Human and Animal Health

Vol.60: e17160414, January-December 2017
http://dx.doi.org/10.1590/1678-4324-2017160414
ISSN 1678-4324 Online Edition

BRAZILIAN ARCHIVES OF BIOLOGY AND TECHNOLOGY

A N I N T E R N A T I O N A L J O U R N A L

Application of Next-generation Sequencing in Clinical Molecular Diagnostics

Morteza Seifi[1], Asghar Ghasemi[2], Sina Raeisi[3], Siamak Heidarzadeh[4*].
[1]*Department of Medical Genetics, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, AB, Canada;*
[2]*Department of clinical biochemistry, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran;*
[3]*Department of Biochemistry and Clinical Laboratories, Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran;*
[4]*Department of Pathobiology, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.*

ABSTRACT

Next-generation sequencing (NGS) is the catch all terms that used to explain several different modern sequencing technologies which let us to sequence nucleic acids much more rapidly and cheaply than the formerly used Sanger sequencing, and as such have revolutionized the study of molecular biology and genomics with excellent resolution and accuracy. Over the past years, many academic companies and institutions have continued technological advances to expand NGS applications from research to the clinic. In this review, the performance and technical features of current NGS platforms were described. Furthermore, advances in the applying of NGS technologies towards the progress of clinical molecular diagnostics were emphasized. General advantages and disadvantages of each sequencing system are summarized and compared to guide the selection of NGS platforms for specific research aims.

Key words: Bioinformatics, clinical molecular applications, ethical aspects, next generation sequencing

[*]Author for correspondence: heidarzadehsiamak@gmail.com

INTRODUCTION

The aim of the Human Genome Project (HGP) was the sequencing of 3 billion nucleotides of DNA and identifying the genes. Moreover, it was reasonable that the data originated from the genome would result in the development of diagnostic tools, novel therapies and more ability to predict the onset, severity as well as diseases progression [1]. The "original" sequencing technology or Sanger chemistry is based on the uses of labeled nucleotides to read a DNA template during its synthesis. This sequencing technology uses a specific primer to start the read at a specific location along the DNA template, and record the different labels for each nucleotide within the sequence. With some technical improvements, the Sanger method has increased the capacity to read through 1000–1200 base pair (bp). However, it is still not able to overcome 2 kilo base pair (Kbp) beyond the specific sequencing primer. Shotgun sequencing, a new approach to sequence longer sections of DNA, was developed during HGP. In this new approach, genomic DNA using specific enzymes or mechanical manner isbroken down into smaller sections and each section iscloned into sequencing vectors in which cloned DNA sections can be sequenced separately. Based on partial sequences overlaps, the complete sequencing of longer DNA fragments can be achieved by alignment and reassembly of sequence fragments. Shotgun, a significant advantage of HGP, made a possible system to sequence the entire human genome [2]. New next generation sequencers (NGS) technologies randomly sequence the DNA templates in the complete genomes. In this technology the entire genome is broken down into small fragments of DNA which will be ligated to designated adapters for random read during DNA synthesis. Therefore, NGS technology is oftenreferred as a massively parallel sequencing system. The bases of massive parallel sequencing used in NGS is acquired from shotgun sequencing [2].

The next generation sequencers, first launched in 2005, generated short sequences (35–500 bp) by immobilizing millions of amplified DNA fragments onto solid surface and then performing the sequencing reaction [3]. Over the last few years, NGS has been developed into a valuable tool for research applications and it shows tremendous potential in clinical genetic diagnostics [4]. NGS represents an entirely new principle of sequencing technology following Sanger sequencing, which was first described in 1977 [5]. These newer technologies constitute various strategies that rely on a combination of template preparation, sequencing and imaging, genome alignment and assembly methods [6].Technical improvements of this sequencing technology such as the introduction of fluorescent dyes (to replace radiolabeling) and capillary array electrophoresis (to replace gel-based polyacrylamide gel electrophoresis), enabled automation of this technique, thereby increasing the sequencing capacity from a few hundred base pairs to several thousands of them within a single analysis [7]. The NGS technologies are different from the Sanger method in that they provide massively parallel analysis and extremely high-throughput from multiple samples at much reduced cost. Millions to billions of DNA nucleotides can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that were used with Sanger sequencing [8]. Owing to these advantages, NGS technologies have been widely used for many applications, such as rare variant discovery by whole genome re-sequencing or targeted sequencing, transcriptome profiling of cells, tissues and organisms, and identification of epigenetic markers for disease diagnosis [9]. Since the generation of first NGS platform, a number of technological improvements have been made, such as advanced sequencing chemistry and novel methods to detect signals. This has resulted in the generation of the new small NGS systems, bench top sequencers, such as MiSeq, 454 GS Junior and Ion Torrent PGM. Bench top

sequencers are low cost and have easy sample processing protocol and simple data analysis workflow [7]. Commercially available NGS platforms that are suitable for clinical applications include the Roche 454 GS FLX Titanium and Junior systems (Roche Applied Sciences, Penzberg, Germany); Life Technology SOLiD, personal genome machine (PGM), Proton systems (Life Technologies, Carlsbad, CA); and Illumina HiSeq and MiSeq systems (Illumina, San Diego, CA) [10].

OVERVIEW OF NGS TECHNOLOGIES

Roche/454

The Roche GS-FLX 454 Genome Sequencer was the first commercial system launched as the 454 Sequencer in 2004. Using this platform, the second complete genome of an individual (James D. Watson) was sequenced. The upgraded 454 GS FLX Titanium system introduced by Roche in 2008 enhanced the average read length and accuracy to 700 bp and 99.997%, respectively. This platform improved an output of 0.7 Gb of data per run within 24 h. The GS Junior bench-top sequencer system produced the average read length of 700 bp, throughput of 70 Mb and runtime of 10 to 18 h [7,9]. The 454 sequencing system uses two applications: emulsion the polymerase chain reaction (PCR) and pyrosequencing technology. Using emulsion PCR the fragment to be sequenced is loaded on a bead and amplified. The beads bearing amplified fragments are deposited into wells of a picotiter plate where solid-phase pyrosequencing is carried out [7,9]. Additional beads, specific enzymes such as polymerase, sufurylase and luciferase, and other required reagents are added to the wells. Then unlabeled nucleotides are added to flood the picotiter plate in a predetermined order. When a nucleotide found owns complementary nucleotide in the fragment to be sequenced gets incorporated in the strand. Finally, light or luminescence emission from the release of pyrophosphate upon template-directed nucleotide incorporation is monitored in real time. The strength of the Roche/ 454 technology depends on its ability to sequence long fragments [8]. 454 systems with the maximum read length of ~600 bp approaches the halfway of current Sanger sequencing capacities with maximum read length of ~1200 bp. The 454 system has its longest short reads among all other NGS platforms at 600 bp; and reads ~400–600 Mb of sequence per run, this property is critical for some applications such as identification of RNA isoform in RNA-seq and de novo assembly of microbes in metagenomics. Despite the higher costs of 452 system compared to other NGS systems this system seems to be suitable for several applications, such as metagenomics [4].

Illumina/Solexa

The Illumina sequencing system uses a reversible terminator chemistry and an array-based DNA sequencing-by-synthesis technology. In this system, DNA to be sequenced is fragmented and hybridized to a reaction chamber on an optically transparent solid surface (or flow cell). Then step by step DNA synthesis is mediated using reversible terminators which is a series of four fluorescently labeled nucleotides at the 3'-hydroxyl terminus. The first Illumina sequencing system, Genome Analyzer (GA), produced 35-bp reads and generate more than 1 gigabase (Gb) of high-quality sequence per run in 2–3 days [5]. Illumina sequencing system provides six industrial-level sequencing machines (NextSeq 500, HiSeq series 2500, 3000, and 4000, and HiSeq X series five and ten) with mid to high output of sequencing data per run (120–1500 Gb) as well as a compact laboratory sequencer called the MiSeq. This system although small in size, has an output per run of 0.3 to 15 Gb and fast turnover rates suitable for targeted sequencing for clinical and small

laboratory applications. The Illumina/Solexa system advantages are its low cost instrument, short run time, and low error rate. The output of sequencing data per run is 600 Gb, the read lengths are ~100 bp, the cost is lower, and the run times are 3-10 days which are much longer than most other systems. However, this system has some disadvantages including complex data analysis and higher cost of data generation [8,10].

Life Technologies/SOLiD

Sequencing by Oligo Ligation Detection (SOLiD) was initially obtained by Applied Biosystems (ABI) in 2006 [11]. In this system, a unique sequencing method is used by ligation approach. It is carried out through sequential cycles of ligation, in which each sequencing primer is ligated to a specific fluorescence-labeled octamer (eight base) probe due to the complementarity between the di-bases of the template and the probe [2,9]. In this sequencing system, the libraries can be sequenced on a flow cell by the probe ligation that it includes ligation site (the first base), cleavage site (the fifth base), and four different fluorescent dyes (linked to the last base) [11]. As each four di-bases (e.g., AG, GA, TC, CT) are labeled with one of the four fluorescent dyes, the di-nucleotides at the same positions of each template emit a unique fluorescent color [3]. Therefore, each cycle consists of a ligation step which is followed by fluorescence detection and it is repeated [2]. The fluorescent signals are recorded during the probes complementary to the template strand and disappear by the cleavage of probes' last three bases. The sequence of the target can be deduced by ladder primer sets after five sequencing cycle [11]. After these five ligation cycles, each nucleotide in the template is read twice by two fluorescent signals, greatly improving base-calling accuracy [4]. After these steps, the original sequence of color coding is accumulated. According to the di-base coding matrix, the original color sequence can be decoded to obtain the base sequence if the base types for one of any position in the sequence were known. Since a unique color corresponding four base pair, the color coding of the base will directly affect the decoding of its following base. Furthermore, a wrong color coding will cause a chain decoding mistakes [11]. ABI released the first SOLiD system at the end of 2007. Initially, the read length of SOLiD was 35 bp and its output was 3G data/run. Due to two-base sequencing method, SOLiD could reach a high accuracy of 99.85% after filtering [11]. SOLiD4 analyzer has a read length of up to 50 bp and can produce 80-100 G bp of sequences per run [2]. In late 2010, the SOLiD 5500xl sequencing system was released. From the first SOLiD to SOLiD 5500xl, five upgrades were released by ABI in just three years. The SOLiD 5500xl realized improved read length, accuracy, and data output of 85 bp, 99.99%, and 30G/run, respectively. A complete run can be finished approximately within seven days. Operating system used by most researchers is GNU/LINUX. Application of SOLiD includes whole genome resequencing, targeted resequencing, transcriptome research (including gene expression profiling, small RNA analysis, and whole transcriptome analysis), and epigenome (likeChromatin immunoprecipitation [ChIP] followed by high-throughput DNA sequencing [ChIP-seq] and methylation). Like other NGS systems, SOLiD's computational infrastructure is expensive and not trivial to use; it requires an air-conditioned data center, computing cluster, skilled personnel in computing, distributed memory cluster, fast networks, and batch queue system. More data will be generated after bioinformatics analysis [11]. The SOLiD system has the lowest error rate among the current NGSs. Its most common error type is substitution. In addition, an underrepresentation of AT-rich regions has also been shown in the SOLiD data [11].

Emerging Technologies

The advent of single-molecule sequencing (SMS) has made a technological leap forward in the development of NGS. The advantage of SMS lies in its ability to directly sequence single nucleic acid molecules (NA) in biological samples without amplification [12]. It can minimize sample handling, decrease sample input requirements, prevent amplification-induced bias and errors, enhance read length flexibility and enable accurate quantitation of NAs [12]. The Helicos Genetic Analysis System is the first commercially available SMS platform [13] In this system, poly-(A)-tailed single-stranded DNA is captured by poly-(T) oligonucleotide primers tethered to the surface of a flow cell. Sequencing is executed via repetitive cycles of DNA polymerase-mediated single-base primer extension that uses a highly sensitive fluorescence detection system to directly detect each nucleotide as it is synthesized [2]. The distinct characteristic of this technology is its ability to sequence single DNA molecules without amplification, defined as Single-Molecule Real Time (SMRT) DNA sequencing. The short-read length ranges from 30 bp to 35 bp at present time, with a raw base accuracy greater than 99%, and 20–28 Gbp of potential sequence reads per run in the near future [11,12]. Helicos system may also have some disadvantage. In this system, there is no GC-content bias in read coverage. Furthermore, the current error rate of this system is relatively high (3-5%). The main error type is deletion which probably results from detection errors and/or incorporation of unlabeled nucleotides [10]. Other SMS technologies with higher sequencing speed, longer read lengths, or lower overall cost are also emerging.

Pacific BioSciences is developing a SMRT DNA sequencing technology. This approach performs single-molecule sequencing by identifying nucleotides which are phosphor-linked with distinctive colors. During the synthesis process, fluorescence emitted as the phosphate chain is cleaved and the nucleotide is incorporated by a polymerase into a single DNA strand [5]. This system presently offers a throughput of approximately 50-100 Mb/run, which is much lower than current NGS platforms. Moreover, the single-read error rate is typically 15%, exceeding the error tolerance of many applications [5,10]. Nano-technology has long been considered a cutting-edge equipment for single-molecule DNA sequencing. The concept of this technology is based on the observation that when a DNA strand is pulled through a nanopore by an electrical current, each nucleotide base creates a unique pattern in the electrical current. This unique nanopore electrical current fingerprint can be used for nanopore sequencing[2]. Nanopore sequencing potentially performs long read lengths of up to tens of kilobases, minimal requirements of reagent and sample preparation, and high sequencing pace at low cost. However, several problems remain to be solved before the use of this sequencing system. The high speed of DNA translocation through nanopores makes it difficult to identify base signals from background noises by an electronic sensor. The random motion of molecules during translocation also can increase the difficulty in reaching single-base resolution [9]. Reveo, an electronic detection for SMS, is developing a technology to stretch out DNA molecules on conductive surfaces for electronic base detection. A stretched and immobilized DNA strand is read via multiple nano-knife edge probes. Each nano-knife edge probe specifically identifies only one nucleotide for SMS [2]. Electron microscopy (EM)for SMS has recently been reassessed with the developing of new technologies. Since scanning tunneling microscopy (STM) can reach atomic resolution, STM for SMS is being explored [10]. Light Speed Genomics is improving a microparticle approach by capturing sequence data with optical detection technology and new sequencing chemistry from a large field of view to decline the time consuming sample and detector rearrangement [2,9]. Halcyon Molecular is emerging a DNA sequencing technology by atom-by-atom recognition and EM analysis. The main advantage of this system is very long read lengths [6]. Other novel sequencing systems are also

under progress, such as fluorescence resonance energy transfer (FRET)-based SMS technology from VisiGen Biotechnologies, Ion semiconductor sequencing technology from Ion Torrent, now part of Life Technologies, and DNA nanoball sequencing technology from Complete Genomics. Despite remarkable advantages of these new systems, there remains much room for development before introducing them into clinical practice [8]. For advantages/disadvantages of NGS and for the comparison of NGS systems see Table 1 and 2.

Table 1. Advantages and disadvantages of NGSs

NGS system	Advantages	Disadvantages
Roche/ 454 Life Sciences	Small sized, low cost instrument, Proven sequencing technology, Short run time, long read length	Low throughput, Complex data analysis.
Illumina	Small sized, low cost instrument, Low error rate, Proven sequencing technology, Short run time, Acceptable read length, High level of multiplexing	Complex data analysis, high cost of data generation.
Life Technologies/ Applied Biosystems SOLiD	Low error rate, Proven sequencing technology, Flexibility and scalability	Complex data analysis, high cost of instrument and sequence generation.
Life Technologies/ IonTorrent	Small sized, low cost instrument, Direct signal detection, Short run time, Acceptable read length	Low throughput.
HelicosBioSciences	Sequencing without amplification, High throughput, High raw base accuracy	High cost of instrument, No GC-content bias in read coverage, High error rate, Long time per run

Table 2. The comparison of NGS systems.

NGS system	Company	Amplification	Sequencing	Read Length (bp)	Throughput	Time /run	Dominant error type	Error rate %
GS FLX Titanium XL+	Roche/ 454 Life Sciences	Emulsion PCR	Pyrosequencing	Up to 1000	700 Mb	23 h	Indel	0.5
GS FLX Titanium XLR70	Roche/ 454 Life Sciences	Emulsion PCR	Pyrosequencing	Up to 600	450 Mb	10 h	Indel	0.5
GS Junior	Roche/ 454 Life Sciences	Emulsion PCR	Pyrosequencing	400	35 Mb	10h	Indel	0.5
HiSeq 2000	Roche/ 454 Life Sciences	Emulsion PCR	Pyrosequencing	36-100	105-600 Gb	2-11 days	Indel	0.5
Genome Analyzer IIx	Illumina	Bridge PCR	Synthesis with reversible terminator	35-150	10-95 Gb	2-14 days	Substitution	0.2
MiSeq	Illumina	Bridge PCR	Synthesis with reversible terminator	36-250	540 Mb-8.5 Gb	4-39 h	Substitution	0.2
5500xl SOLiD	Life Technologies/ Applied Biosystems	Emulsion PCR	Ligation	35-75	10-15 Gb	1 day	Substitution	0.1
SOLiD™ 4	Life Technologies/ Applied Biosystems	Emulsion PCR	Ligation	25-50	25-100 Gb	3.5-16 days	Substitution	0.1
Ion Proton™	Life Technologies/ IonTorrent	Emulsion PCR	Ion semiconductor	Up to 200	Up to 10 Gb	2-4 h	Indel	1
Ion PGM™	Life Technologies/ IonTorrent	Emulsion PCR	Ion semiconductor	35-200	300 Mb-1 Gb	0.9-4.5 h	Indel	1

HeliScope™	HelicosBioSciences	-	SMS	25-55	21-35 Gb	8 days	Deletion	5
PacBio RS	Pacific Biosciences	-	SMS	250-10000	50-100 Mb	0.5-4 h	Indel	13
Nanopore	Oxford Nanopore Technologies	-	SMS	2000-tens of kb	300-400Gb	50 h	Indel	38

APPLICATIONS OF NGS

Mendelian diseases

Mendelian or monogenic disorders result from a mutation at a single genetic locus. A locus may be present on an autosome or on a sex chromosome, and it may manifest in a dominant, or a recessive or a X-linked mode. Phenylketonuria (PKU), cystic fibrosis, sickle-cell anemia, and oculocutaneous albinismare examples of single-gene diseases with an autosomal recessive inheritance pattern and are associated with recessive mutations in the phenylalanine hydroxylase (*PAH*), cystic fibrosis conductance regulator (*CFTR*), beta hemoglobin (*HBB*) and *OCA2* genes, respectively [14]. Huntington's, myotonic dystrophy, polycystic kidney, familial hypercholesterolemia, and neurofibromatosis diseases are instances of an autosomal dominant single-gene diseases. Huntington's, myotonic dystrophy and neurofibromatosis are associated with mutations in the mutant huntingtin gene (*HTT*) and dystrophiamyotonica protein kinase (*DMPK*), neurofibromin (*NF1*) genes respectively. Polycystic kidney disease is associated with mutations in either the polycystic kidney disease 1 (*PKD1*) or the polycystic kidney disease 2 (*PKD2*) genes. Familial hypercholesterolemia can be caused by mutations in both the apolipoprotein B (*APOB*) and the low-density lipoprotein receptor (*LDLR*) genes [14]. Duchenne muscular dystrophy and hemophilia A are examples of single-gene diseases that exhibit an X chromosome-linked recessive pattern of inheritance. Duchenne muscular dystrophy due to mutations in the dystrophin gene (*DMD*) [14]. X-linked dominant hypophosphatemic rickets and Rett syndrome are examples of X chromosome-linked dominant diseases and can be caused by mutations inthe phosphate-regulating endopeptidase (*PHEX*)and the methyl-CpG-binding protein 2 gene (*MECP2*) genes, respectively [14]. Non-obstructive spermatogenic failure is one example of a Y-linked disorder that result from mutations in the ubiquitin-specific protease 9Y gene (*USP9Y*) on the Y chromosome [14].

The gold standard of molecular diagnosis in Mendelian diseases has been Sanger sequencing (dideoxy method) and the technique remains the choice method for clinical genetic study; the purpose of which is to confirm a suspected diagnosis and allow more accurate genetic counseling. Frederick Sanger introduced DNA sequencing assay which was based on Sanger sequencing method (also known as dideoxy method), and also Walter Gilbert developed one other sequencing technology based upon chemical modification of DNA molecule and subsequent cleavage at specific bases. The automatic sequencing instruments and associated software using the capillary sequencing technologies and Sanger sequencing methods became the main procedures for the completion of human genome project in 2001. This project greatly accelerated the development of powerful novel sequencing instrument to enhance speed and accuracy, while simultaneously reducing cost and manpower. Not only this, X-prize also increased the development of NGS. The NGS technologies are different from the Sanger method in sight of massively parallel analysis, high throughput, and reduced cost [15]. In 2004, the NGS methods, highly parallel sequencing platforms, were introduced and the next development in molecular genetics (such as detection the disease-causing genes) is expected [16].

NGS technology has high speed and throughput, both quantitative and qualitative sequence data, equivalent to the data from human genome project, in 10–20 days. Numerous different methods are employed in which NGS is being applied for identifying causal gene variant in the rare diseases. Whole-exome sequencing (WES), Whole-genome sequencing (WGS), methylome sequencing,transcriptome sequencing, and other sequencing ways are used in NGS methods [16]. Miller syndrome is the first rare Mendelian disorder from which its causal variants were identified, due to the development of WES. These investigators explained dihydroorotate dehydrogenase (*DHODH*) mutations in 3 affected pedigrees following filtering against public single nucleotide polymorphism (SNP) databases and eight HapMap exomes [16]. There are increasing numbers of reports identifying the causal variants of the diseases. Over 100 causative genes in various Mendelian diseases have been identified by method of exome sequencing. In addition to gene discovery of diseases that are dominant and recessive, WES has been applied for determining somatic mutations in tumors and rare mutations with moderate effect in common disorders as well as clinical diagnoses [16]. One of the biggest challenges for clinicians is deciding between applications of targeted versus WES. As the cost of sequencing decreases, WES appears to be a more cost-effective approach. However, there are special considerations before embarking on one over the other [17,18]. Although exomes are proposed to cover the protein-coding regions of the genome, the complete coverage can be between 85 to 95 % only. This means that a specific gene of interest with respect to a specific phenotype may not be covered, either completely or partially. Reasons include poorly performing capture probes due to sequence homology, repetitive sequences, or high GC content. In addition, a targeted method has a much higher or even complete coverage of all the phenotype-specific genes by filling in the gaps with complementary methods including Sanger sequencing or long range PCR. Moreover, a targeted testing allows for deeper coverage of these genes compared to WES, which provides more confidence in the variants detected. However, both are still disposed to sequencing artifacts, and Sanger sequencing of candidate variants is suggested in both methods before returning the results to the patients [17,18]. Finally, laboratories that offer targeted method can have expert knowledge of the given phenotype and may be in a better location to prioritize variants detected via NGS. They may also be able to necessitate specific evaluations to define the significance of certain variants [17,18].Currently, whole genome association surveys aim to identify the genetic basis of traits and disease susceptibilities by SNP microarrays that capture most of the common genetic variation in the human population. Risk variants for many diseases have been identified. However, with only a few exceptions (e.g., age-related macular degeneration, type 1 diabetes), the risk variants usually describe only a minor fraction of the genetic risk that is known to exist. There are several factors that are likely to contribute to this observation [19]. Usual variants may have only minor effects on a phenotype, or have variable penetrance due to epigenetic or epistatic influences. Two additional factors are rare variants and copy number variants (CNVs). It is known that these types of genomic variations might have important influences on disease phenotypes. However, the evaluation of these variants cannot be achieved readily using present genotyping microarray technologies [19,20]. WGS gives a potential solution by providing the most comprehensive collection of unusual variants and structural variation for sequenced peoples [19].

Common diseases

Common complex diseases, in opposite to single gene disorders, are caused by the interaction of genetic and environmental factors, each one with a small effect, and a few sometimes acting individually as a necessary, although on their own insufficient,

initiate the disease to occur. For studying the genetic background of these phenotypes, principles of genetic mapping have been developed by populations rather than families [21]. According to the "common disease-common variant" (CD-CV) hypothesis, it is assumed that a massive number of polymorphisms, classically defined as having an allele frequency >1%, are pathogenically associated with common complex diseases [21,22]. Accordingly, testing of all these variants shed light on the underlying heritability and clearly identify the related key susceptibility genes. Recently, the CD-CV hypothesis has been tested following the extension of catalogues of common variants, genotyping arrays, haplotype maps and innovative and more accurate statistical methods. Genome-wide association surveys (GWAS) involve the study of a comprehensive inventory of hundreds of thousands of SNPs in hundreds of thousands of cases and controls from a population to find the variants associated with a traits or disease [23]. Technical progresses and quality control now permit to obtain valid and inexpensive genotyping of more than 1 million SNPs of a person's DNA in a single scan [21,24]. Since 2006, 1628 GWAS have been documented, and identified hundreds loci associated with more than 250 common traits or diseases [21]. Overall, only 12% of SNPs related to complex phenotypes are placed in or occur in tight linkage disequilibrium with protein-coding regions of genes, approximately 40% falls in the intergenic regions and another 40% in noncoding introns, supposing a role of these latter regions in the regulation of gene expression [21]. A few major findings emerge from this great amount of data. The detection of hundreds of loci involved in regulation of the phenotype of complex diseases and traits provides clues to detect the underlying cellular pathways, and in some cases also gives new hints concerning therapeutic methods which has recently been supported by the results of several research studies[21].

Cancer

As cancer is a genetic disease caused by heritable or somatic mutations, new DNA sequencing technologies will have a significant effect on the detection, management and treatment of disease. NGS is empowering worldwide collaborative efforts, including The Cancer Genome Atlas (TCGA) project and the International Genome Consortium (ICGC),to catalogue the genomic landscape of thousands of cancer genomes across many disease types. [25,26]. Numerous preliminary reports from individual studies leading to these consortia have already been published [25,27,28]. These discoveries will ultimately cause a better understanding of disease pathogenesis, bridging to a new era of molecular pathology and personalized medicine. It is easy to imagine that every patient will soon have both their constitutional and cancer genomes sequenced, the latter perhaps several times for monitor disease progression, therefore allowing a proper molecular subtyping of disease and the rational usage of molecularly guided treatments. Several molecular pathology laboratories are now considering the sequencing platforms, methods and additional tools required for making the transition to NGS [25]. The application of NGS, mostly via WGS and WES, has made an explosion in the context and complexity of cancer genomic modifications, including point mutations, deletions or small insertions, copy number alternations and structural variations. By comparison of these changes to matched normal samples, researchers have been able to separate two categories of variants: germ line and somatic. The whole transcriptome approach (RNA-Seq) can not only quantify gene expression templates, but moreover detect RNA editing, alternative splicing and fusion transcripts. In addition, epigenetic alterations, histone modifications and DNA methylation change can be investigated using other sequencing approaches including ChIP-seq and Bisulfite sequencing (Bisulfite-Seq). The combination of these NGS technologies gives a high-resolution and global view of the cancer genome. Application of powerful bioinformatics

equipments, researchers aim to detect the massive amount of data to improve our understanding of cancer biology and to progress personalized treatment strategy [25,29]. In the past few years, many NGS-based studies have been conducted to provide a comprehensive molecular diagnosis of cancers, to identify novel genetic alterations leading to oncogenesis, metastasis and cancer progression, to survey heterogeneity, evolution and tumor complexity [25]. These efforts have provided significant achievements for many diseases such as melanoma, acute myeloid leukemia (AML), breast, lung, liver, kidney, ovarian, colorectal, head and neck cancers [6,30-37].

Epigenetic

Epigenetics is definite as heritable modifications in gene activity and expression that occur without alteration in DNA sequence [38-40]. It is known these non-genetic changes are strictly regulated by two major epigenetic modifications: histone modifications and DNA methylation [38-40]. Functionally, the profiles of epigenetic modifications can serve as epigenetic markers to show expression and gene activity as well as chromatin state [41-43]. Epigenetic modifications are critical for packaging and illustrating the genome under the influence of physiological factors [44]. Epigenetics is one of the most rapid-growing areas of science and has now become a central aspect in biological studies of development and disease [45-47]. Recently, there have been quick progresses in the perception of epigenetic mechanisms, which include DNA methylation, small and non-coding RNAs, histone modifications and chromatin architecture [48-51]. These mechanisms, in addition to other transcriptional regulatory events, ultimately regulate gene function and expression in development and differentiation, or in reply to environmental causes [49-52]. Epigenetic research can help show how cells carrying identical DNA differentiate into various cell types and how they conserve differentiated cellular states [52]. Epigenetics is hence considered a relation between phenotype and genotype [38,39]. While epigenetics denotes to the variations of single or sets of genes, the term epigenome represents the complete epigenetic situation of a cell, and indicatives to total analyses of epigenetic markers across the whole genome [39]. It is therefore very important to map the epigenetic alteration patterns or profile the epigenome in a given cell, which afterwards can be used as epigenetic biomarkers for prognosis, diagnosis, and therapeutic development [44,47,50,53]. Human epigenome projects are currently active to catalog all the epigenetic markers in all major tissues across the whole genome. The resulting reference patterns will usher in epigenetics as an exciting new period of medical science [44,47].The NGS technologies offer the potential to substantially fasten epigenomic research, including posttranslational changes of histones, the interaction among transcription factors and their direct targets, nucleosome positioning on a genome-wide scale and the diagnosis of DNA methylation maps [54-57]. Using methylated DNA immunoprecipitation (meDIP) and bisulphite methods can be studied at the methylation of DNA whereas ChIP-Seq technology, the location of transcription factors, post-translational and modifications of histones can be used to study the whole-genome level [58,59]. The ChIP-Seq on NGS platform allows nowadays investigators to develop both quality and quantity of produced data. Amongst other prevalent high-throughput ways, protein-DNA interactions have been evaluated through the combination of chromatin immunoprecipitation with DNA microarray (ChIP-chip). Contrarily, ChIP-seq method prevents two advantages from the NGS platforms, first, it is not limited by the microarray content and next, it does not rely on the efficiency of probe hybridization [60,61]. Studies have shown that ChIP-seq had well resolution and required fewer replicates [54].

BIOINFORMATICS FOR NGS DATA

In the past few years the advent of several NGS platforms have been observed that are based on various performance of cyclic-array sequencing [62]. The notion of cyclic-array sequencing can be abbreviated as the sequencing of a dense array of DNA traits via iterative cycles of enzymatic correction and imaging-based data collection. The commercial products that are depend on this sequencing technology include 454/Roche, Illumina/Solexa, ABI/SOLiD and the Heliscope/Helicos. Although these platforms are quite different in sequencing biochemistry as well as in how the array is generated, their work flows are basically very similar [62]. All of them feasible the sequencing of millions of short sequences in a time, and are capable of sequencing a whole human genome per week at a cost 200-fold lower than previous methods. Furthermore, NGS platforms allow the production of many kinds of sequence data: for instance, they are used to make de novo sequencing, to resequence persons when a reference genome already exists, sequence RNA to quantify expression level (RNA-seq) and study the regulation of genes through ChIP-Seq [63-66]. The emergence of NGS platforms has made various opportunities for genomic variant detection [67-69].

ETHICAL ASPECTS

One of the advantages of exome or genome sequencing is to identify all variants within a given individual, so they can be identified variants related to the disease in question and to other diseases. This is also known from other genome-wide screening instruments, including microarray analysis, or from brain-wide neuroimaging, such as magnetic resonance imaging (MRI) scans. Remarkably, in a recent study that investigated 1000 exomes for “actionable” pathogenic single-nucleotide variants, that is those that cause treatable or even preventable conditions, 23 participants (approximately 2%) harbored such substitutions. The American College of Medical Genetic and Genomics (ACMG)recently published a policy statement on clinical molecular analysis underlying the importance of alerting the patient/family to the possibility of such results in pretest counseling discussions, also reporting of results [71]. Furthermore, the ACMG put guidelines for clinical testing laboratories that list 56 genes in which incidentally found known pathogenic or expected pathogenic mutations should be reported to the patients [71]. The selection of these 56 genes is relied on pathogenicity and the possibility of the genetic result leading to a specific therapeutic option (“actionable” findings). However, there is an in time discussion how to best proceed with incidental findings [72,73]. Whereas in the research setting, these incidental findings usually do not absorb much attention, it is notable that there is growing intrigue in receiving information about incidental findings on the patient side [73]. This is also reflected by the increasing providing and popularity of direct-to-consumer genetic testing (DTCGT). DTCGT allows persons to obtain genetic tests and receive results without the interfering of a health professional. Shortly after DTCGT became available, the ACMG provided a statement on DTCGT including the following recommendation: “Companies offering DTCGT should uncover the sensitivity, specificity, and predictive value of the test, and the populations for which this data is known, in a readily understandable and accessible fashion” [74]. Nevertheless, even if provided in a transparent fashion, it is not easy for most individuals and even for several doctors to properly interpret the test results and risk assessment. Importantly, most of these test results associate to low-risk variants for common disorders or traits, and the important difference between a causal mutation and a gene variant that only slightly increases the lifetime

risk for a special condition is often unclear to both patients and their attending physicians. For instance, even if numerous risk variants for Parkinson disease occur to coincide, the risk of developing the disorder will only be increased 2.5-fold [75]. While at first seen this may look considerable, given the low frequency of 0.14% of Parkinson disease in the general population, the related lifetime risk for the disease will still be as low as 0.35% [76]. Thus, the results of DTCGT can make significant uncertainty and anxiety, requiring extensive post-test counseling. According to recent studies, the authors of position statements, policies, and recommendations explained more potential harms than benefits. But, although some noted that DTCGT should be actively discouraged, others supported consumer rights to make autonomous choices [77]. Remarkably, many companies submitting DTCGT have currently suspended their health-related genetic assays to comply with the US Food and Drug Administration's (FDA) directive to cut new user access until they can provide suitable evidence that the results are trustworthy and will not jeopardize consumers' health [7].

ACKNOWLEDGMENTS

The authors would like to thank Dr. Ali Samadikuchaksaraei for his assistance in reviewing of this manuscript and useful comments.

CONFLICT OF INTERESTS

There is no conflict of interest

REFERENCES

1. Desai AN, Jere A. Next-generation sequencing: ready for the clinics? Clinical genetics. 2012;81(6):503-10.
2. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. Journal of genetics and genomics. 2011;38(3):95-109.
3. Chan EY. Advances in sequencing technology. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis. 2005;573(1):13-40.
4. Weiss MM, Zwaag B, Jongbloed JD, Vogel MJ, Brüggenwirth HT, Lekanne Deprez RH, et al. Best Practice Guidelines for the Use of Next-Generation Sequencing Applications in Genome Diagnostics: A National Collaborative Study of Dutch Genome Diagnostic Laboratories. Human mutation. 2013;34(10):1313-21.
5. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. 1977;74(12):5463-7.
6. Metzker ML. Sequencing technologies—the next generation. Nature reviews genetics. 2010;11(1):31-46.
7. Lohmann K, Klein C. Next generation sequencing and the future of genetic diagnosis. Neurotherapeutics. 2014;11(4):699-707.
8. Rabbani B, Nakaoka H, Akhondzadeh S, Tekin M, Mahdieh N. Next generation sequencing: implications in personalized medicine and pharmacogenomics. Mol Biosyst. 2016;12(6):1818-30.
9. Xuan J, Yu Y, Qing T, Guo L, Shi L. Next-generation sequencing in the clinic: promises and challenges. Cancer letters. 2013;340(2):284-95.
10. Chang F, Li MM. Clinical application of amplicon-based next-generation sequencing in cancer. Cancer genetics. 2013;206(12):413-9.
11. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. BioMed Research International. 2012;2012.
12. Milos PM. Emergence of single-molecule sequencing and potential for molecular diagnostic applications. Expert review of molecular diagnostics. 2009;9(7):659-66.

13. Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008;320(5872):106-9.
14. Chial H. Mendelian genetics: patterns of inheritance and single-gene disorders. Nature Education. 2008;1(1):63.
15. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. BioMed Research International. 2012:1-11.
16. Rabbani B, Mahdieh N, Hosomichi K, Nakaoka H, Inoue I. Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. Journal of human genetics. 2012;57(10):621-32.
17. Jamuar SS, Tan EC. Clinical application of next-generation sequencing for Mendelian diseases. Hum Genomics. 2015;9:10.
18. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet. 2013;14(4):295-300.
19. Ng PC, Kirkness EF. Whole genome sequencing. Genetic Variation: Springer; 2010. p. 215-26.
20. Fujimoto A, Nakagawa H, Hosono N, Nakano K, Abe T, Boroevich KA, et al. Whole-genome sequencing and comprehensive variant analysis of a Japanese individual using massively parallel sequencing. Nature genetics. 2010;42(11):931-6.
21. Dallapiccola B, Mingarelli R, Boccia S. Genetic prediction of common complex disorders assessed by next generation sequencing and genome wide analysis. Italian Journal of Public Health. 2012;9(4).
22. Reich DE, Lander ES. On the allelic spectrum of human disease. TRENDS in Genetics. 2001;17(9):502-10.
23. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science. 2002;296(5576):2225-9.
24. Spencer CC, Su Z, Donnelly P, Marchini J. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 2009;5(5):e1000477.
25. Meldrum C, Doyle MA, Tothill RW. Next-Generation Sequencing for Cancer Diagnostics: a Practical Perspective. The Clinical Biochemist Reviews. 2011;32(4):177-95.
26. Hudson TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. Nature. 2010;464(7291):993-8.
27. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell. 2010;17(1):98-110.
28. Puente XS, Pinyol M, Quesada V, Conde L, Ordóñez GR, Villamor N, et al. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. Nature. 2011;475(7354):101-5.
29. Harris TJ, McCormick F. The molecular pathology of cancer. Nature reviews Clinical oncology. 2010;7(5):251-65.
30. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Human molecular genetics. 2010:ddq416.
31. Shendure J, Ji H. Next-generation DNA sequencing. Nature biotechnology. 2008;26(10):1135-45.
32. Glenn TC. Field guide to next-generation DNA sequencers. Molecular ecology resources. 2011;11(5):759-69.
33. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proceedings of the National Academy of Sciences. 2003;100(15):8817-22.
34. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
35. McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, et al. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome research. 2009;19(9):1527-41.
36. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53-9.

37. Ronaghi M, Uhlén M, Nyren P. A sequencing method based on real-time pyrophosphate. Science. 1998;281(5375):363.
38. Goldberg AD, Allis CD, Bernstein E. Epigenetics: a landscape takes shape. Cell. 2007;128(4):635-8.
39. Bird A. Perceptions of epigenetics. Nature. 2007;447(7143):396-8.
40. Epigenetic Modifications Regulate Gene Expression. SABiosciences. 2008(8):2-5.
41. Jenuwein T, Allis CD. Translating the histone code. Science. 2001;293(5532):1074-80.
42. Berger SL. The complex language of chromatin regulation during transcription. Nature. 2007;447(7143):407-12.
43. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693-705.
44. Ozanne SE, Constância M. Mechanisms of disease: the developmental origins of disease and the role of the epigenotype. Nature clinical practice Endocrinology & metabolism. 2007;3(7):539-46.
45. Reik W. Stability and flexibility of epigenetic gene regulation in mammalian development. Nature. 2007;447(7143):425-32.
46. Feinberg AP, Tycko B. The history of cancer epigenetics. Nature Reviews Cancer. 2004;4(2):143-53.
47. Esteller M. Epigenetics in cancer. New England Journal of Medicine. 2008;358(11):1148-59.
48. Xu X, Yang P, Shu Z, Bai Y, WANG C-Y. DNA methylation in the pathogenesis of autoimmunity. Gene Discovery for Disease Models. 2011:31.
49. Li B, Carey M, Workman JL. The role of chromatin during transcription. Cell. 2007;128(4):707-19.
50. Laird PW. The power and the promise of DNA methylation markers. Nature Reviews Cancer. 2003;3(4):253-66.
51. Mattick JS, Makunin IV. Non-coding RNA. Human molecular genetics. 2006;15(suppl 1):R17-R29.
52. Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. Cell. 2008;132(4):567-82.
53. Mulero-Navarro S, Esteller M. Epigenetic biomarkers for human cancer: the time is now. Critical reviews in oncology/hematology. 2008;68(1):1-11.
54. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. Journal of Applied Genetics. 2011;52(4):413-35.
55. Chung CAB, Boyd VL, McKernan KJ, Fu Y, Monighetti C, Peckham HE, et al. Whole methylome analysis by ultra-deep sequencing using two-base encoding. PLoS One. 2010;5(2):e9320.
56. Fouse SD, Nagarajan RP, Costello JF. Genome-scale DNA methylation analysis. Epigenomics. 2010;2(1):105-17.
57. Bhaijee F, Pepper DJ, Pitman KT, Bell D. New developments in the molecular pathogenesis of head and neck tumors: a review of tumor-specific fusion oncogenes in mucoepidermoid carcinoma, adenoid cystic carcinoma, and NUT midline carcinoma. Annals of diagnostic pathology. 2011;15(1):69-77.
58. Neff T, Armstrong S. Chromatin maps, histone modifications and leukemia. Leukemia. 2009;23(7):1243-51.
59. Popp C, Dean W, Feng S, Cokus SJ, Andrews S, Pellegrini M, et al. Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature. 2010;463(7284):1101-5.
60. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods. 2007;4(8):651-7.
61. Euskirchen GM, Rozowsky JS, Wei C-L, Lee WH, Zhang ZD, Hartman S, et al. Mapping of transcription factor binding regions in mammalian cells by ChIP: comparison of array-and sequencing-based technologies. Genome research. 2007;17(6):898-909.
62. Magi A, Benelli M, Gozzini A, Girolami F, Torricelli F, Brandi ML. Bioinformatics for next generation sequencing data. Genes. 2010;1(2):294-307.
63. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Meth. 2008;5(7):621-8.

64. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. Science. 2008;320(5881):1344-9.
65. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57-63.
66. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669-80.
67. Chiang DY, Getz G, Jaffe DB, O'Kelly MJT, Zhao X, Carter SL, et al. High-resolution mapping of copy-number alterations with massively parallel sequencing. Nat Meth. 2009;6(1):99-103.
68. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061-7.
69. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008;40(6):722-9.
70. Dorschner Michael O, Amendola Laura M, Turner Emily H, Robertson Peggy D, Shirts Brian H, Gallego Carlos J, et al. Actionable, Pathogenic Incidental Findings in 1,000 Participants' Exomes. The American Journal of Human Genetics. 2013;93(4):631-40.
71. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, Martin CL, et al. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med. 2013;15(7):565-74.
72. Burke W, Matheny Antommaria AH, Bennett R, Botkin J, Clayton EW, Henderson GE, et al. Recommendations for returning genomic incidental findings? We need to talk! Genet Med. 2013;15(11):854-9.
73. Shahmirzadi L, Chao EC, Palmaer E, Parra MC, Tang S, Gonzalez KDF. Patient decisions for disclosure of secondary findings among the first 200 individuals undergoing clinical diagnostic exome sequencing. Genet Med. 2014;16(5):395-9.
74. Hudson K, Javitt G, Burke W, Byers P, Committee ASI. ASHG statement on direct-to-consumer genetic testing in the United States. American Journal of Human Genetics. 2007;81(3):635.
75. Consortium IPDG. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. The Lancet. 2011;377(9766):641-9.
76. Klein C, Ziegler A. From GWAS to clinical utility in Parkinson's disease. The Lancet.377(9766):613-4.
77. Skirton H, Goldsmith L, Jackson L, O'Connor A. Direct to consumer genetic testing: a systematic review of position statements, policies and recommendations. Clinical Genetics. 2012;82(3):210-8.

Received: February 03, 2016;
Accepted: July 14, 2016