Published in final edited form as:
J Genet Genomics. 2011 March 20; 38(3): 95–109. doi:10.1016/j.jgg.2011.02.003.

The impact of next-generation sequencing on genomics

Jun Zhang[a,b,*], **Rod Chiodini**[c], **Ahmed Badr**[a], and **Genfa Zhang**[d]

[a] COE for Neurosciences, Department of Anesthesiology, Texas Tech University Health Sciences Center El Paso, TX 79905, USA

[b] Department of Biomedical Sciences, Texas Tech University Health Sciences Center El Paso, TX 79905, USA

[c] Internal Medicine, Texas Tech University Health Sciences Center El Paso, TX 79905, USA

[d] College of Life Sciences, Beijing Normal University, Beijing 100875, China

Abstract

This article reviews basic concepts, general applications, and the potential impact of next-generation sequencing (NGS) technologies on genomics, with particular reference to currently available and possible future platforms and bioinformatics. NGS technologies have demonstrated the capacity to sequence DNA at unprecedented speed, thereby enabling previously unimaginable scientific achievements and novel biological applications. But, the massive data produced by NGS also presents a significant challenge for data storage, analyses, and management solutions. Advanced bioinformatic tools are essential for the successful application of NGS technology. As evidenced throughout this review, NGS technologies will have a striking impact on genomic research and the entire biological field. With its ability to tackle the unsolved challenges unconquered by previous genomic technologies, NGS is likely to unravel the complexity of the human genome in terms of genetic variations, some of which may be confined to susceptible loci for some common human conditions. The impact of NGS technologies on genomics will be far reaching and likely change the field for years to come.

Keywords

Next-generation sequencing; Genomics; Genetic variation; Polymorphism; Targeted sequence enrichment; Bioinformatics

1. Introduction

Since the time DNA was discovered as the code to all biological life on earth, man has sought to unravel its mysteries. If the genetic code could be sequenced or “read”, the origins of life itself may be revealed. Although this thought might not be entirely true, the efforts to date made have certainly revolutionized the biological field.

The “original” sequencing methodology, known as Sanger chemistry, uses specifically labeled nucleotides to read through a DNA template during DNA synthesis. This sequencing technology requires 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. After a

* Corresponding author. Tel: þ1 915 545 6560; fax: 1 915 545 6984. jun.zhang2000@gmail.com (J. Zhang)..

series of technical innovations, the Sanger method has reached the capacity to read through 1000–1200 basepair (bp); however, it still cannot surpass 2 kilo basepair (Kbp) beyond the specific sequencing primer.

In order to sequence longer sections of DNA, a new approach called shotgun sequencing was developed during Human Genome Project (HGP). In this approach, genomic DNA is enzymatically or mechanically broken down into smaller fragments and cloned into sequencing vectors in which cloned DNA fragments can be sequenced individually. The complete sequence of a long DNA fragment can be eventually generated by these methods by alignment and reassembly of sequence fragments based on partial sequence overlaps. Shotgun sequencing was a significant advantage from HGP, and made sequencing the entire human genome possible. The core philosophy of massive parallel sequencing used in next-generation sequencing (NGS) is adapted from shotgun sequencing (Venter et al., 2003; Margulies et al., 2005; Shendure et al., 2005).

New NGS technologies read the DNA templates randomly along the entire genome. This is accomplished by breaking the entire genome into small pieces, then ligating those small pieces of DNA to designated adapters for random read during DNA synthesis (sequencing-by-synthesis). Therefore, NGS technology is often called massively parallel sequencing.

The read length (the actual number of continuous sequenced bases) for NGS is much shorter than that attained by Sanger sequencing. At present, NGS only provides 50–500 continuous basepair reads, which is why sequencing results are defined as short reads. These short reads are a major limitation in current technology; however, developing NGS technologies, such as single-molecule sequencing, may surpass Sanger methodologies and have the potential to read several continuous kilo basepairs (Kbps) (Table 1). Since next-generation technologies currently produce short reads, coverage is a very important issue. Coverage is defined as the number of short reads that overlap each other within a specific genomic region. For example, a 30-fold coverage for *CYP2D6* gene means that every nucleotide within this gene region is represented in at least 30 distinct and overlapping short reads. Sufficient coverage is critical for accurate assembly of the genomic sequence. In addition to the need for adequate coverage, short reads create many sequences that cannot be interpreted or “mapped” to any reference DNA or be accurately assembled. This is simply because some of the short reads are too short and may match with many different regions of the genome and are not unique to any specific region of the sequence. Short-read sequences that can be assembled and matched with a reference sequence are generally called “mappable reads”. NGS is a rapidly evolving technology that is changing on an almost daily basis. The purpose of this review is to highlight these advances and bring the reader up to date on the latest technological achievements in DNA sequencing technologies, particularly as related to genomics.

The term “genomics” was used as the name of the first journal in the field of genomics (McKusick and Ruddle, 1987). Genomics is defined as the systematic study on a whole-genome scale for the identification of genetic contributions to human conditions. Progress in our understanding of many fundamental biological phenomena has accelerated dramatically over the last decade, driven by advances in genomic technologies. New genomic technologies have revolutionized our understanding of many genes or genomic regions involved in the pathogenesis of human diseases (Novelli et al., 2010). Recent advances in high throughput genomic technologies, such as microarray technologies, have resulted in great achievements in genetic linkage, association studies, DNA copy number, and gene expression analysis. There is no doubt that the progress of genomics will eventually lead to the birth of genetic medicine which will propel significant advances and improvements in human health (Gonzalez-Angulo et al., 2010). Candidate-gene approaches were initially

used in the genomic studies with focus on the genes known to be involved in well-defined molecular pathways for targeted human conditions through linkage and association studies. Through candidate-gene studies, certain genetic variants among many genetic loci have been successfully identified for their important attribution to specific human diseases. Following completion of the HGP, a new approach, genome-wide association study (GWAS), was widely applied to genomics. Although several early GWAS studies reported potentially promising results, the majority of GWAS studies were disappointing because of inadequate sample size, limitation of arrays for certain genetic variations, and/or heterogeneity in phenotype (Daly, 2010a). These obstacles may be overcome by new genomic technology, i.e., next-generation sequencing (NGS), also known as massively parallel sequencing or multiplex cyclic sequencing. In the last few years, NGS has emerged as a revolutionary genomic tool. Like other new genomic technologies, NGS techniques will provide radical insights and change the landscape of genomics. Since many genetic variants which contribute to many human conditions are still unknown, unbiased whole-genome sequencing will help to identify these genetic variants, including single nucleotide variants (SNVs) or single nucleotide polymorphisms (SNPs), small insertions and deletions (indels, 1–1000 bp), and structural and genomic variants (>1000 bp) (Daly, 2010b).

Previously, DNA sequencing was performed almost exclusively by the Sanger method, which has excellent accuracy and reasonable read length but very low throughput. Sanger sequencing was used to obtain the first consensus sequence of the human genome in 2001 (Lander et al., 2001; Venter et al., 2001) and the first individual human diploid sequence (J. Craig Venter) in 2007 (Levy et al., 2007). Shortly thereafter, the second complete individual genome (James D. Watson) was sequenced using next-generation technology, which marked the first human genome sequenced with new NGS technology (Wheeler et al., 2008). Since then, several additional diploid human genomes have been sequenced with NGS utilizing a variety of related techniques to rapidly sequence genomes with varying degrees of coverage (Ley et al., 2008; Wang et al., 2008; Ahn et al., 2009; Kim et al., 2009; Yngvadottir et al., 2009; Metzker, 2010). A common strategy for NGS is to use DNA synthesis or ligation process to read through many different DNA templates in parallel (Fuller et al., 2009). Therefore, NGS reads DNA templates in a highly parallel manner to generate massive amounts of sequencing data but, as mentioned above, the read length for each DNA template is relatively short (35–500 bp) compared to traditional Sanger sequencing (1000–1200 bp).

Several NGS methods recently developed allow larger-scale DNA sequencing. The number of large short-read sequences from NGS is increasing at exponential rates. Currently, five NGS platforms are commercially available, including the Roche GS-FLX 454 Genome Sequencer (originally 454 sequencing), the Illumina Genome Analyzer (originally Solexa technology), the ABI SOLiD analyzer, Polonator G.007 and the Helicos HeliScope platforms. These NGS instruments generate different base read lengths, different error rates, and different error profiles relative to Sanger sequencing data and to each other. NGS technologies have increased the speed and throughput capacities of DNA sequencing and, as a result, dramatically reduced overall sequencing costs (Mardis, 2006, 2009; Schuster, 2008; Shendure and Ji, 2008; Ng et al., 2009a; Tucker et al., 2009; Metzker, 2010).

2. Next-generation sequencing platforms

Among the five commercially available platforms, the Roche/454 FLX, the Illumina/Solexa Genome Analyzer, and the Applied Biosystems (ABI) SOLiD Analyzer are currently dominating the market. The other two platforms, the Polonator G.007 and the Helicos HeliScope, have just recently been introduced and are not widely used. Additional platforms from other manufacturers are likely to become available within the next few years and bring

new and exciting technologies, faster sequencing speed, and a more affordable price. Methodologies used by each of the current available NGS systems are discussed below.

2.1. Roche GS-FLX 454 Genome Sequencer

The Roche GS-FLX 454 Genome Sequencer was the first commercial platform introduced in 2004 as the 454 Sequencer. The second complete genome of an individual (James D. Watson) was sequenced with this platform (Wheeler et al., 2008). The 454 Genome Sequencer uses sequencing-by-synthesis technology known as pyrosequencing. The key procedure in this approach is emulsion PCR in which single-stranded DNA binding beads are encapsulated by vigorous vortexing into aqueous micelles containing PCR reactants surrounded by oil for emulsion PCR amplification. During the pyrosequencing process, light emitted from phosphate molecules during nucleotide incorporation is recorded as the polymerase synthesizes the DNA strand. Initially, the 454 Sequencer had a read length of 100 bp but now can produce an average read length of 400 bp. The maximum ~600 bp capacity of 454 systems approaches the halfway of current Sanger sequencing capacities (~1200 bp). At 600 bp, the 454 Sequencer has the longest short reads among all the NGS platforms; and generates ~400–600 Mb of sequence reads per run; critical for some applications such as RNA isoform identification in RNA-seq and *de novo* assembly of microbes in metagenomics (Mocali and Benedetti, 2010). Raw base accuracy reported by Roche is very good (over 99%); however, the reported relatively error-prone raw data sequence, especially associated with insertion-deletions, is a major concern. Low yield of sequence reads could translate into a much higher cost if additional coverage is needed to define a genetic mutation.

2.2. Illumina/Solexa Genome Analyzer

The Illumina/Solexa Genome Analyzer was the second platform to reach market, and currently is the most widely used system. The Illumina platform uses a sequencing-by-synthesis approach in which all four nucleotides are added simultaneously into oligo-primed cluster fragments in flow-cell channels along with DNA polymerase. Bridge amplification extends cluster strands with all four fluorescently labeled nucleotides for sequencing. The Genome Analyzer is widely recognized as the most adaptable and easiest to use sequencing platform. Superior data quality and proper read lengths have made it the system of choice for many genome sequencing projects. To date, the majority of published NGS papers have described methods using the short sequence data produced with the Genome Analyzer. At present, the new Illumina HiSeq 2000 Genome Analyzer is capable of producing single reads of 2 × 100 basepairs (pair-end reads), and generates about 200 giga basepair (Gbp) of short sequences per run. The raw base accuracy is greater than 99.5%.

2.3. ABI SOLiD platform

The ABI SOLiD platform uses a unique sequencing-byligation approach in which it uses an emulsion PCR approach with small magnetic beads to amplify the DNA fragments for parallel sequencing. During SOLiD sequencing, DNA ligation is carried out to link a specific fluorescent labeled 8-mer oligonucleotides for “dinucleotide-encoding”, whose 4th and 5th bases are encoded by specific fluorescence. Each fluorescent marker on a 8-mer identifies a two-base combination, which can be further distinguished with a universal primer offsetting scheme. The primer offsetting scheme allows a universal primer that is offset by one base from the adapter-fragment position to hybridize to DNA templates in five cycle sets permitting the entire fragment to be sequenced and each base position sequenced twice during each cycle. Each ligation step is followed by fluorescence detection and another round of ligation. SOLiD4 analyzer has a read length of up to 50 bp and can produce 80–100 Gbp of mappable sequences per run. The latest model, 5500×l solid system (previously known as SOLiD4hq) can generate over 2.4 billion reads per run with a raw

base accuracy of 99.94% due to its 2-base encoding mechanism. This instrument is unique in that it can process two slides at a time; one slide is receiving reagents while the other is being imaged. The SOLiD4 platform probably provides the best data quality as a result of its sequencing-by-ligation approach but the DNA library preparation procedures prior to sequencing can be tedious and time consuming. The newly marketed EZ-Bead system may provide some resolution to this problem.

2.4. Danaher/Dover/Azco Polonator G.007

The Danaher/Dover/Azco Polonator G.007 is a new platform on the market with emphasis on competitive pricing. The Polonator platform employs a sequencing-by-ligation approach using a randomly arrayed, bead-based, emulsion PCR to amplify DNA fragments for parallel sequencing. The short-read length is 26 bp, and 8–10 Gbp of sequence reads are generated per run, with 92% of the reads mappable. The random bead-based array will likely be replaced with their patented rolonies technology (rolling circle colonies) on an ordered array to increase accuracy and improve read length.

2.5. Helicos HeliScope

The Helicos HeliScope platform is the first single molecular sequencing technology available that uses a highly sensitive fluorescence detection system to directly detect each nucleotide as it is synthesized. 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.

The advantage of single-molecule DNA sequencing technology is its potential to read extremely long sequences and fast sequencing speed, which could translate into a dramatic reduction in overall sequencing cost. As such, advanced single DNA molecule sequencing technology has been defined as the next-NGS technology (Pettersson et al., 2009). However, the basic philosophy of massive parallel sequencing is still the same and the term next-generation sequencing (NGS) will only be used in this review. More detailed technical description of these platforms are available elsewhere (Mardis, 2008a,b, 2009; Schuster, 2008; Shendure and Ji, 2008; Ansorge, 2009; Harismendy et al., 2009; Tucker et al., 2009; Voelkerding et al., 2009; Metzker, 2010; Ng and Kirkness, 2010).

3. Next-generation sequencing platforms under development

Since single DNA molecule sequencing technology can read through DNA templates in real time without amplification, it provides accurate sequencing data with potentially long-reads and efforts have focused recently in this new direction. Several unique single-molecule DNA sequencing technologies are currently under development; however, little information has been made publically available (Gupta, 2008; Xu et al., 2009; Metzker, 2010; Treffer and Deckert, 2010).

3.1. Fluorescence-based single-molecule sequencing

Pacific BioSciences is developing a single-molecule real time (SMRT) DNA sequencing technology. This approach performs single-molecule sequencing by identifying nucleotides which are phospholinked 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.

A similar approach using total internal reflectance fluorescence (TIRF) technology to measure the time-dependent fluorescent signals emitted from each single DNA strand in parallel has been developed by Visigen Biotechnology based on fluorescence resonance energy transfer (FRET). Using nucleotides with fluorescent labeled terminal phosphate to incorporate in a single DNA strain, this approach can rapidly detect subsequent nucleotide incorporation events for single-molecule sequencing. Visigen Biotechnology is teaming up with Life Technology to further improve this technology, in which the polymerase is modified with a quantum dot fluorescent donor molecule that enables a fluorescence resonance energy transfer during nucleotide incorporation events. This FRET-based labeling system would allow the platform to measure a fluorescent signal emitted only from labeled nucleotides that are being incorporated into a DNA strand leading to a reduction in background noise. Furthermore, this FRET-based detection platform does not require continuous laser excitation which extends the polymerase lifespan and results in longer reads.

U.S. Genomics is developing a fluorescence-based single-molecule sequencing platform, in which short-universal probes are hybridized to their complementary DNA fragments and proprietary microfluidics stretch the DNA strand into full contour length. The single molecule is read by laser excitation and DNA fragment reads mapped into the reference sequence based on the unique characteristics of the individual DNA molecules. Genovoxx is also developing a technology called AnyGene for fluorescence-based single-molecule sequencing by monitoring the sequential addition of each single nucleic acid during DNA synthesis.

3.2. Nano-technologies for single-molecule sequencing

Thousands of nano-tunnels on a chip can be used to monitor the movement of a polymerase molecule on a single DNA strand during replication to perform single-molecule DNA sequencing-by-synthesis. Nano-technologies have long been considered a cutting-edge technology for single-molecule DNA sequencing (Iqbal et al., 2007; Branton et al., 2008) and several nanopore sequencing concepts and technologies are currently under development. One concept is based on the observation that when a DNA strand is pulled through a nanopore by an electrical current, each nucleotide base (A, T, C, G) creates a unique pattern in the electrical current. This unique nanopore electrical current fingerprint can be used for nanopore sequencing.

Oxford Nanopore Technologies has developed an exonuclease sequencing technology that combines a protein nanopore bioengineered with a covalent attachment of a cyclodextrin molecule to the inside of its surface with an exonuclease for the sequential identification of DNA bases as the processing enzyme passes through the nanopore. Nabsys is developing a nanopore 6-mer oligonucleotides hybridization mapping technology in which electrically addressable nanopore arrays in a solid state can sequence DNA strands in parallel. This hybridization-assisted nano-pore sequencing (HANS) platform, which combines nano-pore sequencing with sequencing-by-hybridization (SBH) technology, can be used to sequence DNA strands without fluorescence labeling.

Using an ion channel measurement platform, Electronic BioSciences is also exploring the nanopore DNA single strand sequencing platform. BioNanomatrix is developing a nano-Analyzer platform for single-molecule sequencing that uses a small nanochannel fluidic chip for long DNA molecule sequencing. GE research developed a “Closed Complex” where stable complexes are formed between primed single DNA molecules, polymerase, and nucleotides on a solid surface in a microfluidic system in an attempt to develop a single DNA molecule sequencing technology using this “Closed Complex” chemistry. Likewise, in collaboration with Roche, IBM is developing a platform based on their sensors described as

a “DNA transistor” which is potentially capable of recording nucleotide sequence as a single strand of DNA is pulled through a nanopore sensor. LingVitae (Digital Sequencing) is attempting to develop a new single-molecule sequencing platform using their “design polymers” technology to slow down the translocation of DNA strands through the nanopore for better read quality during nano-sequencing. Complete Genomics has developed a proprietary ligase-based DNA sequencing platform (CGA) with their DNA nanoball (DNB) technology called combinatorial probe–anchor ligation (cPAL).

Surface Enhanced Raman spectroscopy (SERS) is a technology based on the observation that a molecule on a corrugated metal surface produces a Raman signal which is up to a billion times more intense than the signal it would emit if deposited on a flat surface. Base4 Innovation is developing a technology to combine SERS with nano-resolution to interpret individual bases in a single strand of DNA with their patented nanostructure arrays. CrackerBio developed an approach which relies on the sequential conversion of photons to electrons for sequencing. They are developing a hybrid platform which can sequence long single DNA molecules on top of a photodiode (a nanowell) embedded in a chip by the fluorescence-based single-molecule sequencing-by-synthesis methods within the nanowell.

3.3. Electronic detection for single-molecule sequencing

Reveo is developing a technology to stretch out DNA molecules on conductive surfaces for electronic base detection. A stretched and immobilized strand of DNA will be read through by multiple nano-knife edge probes. Each nano-knife edge probe specifically recognizes only one nucleotide for single-molecule sequencing. Intelligent Biosystems is also developing a platform using the electronic detection approach which will allow for high speed and high sensitivity single-molecule analysis with decreased background noise.

3.4. Electron microscopy for single-molecule sequencing

Electron microscopy (EM) was the first proposed and attempted approach to sequence DNA molecules before the Sanger sequencing was established and this concept has recently been reevaluated with the emergence of new technologies. Since scanning tunneling microscopy (STM) can reach atomic resolution, STM for single-molecule sequencing is being explored. LightSpeed Genomics is developing a microparticle approach by capturing sequence data with optical detection technology and new sequencing chemistry from a large field of view to reduce the time consuming sample and detector rearrangement. Halcyon Molecular is developing a DNA sequencing technology by atom-by-atom identification and EM analysis. The key advantage of this technology is very long read lengths. ZS Genetics is also developing EM-based technologies for single-molecule DNA sequencing.

3.5. Other approaches for single-molecule sequencing

Ion Torrent developed an entirely new approach to sequencing based on the well-characterized biochemistry that when a nucleotide is incorporated into a strand of DNA by a polymerase, a hydrogen ion is released as a byproduct. They have developed an ion sensor that can detect hydrogen ions and directly convert the chemical information to digital sequence information. In essence, their NGS platform can be defined as the world's smallest solid-state pH meter.

Focusing on resequencing specific sections of the human genome combined with genome-region enrichment, Genizon BioSciences is developing a sequencing-by-hybridization technology based on known reference sequences. Avantome (acquired by Illumina) is also exploring the single-molecule sequencing technologies.

4. Road to the personal genome project

Since the initiation of 1000 genome project the cost of sequencing an individual genome has been rapidly decreasing and will likely reach $1000 per person within a short period of time (von Bubnoff, 2008), making personalized medicine become a possible reality (Lunshof et al., 2010; Mardis, 2006; Harismendy et al., 2009). In genomics, the personal genome era made available by NGS technologies will mark a significant milestone in entire genomic research field in the foreseeable future (Zhang and Dolan, 2010; Holmes et al., 2009).

It is not clear which NGS technology will eventually dominate the genomic research field, but it is almost certain that further reductions in cost, rapid increases in sequencing speed with improved accuracy, and the advantages conferred by these new technologies will assure that NGS will become an essential molecular tool affecting all aspects of the biological sciences. Detailed information of the NGS technologies and platform discussed above is summarized in Table 1.

5. Current strategies for the NGS project

To ensure the correct identification of genetic variants, short-read coverage must be sufficient to ensure the complete and accurate sequence assembly. Currently, at least 30× coverage is recommended in whole-genome scans for rare genetic variants in human genomes, which is a burden on computer resources and cost management. Although the cost of whole-genome sequencing has dropped substantially, the cost remains a major obstacle; whole-genome sequencing of a single individual currently costs approximately $100,000.00.

By targeting specific regions of interest, selective DNA enrichment techniques improve the overall cost and efficiency of NGS (Rehman et al., 2010; Volpi et al., 2010; Bau et al., 2009; Levin et al., 2009; Ng et al., 2009b; Ng and Kirkness, 2010); however, targeted enrichment must maintain uniform coverage, high reproducibility, and no allele bias for any genomic region (Stratton, 2008). Targeted sequencing generally focuses on all protein-coding subsequences (the functional exome), which only requires ~5% as much sequencing compared to that required for the entire human genome (Pussegoda, 2010; Senapathy et al., 2010; Teer and Mullikin, 2010). This strategy currently reduces the overall cost to around $10,000 or less for the sequencing of a single individual. An important consideration to the cost of such experiments is the depth of sequence coverage required to achieve a desired sensitivity and specificity of at least 25-fold nominal sequence coverage.

The most common techniques for targeted sequence enrichment are either microarray-based (Summerer et al., 2009, 2010; Zheng et al., 2009; Igartua et al., 2010) or solution hybrid-based (Gnirke et al., 2009; Tewhey et al., 2009; Bainbridge et al., 2010). Several targeted selection technologies have been marketed and successfully applied in different NGS projects with variable success and may become the tools of choice to lower the burden of time and cost. For example, using targeted selection strategy, the mutations in DHODH from four individuals from three unrelated families with Miller syndrome have been successfully identified (Ng et al., 2010a), illustrating that selective DNA enrichment techniques will dramatically reduce overall cost and accelerate discovery of genetic variants that cause rare and yet to be discovered genetic disorders. Other genetic loci for rare diseases have also been successfully identified through exome sequencing (Bilguvar et al., 2010; Ng et al., 2010a,b; Rios et al., 2010; Walsh et al., 2010), further validating this strategy.

Commercially available products for targeted sequence-enrichment include Agilent's SureSelect and NimbleGen's SeqCap/EZ Exome (both array- and solution-based technologies), RainDance and Illumina's TruSeq (solution-based technology), Febit's

HybSelect and LC Sciences (microarray-based strategy), Qiagen and Fluidigm (PCR-based method) (Table 2).

6. Bioinformatics for NGS data

The parallel short-read strategy of NGS opens many challenges for bioinformatics to interpret the short reads and the genetic variations in human genomes (Wold and Myers, 2008; Yang et al., 2009). The full benefit of NGS will not be achieved until bioinformatics are able to maximally interpret and utilize these short-read sequences, including alignment, assembly, etc (Pop and Salzberg, 2008). Typically, tens or hundreds of Gbp short reads can be generated during each run in any given NGS platform. As a result, the average NGS experiment generates terabytes of raw data, making data analysis and management of data problematic. Given the vast amount of data produced by NGS, developing a massive data storage and management solution and creating informatic tools to effectively analyze data will be essential to the successful application of NGS technology. Further adding to the bioinformatics problems, there are differences among the various NGS platforms in term of data format, length of reads, etc., which results in the need for diversity in bioinformatics including sequence quality scoring, alignment, assembly, and data processing.

The benefits of NGS sequencing will not be fully appreciated until extremely high-performance computing and intensive bioinformatics support is available. The information accrued by NGS may lead to a paradigm shift in the way that genetics and bioinformatics converge. Since NGS technology is in an early stage of development, a variety of software tools are under development and many are available online for NGS data analysis. Their functions fit into several general categories: (1) alignment of reads to a reference sequence; (2) *de novo* assembly, (3) reference-based assembly; (4) base-calling and/or genetic variation detection (such as SNV, Indel); (5) genome annotation, and (6) utilities for data analysis.

7. Alignment and assembly

Despite the sequencing power of NGS, the short-read length strategy creates serious limitations in many biological applications (Wold and Myers, 2008). Efforts to date have focused on overcoming the limitation of short reads for genome-wide analysis, but unfortunately, current available bioinformatics ability and computing power is lagging far behind the needs for NGS sequencing data analysis (McPherson, 2009).

In genomics, reference-based assembly is often performed to map the number of short reads to a human reference genome which creates challenges for the algorithms and computing of alignment. Since repetitive sequences are widely distributed across the entire human genome, some short reads will align equally to multiple chromosomal locations. This is one of the reasons multiple-fold coverage of a given region is required for NGS and why further resequencing with Sanger methodology is often needed to ascertain the genetic variant detected in short reads.

The most important step in NGS data analysis is successful alignment or assembly of short reads to a reference genome (Flicek and Birney, 2009). It is a challenge to efficiently align short reads to a reference genome, especially when developing new algorithms to handle ambiguities or lack of accuracy during the alignment (Li and Homer, 2010).

Based on the mapping quality concept, MAQ (Mapping and Assembly with Quality), a very popular NGS software program, was developed that can efficiently map short reads to a reference genome and derive genotype calls to the consensus sequence with quality scores (Li et al., 2008a). MAQ is one of the first reference guided assembly programs. It is

accurate, efficient, versatile, and user-friendly, and has been successfully applied to several NGS projects (Ng et al., 2010a). ELAND (Efficient Large-Scale Alignment of Nucleotide Databases), another NGS program designed to search DNA files for short DNA reads allowing up to 2 errors per match, has also been successfully used in several NGS projects (Jiang and Wong, 2008; Zheng et al., 2009). Benchmarks comparing ELAND with other popular NGS software, such as MAQ, BLAST (Basic Local Alignment Search Tool), SOAP (Short Oligonucleotide Alignment Program) (Li et al., 2008b), and SeqMap (Jiang and Wong, 2008), etc. (Table 3), generally place ELAND as one of the fastest available programs.

Compared to reference-based assembly with very short-read length sequences, *de novo* assembly is even more challenging. Currently *de novo* assembly with NGS data is generally limited to microbial genome projects (Metagenomics) due to the small bacterial genome size (Chistoserdova, 2010; Wooley et al., 2010).

The primary goal of current algorithms and computing for short-read assembly with NGS technologies is to increase read length (Chaisson et al., 2009). This goal will likely be achieved by the development of single-molecule sequencing technologies. Certain improvements in existing NGS technologies, such as mate-paired short reads, may also make this goal attainable. Individual human genomes (one Asian and one African) have been successfully sequenced and assembled using the Illumina Genome Analyzer (read lengths ranged from 35 to 75 basepairs) with a modified SOAP program, SOAPdenovo (Li et al., 2010).

Available bioinformatic tools for short-read alignment, *de novo* and reference-based assembly for NGS are listed in Table 3. Since many of the programs are open source, additional programming may be needed to modify the program to the needs of a specific NGS project. Some online utility programs, such as EagleView (Huang and Marth, 2008) or LookSeq (Manske and Kwiatkowski, 2009), also provide some additional assistance on NGS data analysis and interpretation (Table 3).

8. Annotation and functional prediction

After successful alignment and assembly of NGS data, the next challenge is to interpret the large number of apparently novel genetic variants (or mutations) present by chance in any single human genome, making it difficult to identify which variants are causal, even when considering only non-synonymous variants. Many novel genetic variants/variations have been discovered for each sequenced genome, resulting in approximately 400 function-altering variants for protein-coding sequences per individual genome. Recognition of functional variants is at the center of the NGS data analysis and bioinformatics. It is challenging to develop software with the ability to distinguish low-frequency alleles descendent from ancient ancestors from *de novo* or extremely rare mutations recently introduced into the population (Nakken et al., 2007; van Oeveren and Janssen, 2009).

Available bioinformatics tools for annotation and functional prediction of NGS data are listed in Table 4. SIFT (Sorting Intolerant From Tolerant) is used to predict whether an amino acid substitution affects protein function based on sequence homology and the physical properties of the amino acid can be applied to find non-synonymous polymorphisms within NGS data (Ng and Henikoff, 2001,2002). By considering the physiochemical variations presented in protein sequence alignment and the property of variations, Multivariate Analysis of Protein Polymorphism (MAPP) can predict the impact of all possible amino acid substitutions on the function of the protein (Stone and Sidow, 2005). Predictor of human Deleterious Single Nucleotide Polymorphisms (PhD-SNP) is an optimized program to predict if a given single point protein mutation can be classified as

disease-related or as neutral polymorphism based on protein sequence and profile information (Capriotti et al., 2006). Polymorphism Phenotyping (PolyPhen) and updated PolyPhen-2 are tools which predict the possible impact of an amino acid substitution on the structure and function of a specific protein using straightforward comparative physical methods (Sunyaev et al., 2001;Ramensky et al., 2002).

Variation detection software, which includes screening genomes for structural and single nucleotide variants and the differences between genomes (Li et al., 2009; Medvedev et al., 2009; Thusberg and Vihinen, 2009), are generally integrated with alignment and assembly processes and are listed in Table 4.

9. End-user packages

End-user software packages which provide a user-friendly interface, easy to use data input and output formats, and integrates multiple computing programs into one software package, may be the best solution for most biomedical researchers. Based on our experience, among available end-user packages, Genomic Workbench from CLC Bio appears to be the most widely used. NextGENe from SoftGenetics is excellent for candidate-gene resequencing projects, but it cannot handle very large datasets and may not be suitable for large genome sequencing projects. SeqMan Ngen from DNASTAR is under development but currently unavailable. Although commercial end-user packages tend to carry a hefty price, some are available free online as detailed in Table 5.

As previously mentioned, the fact that high-performance computing and intensive bioinformatic support is needed for NGS, it is difficult for many research laboratories to successfully conduct NGS projects due to the high level of information technology support required. A possible solution is cloud computing. In cloud computing, a user can use a virtual operating system (or “cloud”) to process data on a computer cluster for high parallel tasks (Editorial, 2010). CrossBow is the first cloud computing software capable of performing alignment and single nucleotide polymorphism analysis on multiple whole-human datasets (Langmead et al., 2009). CloudBurst is another new parallel read-mapping cloud algorithm optimized for mapping NGS data to a human reference genome, SNP discovery, genotyping and personal genomics (Schatz, 2009). Data generated on Applied Bio-systems' SoLiD platform uses a two colored system which makes it unsuitable for analysis by many available software packages. The Bioscope package, developed by ABI, is devoted to their SoLiD data and can be used as a single software package or for cloud computing, likewise, CASAVA package developed by Illumina is utilized for Genome Analyzer data. Available NGS cloud computing technologies are listed in Table 5. GenomeQuest, Complete Genomics and Geospiza/GeneSifter provide online customer oriented NGS data analysis services, which is a little different from cloud computing by definition.

10. Conclusion

There are two proposed paradigms on the inherited basis of complex genetic traits: “common disease-common variants hypothesis” which consists of many common alleles of small effect, and “common disease-multiple rare variants hypothesis” which consists of few rare alleles of large effect (Manolio et al., 2009; Schork et al., 2009). Both types of genetic loci likely exist; however, the “common disease-common variants hypothesis” is the theoretical framework for GWAS (Manolio et al., 2009; Schork et al., 2009).

After a great deal of effort, much of the data generated from GWAS has been disappointing. Among successful GWAS studies, most variants identified confer only a small proportion of heritability, indicating that GWAS based on the “common disease-common variants

hypothesis" is not very effective in identifying genetic variants for complex traits, and common genetic variability is unlikely to explain the entire genetic predisposition to disease (Singleton et al., 2010). Results also suggest that rare variants missed by GWAS may account for the "missing" heritability. Such rare variants may have a large effect as genetic risk factors for complex genetic diseases (Frazer et al., 2009; Manolio et al., 2009; Schork et al., 2009; Tsuji, 2010).

Remaining challenges will be to define the genetic basis of "missing" heritability (Manolio et al., 2009). NGS technologies will certainly enable us to identify all the causative variants including "rare variants" within individual human subjects. It is anticipated that whole-genome sequencing (or exome sequencing) will make significant contributions to our understanding of the genetic etiologies that contribute to complex human disease, as well as the genetic basis of genomics.

The rapidly changing and ever evolving field of gene sequencing is shifting the way the medical profession views many diseases by documenting the contributions of the human genomic variation in medicine. Likewise, advances in NGS will redefine the field of genomics. Since this review covers a wide spectrum of current development in NGS, detailed information is still scarce or unavailable due to heated competition in the current race in this emerging field. Certain references from many new NGS technologies are not directly cited in the text, however if interested, reader can always find them through websites provided in the designated tables.

Acknowledgments

This work was supported by NINDS/NIH (JZ), Coldwell Foundation (JZ) and TTUHSC (JZ). We are also grateful to Sean Connery in manuscript editing.

Abbreviations

NGS	next-generation sequencing
bp	basepair
Kbp	kilo basepair
HGP	Human Genome Project
GWAS	genome-wide association study
SNVs	single nucleotide variants
SNPs	single nucleotide polymorphisms
indels	small insertions and deletions
Gbp	giga basepair
rolonies	rolling circle colonies
SMRT	single-molecule real time
TIRF	total internal reflectance fluorescence
FRET	fluorescence resonance energy transfer
HANS	hybridization-assisted nanopore sequencing
SBH	sequencing-by-hybridization
DNB	DNA nanoBall

cPAL	combinatorial probe-anchor ligation
SERS	Surface Enhanced Raman Spectroscopy
EM	electron microscopy
STM	scanning tunneling microscopy
MAQ	Mapping and Assembly with Quality
ELAND	Efficient Large-Scale Alignment of Nucleotide Databases
BLAST	Basic Local Alignment Search Tool
SOAP	Short Oligonucleotide Alignment Program
SIFT	sorting intolerant from tolerant
MAPP	multivariate analysis of protein polymorphism
PhD-SNP	predictor of human deleterious single nucleotide polymorphisms
PolyPhe	polymorphism phenotyping

References

Ahn SM, Kim TH, Lee S, Kim D, Ghang H, Kim DS, Kim BC, Kim SY, Kim WY, Kim C, Park D, Lee YS, Kim S, Reja R, Jho S, Kim CG, Cha JY, Kim KH, Lee B, Bhak J, Kim SJ. The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Res. 2009; 19:1622–1629. [PubMed: 19470904]

Ansorge WJ. Next-generation DNA sequencing techniques. Nat. Biotechnol. 2009; 25:195–203.

Bainbridge MN, Wang M, Burgess DL, Kovar C, Rodesch MJ, D'Ascenzo M, Kitzman J, Wu YQ, Newsham I, Richmond TA, Jeddeloh JA, Muzny D, Albert TJ, Gibbs RA. Whole exome capture in solution with 3 Gbp of data. Genome Biol. 2010; 11:R62. [PubMed: 20565776]

Bau S, Schracke N, Kranzle M, Wu H, Stahler PF, Hoheisel JD, Beier M, Summerer D. Targeted next-generation sequencing by specific capture of multiple genomic loci using low-volume microfluidic DNA arrays. Anal. Bioanal. Chem. 2009; 393:171–175. [PubMed: 18958448]

Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S, Kaymakcalan H, Barak T, Bakircioglu M, Yasuno K, Ho W, Sanders S, Zhu Y, Yilmaz S, Dincer A, Johnson MH, Bronen RA, Kocer N, Per H, Mane S, Pamir MN, Yalcinkaya C, Kumandas S, Topcu M, Ozmen M, Sestan N, Lifton RP, State MW, Gunel M. Whole-exome sequencing identifies recessive *WDR62* mutations in severe brain malformations. Nature. 2010; 467:207–210. [PubMed: 20729831]

Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, Di Ventra M, Garaj S, Hibbs A, Huang X, Jovanovich SB, Krstic PS, Lindsay S, Ling XS, Mastrangelo CH, Meller A, Oliver JS, Pershin YV, Ramsey JM, Riehn R, Soni GV, Tabard-Cossa V, Wanunu M, Wiggin M, Schloss JA. The potential and challenges of nanopore sequencing. Nat. Biotechnol. 2008; 26:1146–1153. [PubMed: 18846088]

Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006; 22:2729–2734. [PubMed: 16895930]

Chaisson MJ, Brinza D, Pevzner PA. De novo fragment assembly with short mate-paired reads: does the read length matter? Genome Res. 2009; 19:336–346. [PubMed: 19056694]

Chistoserdova L. Recent progress and new challenges in metagenomics for biotechnology. Biotechnol. Lett. 2010; 32:1351–1359. [PubMed: 20495950]

Daly AK. Genome-wide association studies in pharmacogenomics. Nat. Rev. Genet. 2010a; 11:241–246. [PubMed: 20300088]

Daly AK. Pharmacogenetics and human genetic polymorphisms. Biochem. J. 2010b; 429:435–449. [PubMed: 20626352]

Editorial Gathering clouds and a sequencing storm: why cloud computing could broaden community access to next-generation sequencing. Nat. Biotechnol. 2010; 28:1. [PubMed: 20062015]

Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. Nat. Methods. 2009; 6:S6–S12. [PubMed: 19844229]

Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat. Rev. Genet. 2009; 10:241–251. [PubMed: 19293820]

Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV. The challenges of sequencing by synthesis. Nat. Biotechnol. 2009; 27:1013–1023. [PubMed: 19898456]

Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T, Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat. Biotechnol. 2009; 27:182–189. [PubMed: 19182786]

Gonzalez-Angulo AM, Hennessy BT, Mills GB. Future of personalized medicine in oncology: a systems biology approach. J. Clin. Oncol. 2010; 28:2777–2783. [PubMed: 20406928]

Gupta PK. Single-molecule DNA sequencing technologies for future genomics research. Trends Biotechnol. 2008; 26:602–611. [PubMed: 18722683]

Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009; 10:R32. [PubMed: 19327155]

Holmes MV, Shah T, Vickery C, Smeeth L, Hingorani AD, Casas JP. Fulfilling the promise of personalized medicine? Systematic review and field synopsis of pharmacogenetic studies. PLoS ONE. 2009; 4:e7960. [PubMed: 19956635]

Huang W, Marth G. EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res. 2008; 18:1538–1543. [PubMed: 18550804]

Igartua C, Turner EH, Ng SB, Hodges E, Hannon GJ, Bhattacharjee A, Rieder MJ, Nickerson DA, Shendure J. Targeted enrichment of specific regions in the human genome by array hybridization. Curr. Protoc. Hum. Genet. 2010; 66:18.3.1–18.3.14.

Iqbal SM, Akin D, Bashir R. Solid-state nanopore channels with DNA selectivity. Nat. Nanotechnol. 2007; 2:243–248. [PubMed: 18654270]

Jiang H, Wong WH. SeqMap: mapping massive amount of oligo-nucleotides to the genome. Bioinformatics. 2008; 24:2395–2396. [PubMed: 18697769]

Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Church GM, Lee C, Kingsmore SF, Seo JS. A highly annotated whole-genome sequence of a Korean individual. Nature. 2009; 460:1011–1015. [PubMed: 19587683]

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J,

Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ. Initial sequencing and analysis of the human genome. Nature. 2001; 409:860–921. [PubMed: 11237011]

Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. Genome Biol. 2009; 10:R134. [PubMed: 19930550]

Levin JZ, Berger MF, Adiconis X, Rogov P, Melnikov A, Fennell T, Nusbaum C, Garraway LA, Gnirke A. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. Genome Biol. 2009; 10:R115. [PubMed: 19835606]

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC. The diploid genome sequence of an individual human. PLoS Biol. 2007; 5:e254. [PubMed: 17803354]

Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature. 2008; 456:66–72. [PubMed: 18987736]

Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010; 11:473–483. [PubMed: 20460430]

Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008a; 18:1851–1858. [PubMed: 18714091]

Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. Bioinformatics. 2008b; 24:713–714. [PubMed: 18227114]

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009; 19:1124–1132. [PubMed: 19420381]

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Yang H, Wang J. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20:265–272. [PubMed: 20019144]

Lunshof JE, Bobe J, Aach J, Angrist M, Thakuria JV, Vorhaus DB, Hoehe MR, Church GM. Personal genomes in progress: from the human genome project to the personal genome project. Dialogues Clin. Neurosci. 2010; 12:47–60. [PubMed: 20373666]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

Manske HM, Kwiatkowski DP. LookSeq: a browser-based viewer for deep sequencing data. Genome Res. 2009; 19:2125–2132. [PubMed: 19679872]

Mardis ER. Anticipating the 1,000 dollar genome. Genome Biol. 2006; 7:112. [PubMed: 17224040]

Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet. 2008a; 24:133–141. [PubMed: 18262675]

Mardis ER. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 2008b; 9:387–402. [PubMed: 18576944]

Mardis ER. New strategies and emerging technologies for massively parallel sequencing: applications in medical research. Genome Med. 2009; 1:40. [PubMed: 19435481]

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

McKusick VA, Ruddle FH. Toward a complete map of the human genome. Genomics. 1987; 1:103–106. [PubMed: 3480265]

McPherson JD. Next-generation gap. Nat. Methods. 2009; 6:S2–S5. [PubMed: 19844227]

Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat. Methods. 2009; 6:S13, S20. [PubMed: 19844226]

Metzker ML. Sequencing technologies – the next generation. Nat. Rev. Genet. 2010; 11:31–46. [PubMed: 19997069]

Mocali S, Benedetti A. Exploring research frontiers in microbiology: the challenge of metagenomics in soil microbiology. Res. Microbiol. 2010; 161:497–505. [PubMed: 20452420]

Nakken S, Alseth I, Rognes T. Computational prediction of the effects of non-synonymous single nucleotide polymorphisms in human DNA repair genes. Neuroscience. 2007; 145:1273–1279. [PubMed: 17055652]

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001; 11:863–874. [PubMed: 11337480]

Ng PC, Henikoff S. Accounting for human polymorphisms predicted to affect protein function. Genome Res. 2002; 12:436–446. [PubMed: 11875032]

Ng PC, Kirkness EF. Whole genome sequencing. Methods Mol. Biol. 2010; 628:215–226. [PubMed: 20238084]

Ng PC, Murray SS, Levy S, Venter JC. An agenda for personalized medicine. Nature. 2009a; 461:724–726. [PubMed: 19812653]

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ. Exome sequencing identifies the cause of a mendelian disorder. Nat. Genet. 2010a; 42:30–35. [PubMed: 19915526]

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC, Lee C, Turner EH, Smith JD, Rieder MJ, Yoshiura K, Matsumoto N, Ohta T, Niikawa N, Nickerson DA, Bamshad MJ, Shendure J. Exome sequencing identifies *MLL2* mutations as a cause of Kabuki syndrome. Nat. Genet. 2010b; 42:790–793. [PubMed: 20711175]

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009b; 461:272–276. [PubMed: 19684571]

Novelli G, Predazzi IM, Mango R, Romeo F, Mehta JL. Role of genomics in cardiovascular medicine. World J. Cardiol. 2010; 2:428–436. [PubMed: 21191544]

Pettersson E, Lundeberg J, Ahmadian A. Generations of sequencing technologies. Genomics. 2009; 93:105–111. [PubMed: 18992322]

Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008; 24:142–149. [PubMed: 18262676]

Pussegoda KA. Exome sequencing: locating causative genes in rare disorders. Clin. Genet. 2010; 78:32–33. [PubMed: 20597920]

Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002; 30:3894–3900. [PubMed: 12202775]

Rehman AU, Morell RJ, Belyantseva IA, Khan SY, Boger ET, Shahzad M, Ahmed ZM, Riazuddin S, Khan SN, Friedman TB. Targeted capture and next-generation sequencing identifies C9orf75, encoding taperin, as the mutated gene in nonsyndromic deafness DFNB79. Am. J. Hum. Genet. 2010; 86:378–388. [PubMed: 20170899]

Rios J, Stein E, Shendure J, Hobbs HH, Cohen JC. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. Hum. Mol. Genet. 2010; 19:4313–4318. [PubMed: 20719861]

Schatz MC. CloudBurst: highly sensitive read mapping with Map-Reduce. Bioinformatics. 2009; 25:1363–1369. [PubMed: 19357099]

Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. Curr. Opin. Genet. Dev. 2009; 19:212–219. [PubMed: 19481926]

Schuster SC. Next-generation sequencing transforms today's biology. Nat. Methods. 2008; 5:16–18. [PubMed: 18165802]

Senapathy P, Bhasi A, Mattox J, Dhandapany PS, Sadayappan S. Targeted genome-wide enrichment of functional regions. PLoS ONE. 2010; 5:e11138. [PubMed: 20585402]

Shendure J, Ji H. Next-generation DNA sequencing. Nat. Biotechnol. 2008; 26:1135–1145. [PubMed: 18846087]

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005; 309:1728–1732. [PubMed: 16081699]

Singleton AB, Hardy J, Traynor BJ, Houlden H. Towards a complete resolution of the genetic architecture of disease. Trends Genet. 2010; 26:438–442. [PubMed: 20813421]

Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005; 15:978–986. [PubMed: 15965030]

Stratton M. Genome resequencing and genetic variation. Nat. Biotechnol. 2008; 26:65–66. [PubMed: 18183021]

Summerer D, Schracke N, Wu H, Cheng Y, Bau S, Stahler CF, Stahler PF, Beier M. Targeted high throughput sequencing of a cancer-related exome subset by specific sequence capture with a fully automated microarray platform. Genomics. 2010; 95:241–246. [PubMed: 20138981]

Summerer D, Wu H, Haase B, Cheng Y, Schracke N, Stahler CF, Chee MS, Stahler PF, Beier M. Microarray-based multicycle-enrichment of genomic subsets for targeted next-generation sequencing. Genome Res. 2009; 19:1616–1621. [PubMed: 19638418]

Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum. Mol. Genet. 2001; 10:591–597. [PubMed: 11230178]

Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. Hum. Mol. Genet. 2010; 19(R2):R145–R151. [PubMed: 20705737]

Tewhey R, Nakano M, Wang X, Pabon-Pena C, Novak B, Giuffre A, Lin E, Happe S, Roberts DN, LeProust EM, Topol EJ, Harismendy O, Frazer KA. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol. 2009; 10:R116. [PubMed: 19835619]

Thusberg J, Vihinen M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum. Mutat. 2009; 30:703–714. [PubMed: 19267389]

Treffer R, Deckert V. Recent advances in single-molecule sequencing. Curr. Opin. Biotechnol. 2010; 21:4–11. [PubMed: 20202812]

Tsuji S. Genetics of neurodegenerative diseases: insights from high-throughput resequencing. Hum. Mol. Genet. 2010; 19:R65–R70. [PubMed: 20413655]

Tucker T, Marra M, Friedman JM. Massively parallel sequencing: the next big thing in genetic medicine. Am. J. Hum. Genet. 2009; 85:142–154. [PubMed: 19679224]

van Oeveren J, Janssen A. Mining SNPs from DNA sequence data; computational approaches to SNP discovery and analysis. Methods Mol. Biol. 2009; 578:73–91. [PubMed: 19768587]

Venter JC, Levy S, Stockwell T, Remington K, Halpern A. Massive parallelism, randomness and genomic advances. Nat. Genet. 2003; 33(Suppl.):219–227. [PubMed: 12610531]

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science. 2001; 291:1304–1351. [PubMed: 11181995]

Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin. Chem. 2009; 55:641–658. [PubMed: 19246620]

Volpi L, Roversi G, Colombo EA, Leijsten N, Concolino D, Calabria A, Mencarelli MA, Fimiani M, Macciardi F, Pfundt R, Schoenmakers EF, Larizza L. Targeted next-generation sequencing appoints c16orf57 as clericuzio-type poikiloderma with neutropenia gene. Am. J. Hum. Genet. 2010; 86:72–76. [PubMed: 20004881]

von Bubnoff A. Next-generation sequencing: the race is on. Cell. 2008; 132:721–723. [PubMed: 18329356]

Walsh T, Shahin H, Elkan-Miller T, Lee MK, Thornton AM, Roeb W, Abu Rayyan A, Loulus S, Avraham KB, King MC, Kanaan M. Whole exome sequencing and homozygosity mapping identify mutation in the cell polarity protein GPSM2 as the cause of nonsyndromic hearing loss DFNB82. Am. J. Hum. Genet. 2010; 87:90–94.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Yang H. The diploid genome sequence of an Asian individual. Nature. 2008; 456:60–65. [PubMed: 18987735]

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM. The complete genome of an individual by massively parallel DNA sequencing. Nature. 2008; 452:872–876. [PubMed: 18421352]

Wold B, Myers RM. Sequence census methods for functional genomics. Nat. Methods. 2008; 5:19–21. [PubMed: 18165803]

Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. PLoS Comput. Biol. 2010; 6:2, e1000667.

Xu M, Fujita D, Hanagata N. Perspectives and challenges of emerging single-molecule DNA sequencing technologies. Small. 2009; 5:2638–2649. [PubMed: 19904762]

Yang MQ, Athey BD, Arabnia HR, Sung AH, Liu Q, Yang JY, Mao J, Deng Y. High-throughput next-generation sequencing technologies foster new cutting-edge computing techniques in bioinformatics. BMC Genomics. 2009; 10(Suppl. 1):I1. [PubMed: 19594867]

Yngvadottir B, Macarthur DG, Jin H, Tyler-Smith C. The promise and reality of personal genomics. Genome Biol. 2009; 10:237. [PubMed: 19723346]

Zhang W, Dolan ME. Impact of the 1000 genomes project on the next wave of pharmacogenomic discovery. Pharmacogenomics. 2010; 11:249–256. [PubMed: 20136363]

Zheng J, Moorhead M, Weng L, Siddiqui F, Carlton VE, Ireland JS, Lee L, Peterson J, Wilkins J, Lin S, Kan Z, Seshagiri S, Davis RW, Faham M. High-throughput, high-accuracy array-based resequencing. Proc. Natl. Acad. Sci. USA. 2009; 106:6712–6717. [PubMed: 19342489]

Table 1

The platforms and the detailed information for the NGS technologies.

Technology	Amplification	Read length	Throughput	Sequence by synthesis
Currently available				
Roche/GS-FLX Titanium	Emulsion PCR	400–600 bp	500 Mbp/run	Pyrosequencing
Illumina/HiSeq 2000, HiScan	Bridge PCR (Cluster PCR)	2 × 100 bp	200 Gbp/run	Reversible terminators
ABI/SOLiD 5500xl	Emulsion PCR	50–100 bp	>100 Gbp/run	Sequencing-by-ligation (octamers)
Polonator/G.007	Emulsion PCR	26 bp	8–10 Gbp/run	Sequencing-by-ligation (monomers)
Helicos/Heliscope	No	35 (25–55) bp	21–37 Gbp/run	True single-molecule sequencing (tSMS)
In development				
Pacific BioSciences/RS	No	1000 bp	N/A	Single-molecule real time (SMRT)
Visigen Biotechnologies	No	>100 Kbp	N/A	Base-specific FRET
U.S. Genomics	No	N/A	N/A	Single-molecule mapping
Genovoxx	No	N/A	N/A	Single-molecule sequencing by synthesis
Oxford Nanopore Technologies	No	35 bp	N/A	Nanopores/exonuclease-coupled
NABsys	No	N/A	N/A	Nanopores
Electronic BioSciences	No	N/A	N/A	Nanopores
BioNanomatrix/nanoAnalyzer	No	400 Kbp	N/A	Nanochannel arrays
GE Global Research	No	N/A	N/A	Closed Complex/nanoparticle
IBM	No	N/A	N/A	Nanopores
LingVitae	No	N/A	N/A	Nanopores
Complete Genomics	No	70 bp	N/A	DNA nanoball arrays
base4innovation	No	N/A	N/A	Nanostructure arrays
CrackerBio	No	N/A	N/A	Nanowells
Reveo	No	N/A	N/A	Nano-knife edge
Intelligent BioSystems	No	N/A	N/A	Electronics
LightSpeed Genomics	No	N/A	N/A	Direct-read Sequencing by EM
Halcyon Molecular	No	N/A	N/A	Direct-read Sequencing by EM
ZS Genetics	No	N/A	N/A	Direct-read Sequencing by TEM
Ion Torrent/PostLight	No	N/A	N/A	Semiconductor-based pH sequencing
Genizon BioSciences/CGA	No	N/A	N/A	Sequencing-by-hybridization

Table 2

The targeted sequence-enrichment technologies for NGS.

Technology	Approach	Platform	Website
Agilent/SureSelect	Array- and solution-based	Illumina/Roche/ABI	http://www.chem.agilent.com/
RainDance	Microdroplet-based	Illumina/Roche/ABI	http://www.raindancetechnologies.com/
NimbleGen/SeqCap/EZ Exome	Array- and solution-based	Illumina/Roche/ABI	http://www.nimblegen.com/products/seqcap/index.html
Febit/HybSelect	Microarray-based	Illumina/Roche/ABI	http://www.febit.com/microarray-sequencing/index.cfm
Fluidigm	PCR-based	Illumina/Roche/ABI	http://www.fluidigm.com/targeted-resequencing.html
Mycroarray/Myselect	Solution-based	Illumina/Roche/ABI	http://www.mycroarray.com/products/myselect.html
LC Sciences	Microarray-based	Illumina/Roche/ABI	http://www.lcsciences.com/applications/genomics/
Qiagen/SeqTarget	Long-range PCR-based	Illumina/Roche/ABI	http://www.qiagen.com/products/seqtargetsystem.aspx
Illumina/TruSeq	Solution-based	Illumina/Roche/ABI	http://www.illumina.com/applications.ilmn

Table 3

The alignment, assembly and utility bioinformatic tools for NGS.

Program	Function	Platform	Website
De novo assembly			
Abyss	Alignment/assembly	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/abyss
ALLPATHS	Alignment/assembly	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd
AMOScmp	Alignment/assembly	Roche	http://sourceforge.net/projects/amos/files/
ARACHNE	Alignment/assembly	Roche	http://www.broadinstitute.org/science/programs/genome-biology/crd
CAP3	Alignment/assembly	Roche	http://pbil.univ-lyon1.fr/cap3.php
consensus/Seq-Cons	Alignment/assembly	Roche	http://www.seqan.de/downloads/projects.html
Curtain	Alignment/assembly	Illumina/Roche/ABI	http://code.google.com/p/curtain/
Edena	Alignment/assembly	Illumina	http://www.genomic.ch/edena
Euler-SR	Alignment/assembly	Illumina/Roche	http://euler-assembler.ucsd.edu/portal/?q=team
FuzzyPath	Alignment/assembly	Illumina/Roche	ftp://ftp.sanger.ac.uk/pub/zn1/fuzzypath/fuzzypath_v3.0.tgz
IDBA	Alignment/assembly	Illumina	http://www.cs.hku.hk/~alse/idba/
MIRA/MIRA3	Alignment/assembly	Illumina/Roche	http://chevreux.org/projects_mira.html
Newbler	Alignment/assembly	Roche	roche-applied-science.com/
Phrap	Alignment/assembly	Illumina/Roche	http://www.phrap.org/consed/consed.html#howToGet
RGA	Alignment/assembly	Illumina	http://rga.cgrb.oregonstate.edu/
QSRA	Alignment/assembly	Illumina	http://qsra.cgrb.oregonstate.edu/
SHARCGS	Alignment/assembly	Illumina	http://sharcgs.molgen.mpg.de/
SHORTY	Alignment/assembly	ABI	http://www.cs.sunysb.edu/~skiena/shorty/
SHRAP	Alignment/assembly	Roche	By request
SOAPdenovo	Alignment/assembly	Illumina	http://soap.genomics.org.cn
SOPRA	Alignment/assembly	Illumina/ABI	http://www.physics.rutgers.edu/%7Eanirvans/SOPRA/
SR-ASM	Alignment/assembly	Roche	http://bioserver.cs.put.poznan.pl/sr-asm-short-reads-assembly-algorithm
SSAKE	Alignment/assembly	Illumina/Roche	http://www.bcgsc.ca/platform/bioinfo/software/ssake
Taipan	Alignment/assembly	Illumina	http://sourceforge.net/projects/taipan/files/
VCAKE	Alignment/assembly	Illumina/Roche	http://sourceforge.net/projects/vcake
Velvet	Alignment/assembly	Illumina/Roche/ABI	http://www.ebi.ac.uk/%7Ezerbino/velvet
Reference-based assembly			
BFAST	Alignment/assembly	Illumina/ABI	http://sourceforge.net/apps/mediawiki/bfast/index.php?title=Main_Page
Bowtie	Alignment/assembly	Illumina/Roche/ABI	http://bowtie-bio.sourceforge.net
BWA	Alignment/assembly	Illumina/ABI	http://bio-bwa.sourceforge.net/bwa.shtml
CoronaLite	Alignment/assembly	ABI	http://solidsoftwaretools.com/gf/project/corona/
CABOG	Alignment/assembly	Roche/ABI	http://wgs-assembler.sf.net
ELAND/ELAND2	Alignment/assembly	Illumina/ABI	http://www.illumina.com/
EULER	Alignment/assembly	Illumina	http://euler-assembler.ucsd.edu/portal/
Exonerate	Alignment/assembly	Roche	http://www.ebi.ac.uk/~guy/exonerate
EMBF	Alignment/assembly	Illumina	http://www.biomedcentral.com/1471-2105/10?issue=S1
GenomeMapper	Alignment/assembly	Illumina	http://1001genomes.org/downloads/genomemapper.html
GMAP	Alignment/assembly	Illumina	http://www.gene.com/share/gmap

Program	Function	Platform	Website
gnumap	Alignment/assembly	Illumina	http://dna.cs.byu.edu/gnumap/
ICON	Alignment/assembly	Illumina	http://icorn.sourceforge.net/
Karma	Alignment/assembly	Illumina/ABI	http://www.sph.umich.edu/csg/pha/karma/
LAST	Alignment/assembly	Illumina	http://last.cbrc.jp/
LOCAS	Alignment/assembly	Illumina	http://www-ab.informatik.uni-tuebingen.de/software/locas
Mapreads	Alignment/assembly	ABI	http://solidsoftwaretools.com/gf/project/mapreads/
MAQ	Alignment/assembly	Illumina/ABI	http://maq.sourceforge.net
MOM	Alignment/assembly	Illumina	http://mom.csbc.vcu.edu/
Mosaik	Alignment/assembly	Illumina/Roche/ABI	http://bioinformatics.bc.edu/marthlab/Mosaik
mrFAST/mrsFAST	Alignment/assembly	Illumina	http://mrfast.sourceforge.net/
MUMer	Alignment/assembly	ABI	http://mummer.sourceforge.net/
nexalign	Alignment/assembly	Illumina	http://genome.gsc.riken.jp/osc/english/dataresource/
Novocraft	Alignment/assembly	Illumina	http://www.novocraft.com/
PerM	Alignment/assembly	Illumina/ABI	http://code.google.com/p/perm/
RazerS	Alignment/assembly	Illumina/ABI	http://www.seqan.de/projects/razers.html
RMAP	Alignment/assembly	Illumina	http://rulai.cshl.edu/rmap
segemehl	Alignment/assembly	Illumina/Roche	http://www.bioinf.uni-leipzig.de/Software/segemehl/
SeqCons	Alignment/assembly	Roche	http://www.seqan.de/projects/seqcons.html
SeqMap	Alignment/assembly	Illumina	http://biogibbs.stanford.edu/*jiangh/SeqMap/
SHRiMP	Alignment/assembly	Illumina/Roche/ABI	http://compbio.cs.toronto.edu/shrimp
Slider/SliderII	Alignment/assembly	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/slider
SOCS	Alignment/assembly	ABI	http://solidsoftwaretools.com/gf/project/socs/
SOAP/SOAP2	Alignment/assembly	Illumina/ABI	http://soap.genomics.org.cn
SSAHA/SSAHA2	Alignment/assembly	Illumina/Roche	http://www.sanger.ac.uk/Software/analysis/SSAHA2
Stampy	Alignment/assembly	Illumina	http://www.well.ox.ac.uk/~marting/
SXOligoSearch	Alignment/assembly	Illumina	http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
SHORE	Alignment/assembly	Illumina	http://1001genomes.org/downloads/shore.html
Vmatch	Alignment/assembly	Illumina	http://www.vmatch.de/
Diagnostics/utilities			
Artemis/ACT	Visualization tool	Illumina/Roche	http://www.sanger.ac.uk/resources/software/artemis/
CASHX	Pipeline	Illumina	http://seqanswers.com/wiki/CASHX
Consed	Visualization tool	Illumina/Roche	http://www.genome.washington.edu/consed/consed.html
EagleView	Visualization tool	Illumina/Roche	http://bioinformatics.bc.edu/marthlab/EagleView
FastQC	Quality assessment	Illumina/ABI	http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/
Gambit	Visualization tool	Illumina/Roche	http://bioinformatics.bc.edu/marthlab/Gambit
Goby	Data management	Illumina/Roche/ABI	http://campagnelab.org/software/goby/
G-SQZ	Data management	Illumina/ABI	http://public.tgen.org/sqz
Hawkeye	Visualization tool	Illumina/Roche	http://amos.sourceforge.net/hawkeye
Hybrid-SHREC	Error Correction	Illumina/Roche/ABI	http://www.cs.helsinki.fi/u/lmsalmel/hybrid-shrec/
IGV	Visualization tool	Illumina	http://www.broadinstitute.org/igv/?q=home
LookSeq	Visualization tool	Illumina/Roche	http://lookseq.sourceforge.net
MagicViewer	Visualization tool	Illumina	http://bioinformatics.zj.cn/magicviewer/

Program	Function	Platform	Website
MapView	Visualization tool	Illumina	http://evolution.sysu.edu.cn/mapview/
NGSView	Visualization tool	Illumina/ABI	http://ngsview.sourceforge.net
PIQA	Quality assessment	Illumina	http://bioinfo.uh.edu/PIQA
Reconciliation	Assembly pipeline	Illumina	http://www.genome.umd.edu/software.htm
RefCov	Sequence coverage	Illumina/Roche	http://genome.wustl.edu/tools/cancer-genomics
SAM Tools	Utilities	Illumina/Roche	http://sourceforge.net/projects/samtools/files/
Savant	Visualization tool	Illumina/Roche	http://compbio.cs.toronto.edu/savant/
ShortRead	Quality assessment	Illumina/Roche	http://bioconductor.org/packages/2.6/bioc/html/ShortRead.html
SHREC	Error Correction	Illumina/Roche	http://www.informatik.uni-kiel.de/jasc/Shrec/
Staden Tools (GAP5)	Pipeline	Illumina/Roche	http://sourceforge.net/projects/staden/files/
Tablet	Visualization tool	Illumina/Roche	http://bioinf.scri.ac.uk/tablet
TagDust	Data cleaning	Illumina	http://genome.gsc.riken.jp/osc/english/software/.
TileQC	Quality assessment	Illumina	http://www.science.oregonstate.edu/~dolanp/tileqc
XMatchView	Visualization tool	Illumina/Roche	http://www.bcgsc.ca/platform/bioinfo/software/xmatchview
Yenta	Visualization tool	Illumina	http://genome.wustl.edu/tools/cancer-genomics
Geneus	Data management	Illumina/ABI	http://www.genologics.com/solutions/research-informatics/

Table 4

The genetic variant prediction and detection bioinformatic programs for NGS data analysis.

Variant prediction/detection	Platform	Website
Functional variant prediction		
B-SIFT		http://research-pub.gene.com/bsift/
MAPP		http://mendel.stanford.edu/supplementarydata/stone_MAPP_2005
PhD-SNP		http://gpcr.biocomp.unibo.it/~emidio/PhD-SNP/PhD-SNP
PolyPhen-2/PolyPhen		http://genetics.bwh.harvard.edu/pph2/
SIFT		http://blocks.fhcrc.org/sift/SIFT.html
SNAP		http://www.rostlab.org/services/SNAP
SNAPper/Pedant		http://pedant.gsf.de/snapper
Variant detection		
Structural/genomic variant		
BreakDancer	Roche/Illumina/ABI	http://genome.wustl.edu/tools/cancer-genomics/
BreakDancer/BD- Mini	Roche/Illumina/ABI	http://seqanswers.com/wiki/BreakDancer
Breakway	Roche/Illumina/ABI	http://sourceforge.net/projects/breakway/files/
CNVSeq	Roche	http://tiger.dbs.nus.edu.sg/CNV-seq/
cnvHMM	Illumina	http://genome.wustl.edu/pub/software/cancer-genomics/cnvHMM/
cnD	Illumina	http://www.sanger.ac.uk/resources/software/cnd.html
GASV/GSV	Illumina	http://cs.brown.edu/people/braphael/software.html
Hydra	Illumina	http://code.google.com/p/hydra-sv/
MoDIL	Illumina	http://compbio.cs.toronto.edu/modil/
mrFAST	Illumina	http://mrfast.sourceforge.net/
NovelSeq	Roche/Illumina/ABI	http://compbio.cs.sfu.ca/strvar.htm
PEMer	Roche/Illumina/ABI	http://sv.gersteinlab.org/pemer/
Pindel	Illumina	http://www.ebi.ac.uk/~kye/pindel/
SegSeq	Illumina/ABI	http://www.broadinstitute.org/
SOAPsv	Roche/Illumina/ABI	http://soap.genomics.org.cn
Solid large Indel tool	ABI	http://solidsoftwaretools.com/gf/project/large_indel/
Solid CNV tool	ABI	http://solidsoftwaretools.com/gf/project/cnv/
SWT	Illumina	http://genome.wustl.edu/pub/software/cancer-genomics/GSTAT/
VariationHunter/VH-CR	Illumina	http://compbio.cs.sfu.ca/strvar.html
VARiD	ABI	http://compbio.cs.utoronto.ca/varid
Single nucleotide variant		
Atlas-SNP2	Roche/Illumina	http://www.hgsc.bcm.tmc.edu/cascade-tech-software-ti.hgsc
BOAT	Illumina	http://boat.cbi.pku.edu.cn/
DNA Baser	Roche	http://www.dnabaser.com/help/manual.html
DNAA	Roche/Illumina/ABI	http://sourceforge.net/projects/dnaa/
Galign	Illumina	http://shahamlab.rockefeller.edu/galign/galign.htm
GigaBayes/PbShort	Roche/Illumina	http://bioinformatics.bc.edu/marthlab/GigaBayes
GSNAP	Roche/Illumina	http://share.gene.com/gmap.
inGAP	Roche/Illumina	http://sites.google.com/site/nextgengenomics/ingap

Variant prediction/detection	Platform	Website
ngs_backbone	Roche/Illumina	http://bioinf.comav.upv.es/ngs_backbone/index.html
Omixon Variant	ABI	http://www.omixon.com/omixon/index.html
PyroBayes	Roche	http://bioinformatics.bc.edu/marthlab/PyroBayes
ssahaSNP	Illumina/Roche	http://www.sanger.ac.uk/Software/analysis/ssahaSNP
Slider	Illumina	http://www.bcgsc.ca/platform/bioinfo/software/slider
SNP-o-matic	Illumina	http://snpomatic.sourceforge.net
SNPSeeker	Illumina	http://www.genetics.wustl.edu/rmlab/
SNVMix	Illumina	http://compbio.bccrc.ca
SOAPsnp	Roche/Illumina/ABI	http://soap.genomics.org.cn
SWA454	Roche	http://www.broadinstitute.org/science/programs/genome-biology/crd
SVA	Illumina	http://www.svaproject.org/
VAAL	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd
VarScan	Roche/Illumina	http://genome.wustl.edu/tools/cancer-genomics
VARiD	Roche/Illumina/ABI	http://compbio.cs.utoronto.ca/varid
Differences between genomes		
DIAL	Illumina	http://www.bx.psu.edu/miller_lab/
SomaticCall	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd
SWAP454	Roche	http://www.broadinstitute.org/science/programs/genome-biology/crd
VAAL	Illumina	http://www.broadinstitute.org/science/programs/genome-biology/crd

Table 5

The end-user software packages and cloud computing software for NGS data analysis.

Software packages	Function	Website
End-user software packages		
Genomic workbench/CLCbio	Multi-task	http://www.clcbio.com/index.php?id=1331
NextGENe/SoftGenetics	Multi-task	http://softgenetics.com/NextGENe.html
Genomatix Genome Analyzer	Multi-task	http://www.genomatix.de/genome_analyzer.html
Zoom	Multi-task	http://www.bioinformaticssolutions.com/products/zoom/index.php
SeqMan Ngen/DNASTAR	Multi-task	http://www.dnastar.com/t-products-seqman-ngen.aspx
JMP Genomics	Multi-task	http://www.jmp.com/software/genomics/index.shtml
RTG/Real Time Genomics	Multi-task	http://www.realtimegenomics.com/RTG-Software
PASS	Multi-task	http://pass.cribi.unipd.it/cgi-bin/pass.pl?action=Download
CASAVA	Multi-task	http://www.illumina.com/software/
Geneus/GenoLogics	Multi-task	http://www.genologics.com/solutions/research-informatics/
Roche Analysis tools	Multi-task	http://454.com/products-solutions/analysis-tools/index.asp
VSRAP	Multi-task	http://sourceforge.net/apps/mediawiki/vancouvershortr/
BING	Multi-task	http://www.dinulab.org/bing
PaCGeE/PGI	Multi-task	http://personalgenomicsinstitute.org/index.php/
GATK	Multi-task	http://www.broadinstitute.org/gsa/wiki/index.php/
Geneious Pro	Multi-task	http://www.geneious.com/default,1246,NGS%20Assembly.sm
Partek GS/Partek	Multi-task	http://www.partek.com/partekgs
Bioscope	Multi-task	https://products.appliedbiosystems.com/ab/en/US/adirect/
Cloud computing		
Crossbow	Mapping and SNP calling	http://bowtie-bio.sf.net
CloudBurst	Reference-based mapping	http://sourceforge.net/apps/mediawiki/cloudburst-bio/
Contrail	De novo assembly	http://sourceforge.net/apps/mediawiki/contrail-bio/
Cloud-MAQ	Modified-Maq for cloud	http://geschickten.com/download.html
Bioscope	Reference-based mapping	https://products.appliedbiosystems.com/ab/en/US/adirect/
Cycle Computing	Cloud computing	http://www.cyclecomputing.com/
GenomeQuest	Customer service	http://www.genomequest.com/
Geospiza/GeneSifter	Customer service	http://www.geospiza.com/Contact/genesiftertrial_ng.shtml