Next-Generation Sequencing: Methodology and Application

Ayman Grada[1] and Kate Weinbrecht[2]

Journal of Investigative Dermatology (2013) **133**, e11; doi:10.1038/jid.2013.248

INTRODUCTION

Nucleic acid sequencing is a method for determining the exact order of nucleotides present in a given DNA or RNA molecule. In the past decade, the use of nucleic acid sequencing has increased exponentially as the ability to sequence has become accessible to research and clinical labs all over the world. The first major foray into DNA sequencing was the Human Genome Project, a $3 billion, 13-year-long endeavor, completed in 2003. The Human Genome Project was accomplished with first-generation sequencing, known as Sanger sequencing. Sanger sequencing (the chain-termination method), developed in 1975 by Edward Sanger, was considered the gold standard for nucleic acid sequencing for the subsequent two and a half decades (Sanger *et al.*, 1977).

Since completion of the first human genome sequence, demand for cheaper and faster sequencing methods has increased greatly. This demand has driven the development of second-generation sequencing methods, or next-generation sequencing (NGS). NGS platforms perform massively parallel sequencing, during which millions of fragments of DNA from a single sample are sequenced in unison. Massively parallel sequencing technology facilitates high-throughput sequencing, which allows an entire genome to be sequenced in less than one day. In the past decade, several NGS platforms have been developed that provide low-cost, high-throughput sequencing. Here we highlight two of the most commonly used platforms in research and clinical labs today: the LifeTechnologies Ion Torrent Personal Genome Machine (PGM) and the Illumina MiSeq. The creation of these and other NGS platforms has made sequencing accessible to more labs, rapidly increasing the amount of research and clinical diagnostics being performed with nucleic acid sequencing.

OVERVIEW OF THE METHODOLOGY

Although each NGS platform is unique in how sequencing is accomplished, the Ion Torrent PGM and the Illumina MiSeq have a similar base methodology that includes template preparation, sequencing and imaging, and data analysis (Metzker, 2010). Within each generalized step, the individual platforms discussed have unique aspects. An overview of the sequencing methodologies discussed is provided in Figure 1.

WHAT NGS DOES

- NGS provides a much cheaper and higher-throughput alternative to sequencing DNA than traditional Sanger sequencing. Whole small genomes can now be sequenced in a day.
- High-throughput sequencing of the human genome facilitates the discovery of genes and regulatory elements associated with disease.
- Targeted sequencing allows the identification of disease-causing mutations for diagnosis of pathological conditions.
- RNA-seq can provide information on the entire transcriptome of a sample in a single analysis without requiring previous knowledge of the genetic sequence of an organism. This technique offers a strong alternative to the use of microarrays in gene expression studies.

LIMITATIONS

- NGS, although much less costly in time and money in comparison to first-generation sequencing, is still too expensive for many labs. NGS platforms can cost more than $100,000 in start-up costs, and individual sequencing reactions can cost upward of $1,000 per genome.
- Inaccurate sequencing of homopolymer regions (spans of repeating nucleotides) on certain NGS platforms, including the Ion Torrent PGM, and short-sequencing read lengths (on average 200–500 nucleotides) can lead to sequence errors.
- Data analysis can be time-consuming and may require special knowledge of bioinformatics to garner accurate information from sequence data.

[1]*Department of Dermatology, Boston University School of Medicine, Boston, Massachusetts, USA and* [2]*School of Forensic Sciences, Center for Health Sciences, Oklahoma State University, Tulsa, Oklahoma, USA*

Correspondence: Ayman Grada, Department of Dermatology, Boston University School of Medicine, 609 Albany Street, Boston, Massachusetts 02118, USA. E-mail: grada@bu.edu

Template preparation

Template preparation consists of building a library of nucleic acids (DNA or complementary DNA (cDNA)) and amplifying that library. Sequencing libraries are constructed by fragmenting the DNA (or cDNA) sample and ligating adapter sequences (synthetic oligonucleotides of a known sequence) onto the ends of the DNA fragments. Once constructed, libraries are clonally amplified in preparation for sequencing. The PGM utilizes emulsion PCR on the OneTouch system to amplify single library fragments onto microbeads, whereas the MiSeq utilizes bridge amplification to form template clusters on a flow cell (Berglund *et al.*, 2011; Quail *et al.*, 2012).

Sequencing and imaging

To obtain nucleic acid sequence from the amplified libraries, the Ion Torrent PGM and the MiSeq both rely on sequencing by synthesis. The library fragments act as a template, off of which a new DNA fragment is synthesized. The sequencing occurs through a cycle of washing and flooding the fragments with the known nucleotides in a sequential order. As nucleotides incorporate into the growing DNA strand, they are digitally recorded as sequence. The PGM and the MiSeq each rely on a slightly different mechanism for detecting nucleotide sequence information. The PGM performs semiconductor sequencing that relies on the detection of pH changes induced by the release of a hydrogen ion upon the incorporation of a nucleotide into a growing strand of DNA (Quail *et al.*, 2012). By contrast, the MiSeq relies on the detection of fluorescence generated by the incorporation of fluorescently labeled nucleotides into the growing strand of DNA (Quail *et al.*, 2012).

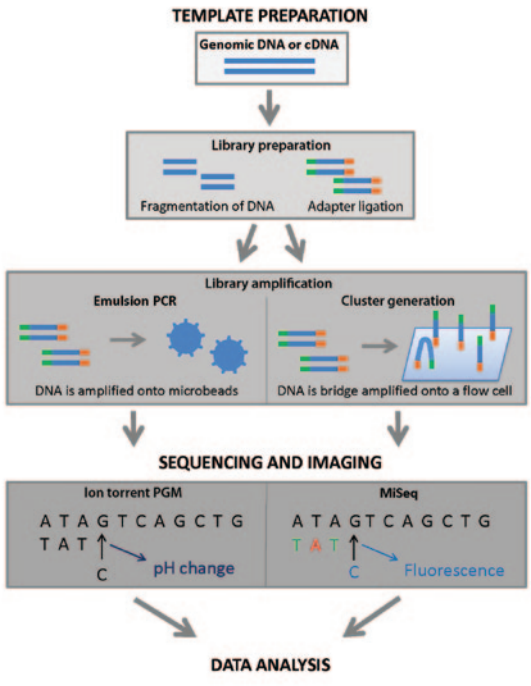

Figure 1. Next-generation sequencing methodology.

Data analysis

Once sequencing is complete, raw sequence data must undergo several analysis steps. A generalized data analysis pipeline for NGS data includes preprocessing the data to remove adapter sequences and low-quality reads, mapping of the data to a reference genome or *de novo* alignment of the sequence reads, and analysis of the compiled sequence. Analysis of the sequence can include a wide variety of bioinformatics assessments, including genetic variant calling for detection of SNPs or indels (i.e., the insertion or deletion of bases), detection of novel genes or regulatory elements, and assessment of transcript expression levels. Analysis can also include identification of both somatic and germline mutation events that may contribute to the diagnosis of a disease or genetic condition. Many free online tools and software packages exist to perform the bioinformatics necessary to successfully analyze sequence data (Gogol-Döring and Chen, 2012).

APPLICATIONS

The applications of NGS seem almost endless, allowing for rapid advances in many fields related to the biological sciences. Resequencing of the human genome is being performed to identify genes and regulatory elements involved in pathological processes. NGS has also provided a wealth of knowledge for comparative biology studies through whole-genome sequencing of a wide variety of organisms. NGS is applied in the fields of public health and epidemiology through the sequencing of bacterial and viral species to facilitate the identification of novel virulence factors. Additionally, gene expression studies using RNA-Seq (NGS of RNA) have begun to replace the use of microarray analysis, providing researchers and clinicians with the ability to visualize RNA expression in sequence form. These are simply some of the broad applications that begin to skim the surface of what NGS can offer the researcher and the clinician. As NGS continues to grow in popularity, it is inevitable that there will be additional innovative applications.

NGS IN PRACTICE

Whole-exome sequencing

Mutation events that occur in gene-coding or control regions can give rise to indistinguishable clinical presentations, leaving the diagnosing clinician with many possible causes for a given condition or disease. With NGS, clinicians are provided a fast, affordable, and thorough way to determine the genetic cause of a disease. Although high-throughput sequencing of the entire human genome is possible, researchers and clinicians are typically interested in only the protein-coding regions of the genome, referred to as the exome. The exome comprises just over 1% of the genome and is therefore much more cost-effective to sequence than the entire genome, while providing sequence information for protein-coding regions.

Exome sequencing has been used extensively in the past several years in gene discovery research. Several gene discovery studies have resulted in the identification of genes that are relevant to inherited skin disease (Lai-Cheong and McGrath, 2011). Exome sequencing can also facilitate the

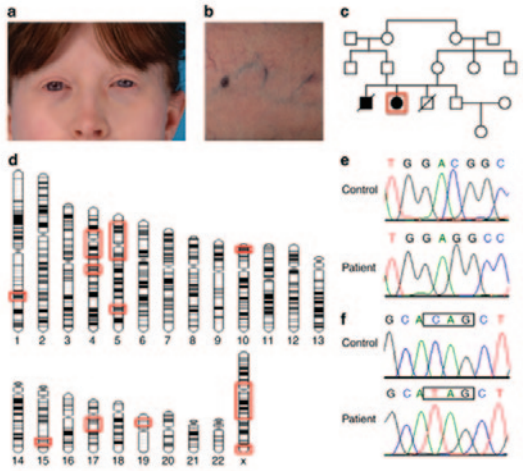

Figure 2. Clinical application of whole-exome sequencing in the detection of two disease-causing mutations. Reprinted from Cullinane *et al.*, 2011.

identification of disease-causing mutations in pathogenic presentations where the exact genetic cause is not known.

Figure 2 (Cullinane *et al.*, 2011) demonstrates the direct effect that NGS can have on the correct diagnoses of a patient. It summarizes the use of homozygosity mapping followed by whole-exome sequencing to identify two disease-causing mutations in a patient with oculocutaneous albinism and congenital neutropenia (Cullinane *et al.*, 2011). Figure 2a and 2b display the phenotypic traits common to oculocutaneous albinism type 4 and neutropenia observed in this patient. Figure 2c is a pedigree of the patient's family, both the affected and unaffected individuals. The idiogram (graphic chromosome map) in Figure 2d highlights the areas of genetic homozygosity. These regions were identified by single-nucleotide-polymorphism array analysis and were considered possible locations for the disease-causing mutation(s). Figures 2e and 2f display chromatograms for the two disease-causing mutations identified by whole-exome sequencing. Figure 2e depicts the mutation in *SLC45A2,* and Figure 2f depicts the mutation in *G6PC3*. This case portrays the valuable role that NGS can play in the correct diagnosis of an individual patient who displays disparate symptoms with an unidentified genetic cause.

Targeted sequencing

Although whole-genome and whole-exome sequencing are possible, in many cases where a suspected disease or condition has been identified, targeted sequencing of specific genes or genomic regions is preferred. Targeted sequencing is more affordable, yields much higher coverage of genomic regions of interest, and reduces sequencing cost and time (Xuan *et al.*, 2012). Researchers have begun to develop sequencing panels that target hundreds of genomic regions that are hotspots for disease-causing mutations. These panels target only desired regions of the genome for sequencing, eliminating the majority of the genome from analysis. Targeted sequencing panels can be developed by researchers or clinicians to include specific genomic regions of interest. In addition, sequencing panels that target common regions of interest can be purchased for clinical use; these include panels that target hotspots for cancer-causing mutations (Rehm, 2013). Targeted sequencing—whether of individual genes or whole panels of genomic regions—aids in the rapid diagnosis of many genetic diseases. The results of disease-targeted sequencing can aid in therapeutic decision making in many diseases, including many cancers for which the treatments can be cancer-type specific (Rehm, 2013).

QUESTIONS

Answers are available as supplementary material online and at http://www.scilogs.com/jid/.

1. The basic methodological steps of NGS include the following:

A. Template preparation, emulsion PCR, sequencing, data analysis.

B. Template preparation, sequencing and imaging, data analysis.

C. Template amplification, sequencing and imaging, data analysis.

D. Template preparation, sequencing and imaging, alignment to a reference genome.

E. DNA fragmentation, sequencing, data analysis.

2. Advantages of targeted sequencing as opposed to full-genome, exome, or transcriptome sequencing include the following:

A. Affordable and efficient for quickly interrogating particular genomic regions of interest.

B. Provides a deeper coverage of genomic regions of interest.

C. Can be utilized in deciding a therapeutic plan of action for both germline and somatic cancers.

D. Detects and quantifies low-frequency variants such as rare drug-resistant viral mutations (e.g., HIV, hepatitis B virus, or microbial pathogens).

E. All of the above.

3. Applications of NGS in medicine include the following:

A. Detecting mutations that play a role in diseases such as cancer.

B. Identifying genes responsible for inherited skin diseases.

C. Determining RNA expression levels.

D. Identifying novel virulence factors through sequencing of bacterial and viral species.

E. All of the above.

CONFLICT OF INTEREST

The authors state no conflict of interest.

SUPPLEMENTARY MATERIAL

Answers and a PowerPoint slide presentation appropriate for journal club or other teaching exercises are available at http://dx.doi.org/10.1038/jid.2013.248.

REFERENCES

Berglund EC, Kiialainen A, Syvänen AC (2011) Next-generation sequencing technologies and applications for human genetic history and forensics. *Invest Genet* 2:23

Cullinane AR, Vilboux T, O'Brien K *et al*. (2011) Homozygosity mapping and whole-exome sequencing to detect SLC45A2 and G6PC3 mutations in a single patient with oculocutaneous albinism and neutropenia. *J Invest Dermatol* 131:2017–25

Gogol-Döring A, Chen W (2012) An overview of the analysis of next generation sequencing data. *Methods Mol Biol* 802:249–57

Lai-Cheong JE, McGrath JA (2011) Next-generation diagnostics for inherited skin disorders. *J Invest Dermatol* 131:1971–3

Metzker ML (2010) Sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46

Quail MA, Smith M, Coupland P *et al*. (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom* 13:341

Rehm HL (2013) Disease-targeted sequencing: a cornerstone in the clinic. *Nat Rev Genet* 14:295–300

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* 74:5463–7

Xuan J, Yu Y, Qing T *et al*. (2012) Next-generation sequencing in the clinic: promises and challenges. *Cancer Lett*; e-pub ahead of print 19 November 2012