Top Curr Chem (2012)
DOI: 10.1007/128_2012_329

Next Generation Sequencing: Chemistry, Technology and Applications

Pei Hui

Abstract High-throughput next generation sequencing (NGS) has been quickly adapted into many aspects of biomedical research and begun to engage with the clinical practice. The latter aspect will enable the application of genomic knowledge into clinical practice in this and next decades and will profoundly change the diagnosis, prognosis and treatment of many human diseases. It will further demand both philosophical and medical curriculum reforms in the training of our future physicians. However, significant huddles need to be overcome before an ultimate application of NGS in genomic medicine can be practical and fruitful.

Keywords Next generation sequencing · Genomic medicine

Contents

P. Hui (✉)
Department of Pathology, Yale University School of Medicine, 310 Cedar Street, New Haven, CT 06520-8023, USA
e-mail: pei.hui@yale.edu

1 Introduction

The conclusion of the human genome sequencing project in 2003 established the molecular basis for the understanding of many disease processes at genetic level [1]. As a result, the availability of reference sequence of the human genome has fueled the emergence of a new era of genomic medicine [2–5]. Advance in technology of high throughput next generation sequencing (NGS), also known as massively parallel or multiplex cyclic sequencing, is the key element that will enable the application of genomic knowledge into clinical practice. However, DNA sequencing and related genomic informatics must become more economical, informative, and readily applicable for the ultimate transition from empirical practice to precision medicine [6, 7]. The current speed of evolution of NGS technology is stunning and will soon result in the delivery of low-cost, high-throughput, and even portable DNA sequencing apparatuses to clinical laboratories. In fact, NGS has already begun to produce clinical benefits in some healthcare setting [8–10]. In the next few decades, genomic medicine driven by NGS will profoundly change the diagnosis, prognosis, and therapy of human diseases. It will demand both philosophical changes and curriculum reform in the training of our future physicians as well.

In order to get there, there are obstacles remaining to be overcome, such as developing sophisticated bioinformatics and computational biology techniques for analyzing vast amount of sequencing data, understanding variations of the genome, understanding genetic and non-genetic bases of human diseases, establishing effective ways to deliver evidence-based genomic medicine, and, finally, resolving ethical and legal issues in the practice of genomic medicine. In this chapter, the chemistry and technological background of NGS will be presented. It will be concluded by the direction of future technological development in these aspects.

2 Chemistry

DNA sequencing undertaken by the Human Genomic Project was completed in 2003 almost exclusively by Sanger's method, the first generation sequencing. In 2007, DNA sequencing was taken to the next level when the Illumina genome analyzer was introduced, heralding the era of next generation sequencing. Within 1 year, NGS was used successfully to sequence the first individual human genome (James Watson) in 2008 [11]. Now NGS technology is evolving at an unprecedented pace along with diminishing cost. It is expected that the cost will be approaching less than

$10,000 per human diploid genome in the coming years [12]. At the time of writing, a typical platform could produce up to 600 giga-base data in a sequencing run that lasts for 7–10 days. The data represent about 6,000,000,000 sequencing reads with a length of 100 bases.

Generally speaking, NGS employs DNA synthesis or ligation chemistry (sequencing-by-synthesis) to read through many independent DNA templates at the same time in a highly parallel fashion to produce a tremendous quantity of DNA sequence data. Sequencing-by-synthesis strategies [13] include a single molecule approach or ensemble approach (sequencing of multiple clonally amplified DNA targets on isolated surfaces or beads). Both approaches can be accomplished in either real-time fashion (DNA polymerase synthesizes without interruption) or synchronous-controlled fashion (DNA polymerase synthesizes in "stop-and-go" through controlled delivery of nucleotides or temporarily limiting extension using modified nucleotides or metal catalysts). The detection of signal can be achieved by fluorescent labeling of nucleotides, enzyme-coupled chemiluminescence assays for pyrophosphate, and pH change as result of proton release during each nucleotide incorporation.

One sharp contrast to the first generation Sanger sequencing is that NGS generates short reads of frequently less than 500 bp as opposed to over 1,000 bp. However, the massive depth of coverage, i.e. multiple reads over the same template DNA region, compensates for the limitations of short reads. NGS technologies have drastically increased the speed and throughput capacities over Sanger sequencing while reducing cost, even as we write. NGS may be classified into second and third generations according to their years of availability and chemistry. Second generation sequencing essentially uses DNA synthesis chemistry as employed by the traditional Sanger's sequencing. Third generation sequencing (Ion Torrent of Life Technologies, Inc and single polymerase sequencing platforms of PacBioRS, Inc) employs distinct chemistries, which will be elaborated in the following technology section.

3 General Workflow of NGS

Regardless of various sequencing chemistries, both second and third generations of NGS require highly complex pre-sequencing target preparation procedures and post-sequencing bioinformatics data analysis (Fig. 1). The pre-sequencing step includes target DNA enrichment and NGS library preparation. Target enrichment can be accomplished by amplification methods (PCR, Long-ranger PCR, or Raindance fluidigm PCR) or hybridization capture methods (by solid phase or in solution). NGS library preparation generally involves (1) fragmentation of the enriched target DNA by physical methods (sonication, acoustic wave or nebulization) into generally a length of 150–500 base pairs (library sequences), (2) ligation of the fragments to adaptor primers, and (3) clonal amplification of the library by either emulsion bead PCR or surface cluster amplification.

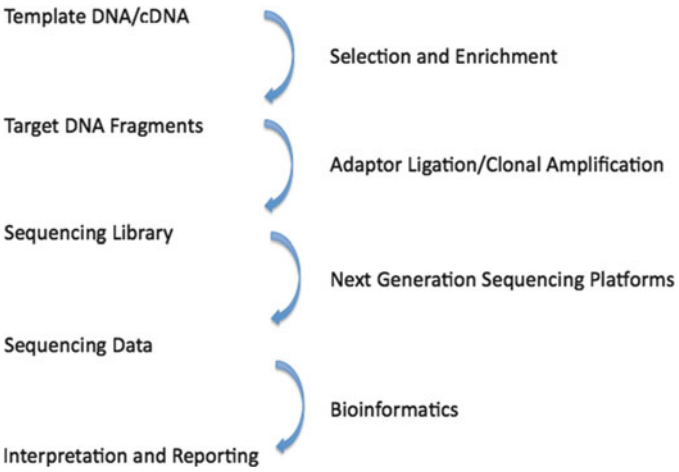


Fig. 1 Workflow of next-generation sequencing. This target enrichment and library sequencing approach are typically used in second and third generation sequencing approaches

The performance of sequencing reaction and capture of sequence data are the subject of a subsequent section. After a sequencing reaction, billions of reads are generated. Each read contains the sequence, typically ~100 bases in length, of a single template clone. Post-sequencing bioinformatics analysis generally involves sequence image processing to generate base sequences, sequence file conversion to readable files, and sequence alignment with reference DNA sequence for final variant identification and annotation. Adequacy of NGS relies on the sequence coverage and the depth. Sufficient coverage of DNA regions of interest is essential and sufficient depth of coverage (how many reads of the same region) is critical for accuracy and interpretation. Some common problems associated with NGS include sequencing reads are too short, resulting in difficulties in final sequence assembly or mapping; not all sequences are equally processed at high GC rich regions and homopolymers, amplification bias is inherent to some target enrichment processes, and sequencing errors (particularly longer reads) occur from 0.01 to 16 per 100 base read [14].

4 Evolution of Sequencing Technology

4.1 First Generation Sequencing

First generation sequencing technologies include sequencing by synthesis developed by Sanger [15] and sequencing by cleavage pioneered by Maxam and Gilbert [16]. Sanger sequencing dominated the biomedical research field before 2008. Standard four-color fluorescent labeling, where each color relates to one of the four DNA

bases, has been the method of choice for detection by automated capillary electrophoresis (CE) platforms, commercially available from Applied Biosystems Inc., Life Technologies Inc., and Beckman Coulter Inc. The first complete human diploid genome (Craig Venter) was sequenced by Sanger's method in 2007 [17]. Although sequencing tasks in large comprehensive research projects have now shifted to NGS platforms, Sanger sequencing-CE platform will likely remain in significant use for targeted sequencing projects (biomarker identification and pathway analysis) and clinical diagnostic applications until small-scale NGS platforms become cheap and fast enough; this is an rapidly evolving area of industrial development (see below).

4.2 Second Generation Sequencing

Second generation sequencing is represented by Roche 454 pyrosequencing, reversible terminator sequencing by Illumina, sequencing by ligation of ABI/SoLiD, and single-molecule sequencing by Helicos. Using DNA polymerase or DNA ligase as their core chemistry, these platforms provide significant performance in large comprehensive whole genome sequencing projects [18]. Roche454 uses emulsion PCR to achieve clonal amplification of target sequence. The sequencing machine contains many picoliter-volume wells each containing a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotide incorporated into the nascent DNA [19–21]. Illumina (Solexa) uses cluster target sequence amplification on solid surface (bridge amplification). Sequencing is performed by adding four types of nucleotides, each labeled by one of four fluorophores and containing a 3′ reversible terminator. In contrast to pyrosequencing, DNA can only be extended one nucleotide at a time in the Illumina approach. After a fluorescent image of the incorporated nucleotide is recorded, the fluorophore along with the 3′ reversible terminator is chemically removed from the DNA molecule, allowing the next cycle to occur [12, 22].

Applied Biosystem/Life Technologies' SOLiD technology employs ligation reaction for sequencing using a repertoire of all possible oligonucleotides of a fixed length that are labeled according to the sequence position. Oligonucleotides are ligated after annealing. The preferential ligation by DNA ligase for matching sequences records the nucleotide position. DNA is clonally amplified by emulsion PCR on beads, leading to each bead containing only copies of the same DNA molecule. The beads are deposited on a glass slide [23] and sequenced. The sequences in terms of quantities and lengths are comparable to Illumina sequencing [20, 24].

HeliScope sequencer employs "true single molecule sequencing" technology [25, 26]. DNA fragments along with added polyA tail adapters are attached to the flow cell surface, followed by extension-based sequencing with cyclic washes of the flow cell with fluorescently labeled nucleotides similar to Sanger sequencing. Although the reads are short, recent improvement of the methodology provides

enhanced accuracy of reading through homopolymers (stretches of one type of nucleotides) and also allows for RNA sequencing [18, 26, 27].

The second generation sequencing platforms vary significantly with regard to their throughput, read-length, and operating cost [12]. They are generally high throughput but highly expensive machines of scales of 0.5–1.0 million US dollars. The signal recording method is either fluorescence labeling or pyrophosphate chemical conversion, both requiring optical detection. The second generation sequencing platforms, although being successful in many research applications, suffer variably from high cost of instrument, complexities of sample preparation and chemistry (fluorescent labeling and enzyme-substrate reaction), complexities of optics and instrumentation, and read-length limitations [13].

4.3 Third Generation Sequencing

Sequencing technology evolves with the high demand for a low cost of technology. In line with the ultimate target goal of $1,000 per genome aimed at by the NIH/NHGRI invited grant challenge in 2004 for developing novel technologies, the third generation sequencing platforms are characterized by new chemistry, less operation time, desktop design, and lower operation cost. At the time of writing, three leading third-generation sequencers have emerged, which include Pacific Biosciences' real-time single molecule sequencing (PacBioRS), Compete Genomics' combined pro-anchor hybridization and ligation (cPAL), and Ion Torrent of Life Technologies, Inc.

PacBioRS is a real-time single molecule-single polymerase sequencing platform that can produce 1,000 bp read. Each chip has so-called zero-mode wave guided (ZMW) nanostructures of 100-nm holes, inside which DNA polymerase performs sequencing by synthesis with phospholinked nucleotides labeled with fluorophores which are introduced sequentially (Fig. 2) [28–32]. In addition to producing DNA sequence, monitoring the kinetics of nucleotide incorporation may help in the future to extract epigenetic information (e.g., methylation pattern) of the native DNA strands [33]. The platform has the ability to sequence mRNA by replacing DNA polymerase with ribosome [34]. The instrument of such configuration will, however, be expensive.

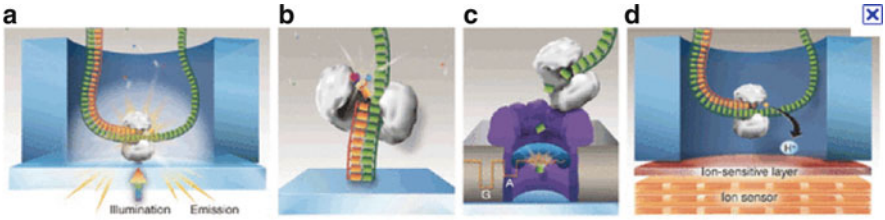


Fig. 2 PacBioRS real-time single molecular-single polymerase sequencing single stranded DNA template is sequenced by synthesis in nanostructure hole. (Copyright permission obtained from PacBioRS) [32]

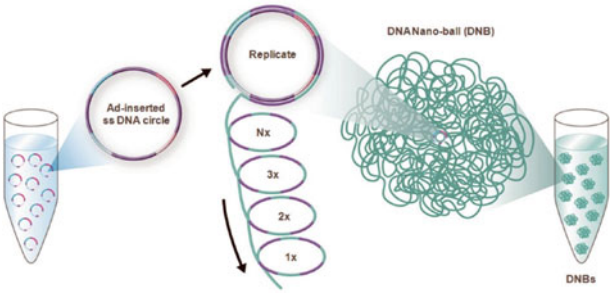
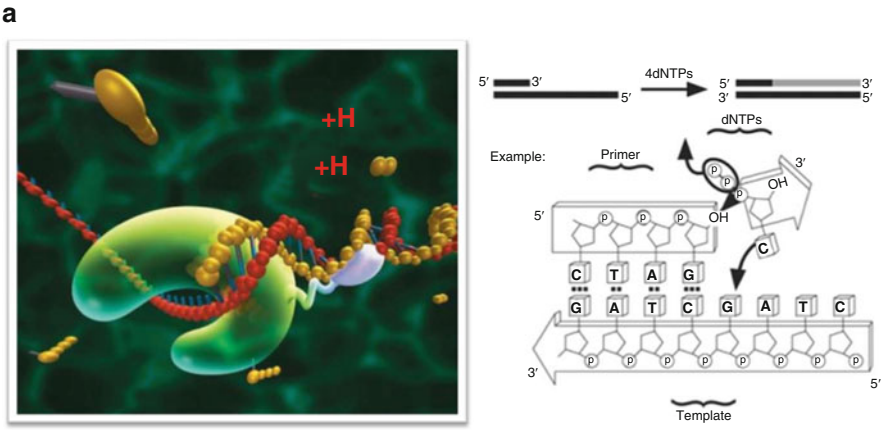


Fig. 3 Complete Genomics' Nanoball formation after rolling cycle amplification. These DNB are sequenced by ligation (©2010 Complete Genomics, Inc. Used with permission)

Complete Genomics announced a combinatorial approach of probe-anchor hybridization and ligation (cPAL) sequencing with the claimed highest throughput among third generation sequencers [35] (Fig. 3). The method uses rolling circle amplification of small DNA sequences into so-called nanoballs. Unchained sequencing by ligation is then used to determine the nucleotide sequence [35]. This method permits large numbers of DNA nanoballs to be sequenced per run and at low consumable costs [36]. The platform has been successfully used in clinical genome sequencing applications such as whole genome sequencing of individuals [37, 38]. However, mapping the short sequencing reads to a reference genomic database can be difficult, especially in the analysis of tumor DNA.

Ion Torrent technology (Life Technologies, Inc), perhaps the current most versatile and low cost method, has been delivered in the form of a personal genomic machine (PGM) as a benchtop instrument to research and clinical laboratories [39]. The sequencing chemistry of Ion Torrent technology involves proton release during each nucleotide incorporation by DNA polymerase. The dense microarray of individual microwell allows DNA polymerase to act on clonally amplified target DNA fragments. Beneath each microwell the Ion-Sensitive Field Effect Transistor (ISFET) detects the pH change as a result of each proton release and a potential change (ΔV) is recorded as direct measurement of nucleotide incorporation events (Fig. 4). The system does not require nucleotide labeling and no optical detection is involved. Ion-Torrent's PGM costs less than 100 K with sequencing capability adequate for small-scaled research projects or clinical diagnostic laboratories. Its introduction into the market hails the beginning of NGS as a commodity for biomedical and clinical applications. As a common feature to many other systems, it has multiplex bar-coding adaptors which allow simultaneous testing of multiple samples. The available chip sizes (314–318) capture 10–1,000 MB of sequence information per run. Although the current Ion Torrent operation is labor intensive, automation with one-step library preparation (one-touch sequencing library preparation kit) has recently become available to simplify the process. The limitations include short read-length (100–200 bp) and technical difficulties in reading through highly repetitive sequences and homopolymers, for which improvement has recently been made.



Sequence is determined by meansuring hydrogen ions released (1per base added per DNA strand) during 2nd strand synthesis when complementary base (A,C,G or T) are sequentially incorporated by DNA polymerase.

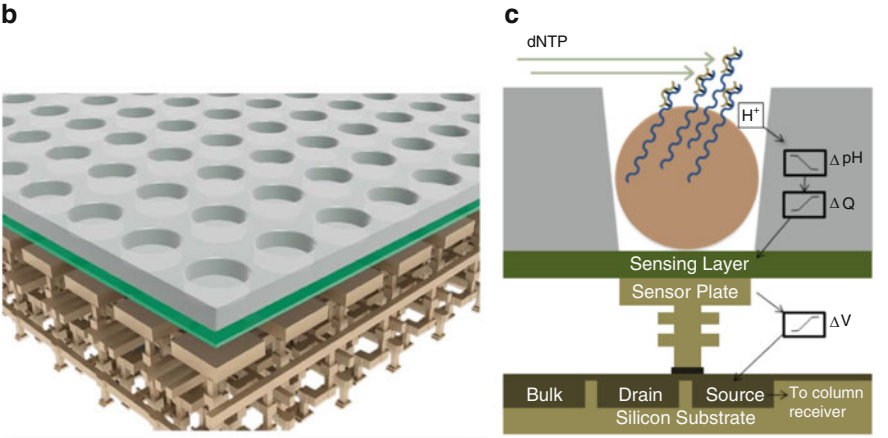


Fig. 4 Ion Torrent Technology. (**a**) Proton release when nucleotide is incorporated by the DNA polymerase into the DNA chain. (**b**) The Ion Torrent proprietary microchip design. (**c**) Cross-section view of a single well that houses ion sphere particles with a clonal amplified DNA template. A hydrogen ion (proton) is released when a nucleotide is incorporated by DNA polymerase. The proton is then detected by the sensing layer due to the change of pH, therefore translating the chemical signal to a digital input. (Copyright permission obtained from Life Technologies, Inc. 2011)

4.4 Emerging Next Generation Sequencing Technology

Oxford's nanopore technology has a different sequencing approach currently in the developmental phase. It uses the scanning tunneling electron microscope (TEM) that measures alterations of conductivity across a nanopore while a single DNA

molecular is passing through. The amount of current that can pass through the nanopore at any given moment varies depending on the shape, size, and length of the nucleotide blocking the ion flow through the pore. The change in current through the nanopore as the DNA molecule passes through represents a direct reading of the DNA sequence. An exonuclease enzyme is used to cleave individual nucleotide molecules from the DNA, and when coupled to an appropriate detection system these nucleotides could be identified in the correct order [40]. Oxford's nanopore technology may also be suitable for integration into a system for analyzing epigenetic modifications. The α-hemolysin nanopore is a promising sensor for ultra-rapid sequencing of DNA strands within nanopores, which may provide additional sequence information using two recognition sites rather than one [41]. Furthermore, nanopore technology is free from some of the drawbacks of other platforms through elimination of the need for optical detection and DNA synthesis and even target DNA amplifications [40, 42–44].

5 Common Problems with NGS Data

There are some common technical problems associated with various NGS platforms. Short reads in many NGS systems result in difficulties with assembling and mapping to the reference sequences, particularly at repetitive regions. Not all sequences are equally processed and sequenced, and DNA regions enriched with GC content are particularly prone to low coverage. For NGS platforms with target amplification or enrichment, amplification bias may be introduced. Last but not least, sequencing errors are present essentially in all NGS platforms. Longer reads are prone to have error readings, particularly towards the ends. Repetitive sequences and homopolymers are also of concern for some third generation sequencers; however, rapid improvement has been made in recent months to overcome these problems. Increase of coverage and deep sequencing are important to correct some of these problems. Table 1 provides a summary of key characteristics of the current NGS platforms.

6 Applications

Genomic medicine driven by the latest development of NGS technologies will have profound impacts on our understanding of the pathogenesis of human disease and many aspects of clinical practice in the future: diagnostics, prognostics, and therapeutics. These clinical applications can be roughly divided according to defined target sequences: whole genome sequencing, targeted sequencing of exomes (selected or whole) or selected genes related to a specific disorder or category of disease, epigenetic mapping, transcriptome sequencing, and microbial population sequencing.

Table 1 Characteristics of current platforms of NGS

NGS	Platform (company)	Chemistry	Read length	Run time [14]	Through-put	Reagent cost/MB [14]	Minimal cost per run ($)	Advantage and limitations
Second	Roche 454	Phosphokinase fluorescent nucleotides	200–400	10 h	High	7–22	1,000–2,000	Long reads, difficulty in homopolymer reading, and expensive
Second	Illumina	Reverse terminator	150	1–14 days	Very high	0.1–0.7	1,000–3,500	Expensive
Second	ABI/SOLiD	Ligation	25–35	8 days	Very high	0.07–0.11	2,000–2,500	Low reagent cost, long turn-around-time, and expensive instrument
Second	Helicos	Single molecular sequencing	25–30	Unknown	High	Unknown	1,100	Expensive instrument
Third	Ion Torrent	Proton/pH detection	100–200	2 h	Moderate	0.93 (318 chip)-50 (314 chip)	750–1,200	Inexpensive, up to 16 bar-coding, Difficulty in homopolymer reading
Third	Complete Genomics	Probe-anchor capture and ligation	10	Unknown	Very high	Unknown	Unknown	Labor intensive, instrument not available in market
Third	Pacific Bio	Real-time single polymerase	1,000–2,000	0.5–2 h	Unknown	11–180	Unknown	Greater base read mistake, instrument not available on market

N.A. Not available

6.1 Whole-Genome Sequencing

Whole genome sequencing can be used to identify germline or somatic mutations, single nucleotide polymorphism (SNPs), indels (insertion and deletion), and copy number variations. In the past 5 years and using NGS, genome-wide association studies have begun to provide unprecedented information of the connection between genetics and disease [45]. For example, such an approach has recently helped to identify new genomic loci for susceptibility to Crohn's disease [46], a chronic debilitating intestinal disease of which the pathogenesis was poorly understood. Not only has the identification of these novel loci improved our understanding of the pathophysiology of the disease but it has also had implications for patient treatment [47]. Genome-wide association studies have also produced data of non-coding sequences implicated in the pathogenesis of complex human disorders [45]. Whole-genome sequencing now permits the compilation of sophisticated databases of full spectrum of germline variants conferring risks for inherited diseases [48] and numerous somatic mutations underlying all aspects of human cancers [47]. The recent $48 million grant from the National Institutes of Health opens the door to the use of NGS to analyze the genomes of thousands of patients who suffer from more than 6,000 rare genetic disorders, many of which follow Mendelian inheritance patterns with mutations involving a single gene (http://www.nih.gov/news/health/dec2011/nhgri-06.htm). It is important to note that as human genomic research has been progressing into the whole-genome sequencing era using NGS, it is imperative to identify and document the genetic variation across human populations to ensure diverse ancestry to be included in our genomic studies so that healthcare disparities introduced by genetics community can be curtailed [49]. Beyond human studies, NGS has been used to study genetic diversities and population structures of endangered animal species [50] and has been found to have significant applications in plant as well [51].

6.2 Targeted Sequencing

NGS has already enabled many forms of clinical diagnostic testing by targeting selected genes or gene exons to accommodate specific clinical needs. Many human disorders are caused by dysfunctions in one of several causative genes. Mutations of genes involved in the same metabolic or signaling pathways may lead to similar disease phenotypes. On the other hand, different mutations involving the same gene may carry subtle to drastically different clinical manifestation of the disease and many diseases may have overlapping mutation profiles. For example, hereditary erythrocyte disorders may involve any of the 27 genes related to the red cell membrane structure, red cell enzyme deficiency, and hemoglobin metabolism [52]. Phenotypic overlap among many involved genes requires precise diagnosis of these

disorders by identifying the corresponding gene mutation(s). Indeed, many medical centers have begun to offer clinical mutation analysis using NGS; examples include extensive panels for detection of mutations in one of the 10–30 genes for the diagnosis of cardiomyopathy [53], X-linked congenital diseases [54–58], comprehensive mutation detection in 24 genes known to cause congenital disorders of glycosylation [59], and various other autosomal disorders [60–62].

Whole exome capture and sequencing by NGS have been successfully applied to identify mutations using various tissue sources [63, 64]. Given the magnitude of carrier burden in human population and increasingly available low-cost NGS platforms, targeted carrier screening is also possible in clinical practice to reduce the incidence and suffering in severe recessive childhood disorders [65]. Targeted gene mutation panel analysis for oncology is becoming increasingly important and cost-effective for cancer diagnosis, prognosis, and precision therapy. Currently dozens of major medical centers in the United States are validating such cancer sequencing panels by NGS.

6.3 Epigenetic Applications

Epigenetic applications of NGS may include platforms such as CHIP-seq-Protein-DNA binding and histone modification [66]. Such technology has been used to map the methylome of the diploid human genome recently [67]. Epigenetic applications of NGS are beginning to provide fundamental insights into human biology and diseases, for example the discovery of widespread allele-specific epigenetic variation in the human genome will likely contribute to our understanding of some common diseases with complex genetic background [68, 69].

6.4 Transcriptome Analysis

Through combination of hybridization capture of cDNAs and next-generation sequencing, targeted RNA-Sequencing (RNA-Seq) provides an efficient and cost-effective method to analyze specific subsets of transcriptome simultaneously for mutation, structural alteration, and expression [70]. NGS technologies with appropriate assembly algorithms have facilitated the reconstruction of the entire transcriptome in the absence of a reference genome [71]. Targeted RNA-Seq is also a powerful tool suitable for a wide range of large-scale tumor-profiling studies to identify sequence variations and novel fusion gene products [72].

6.5 Microbial Population Analysis

NGS is ideal for whole viral, bacterial, and yeast genome sequencing owing to its high throughput, depth of sequencing, and appropriate size of most microbial genomes [39]. Currently large amounts of NGS data have become available that will greatly enhance our understanding of host pathogen interactions with the discovery of new transcripts, splice variants, mutations, regulatory elements, and epigenetic controls [73]. For example, NGS was successfully applied to characterize the genome of the German Enterohemorrhagic Escherichial coli O104:H4 outbreak very recently using the Ion Torrent platform [74]. The ability to perform whole-genome comparisons further allows one to link phenotypic dissimilarities among closely related organisms and their underlying genetic mechanisms, and therefore gain a better understanding of pathogen evolution [75]. Current applications of high-throughput whole viral genome sequencing include detection of viral genome variability and evolution within the host (e.g., human immunodeficiency virus and human hepatitis C virus) and monitoring of low-abundance antiviral drug-resistance mutations. NGS techniques lead to a new field of study called "metagenomics." It is now possible to detect unexpected disease-associated viruses and emerging new human viruses, including cancer-related ones [76]. Other applications include HPV typing because of its high detection sensitivity and its broad spectrum coverage of HPV types, subtypes, and variants [77].

7 Limitations of NGS in Clinical Practice

Several technical limitations of NGS in genomic studies or genomic medicine have already been briefed in the aforementioned sections. To emphasize the clinical aspects of NGS application in medicine the following are important confounding factors that are likely the subjects of many future discussions: (1) quality control/quality assurance programs are likely difficult to be standardized from the initial technical operation to clinical validation; (2) data management and storage (analyzing, managing and instrumentation for storage) require electronic devices of extremely high capacity; (3) daunting challenges in the analysis of sequence data for clinical interpretation (such as previously unknown genetic variants) are major issues to be tackled; (4) reporting complex results may be extremely difficult with regard to clinical implication in disease diagnosis, prognosis and guiding precision therapy; (5) incidental findings with significant biomedical implications may pose ethical responsibilities for pathologists (duty to report); (6) patent infringement may affect laboratories using NGS and reporting genes or DNA sequences under patent protection; and (7) finally how NGS can be adequately reimbursed requires academic institutions, commercial laboratories and regulatory agencies to develop consensus utilities, and fee codes (e.g., CPT codes) in collaboration with and being acceptable by clinicians, service providers and insurance industry.

8 Summary

Defying the Moore's law in computer industry, NGS has been far outperforming its prediction of doubling technical improvement and affordability every 2 years. The exponential improvement of speed and the concomitant astronomical drop in costs are quickly driving NGS from the research arena to the patient bedside. NGS will affect essentially all aspects of clinical care issues enabling many diagnostic tests that have never been considered possible before. In the next 10 years we will likely see the arrival of NGS platforms that are versatile, accurate, affordable, and portable for clinical use. However, a significant hurdle for clinical application of NGS is bioinformatics analysis of the sequencing data. Sequence variant annotation requires data mining into various databases (e.g., locus specific database, HGMD/Biobase, OMIM, SeattleSeq, and 1000 Genome program) and functional prediction programs such as PolyPhen and SIFT are essential for biological and clinical interpretation of new or uncommon sequence variants. Clinical validation of sequence variants in relation to disease or phenotype is more difficult and requires prospective studies of large cohorts of patients. While we transition from single gene analysis in the recent past, to multi-gene panel analysis, to whole exome sequencing, and soon to the whole genome approach, the complexity of the technology and bioinformatics increase dramatically and their clinical applications have been proven far more complicated than previously thought. Navigation from DNA base pairs of the human genome to the bedside of patients will continue to rely on new technologies such as NGS, genomic bioinformatics sciences, large scale collaborative efforts, and a multidisciplinary team approach involving academic institutions, hospitals, government agencies and industries [6].

Acknowledgement and Declaration The author does not have any funding support or financial interest to declare.

References

1. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431(7011):931–945
2. Cancer Genome Atlas Research Network (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216):1061–1068
3. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M et al (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456(7218):66–72
4. Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, Singh B, Heelan R, Rusch V, Fulton L et al (2004) EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A 101(36):13306–13311
5. Dietz HC (2010) New therapeutic approaches to mendelian disorders. N Engl J Med 363(9):852–863

6. Green ED, Guyer MS (2011) Charting a course for genomic medicine from base pairs to bedside. Nature 470(7333):204–213
7. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11(7):476–486
8. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26(10):1135–1145
9. Martinez DA, Nelson MA (2010) The next generation becomes the now generation. PLoS Genet 6(4):e1000906
10. Voelkerding KV, Dames SA, Durtschi JD (2009) Next-generation sequencing: from basic research to diagnostics. Clin Chem 55(4):641–658
11. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT et al (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452(7189):872–876
12. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE (2011) Landscape of next-generation sequencing technologies. Anal Chem 83(12):4327–4341
13. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC et al (2009) The challenges of sequencing by synthesis. Nat Biotechnol 27(11):1013–1023
14. Ekblom R, Galindo J (2011) Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity 107(1):1–15
15. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463–5467
16. Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proc Natl Acad Sci U S A 74(2):560–564
17. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G et al (2007) The diploid genome sequence of an individual human. PLoS Biol 5(10):e254
18. Metzker ML (2010) Sequencing technologies – the next generation. Nat Rev Genet 11 (1):31–46
19. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437(7057):376–380
20. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Methods 5(1):16–18
21. Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. Science 281(5375):363–365
22. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402
23. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K et al (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res 18(7):1051–1063
24. Wu H, Irizarry RA, Bravo HC (2010) Intensity normalization improves color calling in SOLiD sequencing. Nat Methods 7(5):336–337
25. Efcavitch JW, Thompson JF (2010) Single-molecule DNA analysis. Annu Rev Anal Chem (Palo Alto Calif) 3:109–128
26. Thompson JF, Milos PM (2011) The properties and applications of single-molecule DNA sequencing. Genome Biol 12(2):217
27. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW et al (2008) Single-molecule DNA sequencing of a viral genome. Science 320(5872):106–109
28. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323 (5910):133–138

29. Lundquist PM, Zhong CF, Zhao P, Tomaney AB, Peluso PS, Dixon J, Bettman B, Lacroix Y, Kwo DP, McCullough E et al (2008) Parallel confocal detection of single molecules in real time. Opt Lett 33(9):1026–1028
30. Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto GA, Foquet M, Turner SW (2008) Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. Proc Natl Acad Sci U S A 105(4):1176–1181
31. Korlach J, Bjornson KP, Chaudhuri BP, Cicero RL, Flusberg BA, Gray JJ, Holden D, Saxena R, Wegener J, Turner SW (2010) Real-time DNA sequencing from single polymerase molecules. Methods Enzymol 472:431–455
32. Munroe DJ, Harris TJ (2010) Third-generation sequencing fireworks at Marco Island. Nat Biotechnol 28(5):426–428
33. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, Korlach J, Turner SW (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods 7(6):461–465
34. Uemura S, Aitken CE, Korlach J, Flusberg BA, Turner SW, Puglisi JD (2010) Real-time tRNA transit on single translating ribosomes at codon resolution. Nature 464(7291):1012–1017
35. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, Carnevali P, Nazarenko I, Nilsen GB, Yeung G et al (2010) Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science 327(5961):78–81
36. Porreca GJ (2010) Genome sequencing on nanoballs. Nat Biotechnol 28(1):43–44
37. Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, Rowen L, Pant KP, Goodman N, Bamshad M et al (2010) Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328(5978):636–639
38. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D et al (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature 465(7297):473–477
39. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al (2011) An integrated semiconductor device enabling non-optical genome sequencing. Nature 475(7356):348–352
40. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4(4):265–270
41. Stoddart D, Heron AJ, Mikhailova E, Maglia G, Bayley H (2009) Single-nucleotide discrimination in immobilized DNA oligonucleotides with a biological nanopore. Proc Natl Acad Sci U S A 106(19):7702–7707
42. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X et al (2008) The potential and challenges of nanopore sequencing. Nat Biotechnol 26(10):1146–1153
43. Bayley H (2006) Sequencing single molecules of DNA. Curr Opin Chem Biol 10(6):628–637
44. Astier Y, Braha O, Bayley H (2006) Toward single molecule DNA sequencing: direct identification of ribonucleoside and deoxyribonucleoside 5'-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. J Am Chem Soc 128(5):1705–1710
45. Manolio TA (2010) Genomewide association studies and assessment of the risk of disease. N Engl J Med 363(2):166–176
46. Rioux JD, Xavier RJ, Taylor KD, Silverberg MS, Goyette P, Huett A, Green T, Kuballa P, Barmada MM, Datta LW et al (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. Nat Genet 39(5):596–604
47. Van Limbergen J, Wilson DC, Satsangi J (2009) The genetics of Crohn's disease. Annu Rev Genomics Hum Genet 10:89–116
48. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, Cawley S, Chung E, Connell S, Eshragh J et al (2011) Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. Genomics 98(2):79–89

49. Need AC, Goldstein DB (2009) Next generation disparities in human genomics: concerns and remedies. Trends Genet 25(11):489–494
50. Miller W, Hayes VM, Ratan A, Petersen DC, Wittekindt NE, Miller J, Walenz B, Knight J, Qi J, Zhao F et al (2011) Genetic diversity and population structure of the endangered marsupial Sarcophilus harrisii (Tasmanian devil). Proc Natl Acad Sci U S A 108(30):12348–12353
51. Zalapa JE, Cuevas H, Zhu H, Steffan S, Senalik D, Zeldin E, McCown B, Harbut R, Simon P (2012) Using next-generation sequencing approaches to isolate simple sequence repeat (SSR) loci in the plant sciences. Am J Bot 99:193–208
52. Mohandas N, Gallagher PG (2008) Red cell membrane: past, present, and future. Blood 112 (10):3939–3948
53. Hershberger RE, Siegfried JD (2011) Update 2011: clinical and genetic issues in familial dilated cardiomyopathy. J Am Coll Cardiol 57(16):1641–1649
54. Kingsmore SF, Dinwiddie DL, Miller NA, Soden SE, Saunders CJ (2011) Adopting orphans: comprehensive genetic testing of Mendelian diseases of childhood by next-generation sequencing. Expert Rev Mol Diagn 11(8):855–868
55. Tsurusaki Y, Okamoto N, Suzuki Y, Doi H, Saitsu H, Miyake N, Matsumoto N (2011) Exome sequencing of two patients in a family with atypical X-linked leukodystrophy. Clin Genet 80(2):161–166
56. Tsurusaki Y, Osaka H, Hamanoue H, Shimbo H, Tsuji M, Doi H, Saitsu H, Matsumoto N, Miyake N (2011) Rapid detection of a mutation causing X-linked leucoencephalopathy by exome sequencing. J Med Genet 48(9):606–609
57. Schraders M, Haas SA, Weegerink NJ, Oostrik J, Hu H, Hoefsloot LH, Kannan S, Huygen PL, Pennings RJ, Admiraal RJ et al (2011) Next-generation sequencing identifies mutations of SMPX, which encodes the small muscle protein, X-linked, as a cause of progressive hearing impairment. Am J Hum Genet 88(5):628–634
58. Hu H, Wrogemann K, Kalscheuer V, Tzschach A, Richard H, Haas SA, Menzel C, Bienek M, Froyen G, Raynaud M 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
59. 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 13 (11):921–932
60. Doi H, Yoshida K, Yasuda T, Fukuda M, Fukuda Y, Morita H, Ikeda S, Kato R, Tsurusaki Y, Miyake N et al (2011) Exome sequencing reveals a homozygous SYT14 mutation in adult-onset, autosomal-recessive spinocerebellar ataxia with psychomotor retardation. Am J Hum Genet 89(2):320–327
61. Artuso R, Fallerini C, Dosa L, Scionti F, Clementi M, Garosi G, Massella L, Epistolato MC, Mancini R, Mari F et al (2012) Advances in Alport syndrome diagnosis using next-generation sequencing. Eur J Hum Genet 20(1):50–57
62. Bowne SJ, Sullivan LS, Koboldt DC, Ding L, Fulton R, Abbott RM, Sodergren EJ, Birch DG, Wheaton DH, Heckenlively JR 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
63. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S et al (2009) Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A 106(45):19096–19101
64. Choi M, Scholl UI, Yue P, Bjorklund P, Zhao B, Nelson-Williams C, Ji W, Cho Y, Patel A, Men CJ et al (2011) K+ channel mutations in adrenal aldosterone-producing adenomas and hereditary hypertension. Science 331(6018):768–772
65. Bell CJ, Dinwiddie DL, Miller NA, Hateley SL, Ganusova EE, Mudge J, Langley RJ, Zhang L, Lee CC, Schilkey FD et al (2011) Carrier testing for severe childhood recessive diseases by next-generation sequencing. Sci Transl Med 3(65):65ra4

66. Park PJ (2009) ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet 10(10):669–680
67. Fouse SD, Nagarajan RP, Costello JF (2010) Genome-scale DNA methylation analysis. Epigenomics 2(1):105–117
68. Meaburn E, Schulz R (2011) Next generation sequencing in epigenetics: insights and challenges. Semin Cell Dev Biol 23:192–199
69. Ku CS, Naidoo N, Wu M, Soong R (2011) Studying the epigenome using next generation sequencing. J Med Genet 48(11):721–730
70. Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A (2009) Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol Biol Evol 26(12):2731–2744
71. Martin JA, Wang Z (2011) Next-generation transcriptome assembly. Nat Rev Genet 12 (10):671–682
72. Levin JZ, Berger MF, Adiconis X, Rogov P, Melnikov A, Fennell T, Nusbaum C, Garraway LA, Gnirke A (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome Biol 10(10):R115
73. Tripathy S, Jiang RH (2012) Massively parallel sequencing technology in pathogenic microbes. Methods Mol Biol 835:271–294
74. Mellmann A, Harmsen D, Cummings CA, Zentz EB, Leopold SR, Rico A, Prior K, Szczepanowski R, Ji Y, Zhang W et al (2011) Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6(7):e22751
75. Hu B, Xie G, Lo CC, Starkenburg SR, Chain PS (2011) Pathogen comparative genomics in the next-generation sequencing era: genome alignments, pangenomics and metagenomics. Brief Funct Genomics 10(6):322–333
76. Barzon L, Lavezzo E, Militello V, Toppo S, Palu G (2011) Applications of next-generation sequencing technologies to diagnostic virology. Int J Mol Sci 12(11):7861–7884
77. Barzon L, Militello V, Lavezzo E, Franchin E, Peta E, Squarzon L, Trevisan M, Pagni S, Dal Bello F, Toppo S et al (2011) Human papillomavirus genotyping by 454 next generation sequencing technology. J Clin Virol 52(2):93–97