of the analyses, five disease-causing mutations were identified in the 21 families. Three of these mutations are in genes associated with either autosomal dominant RP or autosomal dominant cone–rod dystrophy. Somewhat surprising was the identification of two mutations in *RPGR*, a gene known to cause X-linked RP. This study demonstrates that NGS can be an effective tool for determining the pathogenic mutation in inherited disease families with highly heterogeneous causes. The use of this technology increases the diagnostic yield to be approximately 65 % for autosomal dominant RP cases [98]. To overcome the broad genetic locus heterogeneity, a strategy of DNA pooling with consecutive massively parallel sequencing was performed on NPHP-AC, known to be caused by 18 different genes [36]. For enrichment of the pertinent regions, 120 DNA samples were individually normalized via WGA, combined into five distinct pools each consisting of 24 samples, and all 376 exons were individually PCR amplified using each of the DNA pools as templates (Table 7.3). Following amplification, sample libraries were constructed and sequencing using

the Illumina Genome Analyzer II platform. For proof of principle, DNA from patients with known mutations was used, and detection of 22 out of 24 different alleles (92 % sensitivity) was demonstrated. Significantly, this analysis led to the molecular diagnosis of 30/120 patients (25 %), and 54 pathogenic mutations (27 novel) were identified in seven different NPHP-AC genes. Additionally, in 24 patients, only single heterozygous variants of unknown significance were found. The approach of pooling DNA samples in combination with NGS is a robust, economic, and highly effective method for examining larger patient cohorts for mutations, especially in genetic heterogeneous disorders.

3.6 Quantitative PCR (qPCR) Amplicon Ligation

Using NGS, Coppieters and coworkers published a study designed to provide an accurate, fast, cost-efficient, and comprehensive tool for molecular testing of all known LCA genes to overcome the issue of genetic heterogeneity [12]. In this study, 22 LCA patients, including five positive controls, were screened for mutations in 16 LCA genes based on a novel quantitative PCR (qPCR) amplicon ligation, shearing, and NGS strategy (Table 7.3). The first part of this study aimed at validating the enrichment protocol, whereas the second part of this study consisted of a blind screening of 12 prescreened mutation-negative patients with LCA. For the enrichment, a 152 kb target region composed of 375 qPCR amplicons (252 exons) was designed. In the validation, more than 80 % of amplicons produced a quantification cycle (Cq) value between 23 and 27. NGS validation of 107 variations was performed, and the causal genetic defect and a single heterozygous mutation were identified in patient 3 and 5, respectively, from a total of 17 patients without previously identified mutations. The thorough validation and low-cost characteristic of this approach argue for its implementation in a clinical context.

3.7 Simplex PCR

NGS was evaluated for its HCM diagnostic potential using a long-range PCR to amplify the16 genes implicated in HCM, in a control DNA, followed by 454 GS-FLX and Illumina Genome Analyzer sequencing [102]. The choice of long-range PCR allowed for the design of fewer amplicons (67 total) and the longer term opportunity to investigate potential deep intronic mutations, which to date had not been extensively studied in HCM (Table 7.3). For variant identification in this control individual sample, criteria set to include coverage of 30-fold or greater, and an allelic read percentage of 20 % or greater was set to reduce the false-positive rate. Twenty-seven exon variants were identified by both NGS platforms and were confirmed by Sanger sequencing, demonstrating the NGS advantage in HCM diagnostics.

3.8 From Laboratory Bench to Commercially Available Heterogeneous Disorder Panels

The aforementioned validation studies demonstrated the robustness of NGS, and its clinical use has started to dissipate across clinical laboratories to address the issue of genetic heterogeneity. Recently, a concerted, but independent, effort has been made to launch commercial NGS panels for congenital muscular dystrophies, autism, X-linked intellectual disability, hearing loss, breast cancer, colon cancers, and Marfan and related disorders. The NGS autism panel (Tier 2), offered by the Emory Genetics Laboratory, contains 55 genes that target genetic syndromes that include autism or autism-like features. Similarly, they also offer an NGS X-linked intellectual disability (XLID) panel containing 91 genes to assist in the diagnosis of these patients since no other clinical features are typically observed in non-syndromic XLID. Complementarily, Ambry Genetics also offers the autism and XLID panels. Furthermore, Ambry Genetics offers NGS hereditary cancer panels to examine 14, 14, and 19 genes involved in colon, breast, and ovarian cancers (also includes breast and uterine cancer genes), respectively. In cardiovascular genetic testing, Ambry Genetics has a number of NGS panels for Brugada syndrome (9 genes), LQT syndrome (12 genes), arrhythmia (29 genes), HCM (31 genes), DCM (37 genes), cardiomyopathy (56 genes), and cardiovascular diseases (79 genes). This later comprehensive panel includes 79 genes that have been implicated in multiple cardiovascular diseases including cardiomyopathies, cardiac channelopathies/arrhythmias, and structural heart defects. In addition, the NGS hearing loss panel, with 24 causative genes, is commercially available at Cincinnati Children's Medical Center (CCHMC), and other panels will be offered soon. At Baylor College of Medicine, two NGS-based tests are currently available, namely, the glycogen storage disorders and whole mitochondrial genome panels [105, 106]. The first NGS test includes the massively parallel sequencing of 16 genes known to cause muscle and liver forms of glycogen storage diseases. The second NGS test is for the detection of mtDNA point mutations and large deletions with heteroplasmic levels of the mutations and variants quantified via a novel one step mitochondrial genome method, involving a single amplicons long-range PCR in conjunction with NGS. Other NGS-based tests currently offered by other clinical diagnostic laboratories are listed in the panels section of GeneTests (http://www.ncbi.nlm.nih.gov/sites/GeneTests/?db=GeneTests).

4 Advantages and Challenges of NGS for Diagnosis of Genetically Heterogeneous Disorders

4.1 Sensitivity

It is known that rare variants contribute to pathogenesis of a number of disorders [107]. NGS provides the ability to sequence at a very high depth, thus significantly

improving the chances of identifying rare variants. It has been reported that deep sequencing (average 99X coverage) identified a rare missense variant, not identified by SNP arrays, in the *SLC26A3* gene in a patient with suspected Bartter syndrome [108]. In addition to *SLC26A3*, due to its genetically heterogeneous nature, Bartter syndrome is caused by mutations in *SLC12A1*, *KCNJ1*, *BSND*, *CLCNKA*, *CLCNKB*, and *SLC12A3*. The sensitivity of NGS will aid in the diagnosis of heterogeneous disorders by having the ability to identify rare variants in all genes that are associated with a specific syndrome.

4.2 Identification of Novel Genetic Variations and Genes

Unlike in microarrays where prior knowledge of genetic aberrations is required to generate probes, NGS provides a broader view of the genome and hence opportunities to discover novel genetic aberrations in genetically heterogeneous disorders [109]. The development and application of an NGS approach to detect novel gene defects underlying retinal diseases has been recently reported [110]. Inherited retinal disorders are clinically and genetically heterogeneous with more than 150 gene defects accounting for the diversity of disease phenotypes. A microarray capture of 254 target genes in combination with the Illumina Genome Analyzer IIx was used to analyze 20 samples from 17 families [110]. Three known and five novel mutations were identified in *NR2E3*, *PRPF3*, *EYS*, *PRPF8*, *CRB1*, *TRPM1*, and *CACNA1F*. In addition to discovery novel variants, NGS has the ability to identify disease-causing variants in novel genes as exemplified by the intellectual disability caused by more than 90 gene defects [111]. To expedite the molecular elucidation of autosomal recessive intellectual disability, homozygosity mapping, exon enrichment, and NGS in 136 consanguineous families with autosomal recessive intellectual disability from Iran and elsewhere were performed. In this study, disease-causing variants in 50 novel candidate genes were identified. As the technology continues to improve with a decreasing cost and error rate, it will be possible to sequence regions of interest at high depth, and this will enable identification of greater number of rare variants and novel genes that contribute to the pathogenesis of genetically heterogeneous disorders.

4.3 Sample Requirements

The amount and quality of DNA available for carrying out genetic tests can be a limiting factor. It is not always possible to get a large quantity of DNA for conducting genetic tests, and hence, it is important that techniques used for genetic testing require small amount of DNA [112]. While Sanger sequencing required several micrograms of DNA for a single gene of an average of 25 exons, even larger DNA amounts are needed for large panels (50 ng/exon); sequencing using NGS technologies can be conducted with a range of 50 ng to 3 μg of DNA for current genetically heterogeneous panels with several hundred exons (Table 7.3) [113–115].This offers

a distinct advantage for usage of NGS technologies in the clinic especially in newborns or those with immunodeficiency disorders.

4.4 NGS Challenges of Data Analysis and Reporting

NGS data analysis is an evolving field and possibly the biggest bottleneck for routine adoption of NGS in clinical setting when hundreds of exons are included in a heterogeneous disorder panel [112]. For clinical services, NGS analysis should ideally be simple, fast, and accurate, and the production of an output that is interpretable by medical staff should be a primary goal. NGS data analysis is still largely carried out using open-source tools [116–118]. These open-source tools have been immensely useful in analyzing NGS data in research laboratories, but they do not meet the criteria demanded by clinical laboratories, namely, easy, quick, and accurate analyses. Moreover, commercially available NGS data analysis solutions are geared toward handling the hundreds of gigabyte data generated in research settings but have not been designed from clinical perspective [112]. A possible solution for this necessity is to create powerful clinical software that analyzes an NGS output from multiple gene panels. To aid in the diagnosis of heterogeneous disorders, this software needs to integrate algorithms for data mining that do comparisons between a given clinical sample and public/private databases to include informative variants and exclude uninformative variants. Medical report-like formatting of this information to include only causative variants of unknown clinical significance and rare variants with explanations will help physicians understand this genomic information better and will likely have a direct impact on patient care.

4.5 Challenges of Data Storage, Data Processing, and Computing Infrastructure

The amount of data being generated using NGS systems is outpacing the computational capacity of most systems and requires highly skilled IT personnel and bioinformatics staff to set up, maintain, and run NGS data analysis tools [119]. Due to large number of exons, the raw sequencing data output from heterogeneous disorder panels is substantial and requires specialized computer equipment and algorithms for cost-effective handling. In clinical laboratories, the raw data, downstream of the image file, including fastq, fastq.gz, fasta, and others, must be processed and kept at every step for each patient sample. The challenge is to condense and store these files in an efficient manner to reduce the cost of storage. At the rate that we are generating data, our storage needs will increase, and this demand will increase the cost of storage. The cost of sequencing has decreased, but for this technology to become

popular, we must decrease the cost of storage as well. In terms of process time, clinical laboratories depend on short turnaround time (TAT) to deliver the best possible care for their patients. With larger datasets, multicore processors must be used to provide the advantage to reducing processing time. This computational requirement may be difficult to meet for small laboratories and clinics. Currently, NGS panel processing times vary from several hours to overnight. However, as the panels get larger, increased RAM memory must be used to continue to provide short TATs. In addition, automation of the raw sequencing data processing is a possible solution for efficient handling of multiple data files and may have an impact on the reduction of TATs. Storage and analysis of NGS data demands sophisticated and high-end computing infrastructure (at a minimum 8 quad core, 32 gigabyte RAM, and 10 terabytes of disk space). This kind of computing infrastructure and manpower is impractical for small diagnostic laboratories and clinics. The cost of managing, storing, and analyzing NGS data easily runs into hundreds of thousands of dollars and currently is a severe deterrent for adoption of NGS systems in clinics [120]. In order to benefit from the many advantages of sequencing the genome, the cost of data analysis needs to be made manageable.

4.6 Ethical Issues

Genetic testing plays an important role in the management of inherited heterogeneous diseases, but there are several caveats that must be considered. As sequencing using NGS provides a larger view of the genome, a concern is that genetic testing may reveal unexpected information that is not directly related to the original clinical question [121]. Examples include unexpected nonpaternity and consanguinity. As NGS sequencing costs fall, we may move toward larger and larger panels, and eventually the whole exome and genome; this could lead to the possibility of obtaining incidental findings such as a mutation in a muscular dystrophy-causing gene during investigation for hearing loss. In contrast, a number of variants discovered may not have clinical significance, or their clinical significance may not be known, and sharing of such information with patients should be carefully considered [112]. There are many ethical questions that arise from this topic. How will clinical laboratories report incidental findings to clinicians, and, subsequently, should clinicians communicate these results to patients? Which conditions and/or mutations should be communicated to the clinician and patient? In March of this year, the American College of Medical Genetics and Genomics (ACMG) released a policy statement entitled “Points to Consider in the Clinical Application of Genomic Sequencing” that address these concerns. It is recommended that gene variants known to be associated with a phenotype, but not believed to be related to the condition that led to the testing (“secondary findings”), should be reported. Moreover, it is recommended that laboratories and clinics utilizing NGS data should have clear policies in place related to disclosure of secondary

findings. It states that patients should be informed of those policies and the types of secondary findings that will be reported to them and that patients should be given the option of not receiving secondary findings. However, there will be exceptional cases where the ordering physician and laboratory must decide what is best for the patient.

5 Conclusions

NGS has had a large impact in biomedical research, and it is making its way to the clinical diagnostic arena driven by the reduction of cost and increased flexibility due to the advent of targeted NGS technologies. We have shown that NGS, by employing various enrichment and sequencing methods, provides a solution for the diagnosis of genetic heterogeneous disorders by having the ability to analyze multiple genes in parallel. In fact, we have shown that NGS has already been used to identify mutations underlying well-known genetic heterogeneous disorders, namely, RP, LCA, ciliopathies, CMD, cardiomyopathies, hearing loss, and immunodeficiency disorders. Although this application of NGS along with unique advantages of the technology is encouraging, several technical and ethical challenges need to be addressed as we move forward by using the current and upcoming systems in the clinic. In order to be widely and routinely used in clinical practice, there needs to be further reduction in data storage cost and flexibility in throughput and streamlining of data generation, analysis, interpretation, and reporting. The creation of benchtop sequencers, decreased run times, and reduction of technical complexity by automation alongside bioinformatic infrastructure investments for analysis and interpretation will speed up the adoption process. In the immediate future, it is easy to envisage a more comprehensive use of this powerful technology for the testing of genetic heterogeneous disorders by offering larger and diverse NGS panels.

References

1. Burke W (2002) Genetic testing. N Engl J Med 347(23):1867–1875
2. Tucker T, Marra M, Friedman JM (2009) Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet 85(2):142–154
3. Hartong DT, Berson EL, Dryja TP (2006) Retinitis pigmentosa. Lancet 368(9549):1795–1809
4. Drack AV, Lambert SR, Stone EM (2010) From the laboratory to the clinic: molecular genetic testing in pediatric ophthalmology. Am J Ophthalmol 149(1):10–17
5. Cremers FPM, Collin RWJ (2009) Promises and challenges of genetic therapy for blindness. Lancet 374(9701):1569–1570
6. Simpson DA, Clark GR, Alexander S, Silvestri G, Willoughby CE (2011) Molecular diagnosis for heterogeneous genetic diseases with targeted high-throughput DNA sequencing applied to retinitis pigmentosa. J Med Genet 48(3):145–151
7. Neveling K, Collin RWJ, Gilissen C et al (2012) Next-generation genetic testing for retinitis pigmentosa. Hum Mutat 33(6):963–972

8. Berger W, Kloeckener-Gruissem B, Neidhardt J (2010) The molecular basis of human retinal and vitreoretinal diseases. Prog Retin Eye Res 29(5):335–375
9. Koenekoop RK, Lopez I, den Hollander AI, Allikmets R, Cremers FPM (2007) Genetic testing for retinal dystrophies and dysfunctions: benefits, dilemmas and solutions. Clin Experiment Ophthalmol 35(5):473–485
10. Mandal MNA, Heckenlively JR, Burch T, Chen L, Vasireddy V, Koenekoop RK, Sieving PA, Ayyagari R (2005) Sequencing arrays for screening multiple genes associated with early-onset human retinal degenerations on a high-throughput platform. Invest Ophthalmol Vis Sci 46(9):3355–3362
11. Ávila-Fernández A, Cantalapiedra D, Aller E et al (2010) Mutation analysis of 272 Spanish families affected by autosomal recessive retinitis pigmentosa using a genotyping microarray. Mol Vis 16:2550–2558
12. Coppieters F, De Wilde B, Lefever S et al (2012) Massively parallel sequencing for early molecular diagnosis in Leber congenital amaurosis. Genet Med 14(6):576–585
13. den Hollander AI, Roepman R, Koenekoop RK, Cremers FPM (2008) Leber congenital amaurosis: genes, proteins and disease mechanisms. Prog Retin Eye Res 27(4):391–419
14. Lambert SR, Sherman S, Taylor D, Kriss A, Coffey R, Pembrey M (1993) Concordance and recessive inheritance of Leber congenital amaurosis. Am J Med Genet 46(3):275–277
15. den Hollander AI, Black A, Bennett J, Cremers FPM (2010) Lighting a candle in the dark: advances in genetics and gene therapy of recessive retinal dystrophies. J Clin Invest 120(9):3042–3053
16. Chung DC, Traboulsi EI (2009) Leber congenital amaurosis: clinical correlations with genotypes, gene therapy trials update, and future directions. J AAPOS 13(6):587–592
17. Zernant J, Külm M, Dharmaraj S et al (2005) Genotyping microarray (disease chip) for Leber congenital amaurosis: detection of modifier alleles. Invest Ophthalmol Vis Sci 46(9):3052–3059
18. Coppieters F, Casteels I, Meire F et al (2010) Genetic screening of LCA in Belgium: predominance of CEP290 and identification of potential modifier alleles in AHI1 of CEP290-related phenotypes. Hum Mutat 31(10):E1709–E1766
19. Ferkol TW, Leigh MW (2012) Ciliopathies: the central role of cilia in a spectrum of pediatric disorders. J Pediatr 160(3):366–371
20. Leigh MW, Pittman JE, Carson JL, Ferkol TW, Dell SD, Davis SD, Knowles MR, Zariwala MA (2009) Clinical and genetic aspects of primary ciliary dyskinesia/Kartagener syndrome. Genet Med 11(7):473–487
21. Cardenas-Rodriguez M, Badano JL (2009) Ciliary biology: understanding the cellular and genetic basis of human ciliopathies. Am J Med Genet C Semin Med Genet 151C(4):263–280
22. Satir P, Pedersen LB, Christensen ST (2010) The primary cilium at a glance. J Cell Sci 123(Pt 4):499–503
23. Badano JL, Mitsuma N, Beales PL, Katsanis N (2006) The ciliopathies: an emerging class of human genetic disorders. Annu Rev Genomics Hum Genet 7:125–148
24. Tobin JL, Beales PL (2009) The nonmotile ciliopathies. Genet Med 11(6):386–402
25. Afzelius BA (1976) A human syndrome caused by immotile cilia. Science 193(4250):317–319
26. Noone PG, Leigh MW, Sannuti A, Minnix SL, Carson JL, Hazucha M, Zariwala MA, Knowles MR (2004) Primary ciliary dyskinesia: diagnostic and phenotypic features. Am J Respir Crit Care Med 169(4):459–467
27. Coren ME, Meeks M, Morrison I, Buchdahl RM, Bush A (2002) Primary ciliary dyskinesia: age at diagnosis and symptom history. Acta Paediatr 91(6):667–669
28. Zariwala MA, Leigh MW, Ceppa F et al (2006) Mutations of DNAI1 in primary ciliary dyskinesia: evidence of founder effect in a common mutation. Am J Respir Crit Care Med 174(8):858–866

29. Hornef N, Olbrich H, Horvath J et al (2006) DNAH5 mutations are a common cause of primary ciliary dyskinesia with outer dynein arm defects. Am J Respir Crit Care Med 174(2):120–126
30. Gerdes JM, Davis EE, Katsanis N (2009) The vertebrate primary cilium in development, homeostasis, and disease. Cell 137(1):32–45
31. Hildebrandt F, Otto E (2005) Cilia and centrosomes: a unifying pathogenic concept for cystic kidney disease? Nat Rev Genet 6(12):928–940
32. Gunay-Aygun M (2009) Liver and kidney disease in ciliopathies. Am J Med Genet C Semin Med Genet 151C(4):296–306
33. Yoder BK (2007) Role of primary cilia in the pathogenesis of polycystic kidney disease. J Am Soc Nephrol 18(5):1381–1388
34. Wang S, Luo Y, Wilson PD, Witman GB, Zhou J (2004) The autosomal recessive polycystic kidney disease protein is localized to primary cilia, with concentration in the basal body area. J Am Soc Nephrol 15(3):592–602
35. Hildebrandt F, Attanasio M, Otto E (2009) Nephronophthisis: disease mechanisms of a ciliopathy. J Am Soc Nephrol 20(1):23–35
36. Otto EA, Ramaswami G, Janssen S et al (2011) Mutation analysis of 18 nephronophthisis associated ciliopathy disease genes using a DNA pooling and next generation sequencing strategy. J Med Genet 48(2):105–116
37. Zaghloul NA, Katsanis N (2009) Mechanistic insights into Bardet-Biedl syndrome, a model ciliopathy. J Clin Invest 119(3):428–437
38. Parisi MA (2009) Clinical and molecular features of Joubert syndrome and related disorders. Am J Med Genet C Semin Med Genet 151C(4):326–340
39. Hildebrandt F, Otto E, Rensing C, Nothwang HG, Vollmer M, Adolphs J, Hanusch H, Brandis M (1997) A novel gene encoding an SH3 domain protein is mutated in nephronophthisis type 1. Nat Genet 17(2):149–153
40. Otto E, Hoefele J, Ruf R et al (2002) A gene mutated in nephronophthisis and retinitis pigmentosa encodes a novel protein, nephroretinin, conserved in evolution. Am J Hum Genet 71(5):1161–1167
41. Mollet G, Salomon R, Gribouval O et al (2002) The gene mutated in juvenile nephronophthisis type 4 encodes a novel protein that interacts with nephrocystin. Nat Genet 32(2):300–305
42. Otto EA, Schermer B, Obara T et al (2003) Mutations in INVS encoding inversin cause nephronophthisis type 2, linking renal cystic disease to the function of primary cilia and left-right axis determination. Nat Genet 34(4):413–420
43. Olbrich H, Fliegauf M, Hoefele J et al (2003) Mutations in a novel gene, NPHP3, cause adolescent nephronophthisis, tapeto-retinal degeneration and hepatic fibrosis. Nat Genet 34(4):455–459
44. Otto EA, Loeys B, Khanna H et al (2005) Nephrocystin-5, a ciliary IQ domain protein, is mutated in Senior-Loken syndrome and interacts with RPGR and calmodulin. Nat Genet 37(3):282–288
45. Sayer JA, Otto EA, O'Toole JF et al (2006) The centrosomal protein nephrocystin-6 is mutated in Joubert syndrome and activates transcription factor ATF4. Nat Genet 38(6):674–681
46. Attanasio M, Uhlenhaut NH, Sousa VH et al (2007) Loss of GLIS2 causes nephronophthisis in humans and mice by increased apoptosis and fibrosis. Nat Genet 39(8):1018–1024
47. Delous M, Baala L, Salomon R et al (2007) The ciliary gene RPGRIP1L is mutated in cerebello-oculo-renal syndrome (Joubert syndrome type B) and Meckel syndrome. Nat Genet 39(7):875–881
48. Otto EA, Trapp ML, Schultheiss UT, Helou J, Quarmby LM, Hildebrandt F (2008) NEK8 mutations affect ciliary and centrosomal localization and may cause nephronophthisis. J Am Soc Nephrol 19(3):587–592

49. Otto EA, Tory K, Attanasio M et al (2009) Hypomorphic mutations in meckelin (MKS3/TMEM67) cause nephronophthisis with liver fibrosis (NPHP11). J Med Genet 46(10):663–670
50. O'Toole JF, Liu Y, Davis EE et al (2010) Individuals with mutations in XPNPEP3, which encodes a mitochondrial protein, develop a nephronophthisis-like nephropathy. J Clin Invest 120(3):791–802
51. Valente EM, Logan CV, Mougou-Zerelli S et al (2010) Mutations in TMEM216 perturb ciliogenesis and cause Joubert, Meckel and related syndromes. Nat Genet 42(7):619–625
52. Kyttälä M, Tallila J, Salonen R, Kopra O, Kohlschmidt N, Paavola-Sakki P, Peltonen L, Kestilä M (2006) MKS1, encoding a component of the flagellar apparatus basal body proteome, is mutated in Meckel syndrome. Nat Genet 38(2):155–157
53. Smith UM, Consugar M, Tee LJ et al (2006) The transmembrane protein meckelin (MKS3) is mutated in Meckel-Gruber syndrome and the wpk rat. Nat Genet 38(2):191–196
54. Baala L, Audollent S, Martinovic J et al (2007) Pleiotropic effects of CEP290 (NPHP6) mutations extend to Meckel syndrome. Am J Hum Genet 81(1):170–179
55. Tallila J, Jakkula E, Peltonen L, Salonen R, Kestilä M (2008) Identification of CC2D2A as a Meckel syndrome gene adds an important piece to the ciliopathy puzzle. Am J Hum Genet 82(6):1361–1367
56. Finsterer J, Ramaciotti C, Wang CH, Wahbi K, Rosenthal D, Duboc D, Melacini P (2010) Cardiac findings in congenital muscular dystrophies. Pediatrics 126(3):538–545
57. Cardamone M, Darras BT, Ryan MM (2008) Inherited myopathies and muscular dystrophies. Semin Neurol 28(2):250–259
58. Brockington M, Yuva Y, Prandini P et al (2001) Mutations in the fukutin-related protein gene (FKRP) identify limb girdle muscular dystrophy 2I as a milder allelic variant of congenital muscular dystrophy MDC1C. Hum Mol Genet 10(25):2851–2859
59. Mercuri E, Brockington M, Straub V et al (2003) Phenotypic spectrum associated with mutations in the fukutin-related protein gene. Ann Neurol 53(4):537–542
60. Poppe M, Cree L, Bourke J et al (2003) The phenotype of limb-girdle muscular dystrophy type 2I. Neurology 60(8):1246–1251
61. Harel T, Goldberg Y, Shalev SA, Chervinski I, Ofir R, Birk OS (2004) Limb-girdle muscular dystrophy 2I: phenotypic variability within a large consanguineous Bedouin family associated with a novel FKRP mutation. Eur J Hum Genet 12(1):38–43
62. Jimenez-Mallebrera C, Brown SC, Sewry CA, Muntoni F (2005) Congenital muscular dystrophy: molecular and cellular aspects. Cell Mol Life Sci 62(7–8):809–823
63. van Reeuwijk J, Brunner HG, van Bokhoven H (2005) Glyc-O-genetics of Walker-Warburg syndrome. Clin Genet 67(4):281–289
64. Martin PT, Freeze HH (2003) Glycobiology of neuromuscular disorders. Glycobiology 13(8):67R–75R
65. Muntoni F, Brockington M, Blake DJ, Torelli S, Brown SC (2002) Defective glycosylation in muscular dystrophy. Lancet 360(9343):1419–1421
66. Michele DE, Campbell KP (2003) Dystrophin-glycoprotein complex: post-translational processing and dystroglycan function. J Biol Chem 278(18):15457–15460
67. Endo T (2004) Structure, function and pathology of O-mannosyl glycans. Glycoconj J 21(1–2):3–7
68. Schachter H, Vajsar J, Zhang W (2004) The role of defective glycosylation in congenital muscular dystrophy. Glycoconj J 20(5):291–300
69. Endo T, Toda T (2003) Glycosylation in congenital muscular dystrophies. Biol Pharm Bull 26(12):1641–1647
70. Goldenberg I, Horr S, Moss AJ et al (2011) Risk for life-threatening cardiac events in patients with genotype-confirmed long-QT syndrome and normal-range corrected QT intervals. J Am Coll Cardiol 57(1):51–59
71. Hershberger RE, Lindenfeld J, Mestroni L, Seidman CE, Taylor MRG, Towbin JA (2009) Genetic evaluation of cardiomyopathy–a Heart Failure Society of America practice guideline. J Card Fail 15(2):83–97

72. Ingles J, Zodgekar PR, Yeates L, Macciocca I, Semsarian C, Fatkin D (2011) Guidelines for genetic testing of inherited cardiac disorders. Heart Lung Circ 20(11):681–687
73. Herman DS, Lam L, Taylor MRG et al (2012) Truncations of titin causing dilated cardiomyopathy. N Engl J Med 366(7):619–628
74. Lo YMD, Chan KCA, Sun H et al (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2(61):61ra91
75. Jordan DM, Kiezun A, Baxter SM et al (2011) Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy. Am J Hum Genet 88(2):183–192
76. Lippi G, Montagnana M, Meschi T, Comelli I, Cervellin G (2012) Genetic and clinical aspects of Brugada syndrome: an update. Adv Clin Chem 56:197–208
77. Morton CC, Nance WE (2006) Newborn hearing screening–a silent revolution. N Engl J Med 354(20):2151–2164
78. Hilgert N, Smith RJH, Van Camp G (2009) Forty-six genes causing nonsyndromic hearing impairment: which ones should be analyzed in DNA diagnostics? Mutat Res 681(2–3):189–196
79. Raviv D, Dror AA, Avraham KB (2010) Hearing loss: a common disorder caused by many rare alleles. Ann N Y Acad Sci 1214:168–179
80. Kochhar A, Hildebrand MS, Smith RJH (2007) Clinical aspects of hereditary hearing loss. Genet Med 9(7):393–408
81. Cohen M, Phillips JA 3rd (2012) Genetic approach to evaluation of hearing loss. Otolaryngol Clin North Am 45(1):25–39
82. Van Camp G, Willems PJ, Smith RJ (1997) Nonsyndromic hearing impairment: unparalleled heterogeneity. Am J Hum Genet 60(4):758–764
83. Brownstein Z, Avraham KB (2009) Deafness genes in Israel: implications for diagnostics in the clinic. Pediatr Res 66(2):128–134
84. Vandebona H, Mitchell P, Manwaring N, Griffiths K, Gopinath B, Wang JJ, Sue CM (2009) Prevalence of mitochondrial 1555A→G mutation in adults of European descent. N Engl J Med 360(6):642–644
85. Zbar RI, Ramesh A, Srisailapathy CR, Fukushima K, Wayne S, Smith RJ (1998) Passage to India: the search for genes causing autosomal recessive nonsyndromic hearing loss. Otolaryngol Head Neck Surg 118(3 Pt 1):333–337
86. Shearer AE, DeLuca AP, Hildebrand MS, Taylor KR, Gurrola J 2nd, Scherer S, Scheetz TE, Smith RJH (2010) Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. Proc Natl Acad Sci USA 107(49):21104–21109
87. Pandolfi F, Cianci R, Cammarota G, Pagliari D, Landolfi R, Conti P, Theoharides TC (2010) Recent insights in primary immunodeficiency diseases: the role of T-lymphocytes and innate immunity. Ann Clin Lab Sci 40(1):3–9
88. Barbouche M-R, Galal N, Ben-Mustapha I, Jeddane L, Mellouli F, Ailal F, Bejaoui M, Boutros J, Marsafy A, Bousfiha AA (2011) Primary immunodeficiencies in highly consanguineous North African populations. Ann N Y Acad Sci 1238:42–52
89. Costabile M, Quach A, Ferrante A (2006) Molecular approaches in the diagnosis of primary immunodeficiency diseases. Hum Mutat 27(12):1163–1173
90. Notarangelo L, Casanova J-L, Conley ME, Chapel H, Fischer A, Puck J, Roifman C, Seger R, Geha RS (2006) Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee Meeting in Budapest, 2005. J Allergy Clin Immunol 117(4):883–896
91. Geha RS, Notarangelo LD, Casanova J-L et al (2007) Primary immunodeficiency diseases: an update from the International Union of Immunological Societies Primary Immunodeficiency Diseases Classification Committee. J Allergy Clin Immunol 120(4):776–794
92. Chapel H (2012) Classification of primary immunodeficiency diseases by the International Union of Immunological Societies (IUIS) Expert Committee on Primary Immunodeficiency 2011. Clin Exp Immunol 168(1):58–59

93. Samarghitean C, Ortutay C, Vihinen M (2009) Systematic classification of primary immunodeficiencies based on clinical, pathological, and laboratory parameters. J Immunol 183(11):7569–7575
94. Ishii E, Ueda I, Shirakawa R et al (2005) Genetic subtypes of familial hemophagocytic lymphohistiocytosis: correlations with clinical features and cytotoxic T lymphocyte/natural killer cell functions. Blood 105(9):3442–3448
95. Nagafuji K, Nonami A, Kumano T et al (2007) Perforin gene mutations in adult-onset hemophagocytic lymphohistiocytosis. Haematologica 92(7):978–981
96. Johnson TS, Villanueva J, Filipovich AH, Marsh RA, Bleesing JJ (2011) Contemporary diagnostic methods for hemophagocytic lymphohistiocytic disorders. J Immunol Methods 364(1–2):1–13
97. Ghosh S, Krux F, Binder V, Gombert M, Niehues T, Feyen O, Laws H-J, Borkhardt A (2012) Array-based sequence capture and next-generation sequencing for the identification of primary immunodeficiencies. Scand J Immunol 75(3):350–354
98. Bowne SJ, Sullivan LS, Koboldt DC et al (2011) Identification of disease-causing mutations in autosomal dominant retinitis pigmentosa (adRP) using next-generation DNA sequencing. Invest Ophthalmol Vis Sci 52(1):494–503
99. Daiger SP, Sullivan LS, Bowne SJ, Birch DG, Heckenlively JR, Pierce EA, Weinstock GM (2010) Targeted high-throughput DNA sequencing for gene discovery in retinitis pigmentosa. Adv Exp Med Biol 664:325–331
100. Valencia CA, Rhodenizer D, Bhide S, Chin E, Littlejohn MR, Keong LM, Rutkowski A, Bonnemann C, Hegde M (2012) Assessment of target enrichment platforms using massively parallel sequencing for the mutation detection for congenital muscular dystrophy. J Mol Diagn 14(3):233–246
101. Meder B, Haas J, Keller A et al (2011) Targeted next-generation sequencing for the molecular genetic diagnostics of cardiomyopathies. Circ Cardiovasc Genet 4(2):110–122
102. Voelkerding KV, Dames S, Durtschi JD (2010) Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic cardiomyopathy: a paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. J Mol Diagn 12(5):539–551
103. Hu H, Wrogemann K, Kalscheuer V et al (2009) Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing. Hugo J 3(1–4):41–49
104. Porreca GJ, Zhang K, Li JB et al (2007) Multiplex amplification of large sets of human exons. Nat Methods 4(11):931–936
105. Wang J, Cui H, Lee N-C, Hwu W-L, Chien Y-H, Craigen WJ, Wong L-J, Zhang VW (2012) Clinical application of massively parallel sequencing in the molecular diagnosis of glycogen storage diseases of genetically heterogeneous origin. Genet Med. doi:10.1038/gim.2012.104
106. Zhang W, Cui H, Wong L-JC (2012) Comprehensive one-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem 58(9):1322–1331
107. Need AC, Ge D, Weale ME et al (2009) A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet 5(2):e1000373
108. Choi M, Scholl UI, Ji W et al (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci USA 106(45):19096–19101
109. Welch JS, Westervelt P, Ding L et al (2011) Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA 305(15):1577–1584
110. Audo I, Bujakowska KM, Léveillard T et al (2012) Development and application of a next-generation-sequencing (NGS) approach to detect known and novel gene defects underlying retinal diseases. Orphanet J Rare Dis 7:8
111. Najmabadi H, Hu H, Garshasbi M et al (2011) Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 478(7367):57–63
112. Desai AN, Jere A (2012) Next-generation sequencing: ready for the clinics? Clin Genet 81(6):503–510

113. Venter JC, Adams MD, Myers EW et al (2001) The sequence of the human genome. Science 291(5507):1304–1351
114. Bentley DR, Balasubramanian S, Swerdlow HP et al (2008) Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456(7218):53–59
115. Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, Syed F, Dumas M, Wommack KE (2011) Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl Environ Microbiol 77(22):8071–8079
116. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Res 19(6):1117–1123
117. Gnerre S, Maccallum I, Przybylski D et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108(4):1513–1518
118. Korbel JO, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, Snyder M, Gerstein MB (2009) PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. Genome Biol 10(2):R23
119. Stein LD (2010) The case for cloud computing in genome informatics. Genome Biol 11(5):207
120. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB (2011) The real cost of sequencing: higher than you think! Genome Biol 12(8):125
121. Ware JS, Roberts AM, Cook SA (2012) Republished review: next generation sequencing for clinical diagnostics and personalised medicine: implications for the next generation cardiologist. Postgrad Med J 88(1038):234–239

Chapter 8
Molecular Diagnosis of Congenital Disorders of Glycosylation (CDG)

Melanie Jones and Madhuri Hegde

Abstract Glycosylation is the addition of sugars (glycans) to proteins and lipids. Defective synthesis, assembly, or processing of glycans results in a group of disorders known as congenital disorders of glycosylation (CDG). Next-generation sequencing (NGS) technology is used in many molecular diagnostic laboratories and consists of comprehensive panels of genes associated with particular disorders and whole exome sequencing (WES) which has recently debuted in the diagnostic laboratory. Cautions and challenges with using NGS panels and WES in the clinical setting using CDG as an example are discussed. A comprehensive NGS panel for CDG is being used when there is no indication either biochemically or clinically what gene defect may be present. In the research setting, WES was successfully used to identify the gene defect in several individuals with unknown types of CDG. New gene discoveries for CDG are leading to improved molecular diagnostic testing for CDG, including an updated comprehensive NGS panel. Identification of new CDG genes also provides direction for translational research, which is already occurring for several subtypes of CDG.

Abbreviations

HGMD	Human Gene Mutation Database
OMIM	Online Mendelian Inheritance of Man
WES	Whole exome sequencing

M. Jones, Ph.D
Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

M. Hegde, Ph.D., FACMG (✉)
Department of Human Genetics, Emory University School of Medicine, Atlanta, GA, USA

Whitehead Biomedical Research Building, Emory University School of Medicine,
615 Michael St., Ste. 301, Atlanta, GA 30322, USA
e-mail: mhegde@emory.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_8,

CDG	Congenital disorders of glycosylation
NGS	Next-generation sequencing
GPI	Glycophosphatidylinositol
IEF	Isoelectric focusing
MS	Mass spectrometry
CAP	College of American Pathologists
CLIA	Clinical Laboratory Improvement Amendments
HIPAA	Health Insurance Portability and Accountability Act
OST	Oligosaccharyltransferase complex
VOUS	Variant of unknown clinical significance
NHLBI	National Heart, Lung, and Blood Institute
CGH	Comparative genomic hybridization

1 Introduction

Glycosylation is an essential posttranslational modification and involves eight different biosynthetic pathways located within the ER and Golgi apparatus [1]. Within these pathways, glycans (sugars) are added to proteins and lipids, and after addition, the glycans are further modified. It is estimated that more than half of all proteins are glycosylated [2]. Glycans are added to proteins through the N-linked and O-linked glycosylation pathways. The type of amino acid residue used for the attachment of glycans to proteins differs for these two pathways. For N-linked glycosylation, glycans are added to asparagine residues on proteins. For O-linked glycosylation, glycans are added to serine or threonine residues on proteins [3]. The most abundant type of O-linked glycosylation is N-acetylglucosamine (GalNac), and other types of O-linked glycosylation include O-mannose, O-xylose, O-glucose and O-fucose [4]. Both N- and O-linked glycosylation are important for many different biological processes. N-linked glycosylation is important for many secretory and membrane-bound proteins and is essential for protein folding and stability, inter- and intracellular trafficking, cell signaling and recognition, protein-protein complex formation, and for protease resistance [1]. O-linked glycosylation is important for many cell surface and extracellular proteins and is essential for immunity, serving as lubricants and providing cushioning and stability to the extracellular matrix; receptor-mediated signaling; protein expression, processing, and recognition; and for the determination of blood type [1, 5]. Glycans are added to lipids through the glycosphingolipid and glycophosphatidylinositol (GPI) anchor pathways [1]. Lipid glycosylation is important for proper cell signaling and membrane diffusion and sorting [6]. With all of these processes relying on glycosylation, the proper development and functioning of multiple organ systems in the body is dependent upon correct glycosylation.

2 Defects in the Glycosylation Process Cause Disease

Defects in the synthesis, processing, or transfer of glycans result in a group of over 60 metabolic disorders known as congenital disorders of glycosylation (CDG) (Table 8.1) [4]. Gene defects have been identified in seven of the eight ER-Golgi biosynthetic pathways including the N-linked, O-linked (O-GalNac, O-xylose, O-mannose, and O-fucose), glycosphingolipid, and GPI anchor pathways. Defects have also been identified in additional pathways associated with nucleotide sugar transport into the ER or Golgi apparatus and vesicular transport. The clinical features can differ depending on which pathway the gene defect resides in, but there are also many overlapping features. CDGs caused by defects in the N-linked pathway are characterized by multi-organ dysfunction and symptoms that can appear soon after birth [7]. Developmental delay and failure to thrive are common and can be the first indication of a CDG. The nervous system is impaired in the majority of N-linked defects [4]. Additional organ dysfunction can include liver, gastrointestinal, cardiac, and immune system [8]. Clotting factor deficiencies are also present with N-linked defects and can also be a life-threatening problem by resulting in thrombosis or bleeding tendencies [9]. O-linked defects have more specific organ involvement and are characterized by muscle, bone, cartilage, and extracellular matrix defects [1]. O-mannose defects are categorized under the muscular dystrophies [10]. Patients with combined N- and O-linked defects have symptoms that are suggestive of a metabolic defect but can also present with congenital malformations. Lipid-linked defects primarily impact the nervous system [11–13]. The variation in clinical presentation and severity of disease causes a significant challenge to pediatric healthcare providers. Significant morbidity and mortality is associated with CDG due to system-wide organ dysfunction or severe infections. The majority of CDGs are inherited in an autosomal recessive manner. Exceptions are X-linked recessive MAGT1-CDG and ALG13-CDG and autosomal dominant EXT1/EXT2-CDG and GNE-CDG.

3 Diagnosis and Biochemical Testing for CDG

The first step in determining whether a patient has CDG is biochemical analysis of serum transferrin. Biochemical analysis of serum transferrin gives a characteristic pattern with patients classified as having a type I or type II CDG [14]. Type I designates defects in the synthesis or transfer of the glycan to proteins and results in hypoglycosylation [14]. Type II designates defects in further modification of these glycans after they are attached to protein which results in mis-glycosylation [14]. A type I pattern indicates that the individual has an N-linked glycosylation pathway defect. A type II pattern indicates that the defect can be N-linked, O-linked, or in pathways associated with lipid-linked glycosylation. Defects in genes whose functions are needed in multiple pathways result in a combined type I and type II

Table 8.1 Known CDG types, number of known cases, number of identified mutations, and whether molecular diagnostic testing is available

Gene	CDG type new classification (CDG subtype original classification)	Estimated number of cases or prevalence	Number of reported mutations in Human Gene Mutation Database	Molecular diagnostic testing currently available *Analyzed using NGS
N-linked glycosylation defects				
PMM2	PMM2-CDG (CDG-Ia)	>700	115	Yes*
MPI	MPI-CDG (CDG-Ib)	>20	19	Yes*
ALG6	ALG6-CDG (CDG-Ic)	>30	19	Yes*
ALG3	ALG3-CDG (CDG-Id)	6	9	Yes*
DPM1	DPM1-CDG (CDG-Ie)	14	6	Yes*
MPDU1	MPDU1-CDG (CDG-If)	7	5	Yes*
ALG12	ALG12-CDG (CDG-Ig)	7	11	Yes*
ALG8	ALG8-CDG (CDG-Ih)	5	13	Yes*
ALG2	ALG2-CDG (CDG-Ii)	1	2	Yes*
DPAGT1	DPAGT1-CDG (CDG-Ij)	11	5	Yes*
ALG1	ALG1-CDG (CDG-Ik)	20	9	Yes^
ALG9	ALG9-CDG (CDG-Il)	2	2	Yes*
DOLK	DOLK-CDG (CDG-Im)	16	5	Yes*
RFT1	RFT1-CDG (CDG-In)	7	5	Yes*
DPM3	DPM3-CDG (CDG-Io)	1	1	No
ALG11	ALG11-CDG (CDG-Ip)	5	6	No
SRD5A3	SRD5A3-CDG (CDG-Iq)	14	11	Yes
DDOST	DDOST-CDG (CDG-Ir)	1	2	Yes
ALG13	ALG13-CDG (CDG-Is)	1	1	No
TUSC3	TUSC3-CDG	4	5	Yes*
MAGT1	MAGT1-CDG	14	3	Yes
DHDDS	DHDDS-CDG	18	1	Yes
MAN1B1	MAN1B1-CDG	12	4	No
PGM1	PGM1-CDG	2	2	No
MGAT2	MGAT2-CDG (CDG-IIa)	4	5	Yes*
GCS1	GCS1-CDG (CDG-IIb)	3	2	Yes*
ST3GAL3	ST3GAL3-CDG	12	2	No
O-linked glycosylation defects				
EXT1	EXT1-CDG	1 in 50,000	393	Yes
EXT2	EXT2-CDG	1in 50,000	183	Yes
CHST14	CHST14-CDG	24	12	Yes
CHST3	CHST3-CDG	>18	27	Yes
CHST6	CHST6-CDG	Unknown	162	Yes
CHSY1	CHSY1-CDG	6	6	No
B3GAT3	B3GAT3-CDG	5	1	No
SLC35D1	SLC35D1-CDG	5	7	No
B4GALT7	B4GALT7-CDG	3	3	No
GALNT3	GALNT3-CDG	20	25	Yes
LFNG	LFNG-CDG	1	1	Yes

(continued)

Table 8.1 (continued)

Gene	CDG type new classification (CDG subtype original classification)	Estimated number of cases or prevalence	Number of reported mutations in Human Gene Mutation Database	Molecular diagnostic testing currently available *Analyzed using NGS
B3GALTL	B3GALTL-CDG	>20	9	Yes
POMT1/POMT2	POMT1/POMT2-CDG	1 in 60,500	61/38	Yes*
FKTN	FKTN-CDG	1–9 per 1,000,000	39	Yes*
POMGNT1	POMGNT1-CDG	Unknown	58	Yes*
LARGE	LARGE-CDG	Unknown	12	Yes*
FKRP	FKRP-CDG	Unknown	83	Yes*
Multiple glycosylation defects				
SLC35A1	SLC35A1-CDG (CDG-IIf)	1	1	Yes*
SLC35C1	SLC35C1-CDG (CDG-IIc)	7	5	Yes*
SLC35D1	SLC35D1-CDG	5	7	No
TMEM165	TMEM165-CDG	5	4	No
GNE	GNE-CDG	>100	99	Yes*
ATP6V0A2	ATP6V0A2-CDG	>20	28	Yes*
SEC23B	SEC23B-CDG	>300	59	Yes
B4GALT1	B4GALT1-CDG (CDG-IId)	1	1	Yes*
COG subunit defects				
COG1	COG1-CDG (CDG-IIg)	3	3	Yes*
COG4	COG4-CDG (CDG-IIj)	2	4	No
COG5	COG5-CDG (CDG-Iii)	4	3	No
COG6	COG6-CDG	1	1	No
COG7	COG7-CDG (CDG-IIe)	9	3	Yes*
COG8	COG8-CDG (CDG-IIh)	2	3	Yes*
Lipid defects				
ST3GAL5	ST3GAL5-CDG	8	1	Yes
SIAT9	SIAT9-CDG	8	1	Yes
PIGM	PIGM-CDG	3	1	No
PIGV	PIGV-CDG	8	7	No
PIGA	PIGA-CDG	1 per 500,000	>100	No
PIGL	PIGL-CDG	7	4	No
PIGO	PIGO-CDG	3	3	No

transferrin pattern. Transferrin analysis is relatively straightforward and many laboratories can do this analysis using isoelectric focusing (IEF) [15]. Mass spectrometry (MS) techniques are now being used in the diagnostic setting because they offer greater specificity and sensitivity compared to traditional IEF analysis. If transferrin analysis does not indicate a CDG but the clinical features suggests the patient may have CDG, further biochemical structural analysis of N-glycans and O-glycans can be performed using MS on plasma or serum [16]. This analysis can also provide a clue as to where the defect resides within these pathways due to the characteristic structural patterns. However, structural analysis may only indicate an overall glycosylation defect, and follow-up with molecular genetic analysis will be needed to identify the specific gene defect. There is no single test that can detect all O-linked glycosylation defects, while some tests have been developed to detect certain types of O-linked defects. For O-linked glycosylation defects that involve the biosynthesis of O-glycans, an IEF assay of apoC-III is used [17]. For α-dystroglycanopathies, immunohistochemical staining of a muscle biopsy can be used to detect O-linked mannosylated glycan defects [18]. For lipid-linked defects, flow cytometry can be used to identify GPI anchor deficiencies by specifically analyzing CD59 (protectin) on leukocytes and erythrocytes and CD24 on granulocytes and B cells [11]. Both CD59 and CD24 will be reduced on these cell types if the patient has a GPI anchor defect.

4 Molecular Testing for CDG

Molecular genetic testing has identified CDG disease-causing mutations in more than 60 genes. To date, a total of 17 gene defects have been identified in the N-linked pathway, 17 defects in the O-linked pathway, and 7 defects in the lipid-linked pathway. There are also known combined N-and O-glycosylation defects and defects in other pathways. Classically, CDG nomenclature was alphabetized based on the order of identification of new CDGs. However, with the continuous identification of new types of CDG and defects being identified in new pathways, this nomenclature has recently been updated. The new nomenclature is based on the gene name followed by the suffix -CDG [19]. *This chapter uses the new nomenclature to avoid confusion*. The most common CDG is PMM2-CDG, which is an N-glycosylation defect. More than 700 cases have been identified worldwide and more than 100 different mutations have been identified in the *PMM2* gene [20]. This CDG also has a unique clinical feature of abnormal fat distribution which aids in the diagnosis of this subtype. It is estimated that 60 % of patients with a CDG type I pattern via serum transferrin analysis have a defect in the *PMM2* gene. Only a few families have been identified for the majority of the other CDG subtypes (Table 8.1). Therefore, the natural history and clinical outcome for patients with defects in the majority of CDG subtypes is currently unknown [21]. Treatment is only available for four subtypes with only one subtype MPI-CDG having effective treatment by oral mannose supplementation [22]. This is the only subtype without neurological

involvement, and patients present mainly with hepatic intestinal symptoms. CDG should be considered in any individual presenting with developmental delay or failure to thrive and in individuals that remain with an unknown diagnosis. Even if serum transferrin analysis is normal, CDG must not be ruled out because some subtypes are known to have a normal transferrin pattern including GCS1-CDG, SLC35A1-CDG, and SLC35C1-CDG [1]. Normal transferrin patterns can also be present in adults with CDG due to unknown compensatory mechanisms [23].

The gene defect remains unknown for many patients given a diagnosis of CDG based on clinical features or biochemical testing. Molecular testing for CDG is also not widely available especially for the more rare subtypes, and testing for some CDG-associated genes is only done on a research basis. Given that about 1–2 % (about 200–400 genes) of the genes in the human genome are involved in the process of glycosylation, it is not surprising that there is a high number of unknown cases and it is likely that the majority of these patients will have defects in new genes that are not currently associated with CDG [8].

5 New Technology for Clinical Molecular Testing for CDG and Recent Advancements

Molecular testing has historically relied on the method of Sanger sequencing of individual genes for rare disorders. The strategy of using a stepwise gene-by-gene approach can be very costly and time consuming if there is significant clinical overlap with many candidate genes associated with a certain group of disorders. The delay and uncertainty of the diagnosis also causes significant anxiety for the patient's family if the cause remains unknown. Molecular diagnostic testing by simultaneous analysis of many CDG genes began in 2010 and consisted of Sanger sequencing analysis of individual genes complemented with deletion and duplication analysis using array comparative genomic hybridization (CGH) [24]. Single-gene testing is ideal when a patient either has a clinical phenotype and/or biochemical testing that points to a specific gene defect.

A targeted panel next-generation sequencing (NGS) approach has also been developed for comprehensive analysis of 24 CDG-associated genes [25]. This approach is useful when the clinical phenotype and/or biochemical testing does not provide an indication of which gene defect a patient may have and multiple genes are candidates. NGS technology offers a significant advantage compared to conventional Sanger sequencing because multiple genes can be screened for mutations at the same time leading to a reduced cost and a faster diagnosis if disease-causing mutations can be identified.

This CDG comprehensive panel uses PCR-based enrichment. RainDance™ (Lexington, MA) is used for enrichment for this CDG panel but other PCR-based enrichment options are available including Fluidigm™ (San Francisco, CA) and HaloPlex from Agilent Technologies (Santa Clara, CA). An in-solution hybridization option is also available using Agilent SureSelect™. Limitations of PCR-based

enrichment include poor primer design and difficulty enriching segments with high GC content or sequence complexity which can result in some regions not being amplified. These regions will then have to be covered using an additional optimized amplification condition followed by Sanger sequencing. Genes with pseudogenes are also problematic because target selection technologies (PCR and in-solution hybridization) cannot differentiate between the real gene and the pseudogene. Additional challenges with a targeted NGS panel approach include the need for a strong bioinformatics team that can process the enormous amount of sequencing data once sequencing is completed. Bioinformatic scripts need to be developed to filter through the variants that have low coverage and bad quality because these variants are the least likely to be real. A comprehensive list of all exons that have low coverage or no coverage also needs to be provided and will need to be Sanger sequenced separately. A comprehensive complete test would include every base of every coding exon including the exon/intron boundaries and up to 10 bases into the intron fully analyzed for all genes included in a gene panel.

The CDG NGS panel was validated using 12 CDG patients with mutations in a number of genes identified previously by Sanger sequencing [25]. Validation samples contained different types of mutations including missense, splice site, insertions, and deletions to ensure that NGS technology was sensitive enough to detect a spectrum of different types of mutations. All exons with low or no coverage were determined, and a proactive list was developed. This list includes 15 exons and these exons are Sanger sequenced separately, so a complete comprehensive test is offered to patients [25]. Panel testing has already identified the gene defects in several patients who would likely still remain without a molecular diagnosis if they had been tested using a stepwise gene-by-gene approach. Identifying the mutations in these patients is important for potential future treatment options and for counseling family members with carrier risk information. With the mutations identified, preconception screening is now a possibility, thereby reducing the risk of the parents having another affected child.

Whole exome sequencing (WES) has only very recently debuted in the molecular diagnostic laboratory setting. WES is an extension to the targeted panels by sequencing the majority of genes in the genome. NGS technology is a challenge in the diagnostic setting because laboratories need to be in compliance with both Clinical Laboratory Improvement Amendments (CLIA) and College of American Pathologists (CAP) requirements. However, both CLIA and CAP do not currently have guidelines for NGS in clinical diagnostic laboratories. There is an ongoing effort by experts in the field to determine what guidelines and regulations are needed for this technology. The laboratories also need to determine whether the test is to be performed in-house or off-site with the data delivered to the bioinformaticians in the reporting laboratory. If samples are sent off-site, it also needs to be determined how compliance with the Health Insurance Portability and Accountability Act (HIPAA) can be ensured. Turnaround time and the costs associated with running the test, analysis, Sanger confirmation, and interpretation also need to be considered. Another challenge with offering WES in the diagnostic setting is deciding what types of information from the test should be delivered to the patient and how to generate

reports that are easy for physicians to understand. A disclaimer that lists the regions missed by WES is also important to be included in the reports, especially candidate genes for the patient's phenotype. Pretest and posttest counseling is essential for the patients and their families to make sure that they understand the implications of this testing and how much information the patients want revealed to them about their genetic risk for other conditions independent of the current condition they are being tested for. The psychosocial impact of this test is currently unknown but will be revealed as this testing becomes more widespread. In the molecular diagnostic setting, it is estimated that WES is identifying the gene defect in approximately 20 % of patients that are referred for this testing (unpublished observation).

6 Gene Discovery Approach for CDG

In the research setting, WES has successfully identified the gene defect for a number of disorders. For CDG, all efforts to identify the gene defect in patients without a molecular diagnosis have taken place in the research setting. Researchers involved in CDG have collections of patients who currently do not have a molecular diagnosis, and this number is likely to increase as testing continues in the diagnostic laboratory. Therefore, CDG is an excellent candidate disorder for WES to identify the gene defects in these patients. The majority of glycosylation genes are included in the exome capture design and can be analyzed by WES thereby providing a greater chance of identifying the gene defects in these patients. However, caution needs to be exercised with this approach because genes and exons will be missed including disease-causing mutations, and it is not cost-effective to Sanger sequence all WES missed regions.

In the diagnostic laboratory setting with WES being offered as a diagnostic test, each laboratory will need to decide on a bioinformatics pipeline. Data can be analyzed using two different analysis programs. Variants present in both programs are more likely to be real. One approach for CDG is to generate lists of variants. For example, four lists of variants can be generated: variants that are present in a glycan gene library of 250 genes, all genes in the Human Gene Mutation Database (HGMD) and Online Mendelian Inheritance of Man (OMIM), and all the remaining genes in the exome (Fig. 8.1). The order of analysis will be the glycan gene library, and if no mutations are identified, the focus will turn to all HGMD genes, followed by all OMIM genes and finally all the remaining genes in the exome. An independent map report can be generated for each independent list that provides the percentage of reads that mapped to the exome and the percentage of low-coverage exons. Single-nucleotide changes will be separated into pathogenic and possible pathogenic. Benign variants and variants with a minor allele frequency of 1 % and above will be filtered out as long as the allele frequency data is sufficient. Variants reported at a frequency of less than 1 % will be carefully evaluated. Variants of interest will be confirmed by Sanger sequencing. If a candidate gene is identified, functional studies will commence in the research setting to confirm the functional and pathogenic effects of the variants.

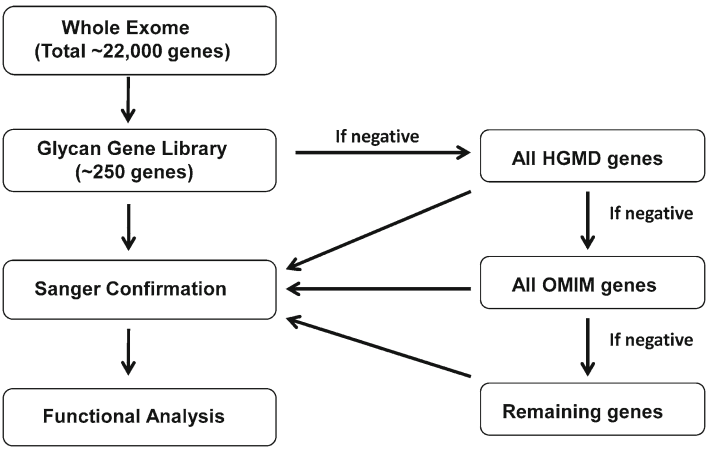

Fig. 8.1 *Whole exome sequencing analysis pipeline*. Whole exome sequencing analysis for CDG first focuses on a glycan gene library of 250 genes. If two nucleotide changes are identified in the same gene that are predicted to result in disease, these will be Sanger confirmed and functional analysis will be performed as appropriate. If the glycan gene library is negative for nucleotide changes that could be causative of disease, analysis will expand out to all HGMD genes, all OMIM genes, and all of the rest of the remaining genes in the exome

With regard to CDG, WES has successfully identified the gene defect in one CDG-Ix patient, a collection of patients given a clinical diagnosis of CHIME syndrome, which stands for colobomas, heart defects, ichthyosiform dermatosis, mental retardation (intellectual disability), and ear anomalies [26]. WES has also identified *PGM1* and *PIGO* mutations in several patients [27, 28]. For the CDG-Ix patient, two mutations were identified in the *DDOST* gene, which is a component of the oligosaccharyltransferase complex (OST). The OST complex is part of the N-glycosylation biosynthesis pathway and is responsible for transferring oligosaccharides to proteins. For the CHIME syndrome cases, mutations were identified in the gene *PIGL*, and for the cases with hyperphosphatasia and mental retardation, mutations were identified in the *PIGO* gene. Both *PIGL* and *PIGO* are part of GPI anchor synthesis pathway for lipid glycosylation. *PIGL* is located at the second step of this pathway and *PIGO* is at step 10 of this pathway [29]. *PGM1* is responsible for both the breakdown and synthesis of glucose.

With the discovery of new CDG genes in the research setting, molecular diagnostic tests can then be set up in the clinical laboratory. For any new genes discovered by WES, adequate evidence is needed before a gene can confidently be associated with a particular disorder, which can include segregation confirmation, functional studies confirming the gene defect, or multiple families identified if possible. With sufficient evidence, a full sequencing test for the newly identified genes can be developed and offered in a diagnostic laboratory. The NGS panels can continually be expanded to include newly discovered genes with confirmed disease association.

7 Variant Interpretation Challenges with Targeted Panels and WES

With NGS panels and WES now being offered in diagnostic laboratories, one significant challenge is the interpretation of detected variants [30]. Online prediction programs are not reliable and cannot be used to come to a final conclusion of whether a variant is benign or pathogenic [31]. The number of variants classified as variant of unknown clinical significance (VOUS) will significantly increase as panel testing and WES increase. Novel variants detected in patients with ethnic backgrounds that are not well represented in dbSNP or the National Heart, Lung, and Blood Institute (NHLBI) exome variant server are difficult to interpret. Since many variants identified using WES are likely not to be previously reported, their effect on protein structure and function and ultimately pathogenicity will be unknown. A conservative approach for variant calling in the diagnostic laboratory is important because family planning can be significantly affected by the interpretation of a variant.

8 The Future of Clinical Molecular Testing for CDG

As new genes are continuously being discovered for CDG, a potential issue to be considered in the near future is whether to refer a patient for CDG panel testing or WES. CDG panels do not include all known CDG genes, whereas WES would allow testing for all known and unknown CDG- and glycosylation-associated genes. However, panel testing allows for complete analysis of all exons for the genes included in the panel, and WES cannot. With this limitation known, it will be important to have a comprehensive list of all exons from genes in the glycan gene library that are not covered, and careful analysis of these genes will be important. If a variant is detected that is likely damaging in a certain gene and several exons of the gene are not covered, Sanger sequencing of the missed exons will be necessary to rule out the gene as the potential cause of the patient's CDG diagnosis. Genes are being discovered faster than the CDG panel can be updated, and it will be impossible to have a panel that includes all known CDG genes. Therefore, if a patient just has panel testing and is negative for nucleotide changes that could be causative for disease, it cannot be ruled out that the patient does not have mutations in known CDG-associated genes that are not currently included in the panel. WES for CDG patients will continue to be necessary because the clinical phenotype and biochemical testing may not be able to indicate the specific gene defect, and there may be no indication of even where to start looking. That is why a full phenotype report will be very important as testing for CDG continues. The patient may have a distinct clinical feature that may indicate a gene defect within a certain pathway or at a certain place within the pathway.

WES will continue to be performed on CDG patients in the research setting because reimbursement issues with WES in the clinical setting may disqualify some patients from having this test. This could change in the future when WES becomes more widespread and accepted by insurance companies. WES will not identify the gene defect in every single CDG patient because some of these patients may have mutations that are outside of the coding regions, which will be missed by a WES approach. It is estimated that 15 % of mutations reside outside of the coding regions and WGS will be necessary to detect these changes [32]. Whole genome sequencing can be used to identify these types of mutations but is not currently feasible in the clinical diagnostic setting. Interpretation of variants outside of the coding regions would be very difficult without functional studies from research laboratories providing evidence that the variants impair gene function. Therefore, clinical laboratories will continue to focus on the exome and on gene panels in the immediate future because variants identified are easier to interpret.

Additional pathways involved in glycosylation may be discovered by using WES. Thus, WES can provide a greater understanding of glycosylation and what phenotypes result in humans when defects are present in these pathways. This in turn will result in research for potential new therapies for defects within these pathways.

9 The Future of Translational NGS for CDG

Identification of the causative genes in CDG patients using NGS technology is a big accomplishment because it is likely that single-gene testing alone would not have picked up the mutations. Individual gene testing usually does not continue when a couple of genes screened are negative for mutations, leaving the patients and their families without an answer. As the causative genes for CDG patients are identified, a better indication on the prognosis for the patient will be possible. As more patients are identified with defects in the same gene, an indication of whether they could possibly develop additional clinical features and organ dysfunction as they age could be predicted. Although an effective treatment only exists for one subtype of CDG and partial treatment is available for a few others, identification of the defective gene in patients can provide the foundation for the investigation of new treatment options in the research setting. Investigation of the effect of potential compounds on patient fibroblasts and the resulting effect on glycosylation can be explored [33, 34]. Recently, it has been discovered that oral mannose is actually helping patients with ALG1-CDG (Miao He, personal communication). This effect was first observed in patient's fibroblasts, and treatment with one severely affected child improved his or her ambulation. Mutations in *ALG1* were identified in the diagnostic laboratory and further investigation of treatment took place in the research setting. Now the treatment is going from bench to bedside, and it is possible that oral mannose may be a partially effective treatment for patients with ALG1-CDG. Recently, it has been discovered that patients with *PGM1* defects have

improvement when they drink 5–6 glasses of milk per day (Morava E, personal communication). Discoveries made with NGS technology will improve the molecular diagnosis rate, patient care and management and will provide opportunities for the research laboratories to investigate possible treatments for CDG that can be brought to the patient's bedside.

References

1. Freeze HH (2006) Genetic defects in the human glycome. Nat Rev Genet 7(7):537–551. doi:nrg1894 [pii] 10.1038/nrg1894
2. Apweiler R, Hermjakob H, Sharon N (1999) On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. Biochim Biophys Acta 1473(1):4–8. doi:S0304-4165(99)00165-8 [pii]
3. Peter-Katalinic J (2005) Methods in enzymology: O-glycosylation of proteins. Methods Enzymol 405:139–171. doi:S0076-6879(05)05007-X [pii] 10.1016/S0076-6879(05)05007-X
4. Freeze HH, Eklund EA, Ng BG, Patterson MC (2012) Neurology of inherited glycosylation disorders. Lancet Neurol 11(5):453–466. doi:S1474-4422(12)70040-6 [pii] 10.1016/S1474-4422(12)70040-6
5. Van den Steen P, Rudd PM, Dwek RA, Opdenakker G (1998) Concepts and principles of O-linked glycosylation. Crit Rev Biochem Mol Biol 33(3):151–208. doi:10.1080/10409239891204198
6. Hancock JF (2004) GPI-anchor synthesis: ras takes charge. Dev Cell 6(6):743–745. doi:10.1016/j.devcel.2004.05.011 S153458070400173X [pii]
7. Jaeken J (2011) Congenital disorders of glycosylation (CDG): it's (nearly) all in it! J Inherit Metab Dis 34(4):853–858. doi:10.1007/s10545-011-9299-3
8. Schachter H, Freeze HH (2009) Glycosylation diseases: quo vadis? Biochim Biophys Acta 1792(9):925–930. doi:S0925-4439(08)00227-5 [pii] 10.1016/j.bbadis.2008.11.002
9. Arnoux JB, Boddaert N, Valayannopoulos V, Romano S, Bahi-Buisson N, Desguerre I, de Keyzer Y, Munnich A, Brunelle F, Seta N, Dautzenberg MD, de Lonlay P (2008) Risk assessment of acute vascular events in congenital disorder of glycosylation type Ia. Mol Genet Metab 93(4):444–449. doi:S1096-7192(07)00602-6 [pii] 10.1016/j.ymgme.2007.11.006
10. Hewitt JE (2009) Abnormal glycosylation of dystroglycan in human genetic disease. Biochim Biophys Acta 1792(9):853–861. doi:S0925-4439(09)00134-3 [pii] 10.1016/j.bbadis.2009.06.003
11. Almeida AM, Murakami Y, Layton DM, Hillmen P, Sellick GS, Maeda Y, Richards S, Patterson S, Kotsianidis I, Mollica L, Crawford DH, Baker A, Ferguson M, Roberts I, Houlston R, Kinoshita T, Karadimitris A (2006) Hypomorphic promoter mutation in PIGM causes inherited glycosylphosphatidylinositol deficiency. Nat Med 12(7):846–851. doi:nm1410 [pii] 10.1038/nm1410
12. Maydan G, Noyman I, Har-Zahav A, Neriah ZB, Pasmanik-Chor M, Yeheskel A, Albin-Kaplanski A, Maya I, Magal N, Birk E, Simon AJ, Halevy A, Rechavi G, Shohat M, Straussberg R, Basel-Vanagaite L (2011) Multiple congenital anomalies-hypotonia-seizures syndrome is caused by a mutation in PIGN. J Med Genet 48(6):383–389. doi:jmg.2010.087114 [pii] 10.1136/jmg.2010.087114
13. Krawitz PM, Schweiger MR, Rodelsperger C, Marcelis C, Kolsch U, Meisel C, Stephani F, Kinoshita T, Murakami Y, Bauer S, Isau M, Fischer A, Dahl A, Kerick M, Hecht J, Kohler S, Jager M, Grunhagen J, de Condor BJ, Doelken S, Brunner HG, Meinecke P, Passarge E, Thompson MD, Cole DE, Horn D, Roscioli T, Mundlos S, Robinson PN (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. Nat Genet 42(10):827–829. doi:ng.653 [pii] 10.1038/ng.653

14. Lefeber DJ, Morava E, Jaeken J (2011) How to find and diagnose a CDG due to defective N-glycosylation. J Inherit Metab Dis 34(4):849–852. doi:10.1007/s10545-011-9370-0
15. Marklova E, Albahri Z (2007) Screening and diagnosis of congenital disorders of glycosylation. Clin Chim Acta 385(1–2):6–20. doi:S0009-8981(07)00369-5 [pii] 10.1016/j.cca.2007.07.002
16. Faid V, Chirat F, Seta N, Foulquier F, Morelle W (2007) A rapid mass spectrometric strategy for the characterization of N- and O-glycan chains in the diagnosis of defects in glycan biosynthesis. Proteomics 7(11):1800–1813. doi:10.1002/pmic.200600977
17. Wopereis S, Grunewald S, Morava E, Penzien JM, Briones P, Garcia-Silva MT, Demacker PN, Huijben KM, Wevers RA (2003) Apolipoprotein C-III isofocusing in the diagnosis of genetic defects in O-glycan biosynthesis. Clin Chem 49(11):1839–1845
18. Muntoni F, Torelli S, Wells DJ, Brown SC (2011) Muscular dystrophies due to glycosylation defects: diagnosis and therapeutic strategies. Curr Opin Neurol 24(5):437–442. doi:10.1097/WCO.0b013e32834a95e3
19. Jaeken J, Hennet T, Freeze HH, Matthijs G (2008) On the nomenclature of congenital disorders of glycosylation (CDG). J Inherit Metab Dis 31(6):669–672. doi:10.1007/s10545-008-0983-x
20. Haeuptle MA, Hennet T (2009) Congenital disorders of glycosylation: an update on defects affecting the biosynthesis of dolichol-linked oligosaccharides. Hum Mutat 30(12):1628–1641. doi:10.1002/humu.21126
21. Vodopiutz J, Bodamer OA (2008) Congenital disorders of glycosylation-a challenging group of IEMs. J Inherit Metab Dis. doi:10.1007/s10545-008-0849-2
22. Jaeken J (2010) Congenital disorders of glycosylation. Ann N Y Acad Sci 1214:190–198. doi:10.1111/j.1749-6632.2010.05840.x
23. Vermeer S, Kremer HP, Leijten QH, Scheffer H, Matthijs G, Wevers RA, Knoers NA, Morava E, Lefeber DJ (2007) Cerebellar ataxia and congenital disorder of glycosylation Ia (CDG-Ia) with normal routine CDG screening. J Neurol 254(10):1356–1358. doi:10.1007/s00415-007-0546-3
24. Tayeh MK, Chin EL, Miller VR, Bean LJ, Coffee B, Hegde M (2009) Targeted comparative genomic hybridization array for the detection of single- and multiexon gene deletions and duplications. Genet Med 11(4):232–240. doi:10.1097/GIM.0b013e318195e191
25. Jones MA, Bhide S, Chin E, Ng BG, Rhodenizer D, Zhang VW, Sun JJ, Tanner A, Freeze HH, Hegde MR (2011) Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. Genet Med. doi:10.1097/GIM.0b013e318226fbf2
26. Ng BG, Hackmann K, Jones MA, Eroshkin AM, He P, Wiliams R, Bhide S, Cantagrel V, Gleeson JG, Paller AS, Schnur RE, Tinschert S, Zunich J, Hegde MR, Freeze HH (2012) Mutations in the glycosylphosphatidylinositol gene PIGL cause CHIME syndrome. Am J Hum Genet 90(4):685–688. doi:S0002-9297(12)00095-X [pii] 10.1016/j.ajhg.2012.02.010
27. Timal S, Hoischen A, Lehle L, Adamowicz M, Huijben K, Sykut-Cegielska J, Paprocka J, Jamroz E, van Spronsen FJ, Korner C, Gilissen C, Rodenburg RJ, Eidhof I, Van den Heuvel L, Thiel C, Wevers RA, Morava E, Veltman J, Lefeber DJ (2012) Gene identification in the congenital disorders of glycosylation type I by whole-exome sequencing. Hum Mol Genet. doi:dds123 [pii] 10.1093/hmg/dds123
28. Krawitz PM, Murakami Y, Hecht J, Kruger U, Holder SE, Mortier GR, Delle Chiaie B, De Baere E, Thompson MD, Roscioli T, Kielbasa S, Kinoshita T, Mundlos S, Robinson PN, Horn D (2012) Mutations in PIGO, a member of the GPI-anchor-synthesis pathway, cause hyperphosphatasia with mental retardation. Am J Hum Genet 91(1):146–151. doi:S0002-9297(12)00260-1 [pii] 10.1016/j.ajhg.2012.05.004
29. Fujita M, Kinoshita T (2010) Structural remodeling of GPI anchors during biosynthesis and after attachment to proteins. FEBS Lett 584(9):1670–1677. doi:doi:S0014-5793(09)00871-0 [pii] 10.1016/j.febslet.2009.10.079

30. Klee EW, Hoppman-Chaney NL, Ferber MJ (2011) Expanding DNA diagnostic panel testing: is more better? Expert Rev Mol Diagn 11(7):703–709. doi:10.1586/erm.11.58
31. Tchernitchko D, Goossens M, Wajcman H (2004) In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. Clin Chem 50(11):1974–1978. doi:doi:50/11/1974 [pii] 10.1373/clinchem.2004.036053
32. Raffan E, Semple RK (2011) Next generation sequencing–implications for clinical practice. Br Med Bull 99:53–71. doi:ldr029 [pii] 10.1093/bmb/ldr029
33. Losfeld ME, Soncin F, Ng BG, Singec I, Freeze HH (2012) A sensitive green fluorescent protein biomarker of N-glycosylation site occupancy. FASEB J. doi:fj.12-211656 [pii] 10.1096/fj.12-211656
34. He P, Ng BG, Losfeld ME, Zhu W, Freeze HH (2012) Identification of intercellular cell adhesion molecule 1 (ICAM-1) as a hypoglycosylation marker in congenital disorders of glycosylation cells. J Biol Chem 287(22):18210–18217. doi:M112.355677 [pii] 10.1074/jbc.M112.355677

Chapter 9
NGS Improves the Diagnosis of X-Linked Intellectual Disability (XLID)

Michael J. Friez and Monica J. Basehore

Abstract X-linked intellectual disability (XLID) is considered to be a collection of conditions that are each caused by mutation in one of the many X-linked genes associated with either a syndromic or nonsyndromic form of intellectual disability. A significant number of XLID conditions have been described, but only approximately 50 % of the causative XLID genes have been discovered. For affected individuals from families with clear or potentially X-linked inheritance, the strategy for next-generation sequencing (NGS) can be tailored appropriately, given the ability to focus sole attention on the X chromosome rather than the entire genome. The primary goal of this chapter is to focus on the principles associated with testing known XLID genes that have been included on various targeted NGS panels. These principles can be extended to other X-linked genes that may be implicated in XLID, as well as to other genes on the X chromosome with relevant medical implications.

1 Introduction

Of the two sex chromosomes, the X chromosome is much larger and more gene dense, and is well known for having many genes associated with a wide variety of clinical conditions. X-linked conditions typically affect males to a greater degree than females given that males have just a single X chromosome (half the genomic dosage of normal females). A sole X chromosome in males leaves this region one of the most vulnerable sectors of the genome; therefore, it is much more susceptible to disease upon mutation, in comparison to females that normally have two X chromosomes [1]. This chapter focuses on X-linked intellectual disability (XLID), previously referred to as X-linked mental retardation (XLMR), as a group of conditions

M.J. Friez, Ph.D., FACMG (✉) • M.J. Basehore, Ph.D., FACMG
Greenwood Genetic Center, 106 Gregor Mendel Circle, Greenwood, SC 29646, USA
e-mail: friez@ggc.org; mbasehore@ggc.org

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_9,

caused by a specific region of the genome that present in either X-linked recessive or X-linked dominant fashion. Taken together, XLID is considered to be a collection of conditions that are individually caused by mutation in one of the many X-linked genes that have been associated with either a syndromic or a nonsyndromic form of intellectual disability (ID). The diagnosis of nonsyndromic ID is typically reserved for those individuals with a nonspecific presentation and no distinguishing features other than ID. In contrast, syndromic ID presents with a variety of additional physical manifestations that often allow different syndromes to be clinically distinct and recognizable by those with proper medical expertise. A particular gene can be associated with both syndromic and nonsyndromic clinical presentations, and for some syndromes, the phenotype is more readily apparent in females due to various factors including lethality in affected males. Additionally, it is clear that different clinical spectrums related to the same gene or condition do occur, which contributes to the extended diagnostic odyssey many families endure before reaching a definitive diagnosis. For affected individuals from families with clear or potentially X-linked inheritance, the strategy for next-generation sequencing (NGS) can be tailored appropriately, given the ability to focus sole attention on the X chromosome rather than the entire genome. Focusing on the X chromosome can be accomplished, as with other NGS applications, at three different levels which include (1) targeted re-sequencing of select genes, (2) re-sequencing of all X-linked genes (X-exome), and (3) genomic re-sequencing of the entire X chromosome.

2 Overview of XLID

2.1 XLID Genes

Intellectual disability (ID) is characterized by significant limitations in intellectual/cognitive function and adaptive behavior with onset prior to 18 years of age [2]. ID is a common cause for referral to a number of clinical specialists including developmental pediatricians and medical geneticists. A broad spectrum of factors play a role in the etiology of ID with underlying genetic disorders being the most frequently identified cause. However, a significant proportion of individuals affected with ID do not have an identifiable cause, and new technologies, including NGS in particular, are expected to profoundly reduce this proportion in today's reality. From a genetic perspective, there are many constitutional abnormalities involved, ranging from common aneuploidies to large copy number variations to single-gene disorders [3–5]. The identification of mutations in X-linked genes has been the most productive area of investigation of ID given the number of families, either proven or assumed, to fall into the clinical spectrum of XLID. In relative terms, XLID genes have been easier to discover than their autosomal counterparts due to the strategies and technologies traditionally utilized. Many of the genes included under the XLID

umbrella are associated with distinguishing features that are recognizable to varying degrees; however, there are also rare XLID conditions that are much more difficult to diagnose clinically, regardless of the clinician's expertise or setting involved. In either case, the cumulative number of conditions and families that map to the X chromosome is substantial, making XLID a relatively common diagnosis among those affected with ID.

Over the last two decades, new XLID genes have been routinely identified, and even though the list of genes is impressive, approximately half of them still remain undiscovered based on estimates from research programs with a substantial number of undiagnosed XLID families. Many of the genes on the X chromosome that encode proteins with recognized functions have not been associated with a specific or nonsyndromic form of ID but continue to serve as viable XLID candidates based on their cellular role. For now, it is assumed that most of the common forms of XLID have been identified, but additional discoveries are expected to yield a more profound understanding of XLID and should be anticipated given current NGS capabilities. The primary goal of this chapter is to focus on the principles associated with testing known XLID genes that have been included in various targeted NGS panels. These principles certainly can be extended to other X-linked genes that have the potential to be implicated in XLID, as well as to all of the other genes on the X chromosome with relevant medical implications.

The *Atlas of X-Linked Intellectual Disability Syndromes* textbook is the most recent and comprehensive review of current XLID conditions [6]. In this text, 156 individual conditions/gene-specific spectrums are described, but interestingly, only 102 XLID genes have been identified to date. The number of XLID conditions/genes is expected to grow as certain phenotypes that have been traditionally lumped together become disentangled along with the significant number of new nonsyndromic presentations that are likely due to mutations in genes not yet implicated in XLID. New candidate genes continuously come to attention but often remain tentative disease-causing entities since they require further scrutiny before being accepted as valid disease-causing XLID genes. Often, there is not a direct correlation between conditions and genes, making the classification of many XLID entities challenging. Complicating matters even further is the issue that a few of the previously reported XLID genes may, in fact, not be disease causing at all. Additionally, one gene can be associated with multiple clinical presentations that have overlapping features, yet each exists as a distinct clinical presentation. Take, for example, *ARX*-associated XLID whereby ten phenotypes are collectively grouped. The phenotypes include subgroups for seizures, brain malformations, dystonia, as well as nonsyndromic presentation that accounts for more than one-third of all *ARX*-related cases. This chapter will not attempt to provide fully detailed phenotypic information for the multitude of XLID conditions as the clinically oriented reports are better suited for this purpose. For the conditions not yet associated with a specific gene, there remains the distinct possibility that the causative alteration is cryptically located within an already-recognized XLID gene (See Appendix: Profiles of the Most Common XLID Syndromes).

2.2 The Most Common XLID Syndromes and Limitations of NGS-Based Testing for XLID

By far the most common form of XLID (and inherited ID in general) is Fragile X syndrome. The clinical features and molecular etiology of this condition have been well described given the reported frequency of 1 in 4,000 males [7]. In most clinical settings, Fragile X testing is commonly pursued when the diagnosis involves ID and/or autism regardless of the gender of the patient, although the clinical manifestations are often not as apparent in females. The condition is caused by expansion of a CGG trinucleotide repeat in the 5′ untranslated region (UTR) of the *FMR1* gene in greater than 99 % of cases and occurs when either a full mutation (>200 CGG repeats) or an unstable premutation allele (55–200 CGG repeats) that has expanded into the full mutation range is inherited from a carrier mother. Expansions of greater than 200 CGG repeats typically lead to hypermethylation of the promoter region of *FMR1*, which in turn plays a key role in repressing expression of this gene. Expansion of premutation alleles occurs only when the allele is transmitted maternally, with larger premutations having the greater probability for expansion. This means that all female offspring of a premutation carrier male will be obligate carriers of the premutation, with this allele having a given propensity for expanding into the full mutation range in each of the daughter's prospective pregnancies. The likelihood for expansion to the full mutation range correlates reasonably well with repeat number in that larger premutations have much higher probabilities of expansion, in comparison to smaller alleles in the premutation size range.

For now, trinucleotide repeat expansions are resistant to detection with most current NGS platforms given the read length needed to fully capture full mutations. Even normal alleles (<55 CGG repeats) are difficult to Sanger sequence in most laboratories due to the technical difficulties associated with repetitive sequences, especially when they have high percentages of GC content. This brief discussion of Fragile X syndrome is made to demonstrate the concern that a common ID condition will routinely be missed by nearly all NGS platforms given the sequence context and read length needed for detection.

A second example of this involves the *ARX* gene, and although the technical difficulties are similar, mutations in this gene are not expansion mutations like those in Fragile X syndrome. For *ARX*, the most frequently reported mutation is a 24-base pair duplication (c.431-454dup24) that elongates a polyalanine tract and is suspected to be the second most common X-linked alteration associated with ID [8]. This region of *ARX* (exon 2) is especially GC rich, and other duplications of similar size have also been reported, including another recurrent mutation consisting of 10 GCG repeats that expands to 17 repeats which also results in the elongation of a polyalanine tract. Limitations such as these involving repetitive and GC-rich regions need to be recognized and accommodated for, given that they serve as impediments to detection by the most current NGS platforms.

PCR-based methods supplemented with Southern blotting continues to be the gold standard for Fragile X testing, and traditional Sanger sequencing needs to be

performed to detect mutations in regions similar to exon 2 of *ARX*. These additional analyses should be performed automatically, in addition to NGS-based testing, in order to have sufficient confidence that the relatively common XLID mutations will be detected. These examples serve as cautionary tales for other disease-related regions of the genome that may be problematic for NGS to correctly analyze.

2.3 The Application of NGS to Molecular Diagnosis of XLID

2.3.1 Diagnostic Strategies: When to Order the NGS-Based XLID Test

The packaging of the known XLID genes into targeted panels was one of the first NGS testing options to be offered diagnostically, and several laboratories in the USA now offer similar panels for clinical purposes. This appears to be an excellent approach for testing affected males or obligate carrier females when there is reasonable justification to focus attention on the X chromosome. This typically occurs when multiple males with ID are present in a pedigree or when only two affected males in a family are recognized and appear to have matching phenotypes and inheritance compatible with being X-linked. The recommended strategy for testing includes ruling out Fragile X syndrome as well as constitutional cytogenetic abnormalities such as aneuploidies and deletions/duplications by traditional cytogenetic and newer array-based testing methods. Additionally, if a specific X-linked syndrome is suspected, it is recommended that single-gene testing be considered prior to proceeding with a comprehensive NGS panel. The decision to pursue single-gene testing is often based on the confidence the referring clinician has in their diagnosis for the patient in question. For example, if a clinician suspects ATRX syndrome with high probability, based on their exam of the patient along with other supportive information, it may be best to target the *ATRX* gene first before moving on to a targeted NGS panel. On the other hand, if the clinician believes the diagnosis of ATRX syndrome is possible, but not necessarily likely, it may be better to utilize a targeted NGS panel from the onset. Consideration needs to be given to the cost and time involved in either approach, and the decision will vary depending on the family, the experience of the clinical team involved, and the perceived degree of urgency for obtaining the test results.

2.3.2 Reasons for Negative Results

Targeted XLID NGS panels are capable of rapidly identifying the disease-causing mutation when the alteration is obviously pathogenic, and this routinely happens in some probands that are tested. Although not specifically documented, there appears to be a direct correlation between larger XLID families and the probability of identifying a mutation that is ultimately deemed pathogenic. However, based on our experience with XLID panel testing, the majority of probands submitted do

not have pathogenic mutations identified, and there are at least several explanations for this. The most straightforward explanation is that many of the probands submitted for testing do not come from X-linked families with multiple affected males. In some cases, clinicians decide to order a targeted XLID panel as a reasonable approach even when there is little certainty about the inheritance pattern. Secondly, in many cases there may be one or more related males with ID in a family (in some cases, distantly related to the proband), but the ID in the family members could be X-linked, while the ID in the proband could be the result of another unrelated cause. With ID being a common condition in all populations and ethnicities, it is not uncommon to find more than one affected male in a typical three-generation pedigree with all affected males clearly not having the same condition. For these, the inheritance pattern is often not X-linked, thus eliminating the possibility of unifying their diagnosis in that regard. Another fraction of cases come from what appear to be sporadically affected males, which may or may not have extenuating circumstances such as recurrent pregnancy loss documented in the family, suggesting the possibility of an X-linked etiology. Lastly, we recognize that targeted NGS panels for XLID are still a relatively new diagnostic option and that many affected individuals have already had their disease-causing mutation identified by single-gene sequencing, particularly for those with more clinically obvious syndromes.

2.3.3 Subsequent Segregation Analyses and Additional Family Member Testing

If a clearly pathogenic alteration is not identified by targeted panel analysis, it does not immediately translate to the notion that further options do not remain. Most males tested continue to have one or more novel X-linked variants reported which require further attention to fully delineate their clinical significance via segregation testing in appropriate family members. The principles of X-linked inheritance make this an easy task to accomplish in large families with a definite X-linked pattern of inheritance, but unfortunately, it leaves relatively small families with very limited options. The first priority is to confirm that any variants in question are maternally inherited. As expected, most X-linked changes are inherited, but occasionally an alteration will be de novo, making it highly probable that it is the causative mutation. This can also be used as additional circumstantial evidence for pathogenicity in genes with no preexisting disease associations. The cautionary note in this regard is that loss of function of approximately 1 % of the genes on the X chromosome is compatible with apparently normal existence [9].

Subsequent or concurrent to confirming maternal carrier status is the necessity to evaluate the genotype of appropriate maternally related males. The objective is to identify an alteration that segregates appropriately in affected males and obligate carrier females and does not appear in phenotypically normal males. The presence of a variant of unknown clinical significance in a male with normal development is considered strong evidence that the change is nonpathogenic. In some families, the

identification of a novel variant that segregates with the disease-causing mutation is possible, even though the variant is ultimately shown to be benign by functional analysis. This occurs when a novel variant is in linkage disequilibrium with the true pathogenic mutation that remains anonymous. For example, a novel intronic substitution may be reported with subsequent follow-up testing demonstrating appropriate segregation in the available family members. However, testing of the mRNA transcript appears to be normal, indicating that the intronic alteration in question does not affect normal splicing and is not likely pathogenic (even though the caveat of abnormal tissue-specific splicing exists). For families enrolled in XLID research programs and for which linkage analysis/localization has been performed, novel variants outside the linkage interval can be excluded from further consideration or de-prioritized, assuming there is confidence in the defined localization interval.

2.3.4 X-Linked Dominant XLID

It is important to keep in mind that not all XLID conditions express themselves in X-linked recessive fashion. A small group of XLID conditions occur exclusively, or nearly so, in females, and are considered as X-linked dominant entities. One of the most well-known examples of this is Rett syndrome (OMIM#312750) which is caused by mutation of the *MECP2* gene (OMIM#300005) located at Xq28 [10]. Almost all pathogenic mutations in females occur spontaneously, with the overwhelming majority being found on the paternally inherited X chromosome. This phenomenon, to a large extent, explains why females are more readily diagnosed with this relatively common condition. Another important component to the story of Rett syndrome is the severity of the condition, which is expected to lead to early lethality in nearly all affected males. Surviving males are extremely rare and often have some additional underlying genetic explanation such as mosaicism of the pathogenic *MECP2* mutation or an additional X chromosome (Klinefelter syndrome) that offsets the predicted phenotypic severity.

Several other examples similar to Rett syndrome exist, but perhaps the most interesting of these is the condition known as epilepsy-intellectual disability limited to females (OMIM#300088). This condition, also referred to as Juberg-Hellman syndrome or early infantile epileptic encephalopathy 9 (EIEE9), is caused by mutations in the *PCDH19* gene (OMIM#300460). Only females are affected with this condition whereby males are left unaffected allowing them to serve as obligate carriers. The mechanism for this intriguing situation has been described and likely involves rescue in males due to the involvement of the *PCDH11Y* homolog (OMIM#400022) located on the Y chromosome [11]. An alternative hypothesis proposes that females have compromised or scrambled cell-to-cell communication due to PCDH19-negative and PCDH19 wild-type tissue mosaicism. One plausible explanation for females who are spared from the effects of a pathogenic *PCDH19* mutation is skewed X-inactivation that preferentially inactivates either chromosome.

2.3.5 Detecting Hemizygous Alterations in Males

In similar fashion to traditional Sanger sequencing, it is easier for NGS to identify X-linked hemizygous changes in males with confidence, in comparison to heterozygous autosomal substitutions in males. This is due in part to the weighted bias towards the abnormal nucleotide frequency for genuine alterations in the hemizygous state. For females, evaluating X-linked alterations is no different than reviewing autosomal data, especially when any given change appears to be heterozygous. When variants have low coverage and/or equally weighted allele frequencies, it is more difficult to judge the correct call; therefore, it is still considered standard practice by diagnostic laboratories, at least for the time being, to Sanger confirm all reported alterations. However, for those working in a discovery-based setting, it is often not practical to confirm every potential alteration by Sanger sequencing, so different algorithms are typically used to prioritize candidate changes based on their rank.

Another advantage that sequencing of X-linked genes in males provides is the more immediate recognition of missing content due to deletions. For males with moderate-sized deletions that might be overlooked by array-based technology (depending on design and level of array resolution), the lack of sequence content for a region expected to be covered by NGS is a strong indicator that a deletion may be present. This is especially true when the missing content maps to consecutive exons or a series of contiguous genes when all other regions are covered as expected. In these cases, further confirmation studies are recommended to ensure that some other underlying technicality is not involved. It is also noted that it may be technically possible to utilize NGS data to infer genomic dosage, but this capability is not yet reliable for diagnostic purposes. This necessitates the need for additional, more quantitative, methods to continue to be performed, especially for females.

3 Representative Scenarios: Application of NGS-Based Testing for XLID

In this section, we present three hypothetical XLID testing scenarios: (1) the case of a sporadic male with ID, (2) a male with ID and clear X-linked inheritance, and (3) a male suspected of having an XLID-related condition. These representative examples will emphasize the concepts presented in the preceding section. The initial testing approach is effectively the same for all three situations and should involve either NGS to simultaneously evaluate many genes or Sanger sequencing if a specific phenotype is present that justifies prioritizing a particular syndrome. As mentioned earlier, the current NGS options include targeted gene panels, whole exome sequencing, or whole genome sequencing. The approach taken for follow-up family member testing will vary depending on the scenario at hand. However, in most instances, follow-up testing of additional family members only requires Sanger sequencing of the identified changes of interest. Given the fact that most pathogenic

mutations and virtually all nonpathogenic alterations are inherited, it is more or less assumed that, for males, every change detected is maternally inherited until proven otherwise.

Pathogenic alterations are those that are expected to cause disease and include most frameshift, nonsense, and consensus splicing alterations and also include other changes, such as missense alterations, that have been previously proven to be pathogenic for one reason or another [12]. An exception is the approximate 1 % of the genes on the X chromosome that are compatible with apparently normal existence, even when there is a loss-of-function mutation [9]. Variants of unknown clinical significance are those in which there is not enough evidence to deem them pathogenic at the time of detection. Publically available databases, such as the Human Gene Mutation Database (HGMD; http://www.hgmd.cf.ac.uk/ac/index.php) and National Center for Biotechnology Information SNP database (dbSNP; http://www.ncbi.nlm.nih.gov/snp/), are important in aiding in the determination of whether an alteration should be considered pathogenic or not. Newer databases which also play key roles in referencing human variation include the 1000 Genomes database (http://browser.1000genomes.org/index.html) and the Exome Variant Server (http://evs.gs.washington.edu/EVS/). Although these databases are informative, it is important to note that they are not always accurate or completely up to date, so one must proceed with caution and reason and not depend on these databases alone to make the final interpretation. Additionally, it is frequent that more than one unclassified variant is identified in the proband, and these databases can assist with assigning priority for follow-up testing. Here, we describe each scenario and present the possible testing options in the event of identifying a pathogenic alteration or a variant of unknown clinical significance.

3.1 Scenario 1: Sporadic Male with Intellectual Disability

In the case of the sporadic male with ID (Fig. 9.1; individual III-3), one has to give strong consideration to the fact that the disability in this individual may not be a result of an X-linked mutation. However, given that the only affected individual in the family is male, it is still worthwhile to investigate the possibility that a mutation in one of the known XLID genes is involved. If a clearly pathogenic alteration is detected, then the cause of the ID in the proband has been identified, and the next step would be to test the proband's mother (individual II-4) to determine if the change is de novo or inherited. If de novo, testing should be considered complete, although there is some residual risk of germline mosaicism in the mother. On the other hand, if the pathogenic mutation is maternally inherited, then other at-risk family members (individuals II-6, II-7, and III-2) may wish to be tested to determine their carrier status.

If a variant of unknown clinical significance is detected, then additional family member testing is necessary in order to determine the significance of the alteration. Testing the proband's mother should be performed first. Demonstrating that a

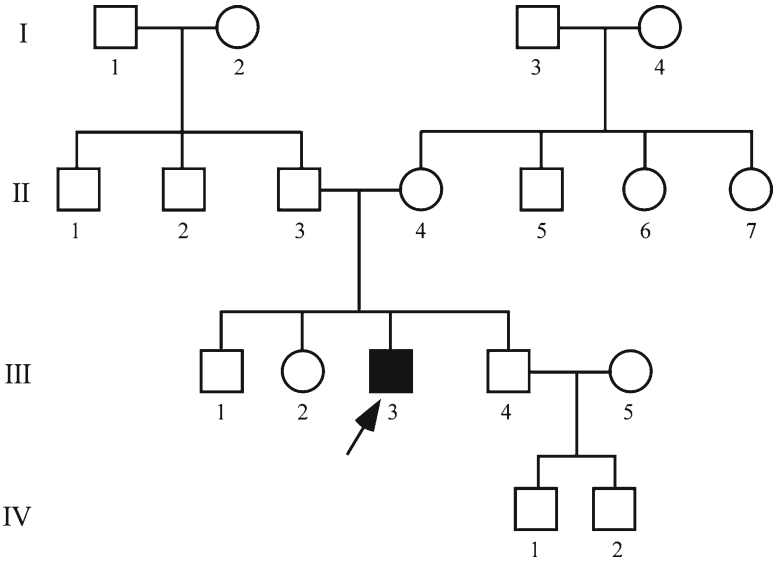

Fig. 9.1 Sporadic male with intellectual disability

change is de novo lends support to the notion that it is disease-causing, but it does not definitively prove its pathogenicity. Additional *in silico* analyses such as the SIFT [13] and PolyPhen2 [14] algorithms are useful in attempting to predict if a particular missense change will have a damaging effect on the protein. The finding of a de novo alteration that is predicted to be deleterious by multiple prediction algorithms is strong collective evidence that a change is pathogenic. However, without functional studies, which are usually not feasible for a diagnostic laboratory, pathogenicity is not absolute.

When a variant of unknown clinical significance is found to be inherited, clinical correlation and additional testing of appropriate maternally related males (individuals II-5, III-1, III-4) are essential to the interpretation of the variant, given that the presence of the variant in a phenotypically normal male relative would provide strong evidence that the change is benign and not clinically relevant. In the case of a sporadic male, there are no other affected males to test in the family, so segregation analysis is not an option and is considered, at best, circumstantial. X-inactivation (XI) studies of the proband's mother is an additional testing option and should be considered informative if she is found to have a highly skewed XI pattern. Obligate carrier females in XLID families commonly have skewed XI identified, and this lends support to the notion that a pathogenic alteration is present on the X chromosome that is preferentially inactivated. At this time, testing for deletions or duplications in XLID genes or mutations in other autosomal intellectual disability genes should be considered. If the phenotype of the patient is not at all compatible with the predicted clinical features of the gene involved, these other testing options become even more attraction options.

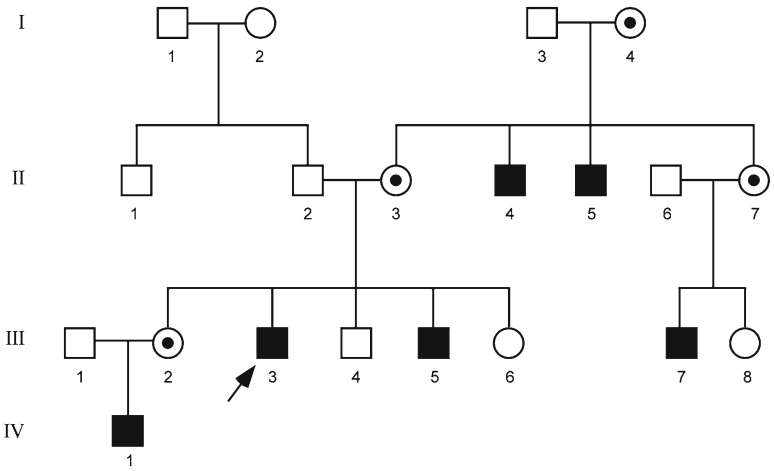

Fig. 9.2 Male with intellectual disability and X-linked inheritance

3.2 Scenario 2: Male with Intellectual Disability and X-Linked Inheritance

In contrast to the case of the sporadic male, one can assume that the ID in a male with a definite X-linked pattern of inheritance (Fig. 9.2; individual III-3) is a result of a mutation on the X chromosome. This also assumes that the proband has a phenotype resembling the other affected males in the family. Testing in this scenario is more straightforward since there are a more discreet number of genes with potential involvement. If an obvious pathogenic alteration is detected, then the cause of the ID in the proband has been identified. The proband's mother (individual II-3) should be tested to prove that she is, as expected, an obligate carrier of the mutation. It is also strongly recommended that additional male family members, affected and unaffected, be tested to demonstrate segregation (affected males II-4, II-5, III-5, III-7, IV-1; unaffected male and III-4). At-risk female family members (individuals III-6, III-8) may also wish to have testing performed to determine their carrier status. Obligate carrier females (individuals I-4, II-7, III-2) might also request to have testing done for confirmation purposes.

If a variant of unknown clinical significance is detected in a male with ID and an X-linked pattern of inheritance, then additional family member testing becomes an essential component in the interpretation of the significance of the alteration. If the variant is inherited, clinical correlation and segregation analysis are imperative. If the variant segregates appropriately in the family (none of the unaffected males carry the variant but all of the affected males and obligate carrier females do), it would greatly support the notion that the variant is, in fact, disease causing. This information, along with deleterious predictions from multiple *in silico* analysis programs, would be strong evidence that the change is pathogenic. However, it does

not completely prove pathogenicity given the presumed pathogenic mutation segregating in the family could be in linkage disequilibrium with the true disease-causing mutation. It is also important to note that segregation analysis requires careful clinical correlation of family members submitted for testing, and this information is not always provided at the time samples are submitted to the testing laboratory.

It should be mentioned that in this scenario, although rare, the possibility of two or more different etiologies for the ID in large families with apparent XLID does exist. As mentioned previously, there may be one or more related males with ID in a family, but the ID in the affected family members could be X-linked while the ID in the proband could be caused by another unrelated or nongenetic cause. Additionally, since ID is a common condition in all populations and ethnicities, it is not uncommon to find more than one affected male in a typical three-generation pedigree with all affected males clearly not having the same condition. In this case, the inheritance pattern is often not compatible with being X-linked, thus eliminating the possibility of unifying their diagnoses.

3.3 Scenario 3: Male with Intellectual Disability and a Pedigree Suggestive of X-linked Inheritance

Similar to the case of the affected male with a definite X-linked pattern of inheritance, it is highly possible that the ID in a male with a suggestive X-linked pattern of inheritance (Fig. 9.3, individual III-1) is the result of a mutation on the X chromosome. However, it is also possible that the disability in this individual results from an autosomal mutation or another nongenetic cause. Nevertheless, even in the case of a suspected pattern of X-linked inheritance, it is logical to start by testing the XLID genes first given the reduced number of genes requiring analysis, compared to other whole exome or whole genome NGS strategies. If a clear pathogenic alteration is detected, then the cause of the ID in the proband has been identified. The proband's mother (individual II-4) should be tested, and this analysis alone will shed light on the pattern of inheritance for the ID in the family. When the variant is inherited, additional affected male family members (individual III-2) and other female family members (individuals I-4, III-3, III-4) may then also wish to be tested for the identified mutation.

If a variant of unknown clinical significance is detected in a male with ID and a suggestive X-linked pattern of inheritance, additional family member testing is also warranted, in order to attempt to delineate the significance of the alteration. If the variant is found to be inherited, clinical correlation and segregation analysis is crucial. As in the scenario with clear X-linked inheritance, if the variant segregates appropriately in the family, it is suggestive that the variant is deleterious. Appropriate segregation analysis, along with deleterious predictions from multiple *in silico* analysis programs, becomes strong evidence that the change is pathogenic. However, when the variant is found in a phenotypically normal male, it would support the notion that the variant is a benign sequence change with no clinical significance.

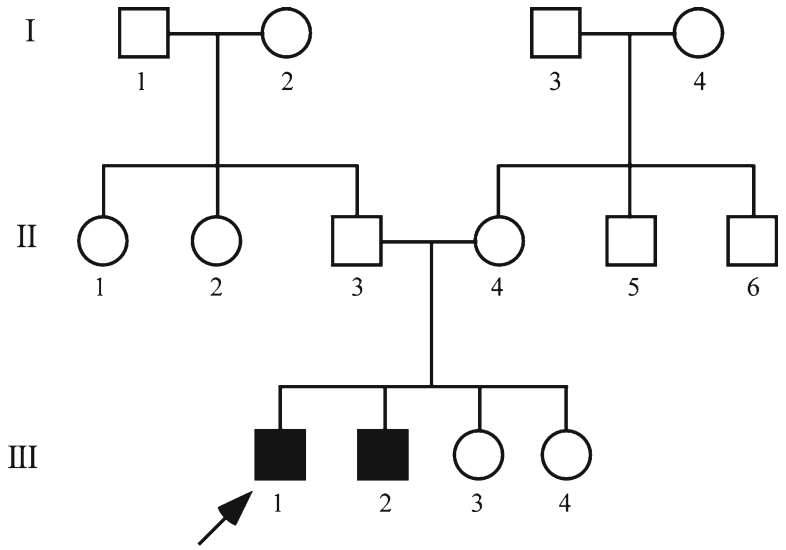

Fig. 9.3 Male with intellectual disability and a pedigree suggestive of X-linked inheritance

Even when the pattern of inheritance in the family is uncertain, determining that the variant is de novo strongly suggests pathogenicity given that a targeted panel was chosen to evaluate the proband. Additionally, when a mutation appears to be de novo in the proband (even though there are multiple affected male siblings), germline mosaicism in the mother, although rare, could be present. An example that best illustrates this concept involves Duchenne muscular dystrophy and its associated gene, *DMD*. While the de novo mutation rate for *DMD* is approximately 33 %, the rate of germline mosaicism in this gene can be as high as 15–20 % (GeneTests, http://www.ncbi.nlm.nih.gov/books/NBK1119/). Germline mosaicism for an X-linked mutation could certainly explain the presence of more than one affected sibling in the event that the mutation or variant appears to be de novo after testing of the proband's mother is complete. Given this fact, the subsequent testing of the additional affected male siblings in the family would confirm or refute gonadal mosaicism in the mother.

4 Clinical Experience with NGS Testing for XLID

In this section, we briefly cover our initial experience with targeted NGS testing in the molecular diagnostic setting using a panel of 90 recognized XLID genes. The information outlined in this section is not meant to fully represent our entire experience with XLID testing, but rather it is intended to provide the reader with an appreciation for the types of referrals and outcomes routinely associated with this line of testing. The primary basis for this section will be the first 100 cases submitted to the

Molecular Diagnostic Laboratory at the Greenwood Genetic Center (GGC) for XLID panel testing.

We note that 71 of the first 100 patients submitted for testing were 10 years of age or younger. A clinical information data sheet was requested for each patient, but only 58 % of the initial patients were submitted with sufficient phenotypic data to be considered useful. Based on this information, the most common physical findings, in addition to ID, were speech delay (94 %), motor milestone delay (88 %), and dysmorphic features (49 %). Other common findings in this group were seizures (41 %), autism (31 %), short stature (23 %), skeletal abnormalities (21 %), macrocephaly (21 %), and microcephaly (18 %). Most other clinical features queried in the data sheet were present in less than 10–15 % of patients, but we do note that some patients had a number of features for which no information was provided by the clinician and this impacts these percentages to some degree. Taken together, it was apparent that the individuals submitted for clinical testing had a great deal of phenotypic heterogeneity even though many of them had the common features often seen in association with ID.

From a family history perspective, most patients were not submitted with a pedigree (44 %). The apparent singletons with no family history accounted for 22 % of cases, and interestingly, only 7 % of cases were submitted with a pedigree that was clearly consistent with X-linked inheritance. Another 10 % of the probands came from families with a pedigree that was suggestive of being X linked. The remainder of the group was classified as having a family history of affected females (6 %), one or more pregnancy losses (5 %), multiple affected males on both sides of the family (2 %), or others (4 %).

Twelve of the first 100 probands submitted had findings believed to be clearly pathogenic based on standard mutation classification. Interestingly, the only gene represented more than once in this group was *ARX* which appeared to be the causative gene for two of the individuals. These two *ARX* mutations involved the common 21- and 24-base pair duplications in exon 2 that were actually identified by supplemental Sanger sequencing given the difficulty in generating sufficient quality NGS data for this exon. Four additional probands had specific gene alterations that appeared to be likely pathogenic based on the clinical features of the patient, but additional studies of appropriate family members is still needed to confirm their status. A normal result was reported for 26 of the probands, while the remaining 58 probands were reported with one or more variants of unknown clinical significance. Most of the variants with unknown significance are unlikely to be pathogenic, but further segregation studies are required to confirm this notion. Based on our experience, it often takes an extended period of time for the appropriate family members to be submitted to perform the necessary segregation studies.

The primary reason for requesting the targeted XLID panel appears to be the coverage it provides across multiple genes compared to the traditional single-gene approach that is often not very productive unless a patient has a clinical presentation fairly consistent with the expected phenotype for a given syndrome. As our experience has grown well beyond these first 100 cases, it has become more routine to analyze the generated NGS data in a timely fashion. Each individual is typically

reported as either clearly abnormal with an obvious pathogenic change, normal with no variants detected, or uncertain with one or more variants of unknown clinical significance detected. To date, the majority of cases continue to be reported as either normal or uncertain, and for the latter cases, there is most often only one or two variants that require further attention. Sanger sequencing is used to confirm all alterations prior to reporting, and our experience demonstrates that a low level of false positives are identified by NGS. For most false-positive results, the primary explanation is low coverage and/or a low quality score for the alteration in question. The potential false-positive results are typically easy to anticipate from the NGS data and often include alterations that appear heterozygous in males with a bias towards the normal allele (a more obvious, hemizygous alteration in a male is expected). For most individuals that are reported as normal on the targeted XLID panel, we do not know if a specific genetic cause has been identified by another means of testing. The most likely reason for a negative result on the XLID panel is that the individual in question does not have an X-linked disorder. However, more time is needed before this assumption can be validated by the use of exome sequencing. In the meantime, targeted panel testing of X-linked genes remains a reasonable approach to testing males suspected of having XLID.

5 Conclusions

When novel variants identified by a targeted NGS panel have appropriate segregation and/or bioinformatic analysis supporting their detrimental nature, the most relevant clinical topic is the phenotype of the individual and/or family compared to those previously reported in the literature. However, this is often complicated since many XLID genes have a limited number of cases attributed to them, and the reported features of the clinical phenotype are not overly comprehensive. Furthermore, even the well-characterized syndromes continue to have their clinical spectrums expanded as more patients are identified and new atypical presentations are published. Situations like these challenge clinicians to determine the likelihood of a match between families and a proposed causative gene. For some XLID conditions, the phenotype is not readily apparent or easily diagnosed during early childhood, and it may take years for the entire clinical picture to develop. These realities are often difficult for clinicians and families alike, given the degree of uncertainty that surrounds them.

In this light, the most pressing matter that follows these uncertainties is the need for families to have accurate information for reproductive decision-making purposes. In many families, this information has the potential to impact a significant number of females that are at risk for carrying the mutation. Knowing this information can have a profound impact for those living with uncertainty. Equally important to family planning is looking at the prospect of potential therapeutic options for those affected by XLID. As more studies disentangle the molecular pathways and protein networks involved in ID, it will become easier to envision new strategies for

developing rational clinical trials designed to test the efficacy of new treatments for various XLID conditions. Correcting or offsetting the functional significance of specific gene mutations, along with the hope of reversing their clinical impact, remains the ultimate goal for many clinicians and scientists working in this field. The uniqueness of the X chromosome in males provides an unparalleled opportunity to do just this, further emphasizing the need to fully capitalize on NGS and its capabilities in order to secure a diagnosis for as many families affected by XLID as possible.

Appendix: Profiles of the Most Common XLID Syndromes

(*Gene, location, and brief clinical findings in addition to intellectual disability*)

Replicated with Permission from Stevenson and Schwartz [1]

Aarskog Syndrome

FGD1, Xp11.21

Short stature, hypertelorism, downslanting palpebral fissures, joint hyperextensibility, shawl scrotum

Adrenoleukodystrophy

ABCD1, Xq28

Variable and progressive vision and hearing loss, spasticity, neurological deterioration associated with demyelination of the central nervous system and adrenal insufficiency

Aicardi syndrome

No gene, Xp22

Agenesis of the corpus callosum, lacunar chorioretinopathy, costovertebral anomalies, seizures in females

Allan-Herndon Syndrome

SLC16A2, Xq13

Generalized muscle hypoplasia, childhood hypotonia, ataxia, athetosis, dysarthria, progressing to spastic paraplegia

ARX-Related Syndromes

(*includes Partington, Proud, West, X-linked lissencephaly with ambiguous genitalia (XLAG) syndromes and nonsyndromic XLID*)

ARX, Xp22.3

Partington: dysarthria, dystonia, hyperreflexia, seizures. West: infantile spasms, hypsarrhythmia. Proud: microcephaly, ACC, spasticity, seizures, ataxia, genital anomalies. XLAG: lissencephaly, seizures, genital anomalies

ATRX Syndrome

(*includes Chudley-Lowry, Carpenter-Waziri, Holmes-Gang, and Martinez spastic paraplegia syndromes and nonsyndromic XLID*)

ATRX, Xq13.3

Short stature, microcephaly, hypotonic facies with hypertelorism, small nose, open mouth and prominent lips, brachydactyly, genital anomalies, hypotonia, in some cases hemoglobin H inclusions in erythrocytes

Christianson Syndrome

SLC9A6, Xq26

Short stature, microcephaly, long narrow face, large ears, long straight nose, prominent mandible, general asthenia, narrow chest, long thin digits, adducted thumbs, contractures, seizures, autistic features, truncal ataxia, ophthalmoplegia, mutism, incontinence, hypoplasia of the cerebellum, and brain stem

Coffin-Lowry

RPS6KA3, Xp22

Short stature, distinctive facies, large soft hands, hypotonia, joint hyperextensibility, skeletal changes

Creatine Transporter Deficiency

SLC6A8, Xq28

Nondysmorphic, autistic, possibly progressive

Duchenne Muscular Dystrophy

DMD, Xp21.3

Pseudohypertrophic muscular dystrophy

Fragile X Syndrome

FMR1, Xq27.3
Prominent forehead, long face, recessed midface, large ears, prominent mandible, macroorchidism

Hunter Syndrome

IDS, Xq28
Progressive coarsening of face, thick skin, cardiac valve disease, joint stiffening, dysostosis multiplex

Incontinentia Pigmenti

IKBKG, Xq28
Sequence of cutaneous blistering, verrucous thickening, and irregular pigmentation. May have associated CNS, ocular abnormalities

Lesch-Nyhan Syndrome

HPRT, Xq26
Choreoathetosis, spasticity, seizures, self-mutilation, uric acid urinary stones

Lowe Syndrome

OCRL, Wq26.1
Short stature, cataracts, hypotonia, renal tubular dysfunction

***MECP2* Duplication Syndrome**

MECP2, Xq28
Hypotonia, progressing to spastic paraplegia, recurrent infections

Menkes Syndrome

ATP7A, Xp13.3
Growth deficiency, full cheeks, sparse kinky hair, metaphyseal changes, limited spontaneous movement, hypertonicity, seizures, hypothermia, lethargy, arterial tortuosity, death in early childhood

Pelizaeus-Merzbacher Disease

PLP1, Xq21.1
Nystagmus, truncal hypotonia, progressive spastic paraplegia, ataxia, dystonia

Renpenning Syndrome

(*includes Sutherland-Haan, cerebropalatocardiac, Golabi-Ito-Hall, Porteous syndromes*)
PQBP1, Xp11.3
Short stature, microcephaly, small testes. May have ocular or genital abnormalities

Rett Syndrome

MECP2, Xq28
XLID in female, cessation and regression of development in early childhood, truncal ataxia features, acquired microcephaly

X-Linked Hydrocephaly

(*includes mental retardation, aphasia, shuffling gait and abducted thumbs (MASA) spectrum*)
L1CAM, Xq28
Hydrocephalus, adducted thumbs, spastic paraplegia

References

1. Stevenson RE, Schwartz CE (2009) X-linked intellectual disability: unique vulnerability of the male genome. Dev Diabil Res Rev 15:361–368
2. Schalock RL, Borthwick-Duffy SA, Bradley VJ, Buntinx WHE, Coulter DL, Craig EM, Gomez SC, Lachapelle Y, Luckasson R, Reeve A, Shogren KA, Snell ME, Spreat S, Tassé MJ, Thompson JR, Verdugo-Alonso MA, Wehmeyer ML, Yeager MH (2009) Intellectual disability: definition, classification, and systems of supports, 11th edn. American Association of Intellectual and Developmental Disabilities, Washington, DC. ISBN 13: 978-1935304043
3. Mefford HC, Batshaw ML, Hoffman EP (2012) Genomics, intellectual disability, and autism. N Engl J Med 366(8):733–743
4. Veltman JA, Brunner HG (2012) De novo mutations in human genetic disease. Nat Rev Genet 13(8):565–575
5. Whibley AC, Plagnol V, Tarpey PS, Abidi F, Fullston T, Choma MK, Boucher CA, Shepherd L, Willatt L, Parkin G, Smith R, Futreal PA, Shaw M, Boyle J, Licata A, Skinner C, Stevenson RE, Turner G, Field M, Hackett A, Schwartz CE, Gecz J, Stratton MR, Raymond F (2010)

Fine-scale survey of X chromosome copy number variants and indels underlying intellectual disability. Am J Hum Genet 87(2):173–188
6. Stevenson RE, Schwartz CE, Rogers RC (2012) Atlas of X-linked intellectual disability syndromes, 2nd edn. Oxford University Press, Oxford/New York. ISBN 13: 978-0199811793
7. Lubs HA, Stevenson RE, Schwartz CE (2012) Fragile X and X-linked intellectual disability: four decades of discovery. Am J Hum Genet 90(4):579–590
8. Shoubridge C, Gardner A, Schwartz CE, Hackett A, Field M, Gecz J (2012) Is there a Mendelian transmission ratio distortion of the c.429_452dup(24 bp) polyalanine tract ARX mutation? Eur J Hum Genet. doi:10.1038/ejhg.2012.61
9. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, Hardy C, O'Meara S, Latimer C, Dicks E, Menzies A, Stephens P, Blow M, Greenman C, Xue Y, Tyler-Smith C, Thompson D, Gray K, Andrews J, Barthorpe S, Buck G, Cole J, Dunmore R, Jones D, Maddison M, Mironenko T, Turner R, Turrell K, Varian J, West S, Widaa S, Wray P, Teague J, Butler A, Jenkinson A, Jia M, Richardson D, Shepherd R, Wooster R, Tejada MI, Martinez F, Carvill G, Goliath R, de Brouwer AP, van Bokhoven H, Van Esch H, Chelly J, Raynaud M, Ropers HH, Abidi FE, Srivastava AK, Cox J, Luo Y, Mallya U, Moon J, Parnau J, Mohammed S, Tolmie JL, Shoubridge C, Corbett M, Gardner A, Haan E, Rujirabanjerd S, Shaw M, Vandeleur L, Fullston T, Easton DF, Boyle J, Partington M, Hackett A, Field M, Skinner C, Stevenson RE, Bobrow M, Turner G, Schwartz CE, Gecz J, Raymond FL, Futreal PA, Stratton MR (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet 41(5):535–543
10. Chahrour M, Zoghbi HY (2007) The story of Rett syndrome: from clinic to neurobiology. Neuron 56(3):422–437
11. Dibbens LM, Tarpey PS, Hynes K, Bayly MA, Scheffer IE, Smith R, Bomar J, Sutton E, Vandeleur L, Shoubridge C, Edkins S, Turner SJ, Stevens C, O'Meara S, Tofts C, Barthorpe S, Buck G, Cole J, Halliday K, Jones D, Lee R, Madison M, Mironenko T, Varian J, West S, Widaa S, Wray P, Teague J, Dicks E, Butler A, Menzies A, Jenkinson A, Shepherd R, Gusella JF, Afawi Z, Mazarib A, Neufeld MY, Kivity S, Lev D, Lerman-Sagie T, Korczyn AD, Derry CP, Sutherland GR, Friend K, Shaw M, Corbett M, Kim HG, Geschwind DH, Thomas P, Haan E, Ryan S, McKee S, Berkovic SF, Futreal PA, Stratton MR, Mulley JC, Gécz J (2008) X-linked protocadherin 19 mutations cause female-limited epilepsy and cognitive impairment. Nat Genet 40(6):776–781
12. Richards CS, Bale S, Bellissimo D, Das S, Grody W, Hegde M, Lyon E, Ward B, Molecular Subcommittee of the ACMG Laboratory Quality Assurance Committee (2007) ACMG recommendations for standards for interpretation and reporting of sequence variations: revisions 2007. Genet Med 10(4):294–300, AC
13. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814
14. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249

Chapter 10
NGS Analysis of Heterogeneous Retinitis Pigmentosa

Rui Chen and Feng Wang

Abstract One of the most significant milestones in biomedical research was the completion of the human genome sequencing project. The subsequent invention of Next-Generation Sequencing (NGS) technology has revolutionized molecular biology, genetics, and genomics. Already, medicine is being tailored to take into account an individual's genome. In particular, diseases with diverse genetic causes, such as retinitis pigmentosa (RP), are poised for the clinical use of NGS.

In this chapter, we discuss the application of NGS technology in the molecular diagnosis of heterogeneous RP. By comparing the current molecular diagnostic methods with the NGS-based approach, we conclude that the NGS method is much more comprehensive and cost effective. We speculate that with further improvement and validation, NGS will become the method of choice for molecular diagnosis of RP in the near future.

1 Retinitis Pigmentosa as a Heterogeneous Disease

Retinitis pigmentosa (RP, OMIM 268000) is a relatively common inherited retina disease affecting about 1 in 4,000 people [1]. RP is a clinically and genetically heterogeneous disease.

R. Chen (✉) • F. Wang
Department of Molecular and Human Genetics, Baylor College of Medicine,
One Baylor Plaza, Houston, TX 77030, USA
e-mail: ruichen@bcm.edu; fengw@bcm.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_10,

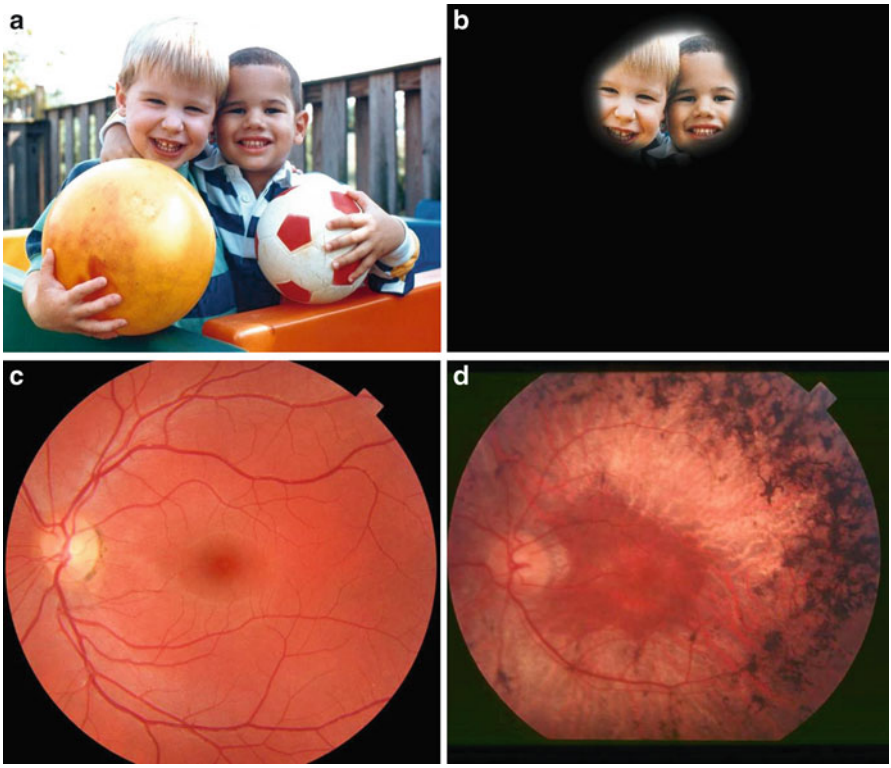

Fig. 10.1 The illustration of retinitis pigmentosa. (**a**) Normal vision, (**b**) tunnel vision from an RP patient, (**c**) normal fundus, (**d**) fundus from an RP patient (Panels A and B are obtained from NIH NEI http://www.nei.nih.gov/photo/keyword.asp?narrow=Eye+Disease+Simulation&match=all. Panel C: published under public domain by Mikael Häggström http://commons.wikimedia.org/wiki/File:Fundus_photograph_of_normal_left_eye.jpg. Panel D: © 2006 Hamel; licensee BioMed Central Ltd [1])

1.1 Clinical Features of RP

In a typical case of RP, rod photoreceptor cells, which are responsible for peripheral vision as well as vision under low light, degenerate, resulting in early onset night blindness and tunnel vision (Fig. 10.1b). As the disease progresses, cone photoreceptors start to degenerate as well, leading to the reduction of central, color, and day vision. At a late stage, the degeneration of photoreceptor cells becomes much more severe, eventually causing complete blindness. In addition, RP presents with pigmentary retinopathy, caused by the release of pigment from degenerating retinal pigment epithelium (RPE) cells. These pigment granules often accumulate in a perivascular fashion, commonly referred to as "bone spicule deposits" (Fig. 10.1d). Variability in pigment deposition leads to hypopigmentation of the retina, translucence, and round pigment deposits. Due to retina atrophy, the retinal vasculature

Table 10.1 Modes of inheritance underlying RP and their proportions

Mode of inheritance	Proportion of all RP cases
Autosomal dominant RP (adRP)	25 %
Autosomal recessive RP (arRP)	20 %
X-linked RP (xlRP)	10 %
Digenic and mitochondrial RP	Very rare
Simplex RP	45 %

becomes attenuated, and the optic nerve head becomes pale as a result of changes in the blood flow to the retina.

Despite these shared clinical features, the severity of RP is variable. First, the age of onset of RP ranges from birth to infancy (juvenile RP) and early adulthood (adult onset RP). Second, the end point of RP differs and a more severe prognosis is associated with earlier onset. Third, syndromic forms of RP exist, in which symptoms are observed in other organ systems. The genetic bases of syndromic and nonsyndromic RP are overlapping as described later.

1.2 Genetic Basis of RP

Consistent with its clinical variability, RP is genetically diverse. As shown in Table 10.1, the most common inheritance mode for RP is autosomal dominant (adRP), which accounts for 25 % of the patients. About 20 % of RP patients have autosomal recessive (arRP) inheritance, while 10 % have X-linked RP (xlRP) [2]. A small fraction of patients have mitochondrial or digenic (controlled by two genes) forms of RP [3–5]. Finally, due to the limited size of patient pedigrees, the inheritance mode cannot be reliably determined for a large portion of patients. These patients are currently classified as simplex RP and represent 45 % of all cases [2]. Additional information, such as molecular diagnosis, is needed to properly assign an inheritance mode for these patients, which is critical for genetic counseling.

The molecular basis for RP is also highly heterogeneous. As of 2012, mutations in 52 genes have been linked to RP [6]. These genes function in strikingly diverse biological pathways, including phototransduction, the retinoid (vitamin A) cycle, gene transcription, RNA splicing, and photoreceptor structure. Defects in the phototransduction cascade are the major causes of RP cases. For example, mutations in phototransduction-related genes Rhodopsin and *PDE* (including both *PDE6A* and *PDE6B*) account for 25 % of adRP cases and 8 % of arRP cases, respectively [2, 7]. Many other pathways are also involved. For instance, *CRX* controls normal development of retinal tissues [8]; *RPE65*, *LRAT*, and *RDH12* are thought to function in the 11-*cis*-retinol metabolic pathway [9–13].

Additional complexity comes from the observation that mutations in the same gene can lead to different retinal diseases and vice versa. For example, mutations in *CRX* can lead to RP, Leber Congenital Amaurosis, and Cone Rod Dystrophy [6]. This may be due to the differences in severity between mutations and modification

by other factors in the genome [14]. In a separate case, two pathogenic mutations in *CYP4V2* were initially found in patients with Bietti Crystalline Corneoretinal Dystrophy [15, 16]. The same mutations were recently found to cause RP as well [17]. Moreover, several genes that are associated with syndromic retinal disease have also been linked to nonsyndromic RP, such as *BBS8* and *USH2A* [18, 19]. As a result, mutations in a large number of genes can lead to RP, making accurate molecular diagnosis of RP very challenging.

2 Molecular Diagnosis of RP

Molecular diagnosis plays an important role in the management and treatment of RP. First, many therapies are gene specific. For example, gene therapy for patients with mutations in *RPE65* [20–22], a known causative gene for both RP and LCA, may soon be available. Without an accurate diagnosis of the causative gene, this treatment cannot be applied. Second, genetic counseling and family planning are based on the information provided by molecular diagnosis, including the disease's inheritance pattern and causative mutation. Therefore, molecular diagnosis is important for patients with RP and may dramatically change the way they are treated and their family plans. Various approaches are currently used for molecular diagnosis of RP. Among them, Sanger sequencing and Array Primer Extension (APEX) are the most commonly used methods for molecular diagnosis of RP.

2.1 Sanger Sequencing

Sanger sequencing, originally developed by Frederick Sanger in the 1970s, is currently the gold standard of DNA sequencing and mutation identification. As shown in Fig. 10.2, specific primers for an exon of the RP causative genes are designed and used to generate PCR amplicons. Each amplicon is then sequenced on the automatic sequencing machine generating reads for both complementary strands of the target DNA. The mutations are then identified by comparing the sequencing results to human reference sequences.

Sanger sequencing is particularly useful and powerful to screen for mutations in a small number of candidate genes because of its fast turnaround time and high accuracy. However, since each Sanger sequencing assay only accurately reads a maximum of about 700–800 base pairs (bp) in length, sequencing all RP causative genes in order to perform a comprehensive diagnosis for one RP patient would take thousands of PCR and sequencing reactions, which are prohibitively expensive, labor intensive, and time consuming. Therefore, although Sanger sequencing is highly accurate, its low throughput and high cost nature prohibits it from being used as a general method for the comprehensive molecular diagnosis of RP.

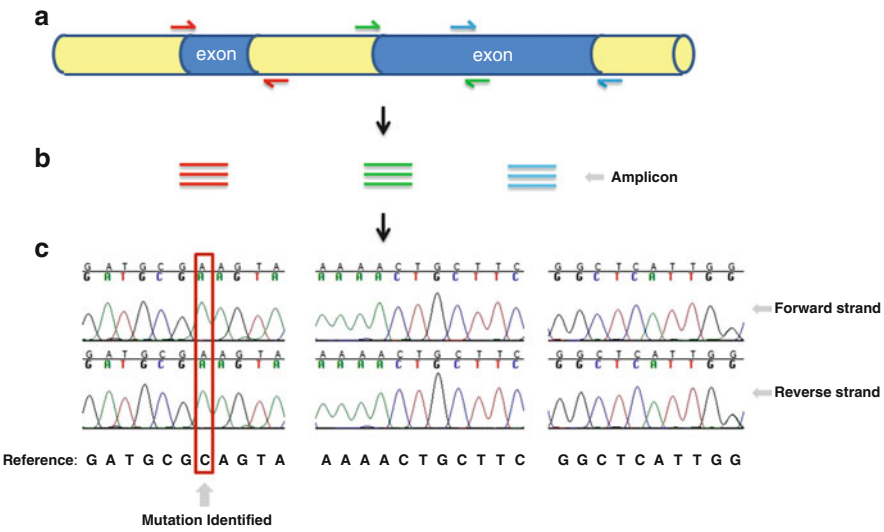

Fig. 10.2 Molecular diagnosis by Sanger sequencing. (**a**) Primers specific to each exon used to amply the exon and its surrounding sequence. (**b**) PCR amplicons will then be sequenced on the capillary sequencer. (**c**) Sequences from each amplicon will be aligned to human reference sequences to identify mutations

3 APEX Method

To address the low throughput issue with Sanger sequencing, a SNP genotyping microarray based method called Arrayed Primer Extension (APEX) has been developed [23]. APEX technology focuses on known mutations with a primer extension-based SNP genotyping approach. As shown in Fig. 10.3, regions of interests are first amplified by PCR. The PCR products are then fragmented and hybridized to a microarray. Each spot on the microarray has oligonucleotides immediately upstream of a known, RP-causing mutation. PCR products hybridized to the oligonucleotides will then serve as the template for subsequent primer extension reaction. Fluorophore-labeled nucleotide terminators are used as substrates, and only the base that is complementary to the PCR fragment is added to the end of the oligonucleotides. Based on the color, the identity of the added base and the genotype of the corresponding position can be determined.

Several panels that are specific to different RP inheritance forms have been developed by Asper biotech [24], including adRP, arRP, and xlRP panels. As shown in Table 10.2, in the latest APEX-based RP panels offered by Asper biotech, 414 known disease-associated variants in 16 genes have been included in the adRP panel, 594 known mutations in 19 genes in the arRP panel, and 184 mutations in two genes in the xlRP panel.

Compared to the Sanger sequencing, APEX method increases the diagnosis throughput and reduces the cost by allowing simultaneous screening of a large

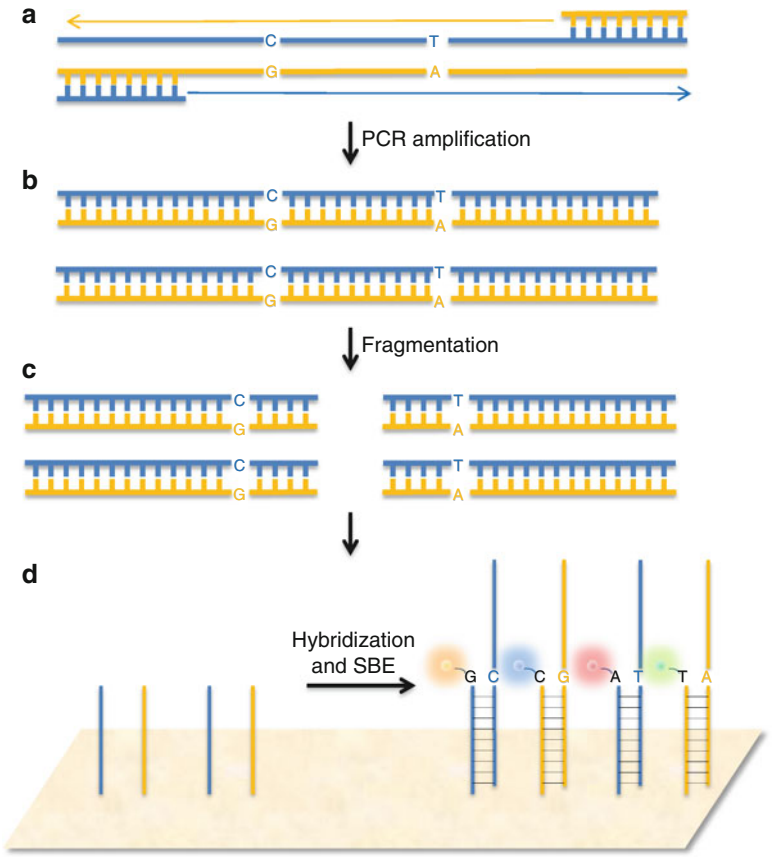

Fig. 10.3 Schematic diagram of APEX-based molecular diagnosis. (**a**) Primers flanking target sites are designed. (**b**) PCR is performed to amplify the region. (**c**) Amplicons are fragmented to the appropriate length. (**d**) DNA fragments containing target sites are hybridized to the oligonucleotides on the microarray. Subsequent single base extension (SBE) reaction adds one fluorophore-labeled terminator nucleotide at the mutation site. The genotype of the target site will then be determined based on fluorescence

Table 10.2 Three latest APEX-based molecular diagnosis panels for RP from Asper biotech [24]

	AD-RP	AR-RP	XL-RP
Total mutations	414	594	184
Total genes	16	19	2
Gene list	*CA4*, *FSCN2*, *IMPDH1*, *NRL*, *PRPF3*, *PRPF31*, *PRPF8*, *RDS*, *RHO*, *ROM1*, *RP1*, *RP9*, *CRX*, *TOPORS*, *KLHL7*, and *PNR*	*CERKL*, *CNGA1*, *CNGB1*, *MERTK*, *PDE6A*, *PDE6B*, *PNR*, *RDH12*, *RGR*, *RLBP1*, *SAG*, *TULP1*, *CRB*, *RPE65*, *USH2A*, *USH3A*, *LRAT*, *PROML1*, and *PBP3*	*RP2* and *RPGR* (ORF15 excluded)

number of mutations from multiple genes. However, since the currently available panels can only test known mutations in a subset of known RP genes, a diagnosis rate of less than 15 % has been achieved using these arrays [25, 26].

4 NGS for Molecular Diagnosis of RP

Current methods for molecular diagnosis of RP are either too expensive or have a low rate of diagnosis. These obstacles can be potentially overcome by the introduction of the NGS technology.

4.1 The Technologies

4.1.1 NGS Technology

Massively parallel sequencing is the principle of NGS technology, which can generate up to billions of reads in a single run. First introduced by the biotech company 454 Life Sciences in 2005, the technology has rapidly evolved and now has multiple platforms that are commercially available. DNA polymerase-based synthesis and sequencing is utilized by 454 and generates reads of up to 1,000 bp in length [27]. SOLiD sequencing adopts a sequence-by-ligation method with a read length of 50–75 bp [28]. Illumina sequencing is based on reversible dye terminators and generates reads of up to 150 bp in length [29]. Despite the difference in chemistry and variations in performance, all technologies share one key common feature, which is the extremely high sequencing efficiency in terms of time and cost compared to the conventional Sanger method. For example, the latest Illumina HiSeq 2500 machine can generate 120 gigabase in 27 h with a cost of only around $0.2 per million bases [29]. It is substantially more efficient than Sanger sequencing, which could take over a year to read through one gigabase at a cost of $100 per million bases [30].

More recently, two miniaturized platforms have been introduced to the market, including MiSeq and Ion torrent PGM. MiSeq use the same technology as Illumina HiSeq [29], while PGM uses a technology based on semiconductor sensors [31]. PGM uses a chip with massively parallel array of semiconductor sensors that are used to detect hydrogen ions which are released when nascent nucleotides are incorporated into the DNA template. The detected signal combined with the nucleotide information is recorded and then converted to the actual sequence. PGM can generate up to 1 gigabase of sequence in less than 4.5 h, while MiSeq is capable of generating 540–610 megabase of sequence in around 4 h. Although the two small platforms have different methodologies, they share a fast turnaround time, making them ideal for low throughput clinical use.

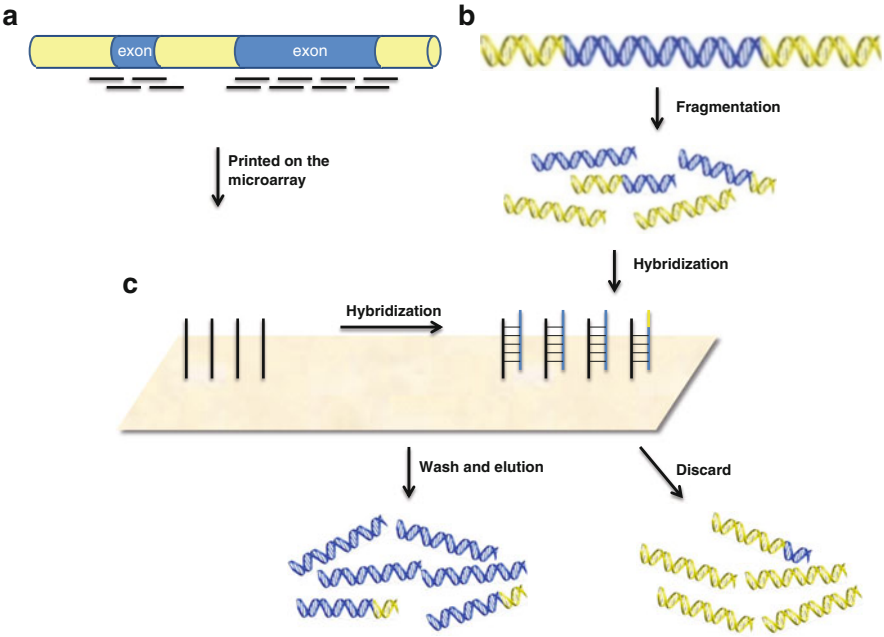

Fig. 10.4 Schematic diagram of array-based DNA capture. (**a**) A tiling probe set covering the targeted region is designed and printed on the microarray. (**b**) Genomic DNA is fragmented to small size. The targeted region is in *blue* and the other regions are in *yellow*. (**c**) DNA fragments from the targeted regions are hybridized to the probes on the microarray. (**d**) After wash and elution, DNA fragments from the targeted regions are recovered

4.1.2 DNA Capture Technology

Despite the high cost efficiency of NGS, sequencing the entire human genome at a deep enough coverage to identify potential mutation is still costly. In the case of RP, only 50–200 genes those are known or likely to cause the disease need to be tested instead of the entire genome. Focusing on the 50–200 genes can further reduce testing cost and turnaround time. In 2007, the first DNA capture technology was reported [32, 33]. The technology is based on hybridization between a probe array, which targets a subset of genomic regions, and the genomic DNA of a patient. As shown in Fig. 10.4, a tiling probe set (either DNA or RNA) that covers the targeted DNA regions is designed and synthesized. After hybridization, targeted patient DNA fragments are bound to the array, while unbound fragments are washed off. Based upon the design of the array, any genomic locus can be specifically isolated. Since the first development of DNA capture technology, it has been further optimized and now it can be performed in solution instead of on a microarray chip, substantially facilitating parallelization and automation of the procedure.

Fig. 10.5 The illustration of multiplex sequencing using the molecular barcode technology. Molecular barcodes are short oligonucleotides which can be added to the end(s) of DNA fragments. (**a**) Unique barcode sequence is added to each sample. (**b**) Barcoded DNA fragments are mixed together and sequenced in parallel. (**c**) Sequence reads are assigned to different samples based on their barcode sequences

4.1.3 Molecular Barcoding Technology

In a typical RP diagnosis panel based on capture-NGS method, a set of 50–200 genes known to cause RP or other retina disease are sequenced per sample. Given the enormous sequencing capacity of current NGS platforms, such a small targeted region (~400 to ~1,500 kb) makes it possible to pool multiple samples together for sequencing. However, in order to sequence multiple samples in parallel, there must be a way to distinguish DNA fragments from different individuals. Molecular barcoding is a widely used technique to label different samples so that they can be pooled together without losing their identity. As shown in Fig. 10.5, a molecular barcode is a short oligonucleotide (6–10 bp) that is added to one or both ends of every DNA fragment. Since a unique sequence is used for each sample, reads can be assigned to different samples based on sample-specific barcode sequences. With molecular barcoding, parallel sequencing of multiple samples is possible. In addition, barcodes can be added before DNA capture so that multiple samples can be enriched in the single DNA capture assay, which reduces the time, labor, and reagent cost.

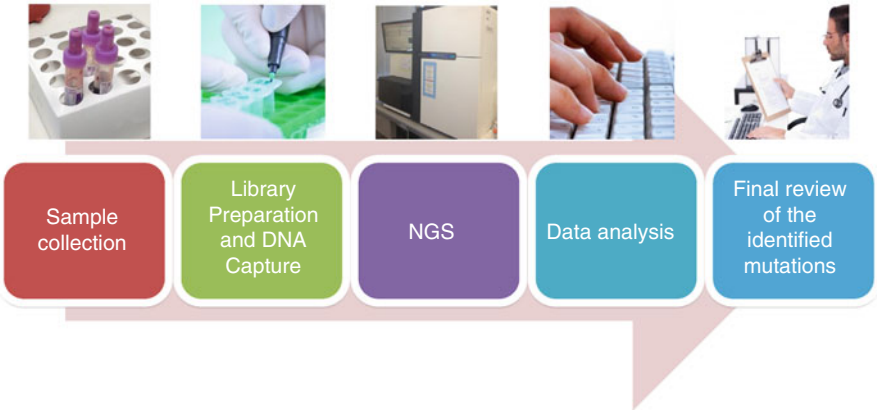

Fig. 10.6 Capture-NGS-based molecular diagnosis workflow

4.1.4 NGS Data Analysis

The high throughput nature of NGS-based molecular diagnosis enables large amounts of sequencing data to be generated in a single test. As a result, automated data analysis is necessary. Fortunately, with the development of NGS, the field of bioinformatics has also evolved rapidly during the last few years. The availability of public domain software greatly facilitates the setup of an automated data analysis pipeline for capture-NGS-based molecular diagnosis.

A typical data analysis pipeline for capture-NGS-based molecular diagnosis of RP includes the following six key steps. First, sequencing reads are mapped to the human reference genome using alignment software such as BWA [34, 35]. Second, aligned reads are de-duplicated, realigned, and recalibrated to improve variant-calling accuracy using software such as GATK [36]. Third, raw variants are called using variant-calling software such as Atlas SNP/INDEL [37]. Fourth, extensive filtering and annotation are performed to remove common polymorphisms, synonymous changes, and variants inconsistent with the inheritance model. Fifth, putative pathogenic mutations are identified based on the several criteria such as whether it was previously reported to be pathogenic, whether it is a severe mutation (e.g., nonsense, splicing, etc.), and whether it is predicted to be detrimental by functional prediction software such as SIFT [38]. Finally, all the identified mutations are subjected to validation through Sanger sequencing as well as segregation test (if applicable), and the results are reviewed carefully by trained scientists and clinicians (Fig. 10.6).

4.2 An Overview of Current Studies on Capture-NGS-Based Molecular Diagnosis of RP

Several studies coupling DNA capture and NGS technology for diagnosis of ocular disease have been reported. In 2010, Gordana Raca et al. first assessed whether NGS

coupled with DNA capture could be used for molecular diagnosis of ocular birth defects [39]. More than 1,000 exons in 100 candidate genes were sequenced using array-based DNA enrichment followed by NGS in their study. Two samples with previously identified mutations were tested. The capture-NGS method detected the known mutations in both samples, suggesting that the new technology could be used in both research and diagnostic settings for genetically heterogeneous diseases.

Subsequently, two studies have shown that capture-NGS technology significantly improves RP diagnosis in terms of diagnostic rate and cost [40, 41]. In the first study, a total of 2011 individual regions (mostly exons) from 111 known retinal diseases genes were targeted and sequenced using array-based DNA capture followed by 454 sequencing. On average, vast majority of the targeted exons were covered by at least 10X, while 15 exons were covered less than 5X due to, according to the authors, either high GC-contents or overwhelming repetitive sequences. Parallel sequencing of 12 positive controls (samples with known causative mutations) revealed a sensitivity rate of 83 %. Similarly, in the second study, 105 known blindness genes were targeted and sequenced by DNA array capture followed by SOLiD sequencing. Specifically, 1,874 individual exons as well as the exon-intron junction regions were targeted. Among them, 92 % were covered by 20X. Analysis of previously Sanger validated samples indicated the NGS assay had a sensitivity rate of 98 % and the only undiscovered mutation was located in a highly repetitive region.

A total of around 150 RP patients with multiple inheritance forms including recessive, dominant, and simplex RP were examined in the two studies. The mutations identified by capture-NGS are accurate and can be confirmed by Sanger sequencing. Segregation tests were further performed to test the pathogenicity of these mutations identified by capture-NGS. An overall diagnosis rate of approximately 50 % was reached, which is more than triple the rate of diagnosis achieved by the APEX method.

4.3 Comparison of Capture-NGS to Current Methods

Based on the above studies, it is clear that the capture-NGS approach is feasible and has several significant advantages over other methods that are currently in use.

First, capture-NGS approach is the most comprehensive method. Instead of only testing a small number of genes, capture-NGS simultaneously sequences all known and suspected RP genes. This is particularly important due to the genetic heterogeneity of RP where most genes only account for a small fraction of patients. In addition, NGS is applicable to the 50 % of RP cases where the inheritance mode cannot be reliably determined. Finally, since sequencing assays capture all positions, both known and novel mutant alleles can be detected. As a result, all RP patients carrying mutations in known disease genes can be successfully diagnosed using NGS, as opposed to a fraction of the cases using conventional, array-based methods.

Second, capture-based target gene enrichment is more robust and sensitive to mutation detection compared to PCR-based gene enrichment method. Capture

enrichment utilizes multiple probes for a single targeted region. As a result, the allelic bias due to DNA polymorphisms at the primer sites is eliminated when compared to either PCR-based Sanger sequencing or APEX. In addition, since an average coverage of 100X or more can be readily achieved, single nucleotide variations (SNVs) and small insertion and deletion (INDEL) can be reliably detected, reducing the false-negative rate.

Third, capture-NGS can potentially detect other types of mutations. In addition to SNV and INDEL, other types of mutations have also been reported in human diseases, such as DNA segmental duplications, large deletions, and chromosomal rearrangements. For example, copy number variations in an arRP gene EYS have been reported as a significant event due to apparent homozygosity [42]. Although no formal survey has been conducted on the percentage of pathogenic alleles due to chromosomal aberrations, it can account for as much as 10 % of all alleles. Therefore, in order to perform comprehensive molecular diagnosis of RP, copy number variations (CNVs) and structure variation (SV) should be addressed as well. Several software tools to detect both of these features through sequencing have been published [43–45]. Both read coverage as well as read pairing information can be used to identify potential CNVs and SVs. However, artifacts introduced from sample preparation and sequencing bias can result in false-positive prediction and false-negative prediction. Additional optimization of both the experiment procedures and the software tools is needed for the purpose of capture-NGS-based CNV/SV detection in the molecular diagnosis of RP.

Fourth, the cost of capture-NGS per base is much lower than that of Sanger sequencing, especially when utilizing barcode technology to run multiple samples in parallel. For example, on an Illumina HiSeq machine, many samples can be multiplexed together in a single lane while still generating sufficiently high coverage for the exons of all known RP genes. These NGS-specific features significantly increase the diagnosis efficiency by supporting batch analysis of multiple genes in multiple samples, which is not achievable by conventional diagnostic methods.

4.4 Current Limitations

NGS-based molecular diagnosis of RP has many advantages; however, some limitations still exist. First, systematic sequencing error is one of the major problems with current NGS technology. Although the error rate is quite low (range from 0.1 % to 1 % depending on the sequencing platforms [46]), the platform-specific error may still have a significant impact on the final diagnosis results especially when the sequencing coverage is low. As a result, high coverage data is necessary to compensate the errors so that accurate diagnosis can be achieved. Second, DNA capture is not 100 % effective. Some of the targeted regions either with high GC content or containing repeat-rich sequences (e.g., *RPGR* ORF15 and exon 1 of *GRM6*) cannot be efficiently captured and enriched [40]. This limitation can be largely solved by performing Sanger sequencing of these regions to reach 100 % coverage.

5 Perspective

The data presented here suggests that capture-NGS is becoming the predominant method for molecular diagnosis of RP in the near future. Nevertheless, additional improvements and studies are still needed in order to interpret the large number of variants quickly and accurately. Novel putative pathogenic mutations currently identified by molecular diagnosis cannot be accurately determined as disease-causing mutations without follow-up experiments. Improvements in data analysis and mutation prioritization are urgently needed. For example, larger and more comprehensive variant databases are essential to filter out common, nonpathogenic variants so that putative pathogenic mutations can be identified with confidence. In addition, novel types of mutations, such as CNVs and SVs, can also be the causative mutations for RP and, therefore, need to be addressed in the molecular diagnosis. Currently, array CGH can potentially serve as the alternative way for CNV detection as the NGS-based detection has not been successfully validated for the implementation in the diagnostic setting. Future investigation of the feasibility of NGS-based CNV/SV detection is needed so that all types of mutations can be identified in a single assay. Furthermore, mutations in noncoding regions can potentially cause disease as well. The lack of annotation and proper filtering criteria for noncoding SNPs makes them difficult to be identified. Functional interpretation of mutations identified in noncoding regions will be necessary in order to broaden the mutation spectrum. Digenic involvement has been proposed to explain the complex genetic inheritance in some RP cases. This area may be better explored with the help of high throughput NGS such that more accurate diagnosis can be achieved. Moreover, additional RP genes need to be identified to improve diagnostic yield. As a result of efforts in all of these areas, molecular diagnosis of RP will become easier, faster, and more accurate leading to better patient treatment and management.

Acknowledgments We would like to thank Eric J. Zaneveld, Hui Wang, and Yumei Li for critical reading and editing for the chapter. R.C. is supported by grants from National Eye Institute (R01EY018571 and R01EY022356).

References

1. Hamel C (2006) Retinitis pigmentosa. Orphanet J Rare Dis 1:40. doi:1750-1172-1-40 [pii] 10.1186/1750-1172-1-40
2. Fahim AT, Daiger SP, Weleber RG. Retinitis Pigmentosa Overview. 2000 Aug 4 2013 Mar 21. In: Pagon RA, Bird TD, Dolan CR, et al., editors. GeneReviews™. Seattle (WA): University of Washington, Seattle; 1993. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1417/
3. Mansergh FC, Millington-Ward S, Kennan A, Kiang AS, Humphries M, Farrar GJ, Humphries P, Kenna PF (1999) Retinitis pigmentosa and progressive sensorineural hearing loss caused by a C12258A mutation in the mitochondrial MTTS2 gene. Am J Hum Genet 64(4):971–985. doi:AJHG981030 [pii]

4. Kajiwara K, Berson EL, Dryja TP (1994) Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. Science 264(5165):1604–1608
5. Dryja TP, Hahn LB, Kajiwara K, Berson EL (1997) Dominant and digenic mutations in the peripherin/RDS and ROM1 genes in retinitis pigmentosa. Invest Ophthalmol Vis Sci 38(10):1972–1982
6. Retnet: http://www.sph.uth.tmc.edu/Retnet
7. Sohocki MM, Daiger SP, Bowne SJ, Rodriquez JA, Northrup H, Heckenlively JR, Birch DG, Mintz-Hittner H, Ruiz RS, Lewis RA, Saperstein DA, Sullivan LS (2001) Prevalence of mutations causing retinitis pigmentosa and other inherited retinopathies. Hum Mutat 17(1):42–51. doi:10.1002/1098-1004(2001)17:1<42::AID-HUMU5>3.0.CO;2-K [pii] 10.1002/1098- 1004 (2001)17:1<42::AID-HUMU5>3.0.CO;2-K
8. Furukawa T, Morrow EM, Cepko CL (1997) Crx, a novel otx-like homeobox gene, shows photoreceptor-specific expression and regulates photoreceptor differentiation. Cell 91(4):531–541. doi:S0092-8674(00)80439-0 [pii]
9. Ruiz A, Winston A, Lim YH, Gilbert BA, Rando RR, Bok D (1999) Molecular and biochemical characterization of lecithin retinol acyltransferase. J Biol Chem 274(6):3834–3841
10. Redmond TM, Yu S, Lee E, Bok D, Hamasaki D, Chen N, Goletz P, Ma JX, Crouch RK, Pfeifer K (1998) Rpe65 is necessary for production of 11-cis-vitamin A in the retinal visual cycle. Nat Genet 20(4):344–351. doi:10.1038/3813
11. O'Byrne SM, Wongsiriroj N, Libien J, Vogel S, Goldberg IJ, Baehr W, Palczewski K, Blaner WS (2005) Retinoid absorption and storage is impaired in mice lacking lecithin:retinol acyltransferase (LRAT). J Biol Chem 280(42):35647–35657. doi:M507924200 [pii] 10.1074/jbc. M507924200
12. Thompson DA, Janecke AR, Lange J, Feathers KL, Hubner CA, McHenry CL, Stockton DW, Rammesmayer G, Lupski JR, Antinolo G, Ayuso C, Baiget M, Gouras P, Heckenlively JR, den Hollander A, Jacobson SG, Lewis RA, Sieving PA, Wissinger B, Yzer S, Zrenner E, Utermann G, Gal A (2005) Retinal degeneration associated with RDH12 mutations results from decreased 11-cis retinal synthesis due to disruption of the visual cycle. Hum Mol Genet 14(24):3865–3875. doi:ddi411 [pii] 10.1093/hmg/ddi411
13. Haeseleer F, Jang GF, Imanishi Y, Driessen CA, Matsumura M, Nelson PS, Palczewski K (2002) Dual-substrate specificity short chain retinol dehydrogenases from the vertebrate retina. J Biol Chem 277(47):45537–45546. doi:10.1074/jbc.M208882200 M208882200 [pii]
14. Sohocki MM, Sullivan LS, Mintz-Hittner HA, Birch D, Heckenlively JR, Freund CL, McInnes RR, Daiger SP (1998) A range of clinical phenotypes associated with mutations in CRX, a photoreceptor transcription-factor gene. Am J Hum Genet 63(5):1307–1315
15. Li A, Jiao X, Munier FL, Schorderet DF, Yao W, Iwata F, Hayakawa M, Kanai A, Shy Chen M, Alan Lewis R, Heckenlively J, Weleber RG, Traboulsi EI, Zhang Q, Xiao X, Kaiser-Kupfer M, Sergeev YV, Hejtmancik JF (2004) Bietti crystalline corneoretinal dystrophy is caused by mutations in the novel gene CYP4V2. Am J Hum Genet 74(5):817–826. doi:10.1086/383228 S0002-9297(07)64351-1 [pii]
16. Lin J, Nishiguchi KM, Nakamura M, Dryja TP, Berson EL, Miyake Y (2005) Recessive mutations in the CYP4V2 gene in East Asian and Middle Eastern patients with Bietti crystalline corneoretinal dystrophy. J Med Genet 42(6):e38. doi:42/6/e38 [pii] 10.1136/jmg.2004.029066
17. Wang Y, Guo L, Cai SP, Dai M, Yang Q, Yu W, Yan N, Zhou X, Fu J, Guo X, Han P, Wang J, Liu X (2012) Exome sequencing identifies compound heterozygous mutations in CYP4V2 in a pedigree with Retinitis Pigmentosa. PLoS One 7(5):e33673. doi:10.1371/journal. pone.0033673 PONE-D-11-23089 [pii]
18. Riazuddin SA, Iqbal M, Wang Y, Masuda T, Chen Y, Bowne S, Sullivan LS, Waseem NH, Bhattacharya S, Daiger SP, Zhang K, Khan SN, Riazuddin S, Hejtmancik JF, Sieving PA, Zack DJ, Katsanis N (2010) A splice-site mutation in a retina-specific exon of BBS8 causes nonsyndromic retinitis pigmentosa. Am J Hum Genet 86(5):805–812. doi:S0002-9297(10)00202-8 [pii] 10.1016/j.ajhg.2010.04.001

19. McGee TL, Seyedahmadi BJ, Sweeney MO, Dryja TP, Berson EL (2010) Novel mutations in the long isoform of the USH2A gene in patients with Usher syndrome type II or non-syndromic retinitis pigmentosa. J Med Genet 47(7):499–506. doi:jmg.2009.075143 [pii] 10.1136/jmg.2009.075143
20. Bainbridge JW, Smith AJ, Barker SS, Robbie S, Henderson R, Balaggan K, Viswanathan A, Holder GE, Stockman A, Tyler N, Petersen-Jones S, Bhattacharya SS, Thrasher AJ, Fitzke FW, Carter BJ, Rubin GS, Moore AT, Ali RR (2008) Effect of gene therapy on visual function in Leber's congenital amaurosis. N Engl J Med 358(21):2231–2239. doi:NEJMoa0802268 [pii] 10.1056/NEJMoa0802268
21. Cideciyan AV, Aleman TS, Boye SL, Schwartz SB, Kaushal S, Roman AJ, Pang JJ, Sumaroka A, Windsor EA, Wilson JM, Flotte TR, Fishman GA, Heon E, Stone EM, Byrne BJ, Jacobson SG, Hauswirth WW (2008) Human gene therapy for RPE65 isomerase deficiency activates the retinoid cycle of vision but with slow rod kinetics. Proc Natl Acad Sci USA 105(39):15112–15117. doi:0807027105 [pii] 10.1073/pnas.0807027105
22. Maguire AM, Simonelli F, Pierce EA, Pugh EN Jr, Mingozzi F, Bennicelli J, Banfi S, Marshall KA, Testa F, Surace EM, Rossi S, Lyubarsky A, Arruda VR, Konkle B, Stone E, Sun J, Jacobs J, Dell'Osso L, Hertle R, Ma JX, Redmond TM, Zhu X, Hauck B, Zelenaia O, Shindler KS, Maguire MG, Wright JF, Volpe NJ, McDonnell JW, Auricchio A, High KA, Bennett J (2008) Safety and efficacy of gene transfer for Leber's congenital amaurosis. N Engl J Med 358(21):2240–2248. doi:NEJMoa0802315 [pii] 10.1056/NEJMoa0802315
23. Kurg A, Tonisson N, Georgiou I, Shumaker J, Tollett J, Metspalu A (2000) Arrayed primer extension: solid-phase four-color DNA resequencing and mutation detection technology. Genet Test 4(1):1–7. doi:10.1089/109065700316408
24. Asper Biotech: http://www.asperbio.com
25. Avila-Fernandez A, Cantalapiedra D, Aller E, Vallespin E, Aguirre-Lamban J, Blanco-Kelly F, Corton M, Riveiro-Alvarez R, Allikmets R, Trujillo-Tiebas MJ, Millan JM, Cremers FP, Ayuso C (2010) Mutation analysis of 272 Spanish families affected by autosomal recessive retinitis pigmentosa using a genotyping microarray. Mol Vis 16:2550–2558. doi:272 [pii]
26. Blanco-Kelly F, Garcia-Hoyos M, Corton M, Avila-Fernandez A, Riveiro-Alvarez R, Gimenez A, Hernan I, Carballo M, Ayuso C (2012) Genotyping microarray: mutation screening in Spanish families with autosomal dominant retinitis pigmentosa. Mol Vis 18:1478–1483
27. 454 Sequencing: http://454.com
28. SOLiD Sequencing: http://www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-next-generation-sequencing.html
29. Illumina Sequencing: http://www.illumina.com
30. Looi M-K (2009) Genomics - the next generation. http://www.wellcome.ac.uk/news/2009/features/wtx056032.htm
31. Ion Torrent Sequencing: http://www.iontorrent.com
32. Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X, Richmond TA, Middle CM, Rodesch MJ, Packard CJ, Weinstock GM, Gibbs RA (2007) Direct selection of human genomic loci by microarray hybridization. Nat Methods 4(11):903–905. doi:nmeth1111 [pii] 10.1038/nmeth1111
33. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME (2007) Microarray-based genomic selection for high-throughput resequencing. Nat Methods 4(11):907–909. doi:nmeth1109 [pii] 10.1038/nmeth1109
34. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760. doi:btp324 [pii] 10.1093/bioinformatics/btp324
35. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, Coarfa C, Milosavljevic A, Gibbs RA, Yu F (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13:8. doi:1471-2105-13-8 [pii] 10.1186/1471-2105-13-8

36. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303. doi:gr.107524.110 [pii] 10.1101/gr.107524.110
37. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, Liu Y, Weinstock GM, Wheeler DA, Gibbs RA, Yu F (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res 20(2):273–280. doi:gr.096388.109 [pii] 10.1101/gr.096388.109
38. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814
39. Raca G, Jackson C, Warman B, Bair T, Schimmenti LA (2010) Next generation sequencing in research and diagnostics of ocular birth defects. Mol Genet Metab 100(2):184–192. doi:S1096-7192(10)00095-8 [pii] 10.1016/j.ymgme.2010.03.004
40. Neveling K, Collin RW, Gilissen C, van Huet RA, Visser L, Kwint MP, Gijsen SJ, Zonneveld MN, Wieskamp N, de Ligt J, Siemiatkowska AM, Hoefsloot LH, Buckley MF, Kellner U, Branham KE, den Hollander AI, Hoischen A, Hoyng C, Klevering BJ, van den Born LI, Veltman JA, Cremers FP, Scheffer H (2012) Next-generation genetic testing for retinitis pigmentosa. Hum Mutat 33(6):963–972. doi:10.1002/humu.22045
41. O'Sullivan J, Mullaney BG, Bhaskar SS, Dickerson JE, Hall G, O'Grady A, Webster A, Ramsden SC, Black GC (2012) A paradigm shift in the delivery of services for diagnosis of inherited retinal disease. J Med Genet 49(5):322–326. doi:jmedgenet-2012-100847 [pii] 10.1136/jmedgenet-2012-100847
42. Pieras JI, Barragan I, Borrego S, Audo I, Gonzalez-Del Pozo M, Bernal S, Baiget M, Zeitz C, Bhattacharya SS, Antinolo G (2011) Copy-number variations in EYS: a significant event in the appearance of arRP. Invest Ophthalmol Vis Sci 52(8):5625–5631. doi:iovs.11-7292 [pii] 10.1167/iovs.11-7292
43. Wang J, Mullighan CG, Easton J, Roberts S, Heatley SL, Ma J, Rusch MC, Chen K, Harris CC, Ding L, Holmfeldt L, Payne-Turner D, Fan X, Wei L, Zhao D, Obenauer JC, Naeve C, Mardis ER, Wilson RK, Downing JR, Zhang J (2011) CREST maps somatic structural variation in cancer genomes with base-pair resolution. Nat Methods 8(8):652–654. doi:nmeth.1628 [pii] 10.1038/nmeth.1628
44. Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat Methods 6(9):677–681. doi:nmeth.1363 [pii] 10.1038/nmeth.1363
45. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics 25(21):2865–2871. doi:btp394 [pii] 10.1093/bioinformatics/btp394
46. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT (2012) Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7(2):e30087. doi:10.1371/journal.pone.0030087 PONE-D-11-17842 [pii]

Chapter 11
Next-Generation Sequencing Analyses of the Whole Mitochondrial Genome

Lee-Jun C. Wong

Abstract Molecular diagnosis of mitochondrial DNA (mtDNA)-related disorders requires the detection and quantification of point mutations and large deletions, including mapping the deletion breakpoints. Currently, comprehensive diagnosis is achieved by employing stepwise procedures. The massively parallel next-generation sequencing (NGS) when appropriately validated, with deep coverage and proper quality controls, can be used as a one-step comprehensive diagnostic approach to simultaneously detect and quantify mtDNA point mutations and deletions in a CLIA-certified clinical laboratory.

1 Introduction

Mitochondria are the only cellular organelles that contain their own genetic material. Most human cell contains hundreds to thousands of mitochondria [1], each of which contains multiple copies of the 16,569 bp circular double-stranded mitochondrial DNA molecule. The number of mitochondria per cell depends on the energy demand of the specific tissue. Since there are multiple copies of mtDNA, if an mtDNA mutation occurs, the mutant mtDNA often coexists with the wild-type mtDNA, a phenomenon called "heteroplasmy." The degree of heteroplasmy of a mutation, nature of the specific mutation, and its tissue distribution determine the clinical phenotype of the affected patient [2, 3]. Phenotype may also be modified by genetic background and environmental factors.

L.-J.C. Wong, Ph.D. (✉)
Medical Genetics Laboratories, Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza NAB 2015, Houston, TX 77030, USA
e-mail: ljwong@bcm.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_11,

Unlike the nuclear genes, mitochondrial genome contains no introns in the protein-coding regions. The entire mitochondrial genome, encoding a total of 37 genes, is efficiently utilized. Polycistronic messages are produced from the mtDNA. Genes reside on both stands of the circular mitochondrial genome. The ATP6 and ATP8 genes even share part of their coding regions in different reading frames [4].

The 13 proteins encoded by the mtDNA are all components of the respiratory chain complexes. The mitochondrial genome also encodes two ribosomal RNAs and 22 tRNAs. Mutations in the rRNA and tRNA may also cause disease (http://www.mitomap.org/MITOMAP). Indeed, the majority of pathogenic mutations reside in the tRNA genes. For example, the common m.3243A>G mutation in the $tRNA^{Leu(UUR)}$ gene is the most frequent cause of MELAS (mitochondrial encephalopathy lactic acidosis and stroke-like episodes) syndrome. The bacterial like rRNAs are similarly sensitive to some antibiotics that target bacterial ribosomes. Thus, the m.1555A>G in the 12S rRNA gene is associated with ototoxicity-induced hearing loss.

In addition to the dense coding regions, there is also an approximately 1.1 kb noncoding displacement loop (D-loop) region where the origins of replication are located.

The purpose of molecular diagnosis of mtDNA disorders is to identify the deleterious changes of mtDNA sequences that contribute to the disease [5]. Two types of mtDNA mutations are usually analyzed; mtDNA point mutations and large mtDNA deletions. While there are common recurrent point mutations, rare or novel pathogenic mutations [5] do occur. Therefore, the diagnostic analysis of the mtDNA usually includes the whole mitochondrial genome.

The mtDNA deletions may be a single large deletion or multiple deletions. Since these mutations lead to malfunction of the electron transport chain, there is frequently multisystem involvement. Historically, different methods have been required for the detection of point mutations and deletions, and for the quantification of mutation heteroplasmy, and the determination of deletion breakpoints [3, 5]. This chapter briefly reviews the conventional molecular diagnostic methods employed in the analysis of mtDNA disorders and then describes the comprehensive one-step approach enabled by the application of next-generation sequencing (NGS) technology.

2 Conventional Methods for the Diagnosis of Mitochondrial DNA Disorders

The methods used for the detection of point mutations are usually PCR-based, while the detection of large deletions is usually achieved by traditional Southern blotting methodology [2]. Since the degree of mutation heteroplasmy is critical in disease diagnosis, prognosis, and genetic counseling, various quantification techniques are used for the measurement of heteroplasmy after the detection of a deleterious

Table 11.1 Current molecular procedures used for the diagnosis of the mitochondrial genome

Method	Area of detection	Limitation
PCR-based ASO[a] dot blot or RFLP[b]	Detection of specific point mutations	1. Use of radioactive material 2. Targeted positions only 3. Nonquantitative
ARMS[c] qPCR	Quantification of mutation heteroplasmy	1. Must be validated first for any given mutation 2. For specific nucleotide positions only 3. Differential PCR efficiency due to mutant and wild-type-specific primers causes large variations
Southern blot analysis	Large deletions, rearrangement	1. Use of radioactive material 2. low sensitivity for detection and quantification 3. Tedious, time consuming
Oligonucleotide array CGH[d]	Large deletions, copy number changes Estimate deletion break points and % heteroplasmy	1. Need PCR/sequencing to obtain exact deletion break points
Sanger sequencing	Detect all known and unknown point mutations and small indels	1. Non-quantitative 2. Detection limit for heteroplasmy is about 15 % 3. Does not detect large deletions

[a]ASO, allele specific oligonucleotide
[b]RFLP, restriction fragment length polymorphism
[c]ARMS, allele specific refractory mutation system
[d]CGH, comparative genome hybridization

mutation [6]. Table 11.1 lists the current stepwise molecular procedures that are required for a comprehensive analysis of the mitochondrial DNA [5]. Pitfalls of each method are also listed (Table 11.1).

2.1 Common Point Mutations: Detection and Quantification

Patients suspected of having maternally inherited mtDNA disorders are usually first screened for the common point mutations by PCR-based assays including RFLP [2, 3, 7] and allele specific oligonucleotide (ASO) dot blot hybridization methods as described in Chap. 2 [7]. Using radioactive probes, the ASO dot blot hybridization method is sensitive enough to detect point mutations at as low as 1 % [7, 8]. If a common point mutation is identified, analysis of the level of heteroplasmy can be carried out, usually by allele refractory mutation system-based quantitative PCR (ARMS qPCR) for the quantitative measurement [6, 8]. These methods require validation and are limited to the analyses of known point mutations [6, 8].

2.2 Detection of Unknown mtDNA Point Mutations

If the common point mutations and large deletions (see Sect. 2.3 below) are not detected, and the maternal inheritance of the disease is still hypothesized, the whole mitochondrial genome is analyzed by Sanger sequencing, which is performed following PCR amplification of the entire mitochondrial genome with multiple pairs of overlapping primers [9, 10]. Sanger sequencing was the gold standard for the identification of unknown mutations for many years before the advent of massively parallel sequence analysis. Sanger sequencing is not a quantitative method, it does not detect mutations at low heteroplasmic levels, and it does not detect large deletions [11–13]. In addition, PCR-based methods will not accurately detect the sequence under the primer binding sites.

2.3 Detection of mtDNA Deletions

Large deletions in mtDNA are detected by Southern blot analysis. Unfortunately, this is a tedious procedure that does not provide deletion breakpoints or the degree of deletion heteroplasmy. These deficiencies can be addressed by array-comparative genome hybridization (aCGH), which not only detects the deletion but also provides deletion breakpoints and an estimate of deletion heteroplasmy [11–13].

2.4 Detection of mtDNA Multiple Deletions

By their very nature, multiple mtDNA deletions are difficult to detect. The individual molecular species are often present at low levels, challenging methods with low sensitivity, such as Southern analysis. Conversely, PCR-based assays will amplify molecules to detectable levels but are very dependent upon choice of primer sites and may fail in the presence of sequence variations. In this method, multiple pairs of primers are employed in order to evaluate suspected regions on the mitochondrial genome. Primer pairs and PCR conditions are selected such that amplification will only occur when a primer pair encompasses a deletion. If multiple deletions are present, it is possible to amplify multiple fragments.

2.5 Determination of Deletion Junctions

Primers outside the approximate deletion regions are designed. Since the exact deletion breakpoints are usually not defined before the junction is sequenced, several pairs of primers covering different ranges of possible deletions have to be tested to

find the approximate limits of the deletion [14]. PCR products are then purified followed by Sanger sequencing [14]. These procedures are time consuming and labor intensive even for single deletions. The presence of multiple deletions exacerbates this situation; more primer pairs must be tested, and more PCR products must be purified and sequenced. Even after extensive efforts, these procedures may not determine all of the different breakpoints. The detection and characterization of multiple deletions is greatly simplified by the adoption of massively parallel sequencing of the entire mitochondrial genome with uniformly deep coverage [15].

3 NGS-Based Analyses

3.1 Target Gene Enrichment

Methods for target gene enrichment including PCR-based and capture-based have been described in previous chapters in this book. Since the mitochondrial genome is small (16.6 kb) and does not contain any introns, the enrichment is usually achieved by PCR, which may use 24–36 pairs of primers [10, 16, 17] to amplify short overlapping regions or 2–3 pairs of primers for long-range PCR (LR-PCR) [18–20]. Recently, we have designed LR-PCR primers to generate a single amplicon of the entire mitochondrial genome [15, 21].

Enrichment of the mitochondrial genome by capture in solution using RNA or DNA probes has been reported [20, 22–24]. However, the coverage profile showed that different parts of the mitochondrial genome are not captured and sequenced uniformly [15, 22]. Therefore, it is not possible to detect large deletions or low heteroplasmic variants from these sequence data. Multiple copies of mitochondrial pseudogenes are stranded on each of the nuclear chromosomes [25–27]. These nuclear mitochondrial sequences (NUMTs) are subject to genetic drift and therefore produce a significant background of sequence variants that must be contended with in order to discern the true mtDNA sequence. In addition, due to the abundance of NUMTs, exome capture/sequencing will co-capture NUMTs even in the absence of mtDNA-specific probes. Thus, interference from NUMT sequences may result in incorrect sequence information and/or errors in the quantification of mtDNA heteroplasmy [25–27].

3.2 Platforms of Massively Parallel Sequencing

Massively parallel sequencing can be performed using various platforms, including 454, SOLiD, Affymetrix re-sequencing chip, Illumina, and IonTorent. The utilization of different MPS sequencing chemistries and machine hardware configurations has been reviewed [21, 28–32]. Each method has its own advantages and

disadvantages [21], and the properties of the mitochondrial genome influence selection of MPS methodology.

The mitochondrial genome contains a number of homopolymeric stretches, high GC content regions, and short tandem repeats. Since low heteroplasmy of deleterious mutations, including small indels in repeat regions, can be clinically significant, it is important to understand the limitations of each different sequencing method. Different platforms may also affect the depth of coverage and the ability to multiplex. Proper controls should be included and analyzed together with each indexed specimen to ensure accuracy [15]. Limit of detection of NGS-based assays should be assessed since quantification of mtDNA mutation heteroplasmy is an important analytical component. Different platforms provide different depth of sequence coverage, which may limit heteroplasmy detection [15, 18–20, 22].

4 Reported Studies of the Mitochondrial Genome by MPS

Although this chapter focuses on the translation of massively parallel sequencing (MPS) to the clinical diagnosis of mtDNA disorders, various studies have utilized high throughput MPS analyses of the mitochondrial genome for different purposes (Table 11.2).

4.1 Detection of Pathogenic Point Mutations and Evaluation of Heteroplasmy

For the purpose of translation to molecular diagnosis, MPS was validated for its ability to simultaneously detect and quantify mtDNA variants of the entire mitochondrial genome by comparing the MPS results to those of Sanger sequencing [15, 19]. MPS has also been used to identify mtDNA variants in mtDNA-related disorders, including left ventricular noncompaction (LVNC) [24], maternally inherited cardiomyopathy [20], and Leigh syndrome [18]. However, these studies were limited to the detection of mtDNA variants at >5 % heteroplasmy in the analyzed tissue.

Due to the nonuniformity of mtDNA coverage, detection of mtDNA deletions was not possible in the previously reported studies [18–20, 24]. An mtDNA single deletion at high heteroplasmy (94 %) and >25,000X coverage was detected by capture using RNA probes followed by MPS [22], in which coding regions of 1,300 nuclear genes involved in mitochondrial production and function were also captured and sequenced [22, 38]. Simultaneous detection of mutations in both the mitochondrial and the nuclear genes is the main goal of MPS-based comprehensive diagnosis for dual genome dysfunction, that is, conditions in which nuclear gene mutations perturb the mitochondrial genome. However, the application of the dual genome MPS approach in one step has not been fully validated for clinical diagnosis [22, 38].

Table 11.2 Analyses of the human mitochondrial genome using massively parallel sequencing (MPS)

	Purpose	Method of enrichment	MPS platform	Average coverage	Uniformity/ deletion	Number of samples and experiment	Multiplex factor	Results	Ref
(a) *Detection of point mutations*									
1	Simultaneous detection and quantification of mtDNA[a]	PCR 2 amplicons equimolar mix	Illumina GAII	1,785	Yes	2, mixed in 1, 5, 10, 20, and 50 %	16/lane	Heteroplasmy >5 % can be detected	[19]
2	To investigate if mtDNA[a] mutation is the primary cause for LVNC[b]	PCR 2 amplicons 9,289 + 7,626 bp	Illumina GAII	634	NA[c]	20	NA[c]	Found a few rare variants, none were proven primarily pathogenic	[24]
3	To evaluate mtDNA mutation as the cause of MI-HCM[d] and MI-DCM[e] Compared to Sanger	PCR 3 amplicons 6,929,7,050,6,866 bp	Roche FLX454	1,300	Variable coverage	20	10	A rare variant m.7501 T>C was found 98 % concordance with Sanger	[20]
4	To evaluate mtDNA mutation in a family with MILS[f]	PCR2 amplicons 9,731 + 12,038 bp	Roche FLX454	182	NA[c]	5 affected members of 1 family	NA[c]	m.8993 T>C mutation at 70–100 % heteroplasmy was detected in family members	[18]
5	To make molecular diagnosis of infantile mitochondrial disease	MitoExome + mtDNA, RNA probe hybridization	Illumina GAII	25,457	Variable coverage	42	1/lane	10 patients had definite AR[g] nuclear gene mutations, 1 had a 7.2 kb mtDNA deletion	[22]

(continued)

Table 11.2 (continued)

	Purpose	Method of enrichment	MPS platform	Average coverage	Uniformity/ deletion	Number of samples and experiment	Multiplex factor	Results	Ref
6	Detection and quantification of point mutations and deletions	16.6 kb single amplicon	Illumina HiSeq2000	>20,000	Yes, break-points mapped	>50[h]	12/lane	100 % sensitivity, specificity, accuracy, and reproducibility LOD[i] =1.3 %	[15]
(b) *mtDNA variability in cancer cells and among different tissues*									
7	To investigate the mtDNA variabilities during embryo-genesis and tumorigenesis	2 sets of 50 pairs each, ssDNA probes	Illumina GAII	16,700	uneven	10 colon cancer, 9 organs of 1 patient, 2 families	Not clear	Heteroplasmies detected in different tissues of the same individual and different individuals of the same matrilineal relatives	[33]
8	To investigate the dynamics of mtDNA hetero-plasmy in maternal transmission	PCR 2 amplicons 8,757 + 9,143 bp	Illumina	1,170	NA	9 individuals 2 tissues each and repeat PCR	NA	Incidence of heteroplasmy may be lower than estimated, heteroplasmy varies among different tissues and between mother and child	[34]

9	To study the effect of radiation therapy on mtDNA alteration	PCR (detail not given)	Illumina GAII	3,981	NA	44 (18mom+26 children) samples	NA	Heteroplasmy positively correlated with mother's age, 9/18 inherited heteroplasmy	[35]
10	Detection of mtDNA heteroplasmy in 131 Eurasian individuals	PCR 2 amplicons 7.3+9.7 kb	Illumina GAII	65–211	NA	131	High multi-plex	37 heteroplasmies at >10 % at 34 sites in 32 individuals detected	[36]
11	Demographic inference from 147 Caucasians and West Asians	PCR 2 amplicons 7.3+9.7 kb	Illumina GAII	87	NA	147	50 per lane	BSP[j] of population size change through time showed history of population expansion	[37]

[a]mtDNA, mitochondrial DNA or mitochondrial genome
[b]LVNC, left ventricular noncompaction
[c]Not available
[d]MI-HCM, maternally inherited hypertrophic cardiomyopathy
[e]MI-DCM, maternally inherited dilated cardiomyopathy
[f]MILS, maternally inherited Leigh syndrome
[g]AR, autosomal recessive
[h]Up-to-date, >800 samples have been analyzed in the author's laboratory
[i]LOD, limit of detection
[j]BSP, Bayesian skyline plots

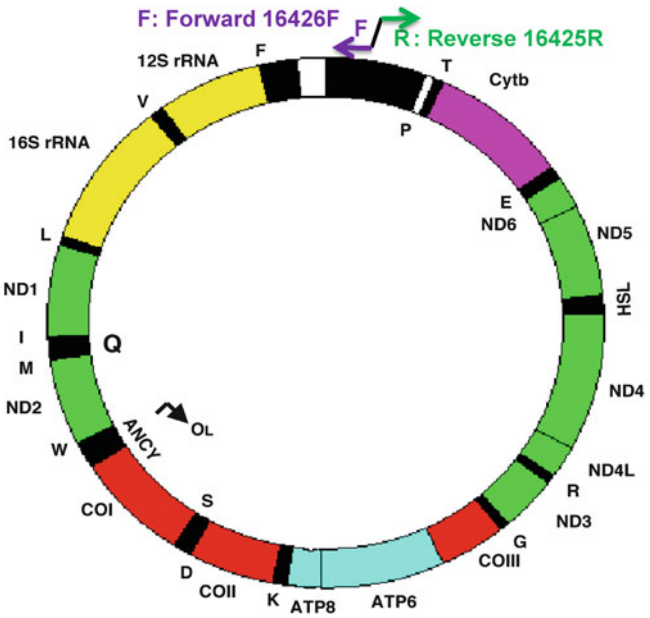

Fig. 11.1 The primer positions for the LR-PCR of the whole mitochondrial genome

Since the human mitochondrial genome is only 16.6 kb, high throughput MPS can usually provide excessive depth of coverage. Therefore, if only the mitochondrial genome is to be sequenced, multiple samples are usually multiplexed in order to efficiently use the capacity of the high throughput instrument (Table 11.2). The read length and depth of coverage vary according to the NGS platform used.

4.2 Comprehensive One-Step Analysis of the Whole Mitochondrial Genome

Our laboratory recently developed a one-step MPS approach that provides quantitative base calls, detection of large deletions, and the exact deletion breakpoints [15]. This approach uses one pair of primers (Fig. 11.1) – mt16426F-5′ccgcacaagagtgctactctcctc3′ and mt16425R-5′gatattgatttcacggaggatggtg3′ – for the long-range PCR (LR-PCR) amplification of the entire mitochondrial genome as a single amplicon, followed by library preparation and sequencing on Illumina HiSeq 2000 [15]. Since the whole mitochondrial genome is amplified as one single amplicon, every base is presumably represented in proportion to its occurrence in the starting population of molecules (Fig. 11.2a). However, molecules containing deletions can have a replicative advantage in limiting conditions. Therefore, while deletions are readily detected (Fig. 11.2b), their heteroplasmy level may be

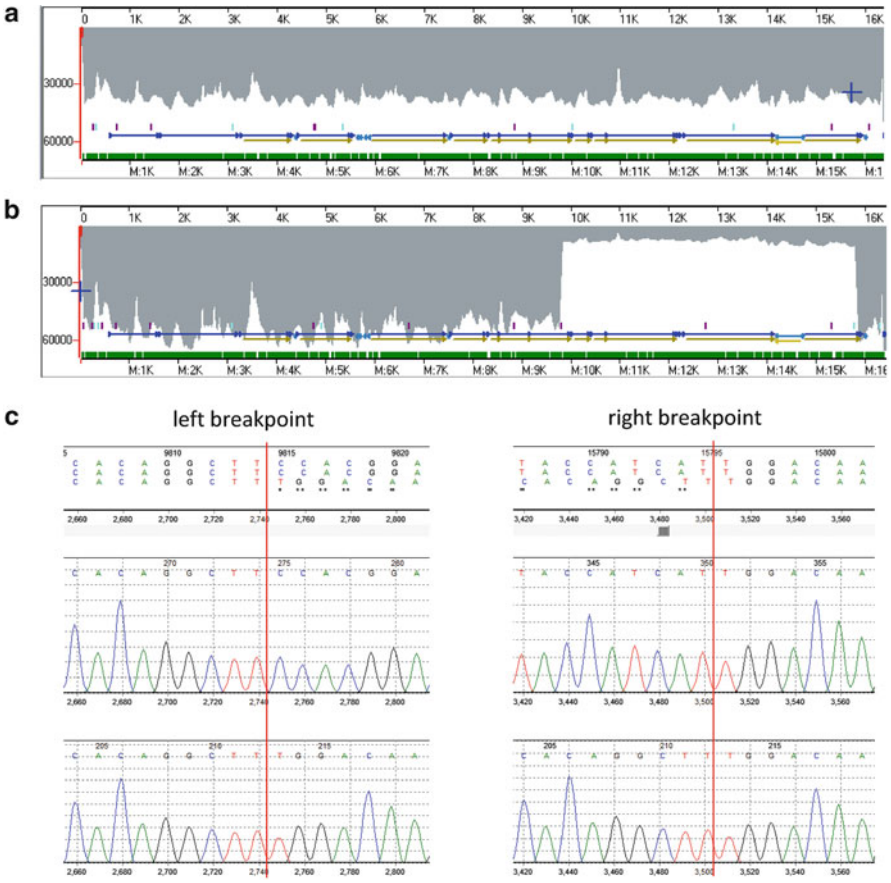

Fig. 11.2 The uniform coverage of the whole mitochondrial genome (**a**), the sharp deletion boundaries detected by MPS (**b**), and the deletion breakpoint sequences (**c**). Provided by Dr. Hui Yu

overestimated. By aligning the unmatched sequences to the mitochondrial reference sequence (rCRS) with less stringent parameters to allow >80 % match in half of the sequence read, the deletion breakpoints can be precisely mapped (Fig. 11.2c). This is a great advantage in contrast to conventional Southern analysis of large deletions where the detection of deletion and determination of the breakpoints are two separate tedious procedures.

By multiplexing 12 samples per lane of the flow cell and 76 cycles of sequencing, an average coverage per base of ~20,000X can be routinely achieved [15]. At this depth of coverage and 0.326 % ± 0.335 % experimental error rate, heteroplasmies >1.5 % are easily detected. For the purpose of quality control, a sample with 1.1 % heteroplasmic m.3243A > G mutation and a series of reference DNA samples containing 1 %, 5 %, 10 %, 20 %, and 50 % of known variants have always been spiked-in and analyzed exactly the same way as the same indexed patient's sample

[15]. It has been demonstrated that the 1.1 % m.3243A>G control has always been detected at 1.14 % ± 0.09 % (unpublished observation). To date, more than 800 samples have been analyzed using this comprehensive one-step approach. Numerous homoplasmic or heteroplasmic variants have been detected. Most of these variants are reported benign SNPs. About 6 % of the samples analyzed harbored reported pathogenic mutations, and only <1 % are novel variants that are likely to be deleterious. Since not all heteroplasmic novel variants are clinically significant, other genetic, biochemical, pedigree, and clinical information, as well as results obtained from in silico analyses using protein structural/functional prediction algorithms, are used to help with the interpretation of these variants [39, 40].

The translation of this next-generation sequencing approach to the diagnosis of mtDNA-related disorders has been fully validated according to the regulatory criteria for the clinical diagnostic laboratories set forth by CLIA (Clinical Laboratory Improvement Amendments) and CAP (College of American Pathologists). All necessary quality and quantity reference samples are incorporated with the analyses of every test sample. All variants detected by Sanger methodology have been detected by this one-step comprehensive MPS method. In addition, this method detects low heteroplasmy changes that are not detected by Sanger sequencing. Thus far, this MPS strategy is the most comprehensive approach for the provision of accurate, reproducible heteroplasmy measurements of variants at every nucleotide position of the entire mitochondrial genome. Furthermore, large mtDNA deletions are detected and the deletion breakpoints are easily mapped [15].

4.3 MPS Investigation of mtDNA Variations in Cancers, Various Tissues, and Among Different Populations

In addition to molecular diagnosis of mtDNA disorders, MPS has also been employed to analyze the whole mitochondrial genome for the purposes of disease prognosis in cancers. Somatic mtDNA alterations in tumor cells can serve as biomarkers for monitoring disease progression [41–44]. Traditionally, mtDNA alterations in cancer cells were analyzed by Sanger sequencing of overlapping short PCR fragments [41–44]. This is tedious if a large number of tumor samples are to be analyzed. A recent report took advantage of massively parallel sequencing to analyze mtDNA of ten colon cancers, nine different tissues of one patient, and members of two CEPH families [33]. Their results revealed variable heteroplasmic mtDNA alterations in colon cancers, in different tissues of the same individual, and in different matrilineal relatives [33]. The degree of heteroplasmic changes varies from 1.6 % to 57 %. These studies provide insights into the nature and variability of mtDNA sequences during embryogenesis and cancer development and demonstrate that human individuals are characterized by a complex mixture of related mitochondrial genotypes rather than a single genotype. However, these studies were performed for research purposes, not for molecular diagnostics.

Similarly, MPS was used to study the effects of radiation therapy on mtDNA alteration [35] and to investigate the dynamics of mtDNA heteroplasmy in maternal transmission [34]. The latter study showed that frequencies of heteroplasmic changes may be lower than previously estimated but agreed with the results of He and coworkers [35] that mtDNA heteroplasmy varies among different tissues of an individual and between mother and child [34]. Furthermore, studies of mtDNA in 131 individuals from five Eurasian populations [36] and 147 individuals from the Caucasus and West Asia [37] revealed that mtDNA heteroplasmies are common and variable among populations. These studies involved a large number of samples and the detection of low level heteroplasmies. Only high throughput, deep coverage sequencing techniques allow these types of studies to be performed in a cost- and time-efficient fashion. These results also suggest caution when mtDNA variants are used for forensic identity verification purposes due to the dynamic occurrence of heteroplasmic changes [34]. However, since the NGS technologies used in these studies have not been evaluated with forensic standards, the application of MPS to forensic investigation requires further assessment [45]. A major concern of these MPS-based applications is that these studies did not discuss the potential interference of nuclear mtDNA homologues (NUMT), which may result in incorrect variant calls or inaccurate heteroplasmy measurements.

5 Regulatory Requirements for the Application of NGS-Based Tests to the Clinical Diagnosis of mtDNA Disorders

Testing of human specimens for diagnostic purposes must follow the regulatory procedures defined by the Clinical Laboratory Improvement Amendments (CLIA), which requires the assessment and documentation of performance characteristics including sensitivity, specificity, accuracy, reproducibility, and any other unique procedures applicable to the analytic validity of the test results. Due to the enormous amount of data produced on each test and the complex laboratory and computational analytical procedures involved, it is difficult to define the standards required for compliance of this newly developed technique with the CLIA regulation. An NGS guideline work-group has been actively exploring these issues. In general, before applying NGS-based tests clinically, the tests must be validated. In particular, for the diagnosis of mtDNA disorders by MPS, the specific parameters for the evaluation of the analytical performance of an NGS run should include depth of coverage, uniformity of distribution of read coverage, poorly covered regions, or base positions (e.g., small indels, repeat and homopolymer regions), quality of base calls, ability to detect mtDNA large deletions, and limit of detection of mutation heteroplasmy. Review of the published papers on the NGS-based analyses of the mtDNA (Table 11.2) revealed that most of these NGS-based assays were not fully validated for clinical diagnosis except for the one-step comprehensive approach reported by Zhang et al. [15].

Zhang and coworkers assessed the performance of NGS-based analysis of mtDNA by comparing the results to those obtained from Sanger sequencing and demonstrated 100 % sensitivity and specificity [15]. Since the measurement of the degree of mutation heteroplasmy is critical in result interpretation and genetic counseling, limit of detection of the NGS approach and the reproducibility of the quantified results from various runs were evaluated. Most importantly, this report described the incorporation of quality and quantity reference specimens with each indexed sample for simultaneous evaluation to ensure the accuracy and reproducibility of the results [15]. Furthermore, a "deep sequencing index" (DSI) formula was developed to evaluate the performance of each sequencing run and to compare the quality of sequencing results among different gene enrichment methods. This equation contains six parameters: (i) the mean number of reads mapped to quality control (QC) DNA, (ii) the mean number of sample reads normalized to the average number of reads of QC DNA, (iii) the correlation coefficient of the expected versus observed proportion of 6 QC DNA variants mixed by known ratios, (iv) the ratio of the standard deviation of the mean number of reads to the average number of reads mapped to sample DNA, (v) the analytical specificity, and (vi) the analytical sensitivity of a run determined from the reads mapped to mtDNA. The analytical specificity of a run was defined as the percentage of reads mapped to target mtDNA reference sequence compared to total reads generated for the sample, which, for the capture-based enrichment, is ~20 % and for the single amplicon LR-PCR based is ~99 %. The analytical sensitivity of a run was defined as the percentage of bases of the reference sequence covered by MPS reads. The analytical sensitivity should be 100 % to achieve 0 % false negative. Thus, it is clear that all performance parameters specific for the novel NGS technology are included in this formula for quality assessment. Each laboratory can define its own minimal passing score that represents the acceptable quality of performance. As a result, this numerical assessment can compare and help to standardize inter-laboratory performance.

6 Caveats in Making Diagnosis

The most difficult tasks in the molecular diagnosis of mtDNA disorders are (i) simultaneous detection and quantification of heteroplasmic mtDNA point mutations, (ii) simultaneous detection and mapping of mtDNA large deletions, and (iii) simultaneous detection and quantification of mtDNA point mutations and large deletions. The MPS approach can simultaneously accomplish each of these goals if it is performed in such a way that it (i) avoids the interference of NUMTs and mtSNPs and (ii) provides even coverage of all nucleotide positions. Since NUMTs are present throughout in the nuclear chromosomes and since mtSNPs are distributed throughout in the mitochondrial genome, the only way to avoid their interferences is to amplify the whole mitochondrial genome as a single piece using a pair of primers that contain the least number of mtSNPs at the lowest frequency. This approach will also provide even coverage of every single nucleotide position [15].

Alternative primers should also be validated for deployment when the first chosen primers fail to amplify efficiently. A set of inwardly facing primers that bind outside of the single amplicon primers is also needed in order to evaluate for SNPs at the main primer binding sites.

Most of the reported studies used at least two pairs of primers to amplify overlapping regions [18, 19, 24, 33, 34, 36, 37]. Therefore, these MPS approaches are not designed to detect large mtDNA deletions. Although the coverage depth may be sufficient to provide heteroplasmic measurements, the variant calls and heteroplasmy measurements must be interpreted cautiously since the presence of SNPs at the primer binding sites and the co-amplification of NUMTs can potentially skew both variant calls and the heteroplasmy quantification.

7 Conclusions

The translation of NGS to the clinical diagnosis of mitochondrial DNA-related disorders has already occurred. One laboratory [15] has fully validated the NGS-based assay by the documentation and implementation of quality control and quality assurance procedures that are required by CLIA and CAP. Experimental errors and limit of detection should be defined before offering the NGS-based quantification of mtDNA mutation heteroplasmy as a clinical test.

Acknowledgment The author would like to thank Dr. W Zhang and Dr. H Cui, who worked together to establish the single amplicon, one-step analyses of the mitochondrial genome by NGS. I would also like to acknowledge doctors J Wang, M Landsverk, and FY Li, for their valuable discussion and assistance in the interpretation of NGS results.

References

1. Dimmock D et al (2010) A quantitative evaluation of the mitochondrial DNA depletion syndrome. Clin Chem 56(7):1119–1127
2. Shanske S, Wong LJ (2004) Molecular analysis for mitochondrial DNA disorders. Mitochondrion 4(5–6):403–415
3. Wong LJ, Boles RG (2005) Mitochondrial DNA analysis in clinical laboratory diagnostics. Clin Chim Acta 354(1–2):1–20
4. Smeitink J, van den Heuvel L, DiMauro S (2001) The genetics and pathology of oxidative phosphorylation. Nat Rev Genet 2(5):342–352
5. Wong L-JC et al (2010) Current molecular diagnostic algorithm for mitochondrial disorders. Mol Genet Metab 100(2):111–117
6. Venegas V, Halberg MC (2012) Quantification of mtDNA mutation heteroplasmy (ARMS qPCR). Methods Mol Biol 837:313–326
7. Tang S et al (2012) Analysis of common mitochondrial DNA mutations by allele-specific oligonucleotide and Southern blot hybridization. Methods Mol Biol 837:259–279
8. Bai RK, Wong LJ (2004) Detection and quantification of heteroplasmic mutant mitochondrial DNA by real-time amplification refractory mutation system quantitative PCR analysis: a single-step approach. Clin Chem 50(6):996–1001

9. Landsverk ML, Cornwell ME, Palculict ME (2012) Sequence analysis of the whole mitochondrial genome and nuclear genes causing mitochondrial disorders. Methods Mol Biol 837:281–300
10. Ware SM et al (2009) Infantile cardiomyopathy caused by a mutation in the overlapping region of mitochondrial ATPase 6 and 8 genes. J Med Genet 46(5):308–314
11. Chinault AC et al (2009) Application of dual-genome oligonucleotide array-based comparative genomic hybridization to the molecular diagnosis of mitochondrial DNA deletion and depletion syndromes. Genet Med 11(7):518–526
12. Wang J et al (2012) Targeted array CGH as a valuable molecular diagnostic approach: experience in the diagnosis of mitochondrial and metabolic disorders. Mol Genet Metab 106(2):221–230
13. Wong LJ et al (2008) Utility of oligonucleotide array-based comparative genomic hybridization for detection of target gene deletions. Clin Chem 54(7):1141–1148
14. Lacbawan F et al (2000) Clinical heterogeneity in mitochondrial DNA deletion disorders: a diagnostic challenge of Pearson syndrome. Am J Med Genet 95(3):266–268
15. Zhang W, Cui H, Wong LJ (2012) Comprehensive 1-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem 58(9):1322–1331, Epub July PMID
16. Brautbar A et al (2008) The mitochondrial 13513G>A mutation is associated with Leigh disease phenotypes independent of complex I deficiency in muscle. Mol Genet Metab 94(4):485–490
17. Wang J et al (2009) Two mtDNA mutations 14487 T>C (M63V, ND6) and 12297 T>C (tRNA Leu) in a Leigh syndrome family. Mol Genet Metab 96(2):59–65
18. Kara B et al (2012) Whole mitochondrial genome analysis of a family with NARP/MILS caused by m.8993 T>C mutation in the MT-ATP6 gene. Mol Genet Metab 107(3):389–393
19. Tang S, Huang T (2010) Characterization of mitochondrial DNA heteroplasmy using a parallel sequencing system. Biotechniques 48(4):287–296
20. Zaragoza MV et al (2010) Mitochondrial DNA variant discovery and evaluation in human Cardiomyopathies through next-generation sequencing. PLoS One 5(8):e12295
21. Zhang W, Cui H, Wong LJ (2012) Application of next generation sequencing to molecular diagnosis of inherited diseases. Top Curr Chem. http://dx.doi.org/a0.1007/128.2012.325
22. Calvo SE et al (2012) Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. Sci Transl Med 4(118):118ra110
23. Gnirke A et al (2009) Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 27(2):182–189
24. Tang S et al (2010) Left ventricular noncompaction is associated with mutations in the mitochondrial genome. Mitochondrion 10(4):350–357
25. Hirano M et al (1997) Apparent mtDNA heteroplasmy in Alzheimer's disease patients and in normals due to PCR amplification of nucleus-embedded mtDNA pseudogenes. Proc Natl Acad Sci USA 94(26):14894–14899
26. Parfait B et al (1998) Co-amplification of nuclear pseudogenes and assessment of heteroplasmy of mitochondrial DNA mutations. Biochem Biophys Res Commun 247(1):57–59
27. Tsuzuki T et al (1983) Presence of mitochondrial-DNA-like sequences in the human nuclear DNA. Gene 25(2–3):223–229
28. Bennett S (2004) Solexa Ltd. Pharmacogenomics 5(4):433–438
29. Bennett ST et al (2005) Toward the $1000 human genome. Pharmacogenomics 6(4):373–382
30. Margulies M et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380
31. Rothberg JM et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352
32. Shendure J et al (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309(5741):1728–1732
33. He Y et al (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 464(7288):610–614

34. Goto H et al (2011) Dynamics of mitochondrial heteroplasmy in three families investigated via a repeatable re-sequencing study. Genome Biol 12(6):R59
35. Guo Y et al (2012) The use of next generation sequencing technology to study the effect of radiation therapy on mitochondrial DNA mutation. Mutat Res 744(2):154–160
36. Li M et al (2010) Detecting heteroplasmy from high-throughput sequencing of complete human mitochondrial DNA genomes. Am J Hum Genet 87(2):237–249
37. Schonberg A et al (2011) High-throughput sequencing of complete human mtDNA genomes from the Caucasus and West Asia: high diversity and demographic inferences. Eur J Hum Genet 19(9):988–994
38. Vasta V et al (2009) Next generation sequence analysis for mitochondrial disorders. Genome Med 1(10):100
39. Zhang VW, Wang J (2012) Determination of the clinical significance of an unclassified variant. Methods Mol Biol 837:337–348
40. Wang J et al (2012) An integrated approach for classifying mitochondrial DNA variants: one clinical diagnostic laboratory's experience. Genet Med 14(6):620–626
41. Bai RK et al (2007) Mitochondrial genetic background modifies breast cancer risk. Cancer Res 67(10):4687–4694
42. Kurtz A et al (2004) Somatic mitochondrial DNA mutations in neurofibromatosis type 1-associated tumors. Mol Cancer Res 2(8):433–441
43. Tan DJ, Bai RK, Wong LJ (2002) Comprehensive scanning of somatic mitochondrial DNA mutations in breast cancer. Cancer Res 62(4):972–976
44. Wong LJ et al (2003) Detection of mitochondrial DNA mutations in the tumor and cerebrospinal fluid of medulloblastoma patients. Cancer Res 63(14):3866–3871
45. Bandelt HJ, Salas A (2012) Current next generation sequencing technology may not meet forensic standards. Forensic Sci Int Genet 6(1):143–145

Chapter 12
Application of Next-Generation Sequencing of Nuclear Genes for Mitochondrial Disorders

Valeria Vasta and Si Houn Hahn

Abstract To date, more than 200 nuclear genes have been found to be linked to mitochondrial disorders [1, 2]. Initial application of next-generation sequencing (NGS) technology has been very successful for the identification of causative genes for mitochondrial disorders [3–7], similarly to other Mendelian diseases. Such advance is of outmost utility for a condition that is characterized by significant genetic and phenotypic heterogeneity with a very difficult diagnostic pathway.

NGS for panels of targeted genes already known to cause mitochondrial disorders has recently become available for clinical testing. In addition, since many genes underpinning these conditions still need to be identified, some of the recent studies targeted the entire coding sequences of the genome (the whole exome), leading to the identification of mutations in novel nuclear genes, adding to the list of loci causing mitochondrial diseases. These studies have been successful even when only single affected families were available. As for many other genetic conditions, NGS is now serving the dual role of discovery and diagnostic tool [8] for mitochondrial disorders, exemplified by the studies described here. These roles, the opportunities, and challenges of NGS in the diagnosis of mitochondrial disorders are discussed in this chapter.

V. Vasta
Seattle Children's Research Institute, Seattle, WA, USA

S.H. Hahn, M.D., Ph.D. (✉)
Seattle Children's Research Institute, Seattle, WA, USA

Department of Pediatrics, University of Washington School of Medicine,
Seattle Children's Hospital, Seattle, WA, USA
e-mail: sihahn@uw.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_12,

1 NGS Application to Mitochondrial Disorders: Validation for Clinical Diagnosis

We initially explored the feasibility of targeted NGS for mitochondrial disorders by sequencing 362 known and candidate genes with promising results [3]. We further expanded the panel to 908 nuclear genes and validated the methodology by analyzing 26 patients with known or highly suspected mitochondrial disease. Overall the sensitivity, specificity, and reproducibility of the test were satisfactory for clinical diagnosis. Analytical sensitivity was above 98 % with average coefficient of variation less than 2 %. The depth of coverage was appropriate for most of the target bases although approximately 8 % of targets did not pass the quality indicator of 20 reads and Q>30 [7]. Calvo and colleagues targeted the "MitoExome" for NGS sequencing, i.e., the mitochondrial genome and the exons of 1034 nuclear genes encoding proteins identified in the human mitochondrial proteome. A high level of sensitivity and specificity of variant detection was also achieved in this study. On average, 96 % of targeted bases were covered and 87 % of targeted bases exceeded the 15X coverage threshold found to ensure 99 % power to detect a variant [5].

In these studies, and most of the ones reported in this review, similar data analysis pipelines were utilized. These follow the general workflow adopted for the 1000 genomes project [9] using Burrows-Wheeler Aligner (BWA) for alignment of the raw base calls to the reference human genome, and GATK, or SAMtools for single nucleotide variants (SNV), small insertions, or deletions. Annotation of the variants was performed using SeattleSeq, GATK Genomic Annotator, or Annovar.

2 Diagnosis by Identification of Mutations in Already Known Pathogenic Genes by NGS

A number of studies, carried out by either targeted or whole-exome sequencing, allowed the identification of mutations in genes that had already been described as causative in the literature. This led to a diagnosis for the patients and often expanded the phenotypic spectrum of the implicated gene and mode of inheritance (Table 12.1).

Whole-exome sequencing identified a homozygous missense mutation in *AFG3L2* as the cause of early-onset spastic ataxia-neuropathy syndrome in two brothers of a consanguineous family [15]. This enzyme encodes a subunit of an m-AAA protease that resides in the mitochondrial inner membrane, responsible for removal of damaged or misfolded proteins and proteolytic activation of essential mitochondrial proteins. *AFG3L2* forms either a homo-oligomeric isoenzyme or a hetero-oligomeric complex with paraplegin, a protein mutated in hereditary spastic paraplegia 7 (SPG7), and its deficiency causes fragmentation of the mitochondrial network [37]. Heterozygous loss-of-function mutations in *AFG3L2* were already known to cause autosomal-dominant Spinocerebellar ataxia 28 (SCA28) [38], a disorder whose phenotype is strikingly different from that of the patients affected by the recessive form described in this study.

Table 12.1 Causative genes identified by NGS in suspected patients with mitochondrial disorders

Causative genes	Disease subtype	Gene/function	Clinical findings	Support of pathogenicity	Ref.
AARS2[a]	Combined RCC deficiency/mtDNA depletion	Mitochondrial alanyl-tRNA synthetase	Hypertrophic cardiomyopathy/ stillborn fetus with mitochondrial myopathy	Conservation, protein degradation	[5, 10]
ACAD8[b]	RCC IV deficiency	Acyl-CoA dehydrogenase protein family	Developmental delay, seizures, hypotonia, inability to fix and follow, cortical atrophy		[5]
ACAD9[a]	RCC I deficiency	Acyl-CoA dehydrogenase protein family, poorly known function	Hypertrophic cardiomyopathy, lactic acidosis, encephalopathy/lethal neonatal mitochondrial disease	Mutations in unrelated patients/lentiviral complement	[5, 6, 11]
ACADSB[b]	Complex III deficiency	Short/branched chain acyl-CoA dehydrogenase	Hypotonia, failure to thrive, developmental delay		[5]
ACO2[a]	Normal RCC, reduced glutamate oxidation	Mitochondrial aconitase	Infantile neurodegenerative disorder	Yeast complement	[12]
ACSF3[a]		Malonyl-CoA synthetase	Combined malonic and methylmalonic aciduria	Mutations in unrelated patients/lentiviral complement	[13, 14]
AFG3L2	Mitochondria variation in shape and size	m-AAA protease, removal of damaged protein, proteolytic activation of essential mitochondrial proteins	Early-onset spastic ataxia-neuropathy syndrome; ptosis, oculomotor apraxia, dystonia, Cbll atrophy, progressive myoclonic epilepsy	Yeast complement	[15]
AGK[a]	mtDNA depletion/ combined RCC deficiency	Acylglycerol kinase, lipid metabolism/role in driving the assembly of the ANT	Sengers syndrome, cardioskeletal myopathy, cataracts, FTT, fatigue, respiratory distress, pulmonary hypertension/ hypertrophic obstructive cardiomyopathy	Mutations in unrelated patients	[5, 16]

(continued)

Table 12.1 (continued)

Causative genes	Disease subtype	Gene/function	Clinical findings	Support of pathogenicity	Ref.
AIFM1	Combined RCC deficiency	Apoptosis-induced factor, affects translation or assembly of mitochondrial proteins	Infantile neurodegenerative disorder; NGS expands phenotype for this gene to include prenatal ventriculomegaly	Family segregation	[17]
AKR1B15[b]	RCC I deficiency	Aldo-keto reductase	Multiple pterygium syndrome, severe fetal hydrops, intrauterine growth restriction		[5]
BCS1L	RCC III deficiency	Assembly factor	Neonatal mitochondrial cytopathy	Conservation, mRNA and protein degradation	[5]
BOLA3	Combined RCC deficiency, defect of pyruvate dehydrogenase complex	Biogenesis of iron-sulfur (Fe-S) clusters, assembly of the mitochondrial RCC	Multiple mitochondrial dysfunctions syndrome -neonatal lactic acidosis, hypotonia, and cardiomyopathy	Lentiviral complement	[18]
C1orf31[b]	Combined RCC deficiency	Putative assembly factor	Hypertrophic cardiomyopathy		[5]
COX6B1	RCC IV deficiency	Complex IV subunit	Neonatal mitochondrial encephalopathy, metabolic acidosis	Protein degradation, family segregation	[5]
CPT2	Combined RCC deficiency	Carnitine palmitoyltransferase 2	Hypotonia, muscle weakness	Enzyme assay in skin fibroblasts and muscle tissues	[7]
EARS2[a]	Combined RCC deficiency	Mitochondrial glutamyl-tRNA synthetase	Infantile neurodegenerative disorder	Mutations in unrelated patients	[19]
FARS2[a]	Abnormal mitochondria	Mitochondrial phenylalanyl-tRNA synthetase	Developmental delay, seizures, and lactic acidosis	Family segregation	[20]
FOXRED1[a]	RCC I deficiency	FAD-dependent oxidoreductase	Leigh syndrome	Lentiviral complement.	[4]
GCDH		Glutaryl-CoA dehydrogenase	Early-onset generalized dystonia	Increased 3-hydroxy glutaric acid in urine	[21]

GFM1	Combined RCC deficiency	Mitochondrial translation elongation factor	Mitochondrial encephalopathy	mRNA degradation, protein degradation, family segregation	[5]
GFM2[a]		Mitochondrial translation elongation factor	Microcephaly, simplified gyral pattern, and insulin-dependent diabetes	Family segregation	[22]
HARS2[a]		Mitochondrial histidyl-tRNA synthetase	Perrault syndrome	Yeast complement.	[23]
LARS[a]	Normal RCC and mtDNA content	Leucine-tRNA synthetase	Infantile hepatopathy – multisystem involvement mimicking mitochondrial disease	Family segregation	[24]
LYRM4[b]	Combined RCC deficiency	Biogenesis of iron-sulfur (Fe-S) clusters	Failure to thrive, metabolic acidosis, hepatomegaly, apnea		[5]
MFF[a]	Normal RCC	Mitochondrial fission factor	Developmental delay	Family segregation; mitochondria histology	[20]
MRPL3[a]	Combined RCC deficiency	Mitochondrial ribosomal protein	Hypertrophic cardiomyopathy, psychomotor retardation	Family segregation, conservation	[25]
MTCH1[b]/ *MNF1*[a]	RCC III deficiency	Mitochondrial carrier homolog 1-apoptosis regulator/mitochondrial nucleoid factor 1	Severe intrauterine growth restriction, lethargy, metabolic acidosis, renal tubular acidosis, dysmorphic facies		[5]
MTERF[b]/ *C7orf10*[b]	RCC IV deficiency	Mitochondrial transcription termination factor/coenzyme-A transferase	Mild global DD, hypotonia, mild cerebral atrophy		[5]
MTFMT[a]	Combined RCC deficiency/Complex I deficiency	Met-tRNAMet formylation; Mitochondrial translation	Leigh syndrome	Lentiviral complement	[6, 26]

(continued)

Table 12.1 (continued)

Causative genes	Disease subtype	Gene/function	Clinical findings	Support of pathogenicity	Ref.
MTHFD1[a]		Synthesis of tetrahydrofolate – nucleotide and homocysteine metabolism	Megaloblastic anemia, atypical hemolytic uremic syndrome, severe combined immune deficiency, elevated blood levels of homocysteine and methylmalonic acid	Family segregation	[27]
MTHFD1L[b]/*UCP1*[b]	RCC III deficiency	Methylenetetrahydrofolate dehydrogenase 1-like, synthesis of tetrahydrofolate, nucleotide and homocysteine metabolism/uncoupling protein	Fetal hypokinesia, Pierre Robin sequence, intra-abdominal calcification		[5]
MTO1[a]	Combined RCC deficiency/complex IV deficiency	Mitochondrial-tRNA Modifier	Hypertrophic cardiomyopathy[28]/Infantile spasm, hypotonia [7]	Lentiviral complement/yeast complement	[7, 28]
NDUFAF2	RCC I deficiency	Assembly factor	Leigh syndrome/neuropathy	Protein degradation, family segregation	[4]
NDUFB3[a]	RCC I deficiency	Complex I subunit	Severe intrauterine growth restriction, failure to thrive, recurrent episodes of metabolic acidosis	Lentiviral complement	[5, 6]
NDUFS1	RCC I deficiency	Complex I subunit	Severe failure to thrive; variant found in an alternative exon not sequenced in previous single gene sequencing test		[20]
NDUFS3	RCC I deficiency	Complex I subunit	Developmental delay, muscular hypotonia, lactic acidosis	Known mutation	[6]
NDUFS4	RCC I deficiency	Complex I subunit	Leigh syndrome/hypertrophic cardiomyopathy, neuropathy, muscular hypotonia	Family segregation, conservation	[4]

NDUFS8	RCC I deficiency	Complex I subunit	Muscular hypotonia/mitochondrial encephalopathy	Family segregation, conservation	[4, 6]
NDUFV1	RCC I deficiency	Complex I deficiency	Lethal infantile mitochondrial disease	Family segregation, conservation	[4]
NUBPL[a]	RCC I deficiency	Assembly factor	Mitochondrial encephalomyopathy	Lentiviral complement	[4, 29]
PDSS1	Coenzyme Q10 deficiency	Coenzyme Q biosynthesis	Developmental delay, nephrotic syndrome	Reduced CoQ10 in white blood cells	[7]
POLG	RCC IV deficiency/ Complex I deficiency	Mitochondrial DNA polymerase	Hypotonia, developmental delay, seizures, ataxia/lethal neonatal mitochondrial encephalopathy	Known mutations	[5, 7]
POP1[a]		Mitochondrial RNA processing	Growth retardation/skeletal dysplasia	Family segregation, conservation	[30]
REEP1		Receptor expression enhancing protein 1, ER shaping and microtubule dynamics	Distal hereditary motor neuropathy; NGS expands phenotype for this gene	Family segregation, conservation, exogenous overexpression, minigene splicing assay	[31]
RRM2B	mtDNA deletions	Ribonucleotide reductase; maintains balanced nucleotide pools for mtDNA replication	Autosomal recessive PEO	Family segregation	[32]
SERAC1[a]	Normal to variable RCC deficiencies	Phosphatidyl-glycerol remodeling	MEGDEL syndrome	Mutations in unrelated patients	[33]
SLC7A13[b]	Combined RCC deficiency	Solute carrier family 7 (anionic amino acid transporter), member 13	Seizures, hypotonia, delayed motor development	Family segregation	[7]
SLC52A2[a]	Normal RCC	Riboflavin transporter 3	Truncal ataxia, optic atrophy, deafness	Family segregation, overexpression study	[34]

(continued)

Table 12.1 (continued)

Causative genes	Disease subtype	Gene/function	Clinical findings	Support of pathogenicity	Ref.
SURF1	RCC IV deficiency	Assembly factor	Leigh syndrome; initially diagnosed as Autosomal recessive ataxia, NGS lead to correct diagnosis		[22]
TK2	mtDNA deletions	Thymidine kinase 2, maintains balanced nucleotide pools for mtDNA replication	Autosomal recessive PEO	TK2 activity of patients fibroblasts and mutated protein	[35]
TSFM	Combined RCC deficiency	Mitochondrial translation elongation factor	Cardio-encephalomyopathy	Family segregation	[5]
UQCR10[b]	RCC III deficiency	Complex III subunit	Ventriculomegaly, apnea, developmental regression, hypotonia, seizures		[5]
WFS1	RCC I deficiency	Wolframin	Atypical Wolfram syndrome	Known mutation	[36]

[a]*Newly identified disease-causing genes*
[b]*Candidate disease-genes found mutated in patients*
RC respiratory chain; *RCC* respiratory chain complex

Linkage analysis, followed by whole-exome sequencing, led to the identification of the cause of prenatal ventriculomegaly observed in three brothers, as a hemizygous change in a gene, *AIFM1* in the X chromosome [17]. While this protein is known to function in apoptosis, a second function appears to be involved in mitochondrial translation. This study expands the clinical spectrum of *AIFM1* mutations to include ventriculomegaly.

Exome sequencing of two affected individuals of a family with dominant distal hereditary motor neuropathy type V identified a variant in *REEP1* [31], a gene previously linked to dominant spastic paraplegia 31 [39]. Altered mitochondrial bioenergetics and abnormal mitochondrial network morphology have been reported for mutations in this gene [40]. It is possible that this protein is involved in mitochondrial fission [40] or dynamic through microtubule interaction [41].

Joint exome analysis on two siblings presenting with severe neonatal lactic acidosis, hypotonia, and intractable cardiomyopathy led to the identification of a single homozygous missense mutation in *BOLA3* in both kids [18]. This protein is predicted to localize to the mitochondria and is postulated to have a role in the biogenesis of iron-sulfur clusters necessary for proper function of respiratory chain and 2-oxoacid dehydrogenase complexes. This gene was recently identified as pathology-associated by traditional sequencing by microcell-mediated chromosome transfer [42].

The genetic background of autosomal recessive progressive external ophthalmoplegia (PEO) is mostly unknown, with *POLG* being the only associated gene thus far [35]. Two exome sequencing studies revealed that *TK2* [35] and *RRM2B* [32] can be additional causative genes for recessive PEO, expanding their genotype-phenotype correlations, thus far limited to mtDNA depletion syndrome and autosomal-dominant PEO. Similarly, targeted exon sequencing of a suspected mitochondrial disease patient revealed a homozygous mutation in *WFS1*, causing an atypical case of Wolfram syndrome, a condition that shares clinical features with mitochondrial disorders [36]. An exome study highlighted a new phenotype for *GCDH* mutations in a patient affected by early-onset dystonia [21], while typical manifestations are infantile encephalopathy and macrocephaly. In a recent targeted exons sequencing study on a single individual, mutations in *MPV17*, a gene usually associated with childhood onset disease, were found as causative of adult-onset multisystemic disorder with mtDNA deletions [43]. In our studies by targeted exons sequencing for suspected mitochondrial patients, we identified mutations in *POLG*, *CPT2*, and *PDSS1* [7]. Of note, CPT2 is involved in fatty acid oxidation pathway and was not previously known to be involved in mitochondrial dysfunction. In another study, several patients were diagnosed by finding mutations in genes already known to be implicated in mitochondrial disorders such as *POLG*, *BCS1L*, *COX6B1*, *GFM1*, and *TSFM* [5]. In other NGS studies, molecular defects in *NDUFS3*, *NDUFS4*, *NDUFS8, NDUFV1*, and *NDUFAF2* were identified in mitochondrial respiratory chain Complex I-deficient patients [5, 6]. In an exome study, a patient was found to present a mutation in *NDUFS1*, a gene for a mitochondrial respiratory chain Complex I subunit, which escaped detection by traditional Sanger sequencing since the alternative exon had not been sequenced [20].

This example highlights the importance to include all exons of alternative transcripts even for NGS sequencing tests (commercial exome capture kits typically include only Refseq genes exons).

3 Identification of Novel Causative Genes for Mitochondrial Disorders by NGS

At least 20 new genes have been added to the list of causative genes of mitochondrial disorders in the last 2 years by NGS technology (Table 12.1), paralleling the rapid rate of discovery for other Mendelian conditions.

In a recent study, whole-exome sequencing of patients from four consanguineous families with suspected mitochondrial encephalomyopathy led to the identification of homozygous mutations in two novel genes [20]. One patient, affected by developmental delay had a homozygous mutation in *MFF*, which encodes the mitochondrial fission factor controlling the dynamic of these organelles that continually divide and fuse [44]. The authors were able to detect a clear shift in patient's mitochondria from the typical punctate to a tubular appearance, indicative of increased fusion and reduced fission. The patient's truncating mutation reduces the protein length from 326 to 64 amino acids, thus removing the transmembrane domain required for proper function [45]. The second gene found mutated in a patient was *FARS2*, which encodes the mitochondrial phenylalanyl-tRNA synthetase, a member of a growing group of mitochondria protein synthesis components to be implicated in mitochondrial disorders, as evidenced from several studies listed below and other recent studies [46].

Exome sequencing of a single proband with hypertrophic mitochondrial cardiomyopathy and combined respiratory chain deficiency led to the identification of a new causative gene, *AARS2*, encoding a putative mitochondrial alanyl-tRNA synthetase [10]. Protein structure modeling suggests that these mutations may lead to incorrect or absent tRNA aminoacylation. The same gene was found to be mutated in two siblings from a second family with the same phenotype, increasing the causative evidence. Moreover, mutations in the same gene have been detected in a stillborn fetus with mitochondrial myopathy in a subsequent study [5].

In a study on a family affected by Perrault syndrome, with ovarian dysgenesis and sensorineural hearing loss, targeted next-generation sequencing of a 4 Mb linkage region confirmed that the only gene with two predicted pathogenic variants was the previously suspected *HARS2* gene [23]. This gene encodes the mitochondrial histidyl-tRNA synthetase and the mutations reduce its aminoacylation activity on tRNA. This finding reveals a role for mitochondria in mammalian ovarian dysgenesis, while it confirms the known role of mitochondria on sensorineural function.

Exome sequencing of a baby affected by mitochondrial leukoencephalopathy led to the identification of causative gene *EARS2*, encoding mitochondrial glutamyl-tRNA synthetase [19]. The authors were able to match the Magnetic Resonance Imaging (MRI) specific features including extensive symmetrical cerebral white

matter abnormalities, symmetrical signal abnormalities of the thalami, midbrain, pons, medulla oblongata, and cerebellar white matter, of this patient to a cohort of 11 patients, also sharing similar symptoms, by screening a database of more than 3,000 cases; remarkably mutations in *EARS2* were found in all these patients. Therefore careful phenotype characterization by MRI will now enable the diagnosis by single gene conventional sequencing in leukoencephalopathy patients with the specific imaging pattern.

Exome sequencing was performed for a family affected by infantile hepatopathy and multisystem manifestations, leading to suspect mitochondrial etiology. The study identified a homozygous missense mutation in *LARS*, encoding a cytoplasmic leucyl-tRNA synthetase [24]. Since the patients in the study presented normal mtDNA content, respiratory function, and mitochondria morphology, it was concluded that mitochondria are not involved in the pathogenicity. This finding exemplifies the difficulty to distinguish between real mitochondrial disorders and conditions presenting with similar symptoms. A similar conclusion can be drawn from an exome study of a patient affected with myopathy, hypotonia, and weakness for which a mutation in *CCDC78*, a gene seemingly unrelated to mitochondria function, has been found [47]. These observations bear weight for the selection of genes to be offered for sequencing. Accordingly, for suspected mitochondria patients affected by myopathy or neurodevelopmental disease, all genes already associated to these phenotypes may need to be included for sequencing, regardless of serving a function in mitochondria. The difficulty of a genetic diagnosis based just on clinical symptoms is not limited to mitochondrial disorders. For example, a recent NGS study on pediatric-onset neurodevelopmental disease has shown that the initial diagnosis was changed upon further medical evaluation based on the mutated genes discovered [22], with some of them actually being mitochondria impacting genes.

Using targeted sequencing of the MitoExome, mutations in *MTFMT* were identified in two unrelated children presenting with Leigh syndrome and combined mitochondrial respiratory chain complex (RCC) enzyme deficiency [26]. The encoded enzyme is methionyl-tRNA formyltransferase, and this study reveals that formylation of the first methionine is essential for mitochondrial protein translation. In a second independent study, mutations in this same gene were identified in two unrelated patients affected with Leigh syndrome but presenting with isolated mitochondrial respiratory chain Complex I deficiency [6]. We also diagnosed a patient with compound heterozygosity in this gene (manuscript in preparation) who presented with hypotonia, failure to thrive, global developmental delay, and seizure. Brain MRI was also reported abnormal.

In a family with microcephaly, simplified gyral pattern, and insulin-dependent diabetes, a presentation overlapping with Wolcott-Rallison syndrome, exome sequencing identified a mutation in *GFM2*, which encodes a mitochondrial translation elongation factor [22].

By sequencing the exons of mitochondrial nuclear genes in a patient with infantile spasms and hypotonia, we detected compound heterozygosity in the gene *MTO1* [7]. This protein is involved in mitochondrial tRNA modification, and mutations in *MTO1* cause respiratory deficiency and impaired mitochondrial RNA metabolism

in *Saccharomyces cerevisiae* [48, 49]. In a separate study, mutations in *MTO1* have been identified by exome sequencing in two siblings and in an unrelated individual affected by cardiomyopathy [28]. In addition to the differences in phenotypes between our study and the latter one, disease progression also shows a large variability.

A study of a family with mitochondrial cardiomyopathy identified several potential regions by linkage mapping, encompassing 710 genes. Considering the large number of genes, the authors opted for exome sequencing. Sequencing a single patient led to the identification of the causative gene, *MRPL3*, which encodes a mitochondrial ribosomal protein of the large ribosomal subunit [25]. This conclusion was corroborated by detecting the same mutations in the affected siblings by traditional DNA sequencing and by proof of mitochondrial translation deficiency in the patient's cultured fibroblasts. It is hypothesized that the mutations affect the interaction of MRPL3 with the other proteins of the large ribosomal subunit or with the 16 s rRNA [25].

An exome sequencing strategy was used for a small family affected by combined malonic and methylmalonic aciduria, a rare recessive inborn error of metabolism, often accompanied by cardiomyopathy. Sequencing of a single proband led to the identification of the causative gene in *ACSF3*, a poorly characterized member of the acyl-CoA synthetases [13]. The same gene was found mutated in another exome study in a child with the same phenotype, increasing the causative evidence [14]. Interestingly, while these studies were published, a paper identified ACSF3 as the longtime sought source of intramitochondrial malonyl-CoA, utilized in de novo fatty acid synthesis in these organelles [50]. This is the first enzyme of mitochondrial type II fatty acid synthesis found mutated; these genetic studies unravel the essential role of this enzyme that leads to the formation of the lipoyl moieties essential for posttranslational modification of several mitochondrial proteins [51].

Exome sequencing of a single patient affected by Sengers syndrome, a disorder characterized by congenital cataracts, hypertrophic cardiomyopathy, skeletal myopathy, exercise intolerance, and lactic acidosis but normal mental development, allowed the identification of the causative gene in *AGK,* a multisubstrate lipid kinase that catalyzes the phosphorylation of diacylglycerol (DAG) and monoacylglycerol [16]. Mutation screening in other individuals from eight families with congenital cataracts and cardiomyopathy identified additional mutations in the same gene, confirming the causal nature of *AGK* deficiency in Sengers syndrome. The products of AGK activity, phosphatidic acid (PA) or lysophosphatidic acid (LPA), act as signaling molecules and take part in the synthesis of phospholipids. It is speculated that AGK, by its effects on phospholipid metabolism in mitochondria, has a role in driving the assembly of proteins within the mitochondrial membranes. A second study also identified *AGK* mutations in two unrelated patients and linked this enzyme deficiency to mtDNA depletion [5].

Another gene involved in phospholipid remodeling in mitochondria, *SERAC1*, has been recently found mutated by exome sequencing in two unrelated patients affected by MEGDEL syndrome, a recessive disorder of dystonia and deafness with Leigh-like syndrome, impaired oxidative phosphorylation, and 3-methylglutaconic

aciduria [33]. Nothing was known about the function of this protein, but the presence of a conserved lipase domain suggested a role in lipid metabolism. The authors studied phospholipid metabolism in affected patients and concluded that this enzyme is a key player in phosphatidylglycerol remodeling that is essential for both mitochondrial function and intracellular cholesterol trafficking.

Exome sequencing of a single individual with cardiomyopathy, encephalopathy, and Complex I deficiency led to the identification of *ACAD9* as causative gene, which encodes a new member of the mitochondrial acyl-CoA dehydrogenase family [11]. Based on this finding the authors were able to detect mutations in the affected sibling and in two unrelated cases. In subsequent studies additional patients with Complex I deficiency were found to have mutations in *ACAD9* [5, 6, 52]. The physiological function of ACAD9 in vivo is poorly understood and no abnormality in β-oxidation could be detected in the patients. Interestingly, riboflavin supplementation has been shown to increase Complex I activity in patient's fibroblasts, as well as to improve patients' clinical condition [11, 52].

Brown-Vialetto-Van Laere syndrome is a neurological disorder that presents symptoms that can be suggestive of mitochondrial disorders and a metabolic profile suggestive of a mild form of the multiple acyl-CoA dehydrogenation defect (MADD). Exome sequencing of a patient for which extensive laboratory testing, including muscle RCC assay, was noncontributory and led to the identification of two mutations in a riboflavin transported *SLC52A3* [34]. As for *ACAD9*, this finding provides a rational for high-dose riboflavin supplementation.

Subjects from two unrelated families with a defect manifesting in infancy with degeneration of the cerebrum, cerebellum, and retina, sharing a 4 Mb homozygous genomic region on chromosome 22, were screened in an NGS study [12]. Within this region there are 65 protein-coding genes, which include a total of 657 exons. Considering the large number of genes, the authors opted for exome sequencing, leading to the identification of a mutation in *ACO2*, the mitochondrial aconitate hydratase which catalyzes the reversible isomerization of citrate to isocitrate in the tricarboxylic acid cycle (TCA). The same single homozygote missense variant in *ACO2* was observed in all eight affected individuals from both unrelated families. Functional validation using a *Saccharomyces cerevisiae* strain lacking the aconitase gene, grown under conditions requiring the TCA cycle and the glyoxylate shunt (growth on ethanol), showed that the yeast strain expressing the mutated human gene exhibits a very clear growth defect. This mitochondrial aconitase hydratase defect joins the growing list of neurodegenerative diseases associated with primary TCA cycle impairment. Similar to the other TCA cycle defects, individuals in this study did not present mitochondrial disease classic biomarkers, such as elevated lactate, pyruvate, and alanine levels in plasma or TCA cycle metabolite levels in urine. In addition, the enzymatic activities of the mitochondrial RCC in muscle were normal, indicating that in clinical settings this enzymatic defect would be easily overlooked.

Mutations in mitochondrial respiratory chain Complex I subunit, *NDUFB3*, have been identified by sequencing the MitoExome in a patient with Complex I and lethal infantile mitochondrial disease [5]. Introduction of wild-type cDNA into subject fibroblasts rescued the defect in Complex I activity, establishing *NDUFB3* as the

causal gene. The same gene was found mutated in an independent study in a patient with muscular hypotonia, developmental delay, lactic acidosis, and Complex I deficiency [6]. This is the 15th nuclear-encoded Complex I subunit gene for which mutations have been shown to cause Complex I deficiency in humans. Mutations in *NUBPL*, a Complex I assembly factor, and in *FOXRED1*, which is an uncharacterized protein containing a FAD-dependent oxidoreductase protein domain, were identified in another study [4].

MTHFD1, encoding a cytoplasmic protein involved in cellular folate metabolism, was found mutated through exome sequencing in an infant with megaloblastic anemia, atypical hemolytic uremic syndrome, severe combined immune deficiency, and elevated blood levels of homocysteine and methylmalonic acid [27]. Considering the importance of folate for mitochondria function, this phenotype could represent an additional mitochondrial disorder spectrum. Mutations in the gene *POP1*, encoding a component of the mitochondrial RNA processing complex, have been detected by whole-exome sequencing, and the associated skeletal dysplasia may be an additional novel manifestation of mitochondria dysfunction [30]. Other genes not previously linked to the disease have been identified in few studies but further proof of pathogenicity is still being sought [5–7].

4 Challenges in Interpretation: Research Versus Clinical Test

4.1 Evaluation of Novel Variants

One major challenge in NGS is in identifying deleterious variants especially when a candidate gene has never been reported in humans as pathogenic. As for any genetic study, the first step in the validation of new pathogenic variants must evaluate the expected segregation of mutations in a family. Secondarily in silico prediction of deleterious effects of variants on protein function or on splicing can be utilized, as these models become more and more refined [53, 54]. However, this type of assessment may not be considered conclusive for the time being [55] and predictions for potentially regulatory variants in noncoding regions (promoters, UTRs, splicing branch-point) are generally lacking. Therefore, more conclusive evidence is generally derived from experimental functional validation utilizing models such as cellular rescue assays by lentiviral gene transfer [56] or yeast models [57]. The evaluation of potentially pathological variants in mitochondrial genes should actually be facilitated by the high conservation of mitochondrial components between human and yeast or other model systems. It is worth noting that implementation of experimental functional validation models for novel variants, in either novel or known genes, will be critical in order to fully exploit the diagnostic power of NGS. In addition, further correlation must be pursued by physicians to substantiate the role of suspected genes in pathogenicity, if other expected markers can be tested (phenotype, biochemistry, and histopathology).

4.2 Choice of Genes to Be Sequenced or Analyzed in Suspected Mitochondrial Patients

We previously explored the clinical use of targeted NGS for mitochondrial disorders including nuclear genes which encode mitochondrial proteins, as well as proteins that do not reside within the mitochondrion, but whose mutations present with similar symptoms [7]. In this study, none of 17 patients with various abnormal mitochondrial RCC activities showed molecular defects in either subunits or assembly factors for mitochondrial RCC enzymes. Some of these suspected mitochondrial patients were found to carry mutations in known pathogenic genes, which encode proteins that are not components of the mitochondrial RCC. Despite their clinical presentations, which were highly suspicious for mitochondrial disorders, and clear deficiencies in mitochondrial RCC enzymes, these cases appeared to be secondary to other genetic defects. This finding indicates that the diagnostic spectrum of mitochondrial disorders is much broader than we thought and this could potentially make the interpretation and diagnosis difficult, unless the targeted genes are thoroughly chosen. Indeed, in our subsequent study on 148 patients submitted for clinical testing (manuscript submitted), the highest percentage of patients had variants in genes that cause secondary mitochondrial defects or in genes that lead to similar clinical presentations to mitochondrial disease. This has also been evidenced by a few other studies described here [24, 36, 47] as conditions unrelated to mitochondria function were found to be the real cause of disease in "suspected mitochondrial patients"; these included for instance lysosomal storage disease (manuscript submitted) and Wolfram syndrome [36]. As an example of the challenges for differential diagnosis, degeneration of the cerebrum and/or cerebellum in infancy could be part of the clinical spectrum of lysosomal storage disorders, mitochondrial RCC defects, carbohydrate glycosylation defects, or infantile neuroaxonal dystrophy.

For a clinical test, the interpretation should be limited to the variants found in the genes that have already been implicated in human diseases. Thus, if a whole exome is sequenced, the analysis should be limited to a defined set of genes.

4.3 Limitations of NGS Sequencing Tests

The following limitations for the completeness of any NGS sequencing tests and interpretation of results should be considered:

(1) The selection and coverage of the exons to be sequenced may be incomplete, since it may not be possible to target repetitive or GC-rich regions or reach sufficient sequencing coverage across all meant targets. This has to be taken into account if a single variant is detected in a gene with recessive inheritance, while part of this gene is not covered by reads.
(2) Limitations still exist in the ability of data analysis tools to detect small insertions/deletions and structural variants.

(3) Both targeted and whole-exome capture may not include all transcript variants of genes, as newer exons are still being identified [58–60], and may not detect potential intronic mutations.
(4) Presence of pseudogenes or paralogs, nonspecifically captured and sequenced, or even mosaicism whose prevalence is still uncertain [61–64], may dilute the reads containing pathogenic variants to a level that is not flagged by the variant calling tools.
(5) Potential pathogenic coding synonymous, intronic, promoter, and UTR variants are difficult to identify, given the current predicting tools.
(6) The minor allele frequency threshold to be considered for a potentially pathogenic variant is still a subjective value. Several human variation databases, in their most updated version, should be searched to derive the population frequency of variants (dbSNP, 1000 Genomes, ClinSeq, NHLBI Exome sequencing project).

5 Future of NGS in the Diagnosis of Mitochondrial Disorders

The arrival of NGS in the field of mitochondrial diagnosis has already shown several tremendous advantages, despite the fact that the clinical sensitivity is still lower than ideal. Several patients have received a molecular diagnosis in a more readily and conclusive way than the traditional diagnostic path, being spared from long, invasive, and often inconclusive workup. Many studies have led to identify new and essential gene functions and mechanisms of pathogenicity for mitochondrial disorders. In some cases, the identification of the mutations has even resulted in the correction of the initial diagnosis or expanded the phenotypes associated to mutations in certain genes. Some NGS studies have brought to the attention that novel causative genes may act by impairment of mitochondria function, even though the patients are not showing typical phenotype of mitochondrial disorder.

The importance of achieving a molecular diagnosis resides also in allowing reproductive counseling, therapies access, or unnecessary therapies avoidance. Moreover, it can be foreseen that the identification of the molecular defects in more patients will expand and help develop the targeted therapies. Currently, many providers are offering the NGS test to patients, with differences in the number of genes being tested (Transgenomic, Courtagen, Arup, Baylor, GeneDx). The full implementation of this test will require further evaluation for clinical sensitivity, detection rate, and ultimately cost-effectiveness and benefits for the patients.

The initial results of the NGS studies demonstrated that the mutations can be present in genes not directly implicated in the mitochondrial respiratory chain complex function. For instance, NGS has drawn attention to genes involved in lipid metabolism and in mitochondrial protein synthesis as additional major group of genes involved in pathogenicity; thus, the genes serving these functions should be included in the test. In addition, considering the similarity of symptoms with other genetic conditions, and driving from our experience of molecular defects

found in presumed mitochondrial patients, genes for non-mitochondrial conditions presenting with similar symptoms should be included in the test. The findings of these studies should be reflected in the revision of recommendations for the diagnosis of mitochondrial disorders [65]. Sequencing a targeted pool of genes, rather than whole-exome sequencing, can still be the first choice for a diagnosis of suspected mitochondrial disease patients until the cost, sequencing coverage, and interpretation make the targeted approach more advantageous.

References

1. Scharfe C et al (2009) Mapping gene associations in human mitochondria using clinical disease phenotypes. PLoS Comput Biol 5(4):e1000374
2. Koopman WJ, Willems PH, Smeitink JA (2012) Monogenic mitochondrial disorders. N Engl J Med 366(12):1132–1141
3. Vasta V et al (2009) Next generation sequence analysis for mitochondrial disorders. Genome Med 1(10):100
4. Calvo SE et al (2010) High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency. Nat Genet 42(10):851–858
5. Calvo SE et al (2012) Molecular diagnosis of infantile mitochondrial disease with targeted next-generation sequencing. Sci Transl Med 4(118):118ra10
6. Haack TB et al (2012) Molecular diagnosis in mitochondrial complex I deficiency using exome sequencing. J Med Genet 49(4):277–283
7. Vasta V et al (2012) Next-generation sequencing for mitochondrial diseases reveals wide diagnostic spectrum. Pediatr Int 54(5):585–601
8. Ku CS et al (2012) Exome sequencing: dual role as a discovery and diagnostic tool. Ann Neurol 71(1):5–14
9. DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498
10. Gotz A et al (2011) Exome sequencing identifies mitochondrial alanyl-tRNA synthetase mutations in infantile mitochondrial cardiomyopathy. Am J Hum Genet 88(5):635–642
11. Haack TB et al (2010) Exome sequencing identifies ACAD9 mutations as a cause of complex I deficiency. Nat Genet 42(12):1131–1134
12. Spiegel R et al (2012) Infantile cerebellar-retinal degeneration associated with a mutation in mitochondrial aconitase, ACO2. Am J Hum Genet 90(3):518–523
13. Alfares A et al (2011) Combined malonic and methylmalonic aciduria: exome sequencing reveals mutations in the ACSF3 gene in patients with a non-classic phenotype. J Med Genet 48(9):602–605
14. Sloan JL et al (2011) Exome sequencing identifies ACSF3 as a cause of combined malonic and methylmalonic aciduria. Nat Genet 43(9):883–886
15. Pierson TM et al (2011) Whole-exome sequencing identifies homozygous AFG3L2 mutations in a spastic ataxia-neuropathy syndrome linked to mitochondrial m-AAA proteases. PLoS Genet 7(10):e1002325
16. Mayr JA et al (2012) Lack of the mitochondrial protein acylglycerol kinase causes Sengers syndrome. Am J Hum Genet 90(2):314–320
17. Berger I et al (2011) Early prenatal ventriculomegaly due to an AIFM1 mutation identified by linkage analysis and whole exome sequencing. Mol Genet Metab 104(4):517–520
18. Haack TB et al (2012) Homozygous missense mutation in BOLA3 causes multiple mitochondrial dysfunctions syndrome in two siblings. J Inherit Metab Dis 36(1):55–62
19. Steenweg ME et al (2012) Leukoencephalopathy with thalamus and brainstem involvement and high lactate 'LTBL' caused by EARS2 mutations. Brain 135(Pt 5):1387–1394

20. Shamseldin HE et al (2012) Genomic analysis of mitochondrial diseases in a consanguineous population reveals novel candidate disease genes. J Med Genet 49(4):234–241
21. Marti-Masso JF et al (2012) Exome sequencing identifies GCDH (glutaryl-CoA dehydrogenase) mutations as a cause of a progressive form of early-onset generalized dystonia. Hum Genet 131(3):435–442
22. Dixon-Salazar TJ et al (2012) Exome sequencing can improve diagnosis and alter patient management. Sci Transl Med 4(138):138ra178
23. Pierce SB et al (2011) Mutations in mitochondrial histidyl tRNA synthetase HARS2 cause ovarian dysgenesis and sensorineural hearing loss of Perrault syndrome. Proc Natl Acad Sci USA 108(16):6543–6548
24. Casey JP et al (2012) Identification of a mutation in LARS as a novel cause of infantile hepatopathy. Mol Genet Metab 106(3):351–358
25. Galmiche L et al (2011) Exome sequencing identifies MRPL3 mutation in mitochondrial cardiomyopathy. Hum Mutat 32(11):1225–1231
26. Tucker EJ et al (2011) Mutations in MTFMT underlie a human disorder of formylation causing impaired mitochondrial translation. Cell Metab 14(3):428–434
27. Watkins D et al (2011) Novel inborn error of folate metabolism: identification by exome capture and sequencing of mutations in the MTHFD1 gene in a single proband. J Med Genet 48(9):590–592
28. Ghezzi D et al (2012) Mutations of the mitochondrial-tRNA modifier MTO1 cause hypertrophic cardiomyopathy and lactic acidosis. Am J Hum Genet 90(6):1079–1087
29. Tucker EJ et al (2012) Next-generation sequencing in molecular diagnosis: NUBPL mutations highlight the challenges of variant detection and interpretation. Hum Mutat 33(2):411–418
30. Glazov EA et al (2011) Whole-exome re-sequencing in a family quartet identifies POP1 mutations as the cause of a novel skeletal dysplasia. PLoS Genet 7(3):e1002027
31. Beetz C et al (2012) Exome sequencing identifies a REEP1 mutation involved in distal hereditary motor neuropathy type V. Am J Hum Genet 91(1):139–145
32. Takata A et al (2011) Exome sequencing identifies a novel missense variant in RRM2B associated with autosomal recessive progressive external ophthalmoplegia. Genome Biol 12(9):R92
33. Wortmann SB et al (2012) Mutations in the phospholipid remodeling gene SERAC1 impair mitochondrial function and intracellular cholesterol trafficking and cause dystonia and deafness. Nat Genet 44(7):797–802
34. Haack TB et al (2012) Impaired riboflavin transport due to missense mutations in SLC52A2 causes Brown-Vialetto-Van Laere syndrome. J Inherit Metab Dis 35(6):943–948
35. Tyynismaa H et al (2012) Thymidine kinase 2 mutations in autosomal recessive progressive external ophthalmoplegia with multiple mitochondrial DNA deletions. Hum Mol Genet 21(1):66–75
36. Lieber DS et al (2012) Atypical case of Wolfram syndrome revealed through targeted exome sequencing in a patient with suspected mitochondrial disease. BMC Med Genet 13:3
37. Maltecca F et al (2012) Respiratory dysfunction by AFG3L2 deficiency causes decreased mitochondrial calcium uptake via organellar network fragmentation. Hum Mol Genet 21(17):3858–3870
38. Di Bella D et al (2010) Mutations in the mitochondrial protease gene AFG3L2 cause dominant hereditary ataxia SCA28. Nat Genet 42(4):313–321
39. Zuchner S et al (2006) Mutations in the novel mitochondrial protein REEP1 cause hereditary spastic paraplegia type 31. Am J Hum Genet 79(2):365–369
40. Goizet C et al (2011) REEP1 mutations in SPG31: frequency, mutational spectrum, and potential association with mitochondrial morpho-functional dysfunction. Hum Mutat 32(10):1118–1127
41. Park SH et al (2010) Hereditary spastic paraplegia proteins REEP1, spastin, and atlastin-1 coordinate microtubule interactions with the tubular ER network. J Clin Invest 120(4):1097–1110
42. Cameron JM et al (2011) Mutations in iron-sulfur cluster scaffold genes NFU1 and BOLA3 cause a fatal deficiency of multiple respiratory chain and 2-oxoacid dehydrogenase enzymes. Am J Hum Genet 89(4):486–495

43. Garone C et al (2012) MPV17 mutations causing adult-onset multisystemic disorder with multiple mitochondrial DNA deletions. Arch Neurol 69(12):1648–1651
44. Gandre-Babbe S, van der Bliek AM (2008) The novel tail-anchored membrane protein Mff controls mitochondrial and peroxisomal fission in mammalian cells. Mol Biol Cell 19(6): 2402–2412
45. Otera H et al (2010) Mff is an essential factor for mitochondrial recruitment of Drp1 during mitochondrial fission in mammalian cells. J Cell Biol 191(6):1141–1158
46. Rotig A (2011) Human diseases with impaired mitochondrial protein synthesis. Biochim Biophys Acta 1807(9):1198–1205
47. Majczenko K et al (2012) Dominant mutation of CCDC78 in a unique congenital myopathy with prominent internal nuclei and atypical cores. Am J Hum Genet 91(2):365–371
48. Colby G, Wu M, Tzagoloff A (1998) MTO1 codes for a mitochondrial protein required for respiration in paromomycin-resistant mutants of Saccharomyces cerevisiae. J Biol Chem 273(43):27945–27952
49. Wang X, Yan Q, Guan MX (2009) Mutation in MTO1 involved in tRNA modification impairs mitochondrial RNA metabolism in the yeast Saccharomyces cerevisiae. Mitochondrion 9(3): 180–185
50. Witkowski A, Thweatt J, Smith S (2011) Mammalian ACSF3 protein is a malonyl-CoA synthetase that supplies the chain extender units for mitochondrial fatty acid synthesis. J Biol Chem 286(39):33729–33736
51. Witkowski A, Joshi AK, Smith S (2007) Coupling of the de novo fatty acid biosynthesis and lipoylation pathways in mammalian mitochondria. J Biol Chem 282(19):14178–14185
52. Gerards M et al (2011) Riboflavin-responsive oxidative phosphorylation complex I deficiency caused by defective ACAD9: new function for an old gene. Brain 134(Pt 1):210–219
53. Lopes MC et al (2012) A combined functional annotation score for non-synonymous variants. Hum Hered 73(1):47–51
54. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12(9):628–640
55. Gray VE, Kukurba KR, Kumar S (2012) Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. Bioinformatics 28(16): 2093–2096
56. Danhauser K et al (2011) Cellular rescue-assay aids verification of causative DNA-variants in mitochondrial complex I deficiency. Mol Genet Metab 103(2):161–166
57. Rinaldi T et al (2010) Mitochondrial diseases and the role of the yeast models. FEMS Yeast Res 10(8):1006–1022
58. Mercer TR et al (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol 30(1):99–104
59. Lerch JK et al (2012) Isoform diversity and regulation in peripheral and central neurons revealed through RNA-Seq. PLoS One 7(1):e30417
60. Halvardson J, Zaghlool A, Feuk L (2013) Exome RNA sequencing reveals rare and novel alternative transcripts. Nucleic Acids Res 41(1):e6
61. Lindhurst MJ et al (2011) A mosaic activating mutation in AKT1 associated with the Proteus syndrome. N Engl J Med 365(7):611–619
62. Riviere JB et al (2012) De novo germline and postzygotic mutations in AKT3, PIK3R2 and PIK3CA cause a spectrum of related megalencephaly syndromes. Nat Genet 44(8):934–940
63. Lee JH et al (2012) De novo somatic mutations in components of the PI3K-AKT3-mTOR pathway cause hemimegalencephaly. Nat Genet 44(8):941–945
64. Lindhurst MJ et al (2012) Mosaic overgrowth with fibroadipose hyperplasia is caused by somatic activating mutations in PIK3CA. Nat Genet 44(8):928–933
65. Haas RH et al (2008) The in-depth evaluation of suspected mitochondrial disease. Mol Genet Metab 94(1):16–37

Chapter 13
Noninvasive Prenatal Diagnosis Using Next-Generation Sequencing

Nancy Bo Yin Tsui and Yuk Ming Dennis Lo

Abstract Noninvasive prenatal diagnosis could be carried out by analyzing cell-free fetal DNA in the plasma of pregnant women. The clinical applications of circulating fetal DNA have been continuously expanded due to the advancement of molecular detection technology. Next-generation sequencing has provided a powerful means to comprehensively analyze cell-free DNA fragments in maternal plasma. Using this technology, such cell-free DNA fragments can be qualitatively and quantitatively analyzed precisely. The application of next-generation sequencing on maternal plasma DNA analysis has allowed researchers to noninvasively detect fetal chromosome abnormalities with high accuracy. The fetal mutational and polymorphic status could also be revealed in a genome-wide scale. Hence, next-generation sequencing would be expected to play an increasingly important role in noninvasive prenatal investigations.

Abbreviations

SNP	Single nucleotide polymorphism
NGS	Next-generation sequencing
GC	Guanine and cytosine
fetal %	Fractional fetal DNA concentration

N.B.Y. Tsui • Y.M.D. Lo (✉)
Li Ka Shing Institute of Health Sciences and Department of Chemical Pathology, Centre for Research into Circulating Fetal Nucleic Acids, The Chinese University of Hong Kong, Prince of Wales Hospital, Shatin, New Territories, Hong Kong SAR, China
e-mail: nancytsui@cuhk.edu.hk; loym@cuhk.edu.hk

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_13,

1 Introduction

Prenatal diagnosis is conventionally performed by collecting fetal genetic material through invasive procedures such as chorionic villus sampling and amniocentesis. These procedures, however, pose a risk of miscarriage [1]. In 1997, the discovery of the presence of fetal DNA in the plasma of pregnant women suggested that prenatal diagnosis could be performed by simply analyzing the mother's blood [2]. Fetal DNA can be robustly detected in maternal plasma even as early as the first trimester of pregnancy [3]. Fetal DNA molecules contribute an average of some 10 % of the total cell-free DNA in maternal plasma during the first and second trimesters and increase to an average of some 20 % for the third trimester [4]. By identifying DNA sequences with placental-specific methylation signatures in maternal plasma, the placenta has been demonstrated to be an important source of circulating fetal DNA [5, 6].

The research on the clinical applications of circulating fetal DNA has gathered increasing momentum [7]. Several noninvasive prenatal diagnostic tests, such as fetal sex determination [3] and fetal rhesus-D blood group genotyping [8], have already been implemented for routine clinical use. Recently, researchers have performed genome-wide analysis of maternal plasma DNA by next-generation sequencing (NGS) and have been able to extract large amount of fetal genetic information noninvasively [9]. This technology has further opened up new possibilities for noninvasive prenatal investigation.

2 Noninvasive Prenatal Diagnosis of Chromosome Aneuploidies

2.1 The Analytical Development

Trisomy 21 (Down syndrome) is the most common reason for pregnant women seeking prenatal diagnosis. The noninvasive prenatal detection of fetal aneuploidies is technically challenging because instead of identifying the presence or absence of a fetal genetic trait, one has to quantify aberration of the aneuploid chromosome dosage of the fetus in maternal plasma. Such quantitative analysis is complicated by the high background of maternal DNA in maternal plasma [4, 10]. One approach to solve the latter problem is to target DNA molecules that bear fetal-specific methylation patterns [5, 11] or RNA molecules that are specifically expressed in the placenta [12]. For a trisomic fetus, an overrepresentation of one of the two alleles of a single nucleotide polymorphism (SNP) on the fetal-specific nucleic acid markers would be observed. This method, however, can only be applied to fetuses that are heterozygous for the SNP markers, and therefore multiple markers are necessary for a broad population coverage. Another approach is to determine the quantitative ratio of a fetal-specific epigenetic marker on an aneuploid chromosome to a fetal-specific genetic marker on an unaffected chromosome [13]. An increased ratio value would be observed for women carrying trisomic fetuses when compared to those carrying

euploid fetuses. The feasibility of this approach has been demonstrated in a proof-of-concept study [13]. In yet another approach, fetal-derived hypermethylated DNA molecules in maternal whole blood samples were enriched by methylated DNA immunoprecipitation, followed by measurement of the loci that were located on chromosome 21 and were hypermethylated specifically in the placenta [14]. For a trisomy 21 fetus, an elevated concentration of these enriched fetal-derived chromosome-21 sequences was expected [14, 15]. However, there is controversy concerning the theoretical basis and reproducibility of this method [16, 17].

An alternative strategy that is independent on the use of fetal-specific nucleic acid markers is to compare the total (fetal plus maternal) DNA copies derived from a locus on the aneuploid chromosome with those derived from another locus on an unaffected chromosome [18]. An overrepresentation of DNA from the aneuploid chromosome would indicate a trisomic fetus. The major challenge of this method is that the degree of chromosome dosage imbalance is small and is dependent on the fractional fetal DNA concentration (fetal %) in maternal plasma. For example, for a maternal plasma sample with 10 % of cell-free DNA contributed by the fetus, the carrying of a trisomy 21 fetus would lead to a 5 % increment of chromosome-21 dosage. In order to discriminate such a small difference, quantification based on digital molecular counting, such as digital PCR, has been investigated [18]. In fact, the precision of this approach is related to both the fetal % and the number of DNA molecules being counted. Researchers have used computer simulations to estimate that in order to detect a trisomic fetus in a maternal plasma sample with 25 % of fetal DNA, around 8,000 target molecules would need to be analyzed [18].

2.2 NGS as a Promising Tool for Noninvasive Diagnosis of Fetal Aneuploidies

NGS allows the digital counting of millions of DNA molecules in a sequencing run and hence provides a high quantitative precision for detecting the relative amounts of different DNA species. The technology has successfully been applied to the noninvasive prenatal diagnosis of trisomy 21 [19, 20]. Researchers have used NGS to sequence maternal plasma DNA fragments derived from virtually all chromosomes. The number of sequenced fragments derived from each chromosome was counted, and the proportional representation of chromosome-21 DNA among all the cell-free DNA in maternal plasma was calculated. The calculated value of the tested sample was then compared with the values of a group of women known to be carrying euploid fetuses to determine if there was a significant increase.

Following two proof-of-principle studies [19, 20], the accuracy of the NGS approach for noninvasive prenatal trisomy 21 detection has been confirmed in many large-scale clinical studies [21–26]. The reported sensitivities range from 99 % to 100 % and the specificities range from 98 % to 100 %.

Theoretically, the NGS method could measure the proportional representations of any potentially aneuploid chromosomes carried by the fetus. The noninvasive diagnosis of fetal trisomies 13 and 18 has subsequently been achieved [23, 27, 28].

In addition to detection of trisomies, the method has been demonstrated to be useful for identifying an abnormally low proportional representation of an aneuploid chromosome, such as for the noninvasive prenatal diagnosis of monosomy X [26]. This method has also been shown to be able to detect trisomies 21 and 13 caused by Robertsonian translocations [26, 29], microdeletions on chromosomes 12 and 22 [30, 31], as well as mosaicism for trisomies 21 and 18 [26]. The applicability of this method in twin pregnancies has also been tested and shown to be clinically useful [32].

3 Noninvasive Fetal Mutation Detection by NGS

When both the father and mother are carriers for a monogenic recessive disease mutation, it is important to determine if the fetus has inherited two mutant alleles from each of the parents. As maternal plasma contains both fetal and maternal DNA fragments, the determination of a fetal allele inherited from the mother is more difficult than that from the father. Researchers have therefore used different approaches to identify paternal- and maternal-inherited alleles of the fetus. To detect which of the paternal alleles is inherited by the fetus, the paternal allele that is absent in the mother's genome is targeted. The presence or absence of this allele in maternal plasma would indicate whether the fetus has inherited this allele or not [33]. To detect the inheritance of the maternal allele, an approach termed relative haplotype dosage analysis has been developed to determine if the mutant- or the wild-type-linked haplotype is inherited by the fetus. This could be deduced by comparing the relative abundance of all SNP alleles between the two haplotypes in maternal plasma. The overrepresented haplotype would be the one inherited by the fetus [33]. This integrated noninvasive fetal genotyping approach has been successfully demonstrated in a prenatal case in which both the father and mother were carriers of β-thalassemia mutations. The mutational status of the fetus was accurately detected using maternal plasma [33].

NGS analyzes plasma DNA molecules sampled from virtually the whole genome in a single run. Hence, by using this fetal genotyping approach, researchers have succeeded in deducing the genome-wide genetic map of a fetus [33, 34]. It is therefore highly feasible to noninvasively detect multiple fetal genetic diseases in a single NGS assay.

4 NGS Analysis of Maternal Plasma DNA: Factors to Be Considered

4.1 Algorithm for Determining Fetal Aneuploidies

The z-score calculation is the most commonly used statistical method to determine if there is a significant increase in the aneuploid chromosome representation in the tested sample when compared to the reference group [19, 21, 22]. For several

large-scale studies, around 100 euploid pregnancies have been included in the reference groups [21, 24, 26]. A z-score of >3 has been used as a cutoff for classifying the presence of an aneuploid fetus, meaning that the proportion of DNA sequences from an at-risk chromosome in tested sample is greater than the 99.9th percentile when compared with that of a reference group [19, 21, 22]. To further reduce the analytical variation, adjustments to the z-score calculation algorithm have been investigated. For example, the read counts derived from the at-risk chromosome could be normalized by the counts from one or a few selected chromosomes, rather than the total counts from all chromosomes, such that the intra- and inter-sequencing run variations would be minimized [23, 26]. Different cutoff values of z-score have also been evaluated [22, 23, 26].

4.2 Guanine-Cytosine Bias Correction

It has been reported that the number of sequenced reads generated by NGS was influenced by guanine and cytosine (GC) contents of DNA fragments [27]. This uneven representation of sequenced reads is possibly related to the GC bias of library amplification before sequencing [35]. In the initial studies of noninvasive detection of fetal trisomies 13 and 18, the GC bias was shown to reduce the precision of measuring the proportional representations of chromosomes 13 and 18 [19, 20]. Following the correction of GC bias with bioinformatics, the accuracy of detection for fetal trisomies 13 and 18 has greatly improved [27, 28].

The amplification-associated GC bias could also be eliminated by using a single molecule sequencing platform, in which DNA fragments are sequenced with no prior amplification [35]. With the advance of single molecule sequencing technology [36], the analysis of maternal plasma would become simpler and more accurate.

4.3 Abundance of Fetal DNA in Maternal Plasma

As discussed above, the noninvasive detection of fetal aneuploidies and maternally inherited alleles is based on determining the quantitative difference of the aneuploid chromosome representation and the haplotype dosage, respectively [19, 33]. The magnitude of these quantitative differences is dependent on the fetal % in maternal plasma. If the fetal % in a maternal plasma sample is very low, the dosage difference would be too small to be precisely identified by NGS, and hence false-negative result would be obtained. In some of the clinical trial studies, maternal plasma samples with fetal % of less than 3.9 % were rejected for NGS analysis [22, 24]. The fetal % is also an important factor for deducing mutational status of the fetus [37].

To measure the fetal %, one approach is to use fetal-specific epigenetic markers such as methylated *RASSF1A* sequences [6]. The NGS reads containing

fetal-specific SNP alleles could also be used for estimating fetal % [33]. In a clinical diagnostic setting, it would be ideal to be able to obtain the fetal % with a minimum number of additional steps. For pregnancies involving male fetuses, one could determine the proportional representation of chromosome-Y DNA from the plasma NGS data [19]. Recently, researchers have shown that with sufficiently high sequencing coverage, such as in the case of targeted sequencing (Sect. 4.4), the fetal % could be deduced directly from NGS data of maternal plasma without the need of prior genotyping information [38]. This method is applicable to pregnancies involving both male and female fetuses.

4.4 Targeted Enrichment of Selected Regions

The precision of the noninvasive measurement of fetal chromosome and haplotype dosage could be enhanced by sequencing more DNA fragments in maternal plasma. A cost-effective way to do this is to selectively enrich DNA molecules originating from the target regions before sequencing. The feasibility of targeted NGS for maternal plasma analysis was first demonstrated by enriching exonic regions on the X chromosome [39]. Capture probes with nucleic acid sequences specific to the targeted regions were used for the enrichment. While the reads covering the targeted regions were enriched by over 200-fold, the proportional amount of fetal- and maternal-derived DNA remained unchanged after enrichment [39]. This finding implied that the quantitative DNA dosage information in maternal plasma was less likely to be altered by the targeted enrichment process.

To further explore the use of targeted NGS for noninvasive prenatal aneuploidy detection, DNA molecules originating from chromosomes 13, 18, and 21 were enriched for analysis in maternal plasma [40]. By enriching *β-globin* gene sequences in maternal plasma, the fetal mutational status for β-thalassemia mutations has been successfully determined by NGS [41]. Hence, this method requires much fewer sequencing resources when compared to the protocol without enrichment [33].

In another targeted NGS approach, the targeted regions were selectively amplified prior to NGS. This strategy has been investigated for the noninvasive diagnosis of fetal trisomies 18 and 21 [42, 43]. However, the robustness of the selective amplification procedure remains to be directly compared with that for nonselective or random NGS.

4.5 High-Throughput Sample Multiplexing

Multiplexing samples for NGS could potentially increase the throughput and reduce the sequencing cost for clinical implementation. DNA molecules from different plasma samples could be labeled with unique index sequences before pooling together and sequencing in the same lane. Since the precision of NGS for measuring

DNA dosage is dependent on the number of sequenced reads, a balance between the level of sample multiplexing and the detection precision has to be considered. For the noninvasive detection of fetal trisomy 21, it has been estimated that in order to achieve a 99.9 % confidence in detecting the chromosome-21 dosage increment, the analytical precision of the NGS platform should have a coefficient of variation less than 0.83 % [21]. Researchers have compared the precision for measuring the proportional representation of chromosome-21 DNA with a 2-plex platform (mean number of reads per sample: 2.3 million) and a 8-plex platform (mean number of reads per sample: 0.3 million), and the coefficient of variations were 0.66 % and 1.59 %, respectively [21]. As throughput of DNA sequencers increases with newer models, the above figures are best considered in terms of the mean number of reads, rather than as the level of "plexing." Another study has shown that the accuracy of the fetal aneuploidy detection remains high with four samples sequenced per lane [24]. We envision that with the growth of sequencing throughput and the improvements in sequencing protocols, the degree of multiplexing would be expected to further increase in the future.

4.6 Other Considerations

There remains other unexplored factors that may influence the performance of the NGS analysis. For example, confined placental mosaicism, in which chromosome aneuploidies are observed in the placental cells but not in the fetus [44], may give rise to false-positive result. However, since confined placental mosaicism has been reported in only 2 % of chorionic villus sampling cases, and chromosomes 21, 18, and 13 were not among the most common chromosomes involved in confined placental mosaicism [44], the rate of false-positives due to this phenomenon is expected to be low. Maternal copy number variations [45] are another potential factor that may introduce variation in measuring the relative proportion of various chromosomes in maternal plasma. It would be interesting to test the diagnostic sensitivity and specificity of bioinformatics algorithms that take such variations into consideration.

5 Clinical Implementation of NGS Testing for Noninvasive Fetal Aneuploidy Detection

The noninvasive prenatal diagnosis of chromosome aneuploidies by NGS has been implemented as clinical service in the USA, Mainland China, Hong Kong, and parts of Europe. Discussion is still ongoing regarding the best approach for clinically implementing the NGS test [46, 47]. One option is to offer the NGS test to high-risk women as identified by current screening approaches, and invasive testing is indicated for women tested positive by the NGS test. Researchers have estimated that

some 98 % of the invasive procedures, together with the related procedural costs and the associated miscarriages, could be avoided by this diagnostic pathway [21, 24, 46]. Another option is to replace current screening approaches with the NGS test, and with the latter's higher diagnostic sensitivity, the chance of missing an aneuploid fetus could be reduced. However, besides cost consideration, the clinical performance of the NGS test has to be extensively evaluated in low-risk populations before such a strategy is implemented on a large scale.

6 Conclusion

The advent of NGS technology has allowed researchers to detect chromosome aneuploidies, gene mutations, and even the whole genome map of the fetus noninvasively using maternal plasma. We foresee that noninvasive prenatal diagnosis using NGS will play an increasingly important role in future prenatal testing. Hence, ethical, legal, and social issues concerning the clinical practice of noninvasive prenatal diagnosis should be a research priority [48].

References

1. Morris JK, Waters JJ, de Souza E (2012) The population impact of screening for Down syndrome: audit of 19 326 invasive diagnostic tests in England and Wales in 2008. Prenat Diagn 32:596–601
2. Lo YMD, Corbetta N, Chamberlain PF, Rai V, Sargent IL, Redman CW, Wainscoat JS (1997) Presence of fetal DNA in maternal plasma and serum. Lancet 350:485–487
3. Devaney SA, Palomaki GE, Scott JA, Bianchi DW (2011) Noninvasive fetal sex determination using cell-free fetal DNA: a systematic review and meta-analysis. JAMA 306:627–636
4. Lun FMF, Chiu RWK, Chan KCA, Leung TY, Lau TK, Lo YMD (2008) Microfluidics digital PCR reveals a higher than expected fraction of fetal DNA in maternal plasma. Clin Chem 54:1664–1672
5. Chim SSC, Tong YK, Chiu RWK, Lau TK, Leung TN, Chan LY, Oudejans CB, Ding C, Lo YMD (2005) Detection of the placental epigenetic signature of the maspin gene in maternal plasma. Proc Natl Acad Sci USA 102:14753–14758
6. Chan KCA, Ding C, Gerovassili A, Yeung SW, Chiu RWK, Leung TN, Lau TK, Chim SSC, Chung GTY, Nicolaides KH, Lo YMD (2006) Hypermethylated RASSF1A in maternal plasma: a universal fetal DNA marker that improves the reliability of noninvasive prenatal diagnosis. Clin Chem 52:2211–2218
7. Lo YMD, Chiu RWK (2012) Genomic analysis of fetal nucleic acids in maternal blood. Annu Rev Genomics Hum Genet 13:285–306
8. Finning K, Martin P, Daniels G (2004) A clinical service in the UK to predict fetal Rh (Rhesus) D blood group using free fetal DNA in maternal plasma. Ann N Y Acad Sci 1022:119–123
9. Chiu RWK, Lo YMD (2012) Noninvasive prenatal diagnosis empowered by high-throughput sequencing. Prenat Diagn 32:401–406
10. Lo YMD, Tein MS, Lau TK, Haines CJ, Leung TN, Poon PM, Wainscoat JS, Johnson PJ, Chang AM, Hjelm NM (1998) Quantitative analysis of fetal DNA in maternal plasma and serum: implications for noninvasive prenatal diagnosis. Am J Hum Genet 62:768–775

11. Tong YK, Ding C, Chiu RWK, Gerovassili A, Chim SSC, Leung TY, Leung TN, Lau TK, Nicolaides KH, Lo YMD (2006) Noninvasive prenatal detection of fetal trisomy 18 by epigenetic allelic ratio analysis in maternal plasma: Theoretical and empirical considerations. Clin Chem 52:2194–2202
12. Lo YMD, Tsui NBY, Chiu RWK, Lau TK, Leung TN, Heung MM, Gerovassili A, Jin Y, Nicolaides KH, Cantor CR, Ding C (2007) Plasma placental RNA allelic ratio permits noninvasive prenatal chromosomal aneuploidy detection. Nat Med 13:218–223
13. Tong YK, Chiu RWK, Akolekar R, Leung TY, Lau TK, Nicolaides KH, Lo YMD (2010) Epigenetic-genetic chromosome dosage approach for fetal trisomy 21 detection using an autosomal genetic reference marker. PLoS One 5:e15244
14. Papageorgiou EA, Karagrigoriou A, Tsaliki E, Velissariou V, Carter NP, Patsalis PC (2011) Fetal-specific DNA methylation ratio permits noninvasive prenatal diagnosis of trisomy 21. Nat Med 17:510–513
15. Tsaliki E, Papageorgiou EA, Spyrou C, Koumbaris G, Kypri E, Kyriakou S, Sotiriou C, Touvana E, Keravnou A, Karagrigoriou A, Lamnissou K, Velissariou V, Patsalis PC (2012) MeDIP real-time qPCR of maternal peripheral blood reliably identifies trisomy 21. Prenat Diagn 32:996–1001
16. Tong YK, Chiu RW, Chan KC, Leung TY, Lo YM (2012) Technical concerns about immunoprecipitation of methylated fetal DNA for noninvasive trisomy 21 diagnosis. Nat Med 18:1327–1328
17. Patsalis PC (2012) Reply to: technical concerns about immunoprecipitation of methylated fetal DNA for noninvasive trisomy 21 diagnosis. Nat Med 18:1328–1329
18. Lo YMD, Lun FMF, Chan KCA, Tsui NBY, Chong KC, Lau TK, Leung TY, Zee BC, Cantor CR, Chiu RWK (2007) Digital PCR for the molecular detection of fetal chromosomal aneuploidy. Proc Natl Acad Sci USA 104:13116–13121
19. Chiu RWK, Chan KCA, Gao Y, Lau VYM, Zheng W, Leung TY, Foo CH, Xie B, Tsui NBY, Lun FMF, Zee BC, Lau TK, Cantor CR, Lo YMD (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. Proc Natl Acad Sci USA 105:20458–20463
20. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. Proc Natl Acad Sci USA 105:16266–16271
21. Chiu RWK, Akolekar R, Zheng YWL, Leung TY, Sun H, Chan KCA, Lun FMF, Go AT, Lau ET, To WW, Leung WC, Tang RY, Au-Yeung SK, Lam H, Kung YY, Zhang X, van Vugt JM, Minekawa R, Tang MH, Wang J, Oudejans CB, Lau TK, Nicolaides KH, Lo YMD (2011) Non-invasive prenatal assessment of trisomy 21 by multiplexed maternal plasma DNA sequencing: large scale validity study. BMJ 342:c7401
22. Ehrich M, Deciu C, Zwiefelhofer T, Tynan JA, Cagasan L, Tim R, Lu V, McCullough R, McCarthy E, Nygren AO, Dean J, Tang L, Hutchison D, Lu T, Wang H, Angkachatchai V, Oeth P, Cantor CR, Bombard A, van den Boom D (2011) Noninvasive detection of fetal trisomy 21 by sequencing of DNA in maternal blood: a study in a clinical setting. Am J Obstet Gynecol 204(205):e1–e11
23. Sehnert AJ, Rhees B, Comstock D, de Feo E, Heilek G, Burke J, Rava RP (2011) Optimal detection of fetal chromosomal abnormalities by massively parallel DNA sequencing of cell-free fetal DNA from maternal blood. Clin Chem 57:1042–1049
24. Palomaki GE, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM, Ehrich M, van den Boom D, Bombard AT, Deciu C, Grody WW, Nelson SF, Canick JA (2011) DNA sequencing of maternal plasma to detect Down syndrome: an international clinical validation study. Genet Med 13:913–920
25. Lau TK, Chen F, Pan X, Pooh RK, Jiang F, Li Y, Jiang H, Li X, Chen S, Zhang X (2012) Noninvasive prenatal diagnosis of common fetal chromosomal aneuploidies by maternal plasma DNA sequencing. J Matern Fetal Neonatal Med 25:1370–1374
26. Bianchi DW, Platt LD, Goldberg JD, Abuhamad AZ, Sehnert AJ, Rava RP (2012) Genome-wide fetal aneuploidy detection by maternal plasma DNA sequencing. Obstet Gynecol 119:890–901

27. Chen EZ, Chiu RWK, Sun H, Akolekar R, Chan KCA, Leung TY, Jiang P, Zheng YWL, Lun FMF, Chan LYS, Jin Y, Go AT, Lau ET, To WW, Leung WC, Tang RY, Au-Yeung SK, Lam H, Kung YY, Zhang X, van Vugt JM, Minekawa R, Tang MH, Wang J, Oudejans CB, Lau TK, Nicolaides KH, Lo YMD (2011) Noninvasive prenatal diagnosis of fetal trisomy 18 and trisomy 13 by maternal plasma DNA sequencing. PLoS One 6:e21791
28. Palomaki GE, Deciu C, Kloza EM, Lambert-Messerlian GM, Haddow JE, Neveux LM, Ehrich M, van den Boom D, Bombard AT, Grody WW, Nelson SF, Canick JA (2012) DNA sequencing of maternal plasma reliably identifies trisomy 18 and trisomy 13 as well as Down syndrome: an international collaborative study. Genet Med 14:296–305
29. Lun FMF, Jin YY, Sun H, Leung TY, Lau TK, Chiu RWK, Lo YMD (2011) Noninvasive prenatal diagnosis of a case of Down syndrome due to robertsonian translocation by massively parallel sequencing of maternal plasma DNA. Clin Chem 57:917–919
30. Jensen TJ, Dzakula Z, Deciu C, van den Boom D, Ehrich M (2012) Detection of microdeletion 22q11.2 in a fetus by next-generation sequencing of maternal plasma. Clin Chem 58:1148–1151
31. Peters D, Chu T, Yatsenko SA, Hendrix N, Hogge WA, Surti U, Bunce K, Dunkel M, Shaw P, Rajkovic A (2011) Noninvasive prenatal diagnosis of a fetal microdeletion syndrome. N Engl J Med 365:1847–1848
32. Canick JA, Kloza EM, Lambert-Messerlian GM, Haddow JE, Ehrich M, van den Boom D, Bombard AT, Deciu C, Palomaki GE (2012) DNA sequencing of maternal plasma to identify Down syndrome and other trisomies in multiple gestations. Prenat Diagn 32:730–734
33. Lo YMD, Chan KCA, Sun H, Chen EZ, Jiang P, Lun FMF, Zheng YW, Leung TY, Lau TK, Cantor CR, Chiu RWK (2010) Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. Sci Transl Med 2:61ra91
34. Kitzman JO, Snyder MW, Ventura M, Lewis AP, Qiu R, Simmons LE, Gammill HS, Rubens CE, Santillan DA, Murray JC, Tabor HK, Bamshad MJ, Eichler EE, Shendure J (2012) Noninvasive whole-genome sequencing of a human fetus. Sci Transl Med 4:137ra176
35. van den Oever JM, Balkassmi S, Verweij EJ, van Iterson M, van Scheltema PN, Oepkes D, van Lith JM, Hoffer MJ, den Dunnen JT, Bakker E, Boon EM (2012) Single molecule sequencing of free DNA from maternal plasma for noninvasive trisomy 21 detection. Clin Chem 58:699–706
36. Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19:R227–R240
37. Lun FMF, Tsui NBY, Chan KCA, Leung TY, Lau TK, Charoenkwan P, Chow KCK, Lo WY, Wanapirak C, Sanguansermsri T, Cantor CR, Chiu RWK, Lo YMD (2008) Noninvasive prenatal diagnosis of monogenic diseases by digital size selection and relative mutation dosage on DNA in maternal plasma. Proc Natl Acad Sci USA 105:19920–19925
38. Jiang P, Chan KCA, Liao GJW, Zheng YW, Leung TY, Chiu RWK, Lo YMD, Sun H (2012) FetalQuant: deducing fractional fetal DNA concentration from massively parallel sequencing of DNA in maternal plasma. Bioinformatics. doi:10.1093/bioinformatics/bts549
39. Liao GJW, Lun FMF, Zheng YW, Chan KCA, Leung TY, Lau TK, Chiu RWK, Lo YMD (2011) Targeted massively parallel sequencing of maternal plasma DNA permits efficient and unbiased detection of fetal alleles. Clin Chem 57:92–101
40. Liao GJW, Chan KCA, Jiang P, Sun H, Leung TY, Chiu RWK, Lo YMD (2012) Noninvasive prenatal diagnosis of fetal trisomy 21 by allelic ratio analysis using targeted massively parallel sequencing of maternal plasma DNA. PLoS One 7:e38154
41. Lam KWG, Jiang P, Liao GJW, Chan KCA, Leung TY, Chiu RWK, Lo YMD (2012) Noninvasive prenatal diagnosis of monogenic diseases by targeted massively parallel sequencing of maternal plasma: application to beta Thalassemia. Clin Chem. doi:10.1373/clinchem.2012.189589
42. Sparks AB, Struble CA, Wang ET, Song K, Oliphant A (2012) Noninvasive prenatal detection and selective analysis of cell-free DNA obtained from maternal blood: evaluation for trisomy 21 and trisomy 18. Am J Obstet Gynecol 206(319):e1–e9

43. Ashoor G, Syngelaki A, Wagner M, Birdir C, Nicolaides KH (2012) Chromosome-selective sequencing of maternal plasma cell-free DNA for first-trimester detection of trisomy 21 and trisomy 18. Am J Obstet Gynecol 206(322):e1–e5
44. Kalousek DK, Vekemans M (1996) Confined placental mosaicism. J Med Genet 33:529–533
45. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97
46. Chitty LS, Hill M, White H, Wright D, Morris S (2012) Noninvasive prenatal testing for aneuploidy-ready for prime time? Am J Obstet Gynecol 206:269–275
47. Deans Z, Newson AJ (2012) Ethical considerations for choosing between possible models for using NIPD for aneuploidy detection. J Med Ethics 38:614–618
48. Greely HT (2011) Get ready for the flood of fetal gene screening. Nature 469:289–291

Part IV
Compliance with CAP/CLIA Regulations

Chapter 14
Guidelines and Approaches to Compliance with Regulatory and Clinical Standards: Quality Control Procedures and Quality Assurance

Ira M. Lubin, Lisa Kalman, and Amy S. Gargis

Abstract Clinical laboratories are beginning to use next-generation sequencing (NGS) for the testing of patient samples. This chapter addresses regulatory and professional standards that have been developed and are under consideration to assure the quality of NGS testing in the clinical setting. The major topics addressed include test validation, quality control procedures, proficiency testing, and reference materials. Considerations for the establishment of performance specifications, such as accuracy, precision, analytic sensitivity, and analytic specificity, for NGS applications are discussed. Emphasis is placed on aspects unique to NGS, such as the reliance on an "informatics" pipeline to process the platform-generated data and challenges to the establishment and use of reference materials, quality control procedures, and proficiency testing. While there are significant benefits to clinical testing achievable through the use of NGS, the complexities associated with its use in the clinical laboratory are significant and will require an evolving set of standards to keep up with the rapidly advancing technology.

I.M. Lubin, Ph.D., FACMG (✉) • L. Kalman, Ph.D.
Division of Laboratory Science and Standards, Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Mailstop G-23, Atlanta 30329, GA, USA
e-mail: Ilubin@cdc.gov; LKalman@cdc.gov

A.S. Gargis, Ph.D.
Division of Laboratory Science and Standards, Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Mailstop G-23, Atlanta 30329, GA, USA

Oak Ridge Institute for Science and Education, Oak Ridge, TN, USA
e-mail: AGargis@cdc.gov

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_14,

1 Current Framework of Regulatory Oversight and Professional Guidance

Regulatory and accreditation requirements, professional guidance, and best practices ensure that reliable results are achieved from clinical laboratory testing. Regulatory oversight of clinical laboratories varies worldwide, and many countries base oversight on compliance with standards set by the International Standards Organization (ISO) [1]. In the United States, laboratory practice is federally regulated under the Clinical Laboratory Improvement Amendments (CLIA) regulations published in 1988 [2]. These regulations provide minimal standards for laboratories offering clinical testing. Clinical laboratories in the United States demonstrate compliance with the CLIA regulations through federal certification by the Centers for Medicare and Medicaid Services (CMS) or accreditation by a program with standards deemed comparable to the federal regulatory requirements such as those offered by the College of American Pathologists Laboratory Accreditation Program (CAP LAP) [3]. Certain states, such as New York and Washington, are exempt from the CLIA regulations because their state regulations are comparable to the federal requirements [4, 5].

In the United States, the Food and Drug Administration (FDA) regulates manufacturers of test equipment, devices, and assay reagent kits used for clinical testing [6]. Manufacturers submit documentation to the FDA that describes the intended use, expected performance specifications, and data supporting analytic and clinical validity of their product in a clinical laboratory setting. Following a review process, the product may be cleared, approved, denied, or additional information requested. Tests developed and used within the same clinical laboratory, which are not provided to other users, are referred to as laboratory-developed tests. As of 2012, the FDA has not required their review of these tests. CLIA and FDA regulations both require that clinical laboratories document the performance specifications of the clinical tests that they perform. For tests that are FDA approved or cleared, the CLIA regulations require that clinical laboratories are able to verify the performance specifications established by the manufacturer. For laboratory-developed tests, CLIA requires that laboratories establish the performance specifications of a test prior to patient testing.

Several guidance documents have been developed to assist clinical laboratories to meet regulatory requirements [7–11]. These include the Laboratory Standards and Guidelines for Clinical Genetics Laboratories issued and regularly updated by the American College of Medical Genetics [7, 8], guidance documents from the Clinical and Laboratory Standards Institute [9, 10], and recommendations for good laboratory practices for molecular genetic testing from the US Centers for Disease Control and Prevention [11]. Significant guidance has also been developed by other countries to assure the quality of clinical molecular genetic testing [12, 13].

2 Validation and Quality Control Procedures

2.1 Considerations for Next-Generation Sequencing (NGS)

Clinical tests must be validated prior to patient testing. During the validation process, the laboratory documents the performance specifications of the test. This typically includes a description of the test's accuracy, precision, analytic sensitivity, analytic specificity, reportable range, reference range, and other characteristics that define how well the test performs and the analytical limitations. Assay validation also informs the development of quality control (QC) procedures that are needed to demonstrate adequate performance during clinical specimen testing. For example, methods to assure that adequate depth of coverage is achieved to assure a reliable test result. The depth of coverage refers to the number of independent reads covering a given base position or region of the genome.

Proficiency testing (PT), External Quality Assessment (EQA), or alternate assessment procedures provide an independent and external assessment of test performance after the test is available for clinical testing. These approaches permit a comparison of test results obtained among different laboratories and can help identify problems related to the analytic testing process that are not revealed by the QC procedures of a single laboratory. Validation, QC, and PT/EQA or alternate assessment are some central tenets of regulatory oversight and professional guidance which ensure the quality of clinical laboratory testing, but their application to NGS requires adaptation.

Good laboratory practices typically include the assessment of a test to detect all possible results, but this is not practical for NGS because of the large number of results that are possible. Also, a comprehensive set of reference or QC materials that encompass all possible test results is not practical. New guidance and adaptation of regulatory requirements are needed to account for these challenges and assure high-quality NGS testing in the clinical laboratory setting. Many groups of experts and professional organizations are developing NGS guidance to help laboratories design strategies for NGS assay validation, QC, and PT/AA practices to meet current regulatory requirements [3, 7, 14]. For example, the CAP has recently published a revised version of its molecular pathology checklist, which includes 18 requirements specific to NGS, to promote a uniform framework for validating sequencing-based clinical tests [3]. These groups are also developing recommendations for reference materials, data management tools, result reporting templates, interpretations of test results, and presentation of test results in the electronic medical record.

2.2 Platform, Test, and Informatics Pipeline Validation

NGS assays can be configured using multiple combinations of nucleic acid biochemistry and data analysis tools. Each component of the NGS test, including DNA fragment library preparation, enrichment methods, sequencing, and data analysis informatics methods, should be optimized, documented, and validated for each clinical test within the clinical setting for the appropriate patient population [14]. When the entire workflow is established, the complete testing process should be validated to ensure that the system is suitable for its intended use as a clinical test [14]. Each validation report must also specify the types of sequence variations that are targeted, such as single nucleotide polymorphisms (SNPs), small repeats, and insertions and deletions (indels).

The validation process may be divided into three interconnected components that address platform, test, and informatics processes [14]. Platform validation establishes the ability of the platform to identify a range of genetic variation across a broad range of genomic regions [14]. Test validation documents that the assay can detect clinically significant sequence variations for the intended application [14, 15]. Validation of the informatics pipeline establishes the computational software settings required to provide accurate sequence data and detection of variations within the targeted genomic region(s) [14]. Because the informatics tools used for data analysis are highly specific and separate from the instrumentation components of an NGS test system, the informatics pipeline should be optimized separately during test development prior to validation of the test [14].

Validation establishes the performance specifications of the test system with the quality of testing dependent on the maintenance of these specifications during clinical testing. Reassessment of the test system is necessary when any changes are made [1–3, 14]. For example, replacement of a depleted reagent requires verification that the new lot performs as well or better than the previous lot. An upgrade of the informatics portion of the test also requires a revalidation. This can present a challenge for the laboratory since they may not know when these updates will occur. Changes to downstream processes may not require revalidation of earlier NGS processes. For example, changes to the informatics pipeline may not require the sequencing platform to be revalidated, provided that the specifications for the input files do not change [14].

2.3 Establishing Quality Control Procedures

Quality control procedures are implemented to monitor the performance of the analytical process. Control procedures detect immediate errors caused by test system failure, adverse environmental conditions, and operator performance by monitoring the accuracy and precision of test performance over time [11]. Laboratories should ensure that appropriate QC procedures are designed to monitor all aspects of the

sequencing process. The controls should be designed to confirm that the previously established performance specifications are met for each run of a patient sample [14]. The significant emphases on data analysis and evolving informatics pipelines for NGS require special attention. Controls to monitor the performance of the informatics analysis should be developed and implemented.

The analytic quality of the sequencing run is monitored using characterized reference materials. These reference materials are used in quality control procedures during patient testing [1, 2, 11, 14]. It is not feasible to use or develop control or reference materials that include every possible nucleotide variation with clinical significance in a targeted genomic region. The use of naturally occurring sequence variations of similar type to those that are disease associated, when present, may be useful surrogates [14]. While it is ideal to utilize a variety of controls, the use of a single characterized external control may be sufficient to demonstrate analytical quality. Alternatively, multiple controls which offer a broad range of sequence variations, both disease-associated and natural, may be used. Internal controls may also be used to monitor assay performance. For example, a characterized DNA reference material (e.g., genomic DNA, human DNA clone, or a nonhuman, synthetic control nucleic acid) may be added to patient samples as a "spike-in control." Analysis of an internal control provides an assessment of performance that should not change from run to run [14]. It also monitors conditions in each individual sample. However, the spiked-in control has limited utility because it does not monitor the human DNA extraction process, and the genetic composition of synthetic controls does not represent the genomic complexity of patient samples. It is also possible to use a control sequence that is intrinsic to the patient's sample, but does not reside in regions of the genome targeted by the test (e.g., a highly conserved housekeeping gene or a mitochondrial genome marker) [14]. Combinations of spiked-in and intrinsic controls may be useful for ongoing quality assessment of the testing process.

It is desirable to detect failures in an NGS test as early as possible in the process. This saves both time and other resources that otherwise would be used in taking a failed test to completion. To accomplish this, "quality checkpoints" can be designed into the assay as steps where failures can be detected. These checkpoints may include evaluation of the DNA fragmentation pattern for library preparation and assessment of selected metrics such as quality scores, coverage, transition to transversion (Ti/Tv) ratio, and GC content, during and after the sequencing process [7, 14]. Such stepwise evaluations may identify factors that would predict impending assay failures as well as procedural problems such as errors made during sample preparation and application of sequencing reactions to the instrument.

2.4 Confirmatory Testing and Use of Alternate Test Methods

Current guidelines recommend the use of confirmatory testing for positive results by the same or a different method when deemed necessary, such as for tests that have a propensity for false-positive results [8–10]. Confirmatory testing of all

clinically relevant variants detected by NGS is recommended at this time because NGS is a relatively new technology, clinical laboratory experience is limited, and the current state of technology is prone to false-positives [15–17]. The false-positive rate for various NGS platforms has been estimated with an upper limit of 7.8 % in one study [18]. Confirmatory testing may be achieved with another test method which is clinically validated for the detection of sequence variations. Sanger sequencing has been the method of choice for confirmatory testing because it is well established and versatile; however, SNP panels and other tests are also used. Turnaround time is always a consideration particularly when confirmatory testing is part of the analysis. The confirmatory test should be validated before the primary analysis is available for patient testing to reduce the turnaround time for reporting. Some laboratories performing NGS testing for gene panels have validated Sanger sequencing tests for all exons and other regions that are targeted. This is a greater challenge for exome and whole genome analysis for which a pre-validated confirmatory test may not be practical due to the larger genomic region analyzed. The cost of NGS is decreasing; therefore, using two different NGS platforms to confirm or compare results may be a reasonable approach [14, 19]. Confirmatory testing establishes clinically relevant variant calls and ensures that sample mix-up has not occurred during the analysis when the assay starts with DNA isolated from the primary clinical specimen. In addition to Sanger sequencing, SNP arrays may be used to affirm sample identity.

Some regions of the genome cannot be sequenced accurately using current PCR-based amplification; NGS methods due to genomic complexity, such as high GC content; or areas with repetitive sequences [20]. These regions yield low or uneven sequencing coverage which should be defined during assay validation and excluded from the reportable range of the NGS test [14]. Other methods, such a Sanger sequencing, may be used to complete the coverage of a genomic region or "fill in" areas for which NGS cannot produce a reliable test result [7, 14].

3 Metrics: New Paradigms for Next-Generation Sequencing

3.1 Accuracy

The Clinical and Laboratory Standards Institute defines accuracy as "the closeness of agreement between a measured value and the true value" [9]. The accuracy of a NGS test is influenced by many factors, such as the quality of individual base calls and the depth of coverage [21]. Current guidance recommends the use of quality scores (Q scores) as a quantitative measure of base call accuracy. Quality or PHRED scores have been used for Sanger sequencing to define the likelihood that a base call is accurate [22]. To assign each base a Q score, the quality of each base called by the instrument is evaluated by assessing the strength of a signal relative to the background across a read length (signal-to-noise ratio) and to neighboring

bases. For example, a base assigned to a Q score of 20 has a 1 in100 likelihood of error. For NGS, quality scores are not directly comparable across platforms because of inherent differences in the way that each manufacturer calculates these values based on their unique technology [23]. More accurate Q scores, or confidence scores, can be calculated later in the analysis using base quality recalibration algorithms to correct for covariates. Recalibration takes into account the confidence in alignment to the reference sequence, aspects of the sequencing technology, depth of coverage achieved, the detection of a second allele (heterozygosity), and other criteria [24, 25].

Achieving an adequate depth of coverage is necessary to produce accurate base calls [23]. The required depth of coverage across the region(s) sequenced varies as a consequence of the type of sequence variation present, the flanking sequence context, and zygosity. Depth of coverage includes consideration of both the average depth of coverage and the uniformity of coverage. Average depth of coverage is the average number of overlapping reads within the region of the genome sequenced [14]. The uniformity of coverage is the distribution of coverage across all regions sequenced [14]. The average coverage established during test validation is set to achieve accurate base calling over the region of the genome targeted for analysis. Early adopters of NGS in the clinical laboratory setting have established average coverage thresholds that range from 15X to 100X [14]; however, this depth will be dependent on the assay design and technology [15, 16]. A separate threshold, termed the minimum base coverage threshold, may be established to define the minimum depth at which a base can be called with confidence and identify areas of low coverage in which a variant cannot be reliably called. When the minimum coverage threshold of a targeted area in a gene panel assay is not achieved, or a specific region is problematic, an alternate method such as Sanger sequencing may be used (in place of or in parallel to NGS) to obtain more accurate results for that region of the genome. For example, the first exon of many genes is GC rich, which is challenging for the current NGS platforms to analyze and therefore necessitates the use of an alternate method [15]. It is also helpful to monitor the distribution of forward and reverse strand reads generated during each run and compare them to those documented in the assay validation to minimize false-positive and false-negative calls as a consequence of strand bias [21]. Related to, but distinct from the depth of coverage, is the allelic read percentage or allelic fraction, defined as the proportion of individual reads required to establish an accurate variant call [23]. Ideally, homozygous variants would contain the variant in every read with an allelic read percentage of 100, while a heterozygous variant should contain the variant in 50 % of the reads for an allelic read percentage of 50. However, the allelic read percentage may be misleading due to amplification bias or duplicate reads [14, 23]. Duplicate reads should be removed prior to analysis as they may alter the allelic fraction or incorrectly indicate the presence of strand bias [23]. When the allelic read percentage is beyond the limits of an established performance range, there is an increased risk for generating false-positive or false-negative results.

3.2 Precision

Precision refers to the degree of agreement between replicate measurements of a single analyte or group of similar analytes [7]. In the clinical laboratory, precision is established by testing an "adequate" number of samples and assessing reproducibility (between-run precision) and repeatability (within-run precision). Reproducibility assesses the consistency of results with the same sample under different conditions, within runs, run to run, and day to day taking into account other factors such as different operators. Repeatability may be established by sequencing the same samples multiple times under the same conditions and evaluating the concordance of variant detection and performance. Also, a single sample is not likely to contain the entire spectrum of clinically relevant mutations necessitating the use of multiple controls [14]. When assessing precision it may be useful to evaluate the concordance of other parameters such as the depth of coverage and allelic read percentage.

3.3 Analytic Sensitivity and Analytic Specificity

Analytic sensitivity is the lower limit of detection or the proportion of biological samples that have a positive test result and are correctly classified as positive [11]. For both Sanger sequencing and NGS assays, analytic sensitivity may be expressed as the likelihood that an assay will detect a sequence variation when one is present within the genomic region that was sequenced [14]. Analytic specificity is the probability that an NGS assay will not detect sequence variation(s) when none are present within the genomic region that is analyzed. For NGS, these specifications are typically established through comparison of test results to those obtained from an independently validated method, such as Sanger sequencing or SNP array analysis. However, the usefulness of an array as an alternative method will depend on the location, number, and distribution of SNPs that are included [26]. Discordance between SNP array and NGS data requires additional techniques, such as Sanger sequencing, to resolve discrepancies.

The large number of disease-associated sequence variations that can be detected by both Sanger sequencing and NGS analysis is a challenge to determining the analytic sensitivity and analytic specificity of detecting potential variants. The number of controls to establish analytic sensitivity and analytic specificity will vary according to the type of sequence variation targeted for analysis (e.g., SNP vs. indel). As with other multianalyte tests, it is not reasonable to analyze samples with a complete range of clinical variants due to the lack of characterized reference materials representing the spectrum of possible variants and the unacceptably high cost and time commitment associated with generating and analyzing the large quantity of data that is generated . It is necessary to establish these performance specifications for each type of sequence variation the assay is designed to detect. This rationale parallels similar recommendations for establishing analytic sensitivity and analytic specificity for chromosome microarray analysis (CMA), which examines the whole

genome for constitutional cytogenetic abnormalities [27]. For CMA, a minimum of 30 specimens with disease-associated chromosomal abnormalities should be evaluated during test validation [28].

A loss of sensitivity and specificity may occur when coverage of a targeted sequence falls below the criteria established during the assay validation. Regions with low coverage should be analyzed with another method, such as Sanger sequencing, or excluded from the analysis. Coverage, sensitivity, and specificity are determined at multiple steps of the NGS data analysis process, including initial base calling, quality score assignment, and the informatics processing where the confidence in the alignment and variant calls are calculated.

NGS is prone to both false-positive and false-negative results [17]. The laboratory should document the false-positive and false-negative rates for the regions targeted for analysis during test validation [7, 9, 14]. It is useful to include prevalent disease-associated and problematic variants in the determination of false-positive and false-negative rates because these are not always detected even when quality measures for the surrounding region are satisfactory [29]. An alternate method should be used for genomic regions that are prone to a high false-negative or false-positive rate of base calls [14].

3.4 Reportable and Reference Range

In the United States, the CLIA regulations define the "reportable range" as "the span of test result values over which the laboratory can establish or verify the accuracy of the instrument or test system measurement response" [2]. Reportable range for NGS may be defined as the portion of the genome for which sequence information can be reliably derived for a defined test system [14]. This may not be a contiguous region of the genome, for example, when a gene panel or the exome is sequenced.

The US CLIA regulations define the "reference range" (or reference intervals) as "the range of test values expected for a designated population of persons" [2]. For NGS, the reference range may be defined as the range of normal sequence variations within the population that the assay is designed to detect [14]. The application of this metric to sequence analysis is problematic because a clear definition of "normal variation" is not always evident, may vary among populations, and may inconsistently correlate with a disease association. Efforts to develop databases which map variations within and among populations will be necessary to define the normal variation and its disease association within a defined population.

4 Proficiency Testing and Alternate Approaches

Proficiency testing (PT) or External Quality Assessment (EQA) is an important component of clinical laboratory quality assurance that compares analytical test performance when a sample is independently assessed among multiple laboratories.

It also provides an independent measure of laboratory performance by comparison to a standard reference or consensus result [1, 30–33]. This comparison may identify analytical and interpretive errors as well as problems with QC, calibration, or assay design. In the United States, current regulations require that clinical laboratories perform an assessment of their analytical performance at least twice per year [2]. This requirement can be fulfilled by participation in a PT/EQA program or by performing an alternative assessment. Similar requirements are also described in ISO-15189 [1]. Laboratories enrolled in PT/EQA programs receive samples that they test and analyze using the same procedure(s) used for testing of clinical samples. Participating laboratories are not provided information about the characterization of the test samples. The participants return their test results to the PT/EQA program, which compares and reports the summary data from all laboratories and comments to participants.

Most PT/EQA schemes for molecular genetic tests of germ line or somatic disorders are disease or gene specific. They assess the ability of the participating laboratories to detect mutations in one or a few genes associated with a particular disorder. Participants are evaluated on their ability to detect the expected variants and interpret the diagnostic significance of the resulting genotype within the limitations of their laboratory test method. Some programs offer method-based schemes which assess the ability of the laboratory to correctly utilize a particular technique, such as Sanger sequencing or CMA, independent of the disorder. Method-based schemes are especially useful for evaluation of techniques that examine large regions of the human genome, rather than a specific gene or genes. Method-based proficiency testing can be used to assess the ability to detect the selected variant used in the PT/EQA survey, and the results can be used to infer the successful performance of the test for the detection of sequence variations in other parts of the genome.

Ideally, PT/EQA should evaluate all phases of the testing process, including sample acquisition, DNA preparation, analytical procedures, data analysis, and clinical interpretation [34]. Although a few molecular genetic PT/EQA programs distribute samples, such as whole blood to simulate patient specimens, most utilize lyophilized DNA from human cell lines because this material is more readily available, and a homogenous, stable, and noninfectious sample can be produced and distributed to a large number of participants. It does not, however, permit evaluation of DNA isolation procedures [34, 35].

Laboratories may also use an alternate approach for assessment when a formal PT/EQA program is not available or appropriate for the test that they perform as required by local regulations or accreditation bodies [1, 2, 31, 32, 36, 37]. Alternative assessment may be performed by exchanging blind samples with another laboratory, retesting de-identified patient samples, or testing DNA from cell lines with previously determined genotypes. Several PT/EQA programs including CAP and the United Kingdom National External Quality Assessment Service (UKNEQAS) organize and facilitate sample exchanges among laboratories which perform tests for the same or similar disorders. Alternative assessment may be a useful substitute

for participation in a PT/EQA program; however, there are some shortcomings. For example, retesting samples within one laboratory may not identify systematic errors when only a single testing method is used. Sample exchanges with a limited group of laboratories may not permit a meaningful comparison of the performance of different methods, and it may be difficult to resolve discrepancies and/or maintain anonymity of participants.

A method-based approach is recommended for designing PT/EQA challenges for NGS because there are significant variations among laboratories with respect to the instrumentation, testing algorithms, and the regions of the genome targeted for a given indication for testing [14]. Organizers of method-based schemes need to consider the limitations of each laboratory's implementation of NGS, the methods used for sequence generation and analysis, and the potential for variation in the capacity to generate high-quality sequence data of regions targeted by the PT/EQA challenge. For example, laboratory tests vary with respect to what regions of the genome are targeted, particularly for gene panel testing. Therefore, the difficulty is to develop a means to evaluate data that is derived from different regions of the genome in a way that permits an assessment of each laboratory's testing algorithm. Using this approach, participants would not be penalized for the inability to detect variants that are outside the scope of their validated clinical test(s). Participants could also be required to interpret their findings and provide a report, which can be evaluated for accuracy and completeness [14]. As of 2012, several PT/EQA schemes for NGS are planned, but not yet offered.

Sample materials for PT/EQA include biological specimens and electronic data files. PT/EQA programs for NGS may be most useful when each challenge assesses the capability to identify sequence variations in each region of the genome targeted by the clinical test being evaluated. Development of PT/EQA samples for NGS that have a complete set of disease-associated sequence variations that may occur in all genes included in a panel is not practical. The analysis of naturally occurring sequence variations to augment those that are disease associated may provide a more viable solution. Samples with a variety of genetic variants (indels, SNPs, CNV, etc.) located in relevant genomic regions may be selected from previously tested patient specimens or from the large number of publicly available genomes that have been sequenced previously [38, 39]. Electronic data files of actual or simulated data may also be used as PT/EQA or alternate assessment samples to evaluate the informatics data analysis pipeline and its capability to make correct variant calls. Electronic files may be modified to contain a broad spectrum of sequence variants that can be used to evaluate test performance. The files may evaluate the participants' ability to detect challenging sequence alterations in a variety of clinically important genomic regions. Electronic files should be designed to be compatible with the data analysis pipeline, software capabilities, and test design of participating laboratories [14]. PT/EQA programs will need to assure compatibility of file types with participant's DNA analysis pipelines when electronic files are used.

5 Reference Materials

A reference material is defined as a "material or substance, one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of a measuring system, the assessment of a measurement procedure, or for assigning values to materials" [40, 41]. The appropriate use of reference materials is essential for complying with regulatory standards and professional guidance, and to assure the quality of the analytical phase of the testing process [1–5, 9, 11, 42–44]. Clinical laboratories use reference materials for a variety of quality assurance purposes that include assay development and validation, QC procedures, PT/EQA, and alternative assessment. Genomic DNA from human cell lines or residual patient samples is often used as a reference material because it closely resembles actual patient specimens and is best suited for QC and other procedures designed to establish, monitor, and verify the reliability of the assay [45]. Electronic data files may also be used as reference materials for NGS to develop and monitor the quality of the post-sequencing data analysis steps of the test.

Reference materials must be well characterized, usually by a variety of methods, to assure that they work properly in the clinical test. Characterization of human genomic DNA is complex and often includes analysis on multiple platforms and settings to ensure its usefulness in different clinical environments and to mitigate systematic biases which may be introduced by a particular platform or analysis software. In addition to NGS sequence analysis, characterization may include the use of a variety of analytical methods such as SNP array analysis, CMA, and Sanger sequencing. Synthetic DNA samples should also be characterized to verify their sequence, assess performance with applicable enrichment methods and sequencing technologies, and to identify interference with the detection of sequence variants when spiked into genomic DNA samples.

Reference materials may be created from a variety of starting materials. Genomic DNA from blood or human cell lines is most similar to clinical specimens in its complexity and performance in NGS assays. However, genomic DNA derived from blood may not be available in large amounts, and the source may not be as easily renewable as cell line-derived DNA. Nevertheless, DNA from blood is often preferred because it may be more homogenous and will not have genomic rearrangements or loss of DNA due to viral transformation and/or extended cell culture passages. Synthetic DNA reference materials, including plasmids, yeast, and bacterial artificial chromosome (YAC and BAC) clones, may be manufactured in large batches and designed to contain many genetic variant types. These materials may also be spiked into genomic samples to serve as internal controls. Synthetic DNA reference materials generally do not represent the entire genome, exome, or gene panels and are not similar in complexity to genomic DNA. These limitations must be considered in evaluating their usefulness for NGS [45, 46].

In 2012, a number of efforts are underway to develop reference materials that will be useful for validation, QC, and PT for clinical NGS. In collaboration with the National Center for Biotechnology Information (NCBI) [47], the CDC's Genetic

Testing Reference Materials Coordination Program (GeT-RM) [15, 48] has initiated a collaborative project to coordinate the characterization of reference materials for clinical NGS applications. In addition, the National Institute of Standards and Technology (NIST) has organized the "Genome in a Bottle Consortium" to develop a set of reference materials, reference data, and reference methods needed to assess performance of human genome sequencing [49].

6 Other Considerations

6.1 Next-Generation Sequencing: Interpretation of Sequence Results

Regulatory requirements and professional recommendations include specification for the reporting of molecular clinical genetics laboratory test results [2, 7–10, 12, 13]. Laboratory professionals are best positioned to report the significance of the sequence variations with regard to the indication for testing and the limitations of the test, which is needed by clinicians to make meaningful clinical decisions. Several professional recommendations provide guidance on result reporting [10, 14]. Standard nomenclature is essential and recommended for describing the clinically relevant sequence variations that are reported [10]. The Human Genome Organization Gene Nomenclature Committee has developed a system for the standardization of gene names [29]. The Human Genome Variation Society Ad Hoc Committee on Mutation Nomenclature has developed a uniform system for describing specific sequence variations [50].

The identification of clinically relevant sequence variations using NGS requires an informatics analysis that aligns, annotates, and accurately determines variant calls. Prioritization of the assigned variants is performed to identify those that are most likely to be disease associated. Optimal alignment depends on the use of well-characterized reference sequence(s). Currently, there is no clinical standard reference material or sequence available for NGS applications, although several are in development. In an effort to encourage the use of a common reference sequence for human gene sequencing, the REFSEQ and RefSeqGene databases have been established [51]. Although the number of well-characterized sequences entered into this database is increasing, a clinical standard reference sequence for exome and whole genome analysis remains undefined. One option is the use of the current human genome assembly, maintained by the US National Institutes of Health's National Center for Biotechnology Information (NCBI) website [52]. Genome builds continue to be refined; therefore, it is necessary to document the genome build that was used as a reference sequence when test results are reported.

The American College of Medical Genetics has developed guidance for interpreting and reporting sequence variations [53]. Six categories were established: known pathogenic, likely pathogenic, unknown pathogenicity, likely benign, known

benign, and benign variant associated with clinical presentation [53]. Large regions of the genome are sequenced using NGS, and many sequence variations are frequently identified; thus, laboratories must apply algorithms to identify genes and sequence variations that may be relevant to the clinical question posed. These algorithms take into consideration patient data, knowledge about other family members, published data about the disease association(s) of the gene, and sequence variations implicated by the NGS test. These algorithms continue to be refined, and standards do not yet exist for these types of analyses.

6.2 Management and Reanalysis of the Electronic Data

The large amount of data generated by NGS poses unprecedented challenges for data analysis, management, and storage. Laboratories may store and analyze NGS data in-house or off-site, for example, through cloud-based computing [54]. Privacy and confidentiality considerations are important when patient data is collected, analyzed, stored, and communicated. In the United States, these issues are addressed by the Health Insurance Portability and Accountability Act (HIPPA) Privacy Policy, CLIA regulations, and other state and local standards [2, 11].

Good laboratory practice guidelines recommend that laboratories should retain sufficient data for reanalysis to confirm the accuracy of the initial test result [11]. This includes recording the versions of the reference sequence and analysis software that was used. This is important when questions arise about the reliability of a test result. The challenge in complying with this recommendation for NGS is that it is often cost-prohibitive to retain the enormous quantity of data that is generated, which is on the order of several terabytes for whole genome analysis. This issue has not been resolved, but several approaches have been taken. Most laboratories retain the full data set for a limited time, such as to the completion of the next PT/EQA or alternate assessment challenge. Some laboratories save a limited data set, usually the list of variants identified and various quality parameters. Some laboratories maintain a limited data set and retain the patient specimen or sample for reanalysis if questions arise because it is concluded that this is cheaper than the cost associated with storage of a large data set [54].

Recent recommendations suggest that patient data should be reanalyzed when new scientific findings alter the test result and/or interpretation [11, 14]. For example, refinement of a reference sequence may aid in the identification of disease-associated variants which were not previously identified. New clinical evidence may change the categorization of disease association for variants identified during the testing process with the potential to change what is reported and/or the interpretation of the test result. Reanalysis may be indicated when no clinically significant findings were derived from the initial NGS test or new published literature provides data that might change the interpretation of the reported test result. The infrastructure and logistics that must be created to trigger reanalysis often does not exist, and it may be cost-prohibitive for laboratories to actively monitor and

reanalyze patient data. One approach that may address this challenge is to build the capacity for reanalysis into information technology and clinical decision support systems.

Good laboratory practice dictates retention of the test report. In the United States, the CLIA regulations require retention for at least 2 years following the date of reporting [2]. However, for molecular genetic tests, it is recommended that reports be retained for a longer time, at least 25 years after the date of reporting [11]. This is due to the long-term and potential lifetime implications that genetic results may have for the patient and family members.

6.3 Genomics and the Electronic Health Record

The representation of genomic data in the electronic health record (EHR) and its criteria for accurate communication among providers and patients are changing. National standards for the EHR are in development in the United States, and the inclusion of genomic data presents unique challenges [55]. There are no standards for the representation of genomic data or its linkages to clinical data at the present time. Electronic systems can support active clinical decision algorithms that incorporate emerging professional guidance with new findings from the clinical research community. This capacity may be a powerful tool for enhancing the quality of patient care, but the development, deployment, and QC of such systems is a significant endeavor. The interoperability of data contained within the EHR requires standard mechanisms for communicating clinical data among computer systems. Health Level Seven International (HL7) is the global authority on standards for interoperability of health information technology. The HL7 Clinical Genomics Working Group is tasked with creating standards for the inclusion of genomic data and its linkages to relevant clinical data within the HL7 framework [56].

6.4 Guidance and Oversight of Clinical NGS as Technologies Evolve

Diverse and unique sequencing technologies will continue to enter clinical laboratory practice and will require development of additional professional guidance. There are an increasing number of benchtop instruments entering the marketplace much smaller than the first generation of instruments with costs that are manageable for clinical laboratories. Therefore, many more laboratories may offer tests that utilize NGS. At the same time, software for the analysis of NGS data is evolving. It will be important to have guidance that advances as technology moves forward to assure the continued provision of quality clinical testing.

7 Conclusions

NGS offers the first practical means to sequence large regions of the human genome and identify sequence variations that are clinically relevant for an individual patient. In 2012, regulatory standards and professional guidance relevant to NGS are beginning to emerge. Establishing performance specifications is complex because NGS testing is a multistep process that generates large data sets that are analyzed using sophisticated algorithms. Performance specifications are initially established during test validation and are used to develop quality control procedures that are applied during the testing of patient samples. For positive findings, confirmatory testing is generally recommended because of several factors that include the propensity for false-positive results and limitations inherent in the validation protocol in light of the large number of sequence variations that may be detected. Another limitation is that NGS is not able to provide high-quality sequence data for all regions of the genome, and in these instances alternate methods must be used.

NGS is rapidly evolving and is anticipated to have a major influence on molecular laboratory testing in the near future; however, a number of challenges remain. Clinically characterized reference materials for NGS testing are in development. Proficiency testing programs for NGS are in their infancy and will be critical to establish interlaboratory comparability. Since the majority of informatics pipelines are developed in-house, access to informatics expertise has been important to assure proper implementation and optimization of the software component of NGS testing. The quantity of data generated challenges current storage capacities, but this is expected to be less of a problem as this technology evolves. These challenges significantly influence the level of guidance and standards that can be developed at this time because the technology and methods are not yet mature. Nonetheless, the potential for routine large-scale analysis of the human genome is significant, and it is anticipated that these challenges will be met in the not too distant future.

Acknowledgment We wish to thank Dr. Nazneen Aziz for the information related to the activities of the College of American Pathologists.

The work was supported in part by an appointment of Dr. Amy S. Gargis to the Research Participation Program at the Centers for Disease Control and Prevention administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the US Department of Energy and CDC.

Disclaimer: The findings and conclusions in this chapter are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention/The Agency for Toxic Substances and Disease Registry.

References

1. ISO/IEC 15189 (2007) Medical laboratories—particular requirements for quality and competence. ISO, Geneva
2. Centers for Medicare & Medicaid Services, Centers for Disease Control and Prevention (2003) 42 CFR Part 493. Medicare, Medicaid, and CLIA programs; laboratory requirements relating

to quality systems and certain personnel qualifications. Final Rule:3640–3714. http://www.gpo.gov/fdsys/browse/collectionCfr.action?collectionCode=CFR. Accessed 3 April 2013
3. College of American Pathology Laboratory Accreditation Program (2012) http://www.cap.org/apps/cap.portal. Accessed 11 April 2013
4. New York State Department of Health (2012) Clinical laboratory evaluation program, laboratory standards; 2008. http://www.wadsworth.org/labcert/clep/standards.htm. Accessed 11 April 2013
5. Washington State Office of Laboratory Quality Assurance (2012) http://www.doh.wa.gov/LicensesPermitsandCertificates/FacilitiesNewReneworUpdate/LaboratoryQualityAssurance.aspx. Accessed 11 April 2013
6. Federal Food, Drug, and Cosmetic Act (FD&C Act) (2012) http://www.fda.gov/regulatoryinformation/legislation/federalfooddrugandcosmeticactfdcact/default.htm. Accessed 11 April 2013
7. American College of Medical Genetics Policies and Standards (2012) www.acmg.net. Accessed 29 July 2012
8. American College of Medical Genetics (2008) ACMG standards and guidelines for clinical genetic laboratories. http://www.acmg.net/AM/Template.cfm?Section=Laboratory_Standards_and_Guidelines&Template=/CM/HTMLDisplay.cfm&ContentID=7439. Accessed 11 April 2013
9. CLSI (2012) Molecular methods for clinical genetics and oncology testing; approved guideline, 3rd edn. CLSI document MM01-A3. linical Laboratory Standards Institute, Wayne
10. NCCLS (2004) Nucleic acid sequencing methods in diagnostic laboratory medicine; approved guideline. NCCLS document MM9-A [ISBN 1-56238-558-5]. NCCLS, 940 West Valley Road, Suite 1400, Wayne, Pennsylvania 19087-1898 USA
11. Chen B, Gagnon M, Shahangian S, Anderson NL, Howerton DA, Boone JD (2009) Good laboratory practices for molecular genetic testing for heritable diseases and conditions. MMWR Recomm Rep 58(RR-6):1–37
12. Eurogentest (2012) http://www.eurogentest.org/laboratories/. Accessed 11 April 2013
13. Organization for Economic Cooperation and Development (OECD) (2007) OECD guidelines for quality assurance in molecular genetic testing. OECD, Paris, p. 33–35
14. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T et al (2012) Assuring the quality of next-generation sequencing in clinical laboratory practice. Nat Biotechnol 30(11):1033–1036
15. Jones MA, Bhide S, Chin E, Ng BG, Rhodenizer D, Zhang VW et al (2011) Targeted polymerase chain reaction-based enrichment and next generation sequencing for diagnostic testing of congenital disorders of glycosylation. Genet Med 13(11):921–932
16. Gowrisankar S, Lerner-Ellis JP, Cox S, White ET, Manion M, LeVan K et al (2010) Evaluation of second-generation sequencing of 19 dilated cardiomyopathy genes for clinical applications. J Mol Diagn 12(6):818–827
17. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M et al (2011) Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30(1):78–82
18. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY et al (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol 10(3):R32
19. Clark MJ, Chen R, Lam HYK, Karczewski KJ, Chen R, Euskirchen G et al (2011) Performance comparison of exome DNA sequencing technologies. Nat Biotechnol 29(10):908–914
20. Zhang W, Cui H, Wong LC (2012) Application of next generation sequencing to molecular diagnosis of inherited disease. Top Curr Chem. doi:10.1007/128_2012_325
21. Ajay SS, Parker SCJ, Ozel Abaan H, Fuentes Fajardo KV, Margulies EH (2011) Accurate and comprehensive sequencing of personal genomes. Genome Res 21(9):1498–1505
22. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–185
23. Voelkerding KV, Dames S, Durtschi JD (2010) Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic

cardiomyopathy: a paper from the 2009 William Beaumont Hospital Symposium on Molecular Pathology. J Mol Diagn 12(5):539–551
24. Li H, Homor N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Breif Bioinform 11(5):473–483
25. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498
26. Welch JS, Westervelt P, Ding L, Larson DE, Klco JM, Kulkarni S et al (2011) Use of whole-genome sequencing to diagnose a cryptic fusion oncogene. JAMA 305(15):1577–1584
27. Shaffer LG, Beaudet AL, Brothman AR, Hirsch B, Levy B, Martin CL et al (2007) Microarray analysis for constitutional cytogenetic abnormalities. Genet Med 9(9):654–662
28. Human Gene Organization Human Gene Nomenclature Committee (2012) http://www.genenames.org/. Accessed 11 April 2013
29. CLSI (2007) Using proficiency testing to improve the clinical laboratory; approved guideline, 2nd edn. CLSI document GP27: Clinical Laboratory Standards Institute, Wayne, PA
30. CLSI (2008) Assessment of laboratory tests when proficiency testing is not available; approved guideline, 2nd edn. CLSI document GP29-A2. Clinical and Laboratory Standards Institute, Wayne, PA
31. CLSI (2005) Proficiency testing (external quality assessment) for molecular methods; approved guideline. CLSI document MM14-A. Clinical Laboratory Standards Institute, Wayne, PA
32. ISO/IEC 17043 (2010) Conformity assessment—general requirements for proficiency testing. ISO, Geneva, Switzerland
33. Bellissimo DB (2007) Practice guidelines and proficiency testing for molecular assays. Transfusion 47(1 Suppl):79S–84S
34. Ramsden SC, Deans Z, Robinson DO, Mountford R, Sistermans EA, Grody WW et al (2006) Monitoring standards for molecular genetic testing in the United Kingdom, The Netherlands and Ireland. Genetic Test 10(3):147–156
35. Richards CS, Grody WW (2003) Alternative approaches to proficiency testing in molecular genetics. Clin Chem 49(5):717–718
36. Organization for Economic Cooperation and Development (OECD) (2007) OECD guidelines for quality assurance in molecular genetic testing. http://www.oecd.org/science/biotechnologypolicies/38839788.pdf. Accessed 11 April 2013
37. 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073
38. International HapMap 3 Consortium (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58
39. ISO 15195 (2003) Laboratory medicine – requirements for reference measurement laboratories. ISO, Geneva
40. Emons H, Fajgelj A, van der Veen AMH, Watters R (2006) New definitions on reference materials. Accred Qual Assur 10(10):576–578
41. American College of Medical Genetics (2012) Standards and guidelines for clinical genetics laboratories 2006 edition. http://www.acmg.net/Pages/ACMG_Activities/stds-2002/g.htm. Accessed 11 April 2013
42. Association for Molecular Pathology (1999) Association for molecular pathology statement: recommendations for in-house development and operation of molecular diagnostic tests. Am J Clin Pathol 111(4):449–463
43. Chen B, O'Connell CD, Boone DJ, Amos JA, Beck JC, Chan MM et al (2005) Developing a sustainable process to provide quality control materials for genetic testing. Genet Med 7(8):534–549
44. CLSI (2008) Verification and validation of multiplex nucleic acid assays; approved guideline. CLSI document MM17-A. Clinical and laboratory Standards Institute, Wayne
45. Strom CM, Janeczko RA, Anderson B, Redman J, Quan F, Buller A et al (2005) Technical validation of a multiplex platform to detect thirty mutations in eight genetic diseases prevalent in individuals of Ashkenazi Jewish descent. Genet Med 7(9):633–639

46. Human Sequence Variation Society (2012) http://www.hgvs.org/. Accessed 29 July 2012
47. National Center for Biotechnology Information (2012) http://www.ncbi.nlm.nih.gov/. Accessed 11 April 2013
48. Genetic Testing Reference Materials Coordination Program (GeT-RM) (2012) http://wwwn.cdc.gov/dls/genetics/RMMaterials/. Accessed 11 April 2013
49. National Institute for Standards and Technology, Genome in a Bottle Consortium (2012) http://www.genomeinabottle.org/. Accessed 11 April 2013
50. Pruitt KD, Tatusova T, Klimke W, Maglott DR (2009) NCBI reference sequences: current status, policy, and new initiatives. Nucleic Acids Res 37(Database issue):D32–D36
51. NCBI Human Genome Resources (2012) http://www.ncbi.nlm.nih.gov/genome/guide/human/. Accessed 11 April 2013
52. ACMG Laboratory Practice Committee Working Group (2000) ACMG recommendations for standards for interpretation of sequence variations. Genet Med 2(5):302–303
53. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB (2011) The real cost of sequencing: higher than you think! Genome Biol 12(8):125
54. The Office of the National Coordinator for Health Information Technology (2012) http://healthit.hhs.gov/portal/server.pt/community/healthit_hhs_gov__home/1204. Accessed 11 April 2013
55. Health Level Seven International (2012) http://www.hl7.org/Special/committees/clingenomics/overview.cfm. Accessed 11 April 2013

Chapter 15
Validation of NGS-Based DNA Testing and Implementation of Quality Control Procedures

Victor Wei Zhang and Lee-Jun C. Wong

Abstract The rapid adoption of next-generation sequencing (NGS) (also known as massively parallel sequencing; MPS) techniques has revolutionized the way molecular diagnosis is performed in a clinical diagnostic laboratory. Hundreds to thousands of genes can be analyzed simultaneously, and samples can be multiplexed and sequenced in parallel with deep coverage. To bring such a complex technology to clinical diagnostic laboratories, limitations and challenges should be recognized when designing and implementing assays for routine clinical use. While most NGS platforms can generate enormous amount of data in a massively parallel manner, the complexities of test validation and the implementation of quality control procedures have posed tremendous challenges on quality assurance. A quality assurance program is an integral part of clinical laboratory operations, in particular for laboratories that are adopting new technologies to provide state-of-the-art genetic testing services. In this chapter, we describe the steps taken in implementation of test validation and quality control procedures in our clinical molecular diagnostic laboratory.

1 Introduction

For many years, Sanger dideoxy sequencing has been the method of choice for clinical molecular diagnostic laboratories. It continues to be the preferred method when a genetic disease has a clear clinical diagnosis or when other biochemical, molecular, and imaging evidence supporting a specific genetic disease. Sanger sequence

V.W. Zhang (✉) • L.-J.C. Wong
Medical Genetics Laboratories, Department of Molecular and Human Genetics,
Baylor College of Medicine, One Baylor Plaza, NAB 2015, Houston 77030, TX, USA
e-mail: wzhang2@bcm.edu

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_15,

analysis of a limited number of candidate genes carried out one by one is also the "gold standard" approach for a disease involving several steps of the same developmental or metabolic pathway or where different otherwise unrelated genes cause the same disease, for example, retinitis pigmentosa. However, some diseases are clinically and genetically heterogeneous, making a clear molecular diagnosis difficult due to atypical clinical presentations or ambiguous pathological or biochemical findings that do not give a clear indication regarding the potential defective genes [1]. An extreme situation reflecting this scenario is the diagnosis of mitochondrial disorders, for which there is significant clinical overlap, and where more than 1,300 genes are responsible for the biogenesis and normal function of mitochondria [2–5]. Sequencing candidate genes one by one will be costly, time consuming, and inefficient. The delay in establishing a specific diagnosis and high cost incurred will impede appropriate patient management and delay family counseling and possible future prenatal diagnosis.

The rapid adoption of NGS technologies has revolutionized molecular diagnostics [6]. These advances have facilitated the rapid identification of new disease genes in genetic research. The testing paradigm is shifting from single gene analysis by Sanger sequencing to multiple target genes and even whole exome analyses by MPS [7–10]. Panel testing has gradually become the first line of choice for molecular testing of genetically heterogeneous complex diseases when there is a clear indication of an underlying genetic cause [11–14]. When compared to serial gene-by-gene Sanger sequencing analysis, this approach can significantly reduce costs and the time required to make a molecular diagnosis.

With the steady decrease in sequencing costs and an increase in data throughput, it can be expected that both whole exome and targeted phenotype/pathway-focused gene panels will be the major methods in clinical diagnostic laboratories for the foreseeable future. The rapid adoption of these new technologies for detecting inherited genetic disorders by clinical molecular diagnostic laboratories requires new standards for quality assurance. In order to provide NGS-based clinical service, the performance characteristics of the new technologies must be validated to meet requirements set by College of American Pathologists (CAP) and/or Clinical Laboratory Improvement Amendments (CLIA) [15]. For practical clinical diagnostic applications, the high-throughput nature of the instruments requires an additional layer of quality assurance.

In this chapter, we describe NGS-based test validation and quality control procedures currently used in our clinical molecular diagnostic laboratory in order to satisfy federal regulatory requirements.

2 Overview of NGS-Based Performance Characteristics

Limitations and challenges always accompany the development and implementation of novel technologies in the clinical diagnostic setting. The ability to parallel sequence a large number of genes with deep coverage, i.e., each nucleotide

sequenced numerous times, requires highly specialized analytical tools in order to analyze enormous amounts of data, annotate large numbers of variants, and interpret complex results in the clinical context. High-throughput sequencing data has higher error rates, and the nature of these errors differs depending on the chemistry and detection methods used by the specific instrument. These sequence errors will produce a high false-positive rate of variant calls, which leads to unnecessary work in interpretation, confirmation, and reporting. Recent studies have shown that high rates of discordant variant calls occur among different NGS-based technologies and analytical algorithms when the same individual sample is sequenced [16]. In order to adapt these technologies into clinical diagnostics, stringent validation and confirmation using a second method is necessary to meet clinical standards [1].

In addition to problems with false calls, it is often necessary to pool samples from different individuals in order to lower the sequencing cost. When a large number of clinical samples are prepared at the same time, it increases the likelihood of sample swapping and/or cross contamination. Should these problems occur, the reduced cost due to high throughput may be offset by the effort required for troubleshooting. Therefore, quality control procedures are essential elements to assure accurate and reliable performance of the novel technologies [7, 17–20].

3 Validation of NGS-Based Test

The purpose of a molecular diagnostic test for genetic disorders is to detect pathogenic DNA alterations (mutations) that are responsible for an individual's genetic disease. Unlike other clinical laboratory tests such as immunohistochemistry, the detection of abnormal metabolites, or enzyme assays that measure the biochemical consequences of genetic defects, molecular genetic testing directly searches for the genetic alterations in order to provide a definitive diagnosis of a genetic disease. The superior accuracy of a genetic testing relies on the genetic material of an individual, which for the most part is immutable and not influenced by physiological and environmental conditions. The results obtained by molecular testing of an individual can be very informative for genetic counseling of other family members and may also be applicable to prenatal diagnosis.

The validation process of NGS-based tests should include three major interconnected components: the platform, the specific test, and the informatics pipeline [31]. A full test validation includes the validation of the platform used for performing the test and the bioinformatics pipeline used for the analysis of the sequence results and variant annotation. Platform validation establishes the performance specifications of the sequencing platform and the types of variants detected by the test. Validation of the informatics pipeline establishes the analytical software parameters necessary to provide accurate sequence data and the detection of variations within the targeted regions. Test validation documents the specificity, sensitivity, and limitations of the assay used to detect clinically significant sequence variations.

Ideally, a complete validation of an NGS-based test includes three steps. The first step is to establish the analytic sensitivity (the likelihood of detecting a sequence variation when one is present within the target regions sequenced) and analytic specificity (the likelihood of not detecting sequence variations when they are not present in target regions) by the comparison of test results to those obtained from an independently validated method such as Sanger sequencing or single nucleotide polymorphism (SNP) array analysis. This validation step is also used to optimize analytical parameters such that the highest accuracy, specificity, sensitivity, and reproducibility can be achieved with the particular NGS platform and informatics pipeline. The second step is to establish the range of detectable mutations and the limits of detection. This is performed by the analysis of a broad range of samples with known types of mutations, including nucleotide substitutions, small indels (insertions/deletions), large deletions, or mutations in a DNA repeat region or a homopolymeric nucleotide track. If the assay involves quantitative analysis, for example, measuring the percentage of a heteroplasmic mtDNA mutation, then control specimens must be included for the evaluation of experimental and analytical errors and the assessment of the limits of detection. This step also validates the analytical and annotation pipelines. Finally, a set of samples from patients with a clinical indication of a genetic disease but without a molecular diagnosis are analyzed to validate the ability of the NGS-based test to detect mutations in undiagnosed patients.

4 Quality Control and Quality Assurance Procedures for a NGS-Based Assay

A robust quality assurance program is an integral part of clinical laboratory operations, in particular for those laboratories that are adopting new technologies in order to provide state-of-the-art genetic testing services. The goal of a quality assurance program is to have a coordinated approach to ensure the quality of service and to deliver accurate and reliable results to the medical community, such that the performance of clinical testing can be reevaluated and improved frequently. The use of a continuous quality assurance program is required for a clinical laboratory to be certified by CLIA or CAP. The quality assurance program needs to effectively monitor the quality of testing procedures within a laboratory, and needs three major components: pre-analytical, analytical, and post-analytical. Quality controls are procedures that are incorporated into each step to monitor the performance characteristics of a test and to immediately detect any errors when they occur. Proficiency testing samples often serve as part of a quality control assessment. However, MPS-based genetic testing is relatively new, its workflow is long, complex, and the quality control system used for Sanger sequencing tests may not be adequate when assessing the entire MPS-based test procedures.

4.1 Pre-analytic Phase

This section describes the quality control procedures implemented in our laboratory, a platform-independent and assay-independent quality control system that can be applied for general use in MPS-based genetic testing in a clinical laboratory. The pre-analytic process starts upon receipt of a sample by the laboratory performing the test, and involves entering the required demographic and test ordering information into the Laboratory Information Management System (LIMS). The accuracy of the ordering information, patient demography, and physician's contact information is indispensable and requires rigorous quality assurance when initiating the cascade of testing activities in the clinical laboratory. At this stage, it is important to carefully determine the error sources if a sample does not meet the sample acceptance criteria of the laboratory. Follow-up procedures should be clearly documented in the laboratory's standard operation procedure (SOP). Errors may include the wrong specimen type or specimen integrity, which can be due to an error on the part of the sender or during shipping or handling. In cases where a test may be cancelled, the tracking log should be reviewed regularly.

4.2 Analytic Phase

In addition to the actual laboratory performance of the sequence analysis, other important components, including trained medical technologists, instruments, and reagents required for carrying out the assay, should be fully validated to satisfy clinical standards. This section describes quality controls that are used to monitor MPS test performance in our laboratory.

4.2.1 Performance of Illumina Sequencer

The manufactory's performance specifications usually need to be established at the time of installation and initial evaluation. Each high-throughput sequencing instrument has its own unique error model associated with the sequencing chemistry and detection method. The bacteriophage PhiX used by Illumina to spike the samples used in the test is of great importance in checking the quality of cluster generation, signal detection, and phasing calibration. The advantage of the PhiX genome is its relatively small size and well-balanced nucleotide composition, which can be used to calculate the sequencing error rate at the same time when sequencing is performed.

Besides the general performance evaluation, each individual base call generated from a run is also given a Phred probability score for base call quality. The value of the Phred quality score (Q score) is the normalized logarithmic value of the likelihood for the instrument to make an incorrect base call. Because the Phred score was

derived from a set of sequencing parameters related to the Sanger method, the high-throughput sequencing process was parameterized to calibrate the mathematic formula against a set of well-defined existing empirical sequencing data with a known accuracy. Low-quality scores may indicate a high possibility of making false base calls. For example, the value of Q10 for Illumina sequencing chemistry represents 1 incorrect base call per 10 sequenced bases, Q20 means an error rate of 1 in 100, and Q30 means 1 error per 1,000 sequenced bases. The Q score decreases as the number of sequencing cycles increases. The acceptable Q value of an NGS-based test is determined by each clinical laboratory. Based on our laboratory experience, the use of higher cutoff Q scores can significantly reduce false-positive calls and the downstream work of variant conformation.

After establishing the sequencing performance, the subsequent runs also need to meet the general sequencing quality metrics specified by the manufacturer. In case the operation of the instruments in a clinical laboratory requires a modification in the procedure, a set of laboratory-specific quality metrics should be documented. The general instrument and laboratory-specific parameters that have been determined during instrument and assay validation and optimization also should be reevaluated in order to assure the quality and accuracy of sequence data.

4.2.2 Sequencing Errors and Quality Control for Quantitative Measurements of Variant Calls (ExQC)

In general, sequencing errors associated with a particular instrument are intrinsic to the instrument and stay relatively constant as long as the sequencing and detection chemistry remains the same. However, the error rates may vary depending on the quality of sample, the DNA template preparation method, and bar-coding and demultiplexing methodologies. To monitor these errors, an external quality control (ExQC) sample must go through the same procedures with each individual sample [21]. Our ExQC samples are synthetic DNA fragments with predefined changes at specific positions that are mixed at different ratios to generate a series of control DNAs with 0.1 %, 0.5 %, 2 %, 5 %, 20 %, and 50 % of a particular variant, as depicted in Fig. 15.1. A small aliquot of the ExQCs mixture is added to each individual sample for indexing, and the DNA template library is prepared and sequenced as part of each indexed sample. The application of this ExQC sample is ideal for the determination of sequencing error rates and the limits of detection of each specific sample. This is extremely important for the detection of variants at low levels, such as those seen in mosaicism, particular infectious species, mitochondrial DNA heteroplasmy, or somatic alterations, and, if the presence or absence of a variant is in question.

Unlike the PhiX control, the ExQC can be used to assess the experimental and analytical error for the entire processing of the specific sample and truly represents the quality control for the sample, as individual samples may be subject to varying quality of sample preparation, including human, instrumental, and computational errors. In our laboratory, the overall experimental error for the quality control

	Variant position						
	1	2	3	4	5	6	
Reference allele	A	T	G	C	G	T	Parts in QC mix
ExQC_1	T	A	T	T	T	G	1
ExQC_2		A	T	T	T	G	4
ExQC_3			T	T	T	G	15
ExQC_4				T	T	G	30
ExQC_5					T	G	150
ExQC_6						G	300
ExQC_7	A	T	G	C	G	T	500
Total ExQC mix	A/T	A/T	G/T	C/T	G/T	T/G	1000
Ratio	1/999	5/995	2/98	5/95	20/80	50/50	
Mutant allele	0.1%	0.5%	2%	5%	20%	50%	
Reference allele	99.9%	99.5%	98%	95%	80%	50%	

Fig. 15.1 External quality control and proportion of nucleotide at each specific variant position

specimen has been determined to be 0.326 % ± 0.335 %. Based upon this, the limit of detection is calculated at ~1.33 %, which is 3 standard deviations above the mean error. The error rate of the testing sample can also be determined and is usually slightly higher than the parallel control sample due to the DNA source and extraction procedures.

4.2.3 Internal Sample Tracking System (InQC)

Due to the high throughput of NGS platforms, many samples are often pooled for sequence analysis in order to maximize the capacity of the instrument. To avoid sample swaps and cross contamination, we have designed an internal sample identity tracking system ("InQC") to be incorporated into DNA template library preparation step with each sample, followed by subsequent analytical steps to ensure that the results indeed belong to the original sample [21]. Our InQC system is a set of 14 polymorphic nuclear markers that are PCR-amplified and sequenced along with the target sequences. If mtDNA is to be analyzed, the PCR products of nuclear polymorphic markers are mixed with the mtDNA template library from the same individual. Meanwhile, a separate DNA aliquot is used for genotyping of the InQC markers by a second method, either by Sanger sequencing or by a TaqMan assay. The genotyping results are used to verify the results obtained by MPS analysis. For mtDNA analysis, in addition to the nuclear polymorphic markers, the identity of the mtDNA sample is also verified by Sanger sequencing of the D-loop and LR-PCR primer regions using a separate aliquot of DNA sample that usually contains about 10–12 variants. If the results are not consistent, the erroneous step will be sought and the analysis repeated for clarification. This is an important and necessary step to assure that there is no sample mix-up or cross contamination, especially when tens or hundreds of samples are processed at the same time and sequenced using the same flow cell.

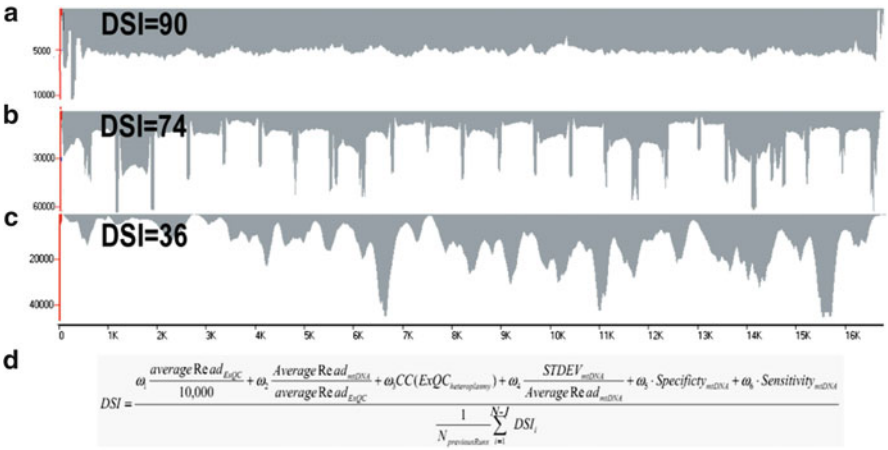

Fig. 15.2 The formula of deep sequencing index (DSI) of MitoNGS

4.2.4 Performance Evaluation by "Deep Sequencing Index" (DSI) Using Mitochondrial Whole Genome Sequencing as an Example

Various target gene enrichment methods and MPS platforms have been used for NGS-based clinical analysis. Each method may have its own unique advantages and disadvantages. Regardless of the method used, the quality of an NGS run can be evaluated by six parameters that define the characteristics of NGS-based assays. We developed a mathematical formula termed the "deep sequencing index" (DSI) that contains these six parameters: (i) the average number of reads mapped to the ExQC DNA, (ii) the average number of sequence reads of the sample normalized to the average number of reads of the ExQC DNA, (iii) the correlation coefficient of the expected versus observed variant percentages of the ExQC DNA, (iv) the ratio of the standard deviation of the average number of sequence reads to the average number of reads mapped to the sample, (v) a specificity function (the percentage of sequence reads mapped to reference mtDNA), and (vi) a sensitivity function (the percentage of sequence variants correctly identified). Using the DSI, different enrichment methods or sequencing platforms can be evaluated and compared in a standardized fashion. For example, the entire mitochondrial genome can be enriched by three different methods: DNA capture by hybridization, multiplex PCR of overlapping DNA fragments, or a long-range PCR (LR-PCR) single amplicon of the whole mitochondrial genome. As shown in Fig. 15.2, the profile of the read coverage clearly demonstrates the uniform coverage of a single amplicon LR-PCR in comparison to the other two gene enrichment methods [21]. The coverage profiles are consistent with the DSI scores (Fig. 15.2). Although all these enrichment methods can achieve reasonable sensitivity and specificity, the best performance also relies on uniformity of coverage and its reproducibility, which is extremely important particularly if large deletions are to be detected [21]. The DSI formula also takes the ExQC into consideration, not only the coverage but also the correlation coefficient of the quantitative controls.

Table 15.1 Checklist of an NGS-based analysis using mitochondrial whole genome analysis as an example

1	Illumina read Q score plot ($Q > 25$)
2	Limit of detection determination
3	Results of positive control (within acceptable range 1.1 % ± 0.3 %)
4	Verification of sample identified by genotyping: internal quality control (InQC)
5	Mean coverage depth and standard deviation
6	Coverage of all targeted regions: identify insufficiently covered regions
7	Correlation coefficient of quantitative ExQCs (1 %, 5 %, 20 %, 50 %, 80 %, 95 %)
8	Overall DSI >80 to troubleshoot and determine the pass/fail of a run
9	Verification of NGS results and reporting.

Based on the DSI, the performance among different gene enrichment methods can be quantitatively assessed [21]. As shown in Fig. 15.2, the DSI for hybridization-based sequence capture is 36, that for multiple overlapping PCR amplicons are 74, and when the LR-PCR single amplicon of whole mtDNA is used, the DSI is 90. The determination of DSI provides a single parameter to gauge the performance of a sequencing run, with emphasis on the data quality for each individual sample. The mathematical performance assessment simplifies the evaluation of complex procedures in a manageable scale. More importantly, the measurement of individual components in the DSI formula is valuable in order to pinpoint the cause of a failed run and to help with troubleshooting.

4.3 Post-analytic Phase

Since both the InQC and ExQC samples are analyzed together with the indexed sample from the beginning to the end, the results of these quality control samples also serve as controls for the post-analytical step. At the end of the analyses, there should be a checklist to check the acceptable range of each step (Table 15.1). The items to be checked include the depth and the completeness of coverage of every single base of the region of interest. Assessing the evenness of coverage profile is useful for the detection of mtDNA large deletions. For NGS, having sufficient coverage is one of the necessary factors in order to make accurate variant calls. Being able to identify the insufficiently covered regions can help with the recognition of copy number changes (deletion and duplication) in the target regions. The final test interpretation and reporting requires qualified professional personnel to explain to clinicians the test results generated through the clinical testing pipeline. Reporting results and archiving reports generated from NGS-based assays should follow the general requirements of other clinical tests performed in the certified laboratory. It is noteworthy that current NGS-based testing has longer turnaround time (TAT) than Sanger-based assays. In cases where corrective actions are needed or where repeating the analysis is required, the client should be notified about the delay in reporting results. Monthly review and summary of the TAT should be performed in order to improve the clinical service.

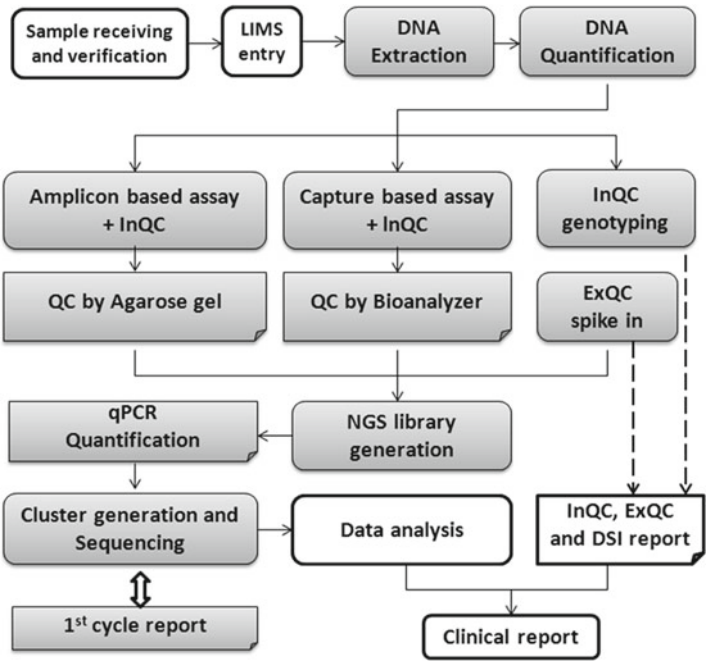

Fig. 15.3 Representative QA/QC scheme of MPS-based clinical testing

5 Conclusion

Application of MPS technologies to clinical molecular diagnostics has become a reality. Due to the high instrument and analytical complexity of NGS-based assays and the enormous amount of data output, careful test validation and stringent quality control procedures are necessary. Detailed procedures for test validation and the necessary quality controls described in this chapter can ensure that accurate and dependable results are delivered to patients and their caregivers. Checkpoints are illustrated in Fig. 15.3. A mathematical formula (Fig. 15.2d) was developed for generalized, non-biased, intra- or inter-run, and inter-laboratory evaluation of NGS performance. With the wide application of NGS technologies and the foreseeable federal regulations in this field, there is a pressing need for a robust quality assurance program.

References

1. Zhang W, Cui H, Wong LJ (2012) Application of next generation sequencing to molecular diagnosis of inherited diseases. Top Curr Chem 2012 May 11 (Epub ahead of print)
2. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong S-E, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK (2008) A mitochondrial protein compendium elucidates complex I disease biology. Cell 134(1):112–123

3. Calvo S, Jain M, Xie X, Sheth SA, Chang B, Goldberger OA, Spinazzola A, Zeviani M, Carr SA, Mootha VK (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet 38(5):576–582
4. Thorburn D (2004) Mitochondrial disorders: prevalence, myths and advances. J Inherit Metab Dis 27(3):349–362
5. Scharfe C, Lu HH-S, Neuenburg JK, Allen EA, Li G-C, Klopstock T, Cowan TM, Enns GM, Davis RW (2009) Mapping gene associations in human mitochondria using clinical disease phenotypes. PLoS Comput Biol 5(4):e1000374
6. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7(2):111–118
7. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9(1):387–402
8. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10): 1135–1145
9. Collins FS (2011) Faces of the genome. Science 331(6017):546
10. Green ED, Guyer MS (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470(7333):204–213
11. Hu H, Wrogemann K, Kalscheuer V, Tzschach A, Richard H, Haas SA, Menzel C, Bienek M, Froyen G, Raynaud M, Van Bokhoven H, Chelly J, Ropers H, Chen W (2009) Mutation screening in 86 known X-linked mental retardation genes by droplet-based multiplex PCR and massive parallel sequencing. Hugo J 3(1–4):41–49
12. Gowrisankar S, Lerner-Ellis JP, Cox S, White ET, Manion M, LeVan K, Liu J, Farwell LM, Iartchouk O, Rehm HL, Funke BH (2010) Evaluation of second-generation sequencing of 19 dilated Cardiomyopathy genes for clinical applications. J Mol Diagn 12(6):818–827
13. Voelkerding KV, Dames S, Durtschi JD (2010) Next generation sequencing for clinical diagnostics-principles and application to targeted resequencing for hypertrophic Cardiomyopathy: a paper from the 2009 William Beaumont Hospital symposium on molecular pathology. J Mol Diagn 12(5):539–551
14. Vasta V, Ng S, Turner E, Shendure J, Hahn SH (2009) Next generation sequence analysis for mitochondrial disorders. Genome Med 1(10):100
15. Halling KC, Schrijver I, Persons DL (2012) Test verification and validation for molecular diagnostic assays. Arch Pathol Lab Med 136(1):11–13. doi:10.5858/arpa.2011-0212-ED
16. Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, Dewey FE, Habegger L, Ashley EA, Gerstein MB, Butte AJ, Ji HP, Snyder M (2011) Performance comparison of whole-genome sequencing platforms. Nat Biotechnol 30:78–82
17. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD, Sheth V, Woodward JE, Peckham HE, Schroth GP, Kim RW, Kingsmore SF (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3(65):65ra64
18. Turner EH, Ng SB, Nickerson DA, Shendure J (2009) Methods for genomic partitioning. Annu Rev Genomics Hum Genet 10(1):263–284
19. Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55(4):641–658
20. Chen B, Gagnon M, Shahangian S, Anderson NL, Howerton DA, Boone JD (2009) Good laboratory practices for molecular genetic testing for heritable diseases and conditions. MMWR Recomm Rep 58(RR-6):1–37
21. Zhang W, Cui H, Wong LJ (2012) Comprehensive 1-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem 58(9):1322–1331

Chapter 16
Frequently Asked Questions About the Clinical Utility of Next-Generation Sequencing in Molecular Diagnosis of Human Genetic Diseases

Ephrem L.H. Chin, Victor Wei Zhang, Jing Wang, Margherita Milone, Susan Pacheco, William J. Craigen, and Lee-Jun C. Wong

Abstract Before the advent of next-generation massively parallel sequencing (MPS), Sanger sequencing for many years has been the gold standard for the identification of unknown mutations in candidate genes. However, since the introduction of high-throughput massively parallel sequencing for clinical diagnostics, the next-generation sequencing (NGS) technology has revolutionized the molecular diagnosis of human inherited disorders. This new technology provides a broad spectrum of clinically relevant features at different levels of mutation detection, cost, and turnaround time. Nevertheless, the complex technologies involved, novel analytical and bioinformatics pipelines, and the challenges of nucleotide variant interpretation are difficult topics for general clinicians to understand fully. Yet, a better grasp of these subjects is needed in order to choose the most effective diagnostic approach and to convey the complex results to their patients appropriately. This chapter describes the frequently asked questions and answers related to the clinical utility of NGS-based molecular diagnostic tests in the hope that clinicians in all medical specialties better understand what NGS technology can deliver for the molecular diagnosis of human genetic disorders.

E.L.H. Chin • J. Wang • W.J. Craigen
Department of Molecular and Human Genetics, Baylor College of Medicine,
One Baylor Plaza, NAB 2015, Houston 77030, TX, USA

V.W. Zhang • L.-J.C. Wong, Ph.D. (✉)
Medical Genetics Laboratories, Department of Molecular and Human Genetics,
Baylor College of Medicine, One Baylor Plaza, NAB 2015, Houston 77030, TX, USA
e-mail: ljwong@bcm.edu

M. Milone
Department of Neurology, Neuromuscular Division, Mayo Clinic,
Rochester 55905, MN, USA

S. Pacheco
Department of Pediatrics, Division of Allergy/Immunology,
University of Texas Health Science Center, Houston 77030, TX, USA

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4_16,

1 Introduction

Since its introduction in the mid-2000s, next-generation sequencing (NGS) has revolutionized the way sequencing is being conducted in many clinical laboratories. NGS technology has transformed molecular diagnosis of human genetic diseases by its ability to sequence thousands to millions of DNA fragments in parallel and generate massive amount of sequencing data from a single sequencer. Thus, clinical laboratories can analyze more genes quickly and in a very cost- and time-effective manner when compared to the universally recognized "gold standard" Sanger sequencing.

The complexity of NGS technology makes it challenging for healthcare professionals such as physicians and genetic counselors to evaluate the benefits of currently available clinical NGS testing objectively. This chapter provides answers to common questions by leading healthcare professionals to assist other healthcare professionals who are looking to utilize this new class of molecular testing for their patients.

2 NGS Technology

2.1 What Is the Next-Generation Sequencing (NGS)?

Next-generation sequencing (NGS) refers to high-throughput massively parallel sequencing of hundreds to millions of DNA fragments simultaneously. There are now second-generation and third-generation sequencing technologies that continue to reduce the cost and enhance the accuracy of DNA sequencing. In this book, NGS primarily refers to second-generation sequencing technology.

2.2 How Does NGS Work?

There are two major steps: target gene enrichment necessary to "capture" the genes of interest and massively parallel sequencing of the captured genes. Please refer to chap. 3 for details. If the analysis is for the whole genome, then there is no need for gene enrichment step.

2.3 How Does NGS Differ from the Traditional Molecular Diagnostic Methods?

The traditional diagnostic methods focus on identifying one mutation at a time by various mutation detection methods (target mutation analysis) for known mutations

or analyze one gene at a time by Sanger sequencing of the candidate gene in order to indentify previously unknown mutations (target gene analysis). The NGS can detect multiple nucleotide changes from many genes at the same time.

2.4 Is There a Limit Number of Genes to Be Analyzed by NGS?

No. The NGS approach may be designed and scaled for a single gene, a panel of genes involved in the same pathway or that confer a recognizable disease phenotype, the whole exome (the protein coding regions of about 20,000–30,000 genes), or the whole genome, including gene regulatory regions that do not directly encode proteins. If the analysis is for the whole genome, there is no need for the target gene enrichment step.

2.5 What Kind of NGS-Based Clinical Tests Are Available Currently?

This information can be found at GeneTests (http://www.ncbi.nlm.nih.gov/sites/GeneTests). The clinically available NGS-based tests include mtDNA depletion syndromes, glycogen storage diseases, Noonan syndrome, cardiomyopathy, hearing loss, Usher syndrome, high and low bone mineral diseases, retinitis pigmentosa, various mitochondrial respiratory chain complexes and combined mitochondrial disorders, the mitochondrial genome, X-linked intellectual disabilities, carrier testing for inherited diseases, and the whole exome. This list will continue to grow with time, as costs and data analysis requirements steadily reduced. It will encompass even more common diseases not considered heritable disorders such as common cancers and infectious diseases.

2.6 Is It Necessary to Know What Enrichment Methods and Sequencing Platforms Used in the Diagnostic Laboratories?

The laboratory should provide you with the necessary information, including the test performance characteristics such as false-positive and false-negative rates of the test method they use. However, knowing the pros and cons of the enrichment methods and the chemistry of each sequencing platform will enable you to judge which method is best suitable for testing purposes. The details and comparison of various methods are described in chap. 3 of this book. In general, there are two major target gene enrichment methods: DNA amplification by the polymerase

chain reaction method (PCR) and capture by hybridization. Generally, if the number of genes to be analyzed is small (<100), gene coverage can be achieved by PCR. However, if the gene number is large, it would be more efficient to use the capture method. The PCR method has several pitfalls, including the limited number of total amplicons (units of DNA amplification), allele "dropout" due to the presence of SNPs at the primer sites, and the cost of PCR primers and necessary optimization of PCR conditions. Occasionally, long range PCR may be used to avoid the interference of pseudogenes and/or SNPs at the primer sites. The capture by hybridization in solution requires the synthesis of RNA or DNA probes to selectively pull down coding exons of the genes of interest. It is scalable from a few genes to thousands of genes or even the whole exome, and it lends itself to automation. The capture method often has difficulties with DNA regions of high GC content and genes with DNA repeat elements or homologous sequences or pseudogenes. However, these hard to capture and/or sequence regions can be filled in by individual gene PCR followed by Sanger sequencing in order to provide 100 % coverage. Several commercial NGS platforms, including Roche 454, Illumina HiSeq, SOLiD, and Ion Torrent, are available. The chemistries and advantages and disadvantages of different platforms have been discussed in detail in Chap. 3.

2.7 What Does Multiplex and Barcode Mean in NGS?

Due to the high throughput of NGS, DNA templates from multiple individuals may be pooled for sequencing. Therefore, DNA samples from each individual must be clearly "bar-coded" using identifiable DNA tags added to the target DNA sequences, and the sequence reads sorted according to the barcode in order to reassemble a subject's unique DNA sequence and thus avoid sample mix-up.

2.8 Are Sequence Data Generated by NGS as Good as Those from Sanger Sequencing or Better?

In general, results generated by NGS are at least as good as those from Sanger sequencing. NGS has also demonstrated its ability to detect low levels of changes that cannot be detected by Sanger sequencing such as germ line mosaicism, somatic changes, and mitochondrial heteroplasmic mutations [1–3]. However, NGS may also give false-positive results due to chemical, platform, and/or bioinformatic biases that necessitate the confirmation of all clinically relevant changes, especially insertion/deletion mutations, by a second method such as Sanger sequencing. Combining NGS and a secondary confirmatory algorithm provides greater accuracy and confidence than Sanger sequencing-based tests.

3 NGS Data Analyses and Variant Interpretation

3.1 How Are NGS Sequence Results Analyzed?

Regardless of the platform used, the NGS results are usually generated in the form of light emission or micro-conductance changes that occur during DNA synthesis according to the DNA template sequence on a highly dense microarray. Images are taken for each cycle of nucleotide incorporation. These image files are converted to nucleotide sequence reads based on the color of light emitted (primary analysis). Millions of short sequence reads generated in this manner are assembled to form stacks of longer reads, typically encompassing the coding exons and at least 20 bp of the flanking intron regions (secondary analysis). The target sequences are then compared to the reference sequences in the Genbank (www.ncbi.nlm.nih.gov/genbank), and differences are annotated as variant calls (tertiary analysis). The final interpretation of these variant calls requires qualified laboratory staff.

3.2 What Is the Minimum Coverage per Base and the Minimum Coverage per Coding Exon? What Is the Minimum Coverage in Order to Provide Dependable and Accurate Variant Calls?

The coverage of a base of an exon is also known as the "sequence depth," meaning that the number of times the base or exon is individually sequenced. In general, if only heterozygous or homozygous variants of two alleles of nuclear genes are considered, a minimum of 20X [4–6] coverage may be sufficient if the quality of "balanced sequence reads" are good (usually quality scores >30) [4–6]. Balanced sequence reads refer to the observation that the forward sequences and reverse sequences, i.e., the sequence of both strands of complementary DNA, are read a similar number of times. However, if quantitative variant calls are of interest, such as in the case of somatic mutations, mosaicism, or mitochondrial DNA (mtDNA) heteroplasmy, then the coverage needs to be deeper in order to provide reliable and reproducible variant calls. Similar to Sanger sequencing, there is background noise or technical variation that can be minimized by increasing the sequence depth (coverage). Coverage is usually inversely proportional to the number of genes to be sequenced. To sequence a small panel of genes, e.g., 20 genes, a coverage of 1000X can be achieved, whereas the average coverage of the whole exome (~20,000–30,000 genes) is about 50–100X. Since the coverage for any given coding exon that is captured and sequenced is variable, the term "average coverage" is often used. Some exons or bases are poorly covered (<20X) or not covered at all due to various reasons such as high GC content, repeat regions, or the presence of homologous sequences like pseudogenes. In this case, a fully validated panel test usually will fill

these gaps by additional PCR/Sanger sequence methods in order to provide "100 %" coverage. However, due to the large number of genes in a whole exome analysis, the "no coverage" coding exons in whole exome sequencing are usually not well characterized, or there are too many to be filled in by PCR/Sanger method. The "poorly covered" exons in the whole exome sequencing may be sufficiently covered in panel testing due to its deeper average coverage. For nuclear gene testing, an overall average coverage of >600X usually would provide sufficient coverage for most of coding exons except for the "no coverage" regions that can never be sufficiently covered despite increasing the coverage depth [1, 7]. For NGS of mtDNA, an average coverage of >20,000 is required in order to reliably detect 1.5 % heteroplasmic variants at an experimental error rate of 1 % [2, 3].

3.3 Can All of the Different Types of Mutations Be Detected by NGS-Based Analysis?

In theory, all types of mutations should be detectable by NGS-based analysis. However, depending on the surrounding sequence environment some nucleotide changes are difficult to detect. Single nucleotide substitutions and small insertion/deletions (indels) are relatively easily detected. However, if these simple changes are embedded in a homopolymeric tract or within short tandem repeat regions, they may not be readily detected. The ability to detect changes in these regions also depends on the sequencing chemistry. For example, Roche 454 and Ion Torrent are less accurate in detecting substitutions or small indels at homopolymeric tracts. Detection of large intragenic deletions (also known as copy number changes) by NGS-based analysis is currently not clinically available. While the fully validated panel NGS analyses can easily detect homozygous exonic deletions, heterozygous large exonic deletions are more difficult to detect. Further optimization of analytical algorithms is needed before clinical use. Individual laboratories should validate their NGS-based methods and the analytical pipelines they use in order to provide information regarding test sensitivity, specificity, error rate, and any other limitations. Detection of larger indels remains challenging, but some laboratories are validating their analytical algorithms for this purpose.

3.4 Will an NGS-Based Test Miss Any Mutations That Sanger Sequencing Detects? What Are the Major Obstacles in the Analysis of NGS Sequence Data?

In general, a fully validated NGS-based test will not miss any mutations that are detected by Sanger sequencing. Similar to Sanger sequencing, NGS may not detect genomic structural rearrangements (e.g., deletions, duplications, and inversions), large insertion (e.g., ALU mediated insertion), mutations within the promoter or

deep intronic regions. Any regions that are missed by NGS would have been characterized. However, any novel changes detected by NGS should be confirmed by a second method such as Sanger sequencing. On the other hand, large mtDNA deletions can be easily detected by NGS but cannot be detected by Sanger sequencing [2, 3]. In addition, Sanger sequencing cannot detect low mtDNA heteroplasmy (<20 %). With high coverage (>20,000X) and proper target gene enrichment methods, NGS can detect mtDNA heteroplasmy as low as 1 % [2, 3]. Therefore, NGS-based tests are at least as good as the "gold standard" Sanger sequencing, if not better. The major obstacles of NGS-based tests are the inability to accurately and adequately capture and sequence the coding regions that contain pseudogenes, repeats, and/or high GC content. These obstacles are often overcome by traditional PCR/Sanger sequencing in order to fill in these regions.

3.5 How Much Confidence Should I Have in NGS-Based Diagnosis? Can I Trust the Sequence Results Obtained from NGS? Are There False-Positive Changes?

NGS provides the laboratory the ability to analyze a large number of genes in parallel at a reasonable cost. While NGS at this time is not a stand-alone test, when combined with a secondary confirmatory method, it will provide a greater analytical power when compared to any single technology alone. NGS technology development is improving daily, and it is a matter of time before NGS technology will be proven as an effective stand-alone test. At this time most clinical laboratories are confirming clinically relevant changes by a secondary method before reporting the results. Therefore, the clinically reported NGS results, if confirmed by a second method, are reliable and can be used for clinical purposes. As such, a clinical laboratory is very unlikely to report erroneous false-positive changes.

3.6 What Is the False-Negative Rate?

An accurate false-negative rate can be determined by the individual testing laboratory based on their validation performed on the specific assay. Such test performance characteristics should be made available by the testing laboratory. Clinicians should obtain these data and understand the utility and limitations of the particular NGS-based test before submitting a patient's sample for testing. The false-negative rate for a fully validated NGS-panel testing is minimal if all possibly causative genes are analyzed; however, it is much higher for the whole exome or whole genome analyses. For example, if a whole exome study reports 99 % gene coverage that would mean an equivalent of $0.01 \times 20{,}000$ genes $= 200$ genes are not fully covered, or at least a 1 % false-negative rate, while for focused gene panel testing, the coverage is expected to be 100 % if all disease causative genes are analyzed.

3.7 Who Are Qualified to Interpret the Variant Results?

For genetic testing, qualified personnel for the interpretation of variant results are geneticists certified by the American Board of Medical Genetics.

3.8 Will NGS Detect Copy Number Changes? Do I Still Need Array CGH for Large Deletions or Duplications?

At this time most NGS-based tests are not able to detect heterozygous intragenic copy number changes, and it is still necessary to use the aCGH assay to detect large copy number aberrations.

3.9 Will I Receive Incidental Findings with Targeted Panels?

You will not receive any incidental findings with a targeted panel because only genes of interest are analyzed. However, medically actionable incidental findings unrelated to the patient's immediate condition do occur in whole exome or whole genome testing.

3.10 How Does the Lab Handle Incidental Findings?

Each clinical laboratory offering NGS testing will have established their policies on reporting incidental findings. Some laboratories do not report incidental findings due to the complexities and uncertainties of these findings. In addition, most clinicians do not feel comfortable explaining incidental findings to patient's families.

4 Regulatory Aspects

4.1 What Are the Regulatory Agencies That Oversee the Clinical Laboratories Performing the NGS-Based Test?

The Clinical Laboratory Improvement Amendments (CLIA) and the College of American Pathologists (CAP) set the relevant guidelines. Laboratories providing clinical service must perform the diagnostic procedures according to the guidelines to receive accreditation from these agencies.

4.2 Are the NGS-Based Tests CLIA and CAP Approved?

NGS-based tests are developed by following the guidelines for CLIA's Laboratory Developed Testing (LDT). Thus, the laboratory offering the NGS test must document procedures that meet all CLIA's requirements for a LDT test. CAP has recently published additional guidelines in the Molecular Pathology checklist dealing specifically with NGS testing, and all CAP accredited laboratories offering NGS testing are expected to comply with these new requirements.

4.3 What Role Will FDA Have in NGS Testing?

The exact role that FDA will play is still unclear at this time: however, clinical laboratories are required to comply with established CLIA and CAP regulations.

4.4 What Exactly Are the CLIA Requirements That Laboratories Offering NGS Need to Follow?

Compliance with the same rules that apply to all clinical tests must occur. In addition, the laboratory must provide specifications for the novel technology by clearly describing the "performance characteristics," including coverage depth, specificity, sensitivity, false-positive, false-negative, accuracy, reproducibility, experimental error rate, quantitative and qualitative limits of detection, and turnaround time.

4.5 What Should I Do with Positive Results?

With appropriate clinical correlation, positive results most likely mean that you have found the definitive diagnosis for your patient. You can proceed to counseling of the family, testing of relevant family members if necessary, execution of proper medical actions including specific treatment or management, and/or offering prenatal diagnostic testing if appropriate. If a positive result contains variants of unknown significance (VUS), then additional functional, molecular/biochemical, and family studies are required to confirm the pathogenicity of the VUS.

4.6 What Should I Do with Negative Results?

If the results are from a fully validated gene panel, it is highly unlikely that the genes analyzed are the cause of the patient's disease condition. However, several caveats need to be considered. It remains possible that pathologic mutations exist within one of the genes analyzed but that the mutation(s) are in regions of the gene

not being analyzed such as transcription promoter elements or deep within an intron. Alternatively, it could be that the disease is caused by another not yet identified gene, or that the clinical features overlap with other disease loci not included in the chosen gene panel. Similarly, if the negative result is from whole exome sequencing (WES), certain mutations may not be detected since the current WES technology, although it covers almost all coding exons in the human genome, may miss regions of interest such as promoter or deep intronic mutations.

5 Test Cost/Reimbursements

5.1 Will the Insurance Companies Pay for These New NGS-Based Tests?

Yes, at this time insurance companies are covering NGS-based testing. The amount covered will be dependent on the individually negotiated insurance plans and may require pre-authorization and a letter of medical necessity.

5.2 For Apparently the Same Test, e.g., Mitochondrial Whole Genome Sequencing, Why Is There Such a Difference in Pricing from One Vendor to Another?

Traditionally, it has been difficult for clinician's and genetic counselors to fully evaluate genetic testing offered by multiple different laboratories in terms of cost-effectiveness. This is essentially true with NGS, as there exist a variety of different NGS platforms as well as gene enrichment and confirmatory technologies. As with any purchase, it is up to the consumer to compare and contrast the capabilities and performance of each vendor's test offering to ensure that the "quality of the products" aligns with what you are seeking. Most importantly, you need to know if the test is fully validated and carefully compare the "performance characteristics" mentioned above, including percentage coverage of the genes of interest, coverage depth, specificity, sensitivity, false-positive, false-negative, accuracy, reproducibility, experimental error rate, quantitative and qualitative limits of detection, turnaround time, confirmation of findings, and the expertise of the laboratory directors.

5.3 The Cost per Base by NGS Is So Much Reduced Now, Why Does It Still Cost So Much to Have a Clinical NGS-Based Test Done?

According to a recent study by Macquarie Equity Research report on DNA Sequencing data from (13 August 2012), an Illumina HiSeq2000 instrument the

cost per megabase was only about US$0.04. Therefore, to sequence three billion bp of the whole human genome once is about $120 at 1X coverage. For clinical diagnosis, the average coverage is about 50–100X, that would bring the cost to $6,000–$12,000 per human genome. This is within the range of current costs per human exome. Compared to the cost of $3,000,000 per human genome by Sanger sequencing on the ABI 3730XL sequencer, the NGS cost is much less.

5.4 Why Is It that the Cost for Target Sequencing of 100–400 Genes Can Be Almost the Same as the Cost for the Whole Exome Analysis of ~20,000 Genes?

The cost mentioned in the previous section does not include the cost for analytical software, the computational time, results filtering process, data storage, and expert interpretation. In fact, the major cost of NGS is not in the DNA sequencing but the data processing, alignment, analysis, variant annotation, and interpretation. Due to the high coverage depth, usually at about 500–1000X, at least 10 times higher than that for the WES, the cost per base for a fully validated, 100 % covered target gene analysis is at least 10 times higher than the cost per base for WES. In addition, for a 100 % covered target gene analysis, the "no coverage" coding exons are to be filled in by PCR/Sanger sequencing method. On average, about 1–2 % of the coding exons require PCR/sequencing due to high GC, repeat regions, or pseudogenes. Therefore, for a panel of 200 genes, this would translate into about 20–40 exons that need to be PCR/Sanger sequenced. This adds to the cost of the NGS-based target gene analysis, which usually is carried out by laboratories with the necessary expertise who are familiar with the molecular, biochemical, and molecular genetics of target genes. In addition, the turnaround time of the target gene analysis is usually significantly shorter than the whole exome analysis.

6 Patient Care and Customer Service

6.1 How Do I Decide Which NGS Test Providing Laboratories to Use?

Although price may be the first consideration, other factors such as test validation, performance characteristics, reliability of the test results performed by the laboratory, the qualifications of laboratory personnel, and turnaround time should also be considered. In addition, the availability of the laboratory personnel in assisting clinicians with the interpretation of data can be a key element in selecting the providing laboratory for NGS-based tests. Moreover, knowing the pros and cons of the

NGS platforms, confirmation procedures, and specialty area of the performing laboratory will help in deciding which laboratory should receive the patient's samples. Turnaround time is another important factor if identifying the correct diagnosis rapidly is critical in treating a patient's condition. Finally, you want to know if all coding regions of genes of interest are 100 % covered, and if not, then which regions or genes are not sufficiently covered, such that a negative result should lead to further testing in these poorly covered regions.

6.2 What Are the Benefits of Using NGS-Based Tests?

In short, the major benefits are a much reduced cost and shorter turnaround time for a definitive diagnosis of a disease.

6.3 What Is the Difference Between Targeted Gene Analysis and Whole Exome Sequencing?

If the patient's clinical diagnosis is clearly defined and is known to be caused by a group of genes, for example, glycogen storage disorders [1], serial Sanger sequencing of the candidate genes is time consuming and cost ineffective. In this case, an NGS-based panel that simultaneously analyzes a group of candidate genes will provide a rapid and definitive diagnosis at a much reduced cost and shorter turnaround time. The target gene approach is suitable for a genetically highly heterogeneous but clinically relatively defined disease that involves 100–200 genes, such as deafness, retinitis pigmentosa, X-linked intellectual disability, and cardiomyopathy. Mitochondrial disorders are a group of clinically and genetically heterogeneous disorders involving as many as 1,500 genes, with only about 200 causative genes identified [8]. In this case, if the diagnosis is confirmed by other biochemical and histochemical method, then NGS-based analysis of 200 known genes or all 1,500 genes may be the appropriate choice. However, if the patient's clinical features are nonspecific, and a clinical diagnosis is not clear, then an NGS-based analysis of the whole exome is warranted. The major difference is that target gene analysis usually provides close to 100 % coverage, while most exomes may miss 5–15 % of coding regions of interest. Most laboratories currently offering clinical exomes report only the genes known to cause disease based on OMIM (www.ncbi.nlm.nih.gov/omim) and/or HGMD (www.hgmd.cf.ac.uk) databases. However, sequence data from a large number of genes not yet recognized to cause human disease are collected by exome sequencing; thus, exomes are capable of discovering new disease genes.

6.4 Can NGS Tests Be Performed on the Same DNA Samples Previously Submitted to the Laboratories for Sanger Sequencing?

Yes, most laboratories will use the same DNA sample previously submitted for Sanger sequencing as long as the DNA is of good quality and sufficient quantity. DNA may be extracted from various specimen types, including blood cells, fibroblasts, and tissues.

6.5 Why Does It Take So Much Longer to Obtain NGS Results Compared to Sanger?

The longer turnaround time is due to DNA template library preparation, the long instrument run, and alignment and analysis of the massive amount of sequencing data. However, through technical advancements these processes will undoubtedly become more streamlined and faster.

Acknowledgment The authors would like to thank doctors Brett Graham, Fernando Scaglia, Eric Schmitt, Megan Landsverk, and Fangyuan Li for their valuable discussion.

References

1. Wang J et al (2013) Clinical application of massively parallel sequencing in the molecular diagnosis of glycogen storage diseases of genetically heterogeneous origin. Genet Med 15(2):106–14. doi: 10.1038/gim.2012.104. Epub 2012 Aug 16
2. Cui H et al (2013) Comprehensive next generation sequence analyses of the entire mitochondrial genome reveals new insights into the molecular diagnosis of mitochondrial DNA disorders. Genet Med
3. Zhang W, Cui H, Wong LJ (2012) Comprehensive 1-step molecular analyses of mitochondrial genome by massively parallel sequencing. Clin Chem 58:1322–1331
4. Asan et al (2011) Comprehensive comparison of three commercial human whole-exome capture platforms. Genome Biol 12(9):R95. doi: 10.1186/gb-2011-12-9-r95
5. McDonald KK et al (2012) Exome analysis of two limb-girdle muscular dystrophy families: mutations identified and challenges encountered. PLoS One 7(11):e48864
6. Parla JS et al (2011) A comparative analysis of exome capture. Genome Biol 12(9):R97
7. Sule G et al (2013) Next generation sequencing for disorders of low and high bone mineral density. Osteoporos Int. 2013 Feb 27. [Epub ahead of print]
8. Koopman WJ, Willems PH, Smeitink JA (2012) Monogenic mitochondrial disorders. N Engl J Med 366(12):1132–1141

Index

L.-J.C. Wong (ed.), *Next Generation Sequencing: Translation to Clinical Diagnostics*,
DOI 10.1007/978-1-4614-7001-4, © Springer Science+Business Media New York 2013