Published in final edited form as:
Methods Enzymol. 2013 ; 529: 201–208. doi:10.1016/B978-0-12-418687-3.00016-1.

Explanatory Chapter: Next Generation Sequencing

Srinivasan Yegnasubramanian[1]
Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA

Abstract

Technological breakthroughs in sequencing technologies have driven the advancement of molecular biology and molecular genetics research. The advent of high-throughput Sanger sequencing (for information on the method, see Sanger Dideoxy Sequencing of DNA) in the mid- to late-1990s made possible the accelerated completion of the human genome project, which has since revolutionized the pace of discovery in biomedical research. Similarly, the advent of next generation sequencing is poised to revolutionize biomedical research and usher a new era of individualized, rational medicine.

The term next generation sequencing refers to technologies that have enabled the massively parallel analysis of DNA sequence facilitated through the convergence of advancements in molecular biology, nucleic acid chemistry and biochemistry, computational biology, and electrical and mechanical engineering. The current next generation sequencing technologies are capable of sequencing tens to hundreds of millions of DNA templates simultaneously and generate >4 gigabases of sequence in a single day. These technologies have largely started to replace high-throughput Sanger sequencing for large-scale genomic projects, and have created significant enthusiasm for the advent of a new era of individualized medicine.

1. THEORY

1.1. Overview of commercialized next generation sequencing platforms

Given the promise of and the demand for next generation sequencing technologies, there has been intense competition for the development of NGS platforms. 454 life technologies, later acquired by Roche, was the first to commercially release an NGS platform. Solexa, now part of Illumina, released the next platform, with Applied Biosystems marketing the third commercialized platform, which was acquired from Agencourt. Helicos was the first company to release a single-molecule sequencing NGS platform, and more recently several new companies have entered the arena, including Complete Genomics, Pacific Biosciences, and Ion Torrents, with others likely to follow in the near future.

The major steps involved in next generation sequencing technologies that are generically applicable to all of the current technologies are library choice/construction, preparation of libraries for sequencing, and massively parallel sequencing. We first discuss some of the terminology used in Next Generation Sequencing experiments and then discuss each of these steps below and highlight the broad similarities and differences between platforms.

[1]Corresponding author: syegnasu@jhmi.edu.

Referenced Protocols in Methods Navigator
Sanger Dideoxy Sequencing of DNA
Preparation of fragment libraries for Next Generation Sequencing on the Applied Biosystems SOLiD platform

1.2. Terminology in next generation sequencing

- Read: refers to a single contiguous stretch of sequence returned from the instrument.
- Fragment read: a read generated from a fragment library; these reads are generated from a single end of a small DNA fragment that is typically in the order of 100–500 bps depending on the sequencing platform.
- Fragment paired-end reads: two reads generated from each end of a DNA fragment from a fragment library.
- Mate-paired read: two reads generated from each end of a large DNA fragment with a predefined size-range.
- Coverage: the average number of times each base pair in the target genome was covered by reads. For example, 30× coverage implies that each base pair in the reference genome was covered by 30 reads on average.

1.3. NGS library choice and construction

Two major types of libraries can be employed depending on the application: fragment libraries, and mate-paired libraries. For fragment libraries, genomic DNA from a sample is randomly fragmented to a small modal size, typically just 1–5 times the size of the sequencing platform's read length. Sequencing adaptors are then attached to these library molecules to allow sequencing from a single end of each DNA fragment in the library. The protocols used to generate such fragment libraries for the Applied Biosystems SOLiD platform are described in detail in the accompanying protocols chapter (see Preparation of fragment libraries for Next Generation Sequencing on the Applied Biosystems SOLiD platform). More recently, it has become possible to sequence from both ends of such library DNA fragments using a process referred to as fragment paired-end sequencing. Fragment libraries are ideal for analysis of single-nucleotide substitutions/variations. Each DNA fragment in the library produces a single read and multiple overlapping fragments are sequenced for each position in the genome. A coverage of >30× is usually needed to confidently distinguish true variation from sequencing errors and for robustly distinguishing homozygous and heterozygous SNPs. Additionally, fragment libraries can also provide information on genomic copy number. This can be done by taking all of the fragment library reads within fixed genomic bins and carrying out analyses to assess whether the number of reads observed is different from the number expected by random chance (e.g., Xie and Tammi, 2009), representing an extension of digital karyotyping analyses (Wang et al., 2002). Fragment libraries can also be target-enriched with microarray- or solution-based hybrid capture strategies for targeted resequencing (Albert et al., 2007; Gnirke et al., 2009). In these analyses, first, a fragment library is prepared. Next, the library is subjected to target sequence enrichment by hybridization to target-complementary oligonucleotides, called 'baits.' The oligonucleotide 'baits' can be immobilized on the surface of a microarray. Agilent and Nimblegen, among other companies, offer this as a standard or custom design product. More recently, the 'bait' oligonucleotides are synthesized in situ on microarrays, then released by cleavage from the microarray, amplified, and modified with biotin and immobilized on magnetic beads to allow solution-based capture of targets (Gnirke et al., 2009). Agilent markets this as their SureSelect solution-capture-based target enrichment process, and kits have been released for use with the SOLiD and Illumina NGS platforms. Such approaches have allowed targeted resequencing of any portion of the genome, such as all exons in the human genome (Maher, 2009).

A mate-paired library is constructed by first randomly shearing or fragmenting genomic DNA to a modal size that is typically >1000 bps, which significantly exceeds the read

lengths produced by most of the currently commercialized platforms. This library is then size-separated on a gel, and the part of the library corresponding to a specific size range, for example, 2–3 kbp, is excised and purified. These fragments are then circularized via ligation of an adapter under conditions that promote circularization of library molecules with the adaptor separating the two ends, as opposed to ligation of different library molecules together. This geometry allows generation of a library consisting of DNA fragments comprised of subfragments from the two ends of the original size-selected DNA library. The two mate-paired subfragments are then sequenced to reveal the sequences at the two ends of each 2–3 kbp library template. Because we know a priori the possible distances between the two sequences comprising the mate-paired read, after alignment to the reference genome, we can assess whether there was likely to be an amplification, deletion, or translocation between the mate-paired sequences. Similarly, the orientation of the sequences can be used to detect potential inversions. Therefore, mate-paired libraries not only provide information on single nucleotide substitutions, but also on genomic structural variation, as has been demonstrated in several recent reports (Korbel et al., 2007; McKernan et al., 2009).

With the advent of more recent NGS platforms, other library types are also possible. Pacific Biosciences, for instance, has developed ultra-long read lengths of >1000 base pairs. They have deployed these highly processive reads to generate repeated serial reads of both strands of double-strand DNA after circularization of a fragment library with hairpin adaptors ligated to each end of the fragments. The resulting ‘SMRT Bell’ libraries allow high-fidelity sequencing where the accuracy increases with the number of times the polymerase traverses the circularized SMRT Bell fragments (http://www.pacificbiosciences.com/). This company is also developing strobe-sequencing, where the progress of the processive polymerase in copying long template DNA is followed in an on-off periodic fashion as a way to generate several mate-tags of sequence from a long DNA template, with all tags oriented in the same direction. Complete Genomics has introduced highly complex library generation strategies involving serial cutting and circularization to fabricate DNA nanoballs for unchained ligation-based sequencing (Drmanac et al., 2010). This strategy has been used for resequencing of whole human genomes (Roach et al., 2010). Other library configurations and geometries are likely to be introduced as the diversity of NGS platforms increases.

1.4. Preparation of libraries for sequencing on different NGS platforms

The steps involved in preparing libraries for sequencing on a specific NGS platform are usually tailor-made. For the Roche 454 and Applied Biosystems SOLiD systems, this involves emulsion PCR (Dressman et al., 2003) to amplify individual template DNA molecules clonally on the surface of a bead. In emulsion PCR, individual DNA templates are sequestered along with PCR reagents, such as nucleotide triphosphates, primers, and Taq polymerase, and a primer-coated bead within an aqueous droplet surrounded by a hydrophobic shell within an oil-in-water emulsion. Subjecting these droplets to PCR allows clonal amplification of each template DNA molecule on the surface of the bead. In the case of Roche 454, the beads are then deposited in picoliter wells of a plate. These beads serve as the substrate for sequencing on the instrument (Margulies et al., 2005). In the case of Applied Biosystems, the clonally amplified DNA molecules on the surface of the bead are end-modified and covalently and randomly attached to the surface of a glass slide (http://www.appliedbiosystems.com). This glass slide is then loaded for sequencing on the instrument. Recent improvements in the automation of the emulsion PCR process have streamlined these otherwise cumbersome steps. For the Illumina/Solexa Genome Analyzer and HiSeq platforms, DNA libraries are clonally bridge amplified to generate clonal clusters of each DNA template in situ on the surface of lanes in a flow cell (http://www.illumina.com). These flow cells are then subjected to massively parallel sequencing as described in the following section. For Helicos, library generation is simpler and does not

require any clonal amplification steps. In their true single-molecule sequencing (tSMS) platform, library fragments are tailed with poly-adenosine using the terminal deoxynucleotidyl transferase (TdT) enzyme and hybridized onto oligo-dT primer-conjugated flow cells, which are then subjected to sequencing via extension from the oligo-dT primers (Harris et al., 2008).

1.5. Massively parallel sequencing of libraries on different NGS platforms

Each of the currently commercialized NGS platforms uses distinct chemistries to allow massively parallel sequencing of many millions to billions of template DNA molecules. The differences in chemistries confer various strengths and weaknesses to each platform. Because these technologies are rapidly evolving, we focus our discussion on the broad characteristics of the chemistries that are likely to remain stable for the currently commercialized platforms and only touch briefly on up-and-coming platforms that have not yet seen widespread adoption.

The Roche 454 system (http://www.454.com/) uses a sequence-by-synthesis strategy in which DNA templates on the surface of a bead are copied by a DNA polymerase which is forced to add a single-nucleotide species one at a time by cycling the flow of each nucleotide in turn and repeating these cycles for several iterations (Margulies et al., 2005). The pyrophosphates released by the polymerase are converted to light by a pyrophosphatase-based pyrosequencing process in which the amount of light emitted can be used to calculate the number of a specific nucleotide added at each cycle. One somewhat persistent problem with this method is that mononucleotide repeat tracks (e.g., a run of 12 adenines in a row) can lead to errors. This method allows sequencing read lengths of 400 or more base pairs in current implementations. However, the overall throughput is limited by the number of picoliter wells on a plate that can be sequenced, and this platform currently has the lowest sequence capacity per time or per dollar compared to other commercialized platforms.

Illumina (http://www.illumina.com/) and Helicos (http://www.helicosbio.com/) also use a sequence-by-synthesis strategy, but avoid errors associated with mononucleotide runs by using fluorescently labeled reversible chain terminator nucleotides allowing controlled addition of only a single nucleotide at a time, even in stretches of mononucleotide repeats. Because these platforms halt at the addition of every single nucleotide, the coupling efficiencies become limiting and read lengths are typically less than 100 bp. In the case of Helicos, which uses tSMS technology, there appears to be a persistent issue of ‘dark bases’ in which the nucleotide incorporation is not associated with fluorescence generation. This will probably be an issue with other emerging tSMS platforms as well.

The Applied Biosystems (http://solid.appliedbiosystems.com) (now Life Technologies) SOLiD platform uses a sequence-by-ligation approach in which a DNA ligase, instead of a DNA polymerase, is used to assess sequence via sequential ligation of fluorescently labeled oligonucleotide probes that can interrogate each combination of two adjacent bases (16 combinations possible). However, there are only four different fluorescent dyes, and each one must interrogate one of four possible dinucleotide combinations. Because of this, an individual ligation reaction does not uniquely identify the corresponding dinucleotide combination. Each base in the sequence is interrogated twice in this degenerate fashion and the combined data across an entire read can be deconvoluted to decipher the final sequence. The first step of the sequencing reaction is to anneal a sequencing primer to the P1 adaptor on the library template (see accompanying fragment library preparation chapter for details) and then to add a mixture of the 16 possible labeled probes. The appropriate di-base probe binds to the first and second base of the template and is ligated to the sequencing primer only if there is a perfect match. The fluorophore associated with this probe is then registered

and the probe is enzymatically processed to allow sequential ligation of another probe to query the sixth and seventh bases. This process is carried out a total of ten times for the first primer. After the last ligation step, the reaction is 'reset' by denaturing and washing away the newly synthesized DNA strand from the template DNA that is covalently linked to the bead (see emulsion PCR description above). A new sequencing primer designed to hybridize to a sequence that is offset by one base from the first primer is then annealed so that the first ligation reaction stemming from this sequencing primer interrogates the last base of the adaptor sequence (position 0) and the first base of the template. This primer also goes through a total of ten ligation steps. There are a total of five different sequencing primers that each undergo ten ligation steps. This results in each base being interrogated twice and a sequencing length of 50 base pairs per read.

1.6. The near- and long-term horizon

Each of the platforms described above is routinely making advancements in sequencing throughput in terms of both time and cost per gigabase pairs of sequence output. In the meanwhile, other platforms such as Complete Genomics (http://www.completegenomics.com/), Pacific Biosciences (http://www.pacificbiosciences.com/), Ion Torrents (http://www.iontorrent.com/), and possibly several other players are preparing to enter the market. As a result of this intense competition, cost and time per gigabase pair of sequence produced is rapidly declining. Another consequence of this is that currently commercialized platforms may lose or fail to gain market share and be in danger of folding.

Nonetheless, with the rapid declines in cost, it will be possible to carry out large-scale genomic projects to elucidate novel biology at an unprecedented scale. Additionally, it may become routine to sequence entire human genomes in the context of health and disease and apply such technologies to entire populations and not just individuals. This information can serve as a source of individualized biomarkers that can provide individualized guidance for therapeutic decision-making. The key will be to develop, in parallel, the computational, biostatistical, and bioinformatics solutions to harness the power of these increasingly cost-effective technologies and deploy them not only to understand novel biology, but also to improve the practice and delivery of health care.

Referenced Literature

Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. Nature Methods. 2007; 4:903–905. [PubMed: 17934467]

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 of the United States of America. 2003; 100:8817–8822. [PubMed: 12857956]

Drmanac R, Sparks AB, Callow MJ, et al. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science. 2010; 327:78–81. [PubMed: 19892942]

Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology. 2009; 27:182–189.

Harris TD, Buzby PR, Babcock H, et al. Single-molecule DNA sequencing of a viral genome. Science. 2008; 320:106–109. [PubMed: 18388294]

Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science. 2007; 318:420–426. [PubMed: 17901297]

Maher B. Exome sequencing takes centre stage in cancer profiling. Nature. 2009; 459:146–147. [PubMed: 19444175]

Margulies M, Egholm M, Altman WE, et al. Genome sequencing in micro-fabricated high-density picolitre reactors. Nature. 2005; 437:376–380. [PubMed: 16056220]

McKernan KJ, Peckham HE, Costa GL, 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; 19:1527–1541. [PubMed: 19546169]

Roach JC, Glusman G, Smit AF, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010; 328:636–639. [PubMed: 20220176]

Wang TL, Maierhofer C, Speicher MR, et al. Digital karyotyping. Proceedings of the National Academy of Sciences of the United States of America. 2002; 99:16156–16161. [PubMed: 12461184]

Xie C, Tammi MT. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. BMC Bioinformatics. 2009; 10:80. [PubMed: 19267900]