Cancer Letters 340 (2013) 284–295

Contents lists available at ScienceDirect

Cancer Letters

journal homepage: www.elsevier.com/locate/canlet

Mini-review

Next-generation sequencing in the clinic: Promises and challenges

Jiekun Xuan [a,b], Ying Yu [a], Tao Qing [a], Lei Guo [b,*], Leming Shi [a,b,*]

[a] School of Pharmacy, Fudan University, 826 Zhangheng Road, Shanghai 201203, China
[b] National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, AR 72079, USA

ARTICLE INFO

Keywords:
Whole-genome sequencing
Exome sequencing
RNA-Seq
Bioinformatics
FFPE
Tumor heterogeneity
Clinical applications

ABSTRACT

The advent of next generation sequencing (NGS) technologies has revolutionized the field of genomics, enabling fast and cost-effective generation of genome-scale sequence data with exquisite resolution and accuracy. Over the past years, rapid technological advances led by academic institutions and companies have continued to broaden NGS applications from research to the clinic. A recent crop of discoveries have highlighted the medical impact of NGS technologies on Mendelian and complex diseases, particularly cancer. However, the ever-increasing pace of NGS adoption presents enormous challenges in terms of data processing, storage, management and interpretation as well as sequencing quality control, which hinder the translation from sequence data into clinical practice. In this review, we first summarize the technical characteristics and performance of current NGS platforms. We further highlight advances in the applications of NGS technologies towards the development of clinical diagnostics and therapeutics. Common issues in NGS workflows are also discussed to guide the selection of NGS platforms and pipelines for specific research purposes.

Published by Elsevier Ireland Ltd.

1. Introduction

Increased awareness that decoding the human genome provides critical clues to the genetics of diseases as well as the development of more specific preventive, diagnostic and therapeutic strategies has driven extensive sequencing and mapping efforts in the past decades. After the completion of the first human genome sequence in 2004 [1], the growing need to sequence a large number of individual genomes in a fast, low-cost and accurate way has directed a shift from traditional Sanger sequencing methods towards new high-throughput genomic technologies. In 2005, the first massively parallel DNA sequencing platforms emerged, ushering in a new era of next-generation sequencing (NGS) [2,3]. To date, the development of NGS technologies has vastly accelerated the pace of data generation—on the order of hundreds of gigabases of nucleotide sequence per instrument run, while reducing sequencing cost by over five orders of magnitude. Owing to these advantages, NGS technologies have been widely used for many applications, such as rare variant discovery by whole genome resequencing or targeted sequencing, transcriptome profiling of cells, tissues and organisms, and identification of epigenetic markers for disease diagnosis.

Here, we first provide a brief overview of the characteristics, strengths and limitations of current NGS platforms (Table 1). We then discuss the major applications of NGS technologies, with a focus on cancer diagnosis and prognosis. Finally, we discuss the bioinformatics tools and challenges in NGS data analysis.

2. Overview of NGS technologies

The technical details of the three major NGS platforms have been well described elsewhere [4]. The following section primarily focuses on the performance of each sequencing system.

2.1. Roche/454

The 454 sequencing system is based on the combination of emulsion PCR and pyrosequencing technology [2]. In emulsion PCR, single-stranded template carrying beads are confined to individual emulsion droplets in which millions of copies of each template are produced by PCR amplification. Amplicon-bearing beads are subsequently enriched and deposited into individual wells of a picotiter plate where solid-phase pyrosequencing is carried out. In this sequencing-by-synthesis method, luminescence emission from the release of pyrophosphate upon template-directed nucleotide incorporation is monitored in real time. The strength of the 454 system lies in its ability to sequence long reads. The latest 454 GS FLX platform with Titanium chemistry can produce approximately one million reads with lengths of up to 1000 bp per instrument run. Owing to this advantage, despite the higher costs compared with other NGS platforms, the 454 platform is best suitable for several applications, including *de novo* assembly [5] and

* Corresponding authors. Address: School of Pharmacy, Fudan University, 826 Zhangheng Road, Shanghai 201203, China.
E-mail addresses: lei.guo@fda.hhs.gov (L. Guo), leming.shi@gmail.com (L. Shi).

0304-3835/$ - see front matter Published by Elsevier Ireland Ltd.
http://dx.doi.org/10.1016/j.canlet.2012.11.025

metagenomics [6]. However, the 454 technology has an inherent problem in the detection of homopolymers, stretches of the same nucleotide. Due to the lack of a terminating moiety, multiple incorporations of identical nucleotides can occur in homopolymeric regions during a single sequencing cycle. This can lead to nonlinearity between the signal intensity and the length of homopolymer stretches when more than three or four nucleotides are consecutively incorporated. Consequently, the 454 system has a relatively high error rate for calling insertions and deletions (indels) in homopolymers [7].

2.2. Illumina/Solexa

The Illumina sequencing system employs an array-based DNA sequencing-by-synthesis technology with reversible terminator chemistry [8]. In this approach, template DNA fragments are hybridized to a reaction chamber on an optically transparent solid surface (i.e., flow cell). Reversible terminators [9], a series of four modified nucleotides each labeled with a different removable fluorescent dye at the 3′-hydroxyl terminus, are used for step-by-step DNA synthesis. Millions of clonal clusters can be generated in each lane of the flow cell, which contains eight independent lanes for multiple libraries to be sequenced in parallel. The Genome Analyzer (GA), the first Illumina sequencing platform, originally produced 35-bp reads and generate more than 1 gigabase (Gb) of high-quality sequence per run in 2–3 days. The upgraded platforms, such as GA IIx and HiSeq 2000, yield much higher sequence output with increased read lengths. Despite its ultra-high-throughput and cost-effective advantages, the utility of the Illumina systems is limited to short-read sequencing. The limitation in read lengths is primarily due to dephasing effects [4]. Decreased or increased efficacy of nucleotide incorporation and failures in removing or adding terminating moieties in any given cycle can cause incomplete extension or overextension of the growing strand along the template, resulting lagging-strand or leading-strand dephasing. Moreover, signal dephasing can be caused by decay in fluorescent signal, incorporation of nucleotides without a fluorescent label (dark nucleotides) or incomplete removal of fluorescent labels, leading to base-calling errors. Consequently, base substitution error rates increase with read length. In addition, uneven read coverage has been revealed across AT-rich and GC-rich regions, with a bias towards the latter.

2.3. Life Technologies/SOLiD

The SOLiD system uses a ligation-based sequencing technology originated from previous work [3]. The sequencing library is prepared by emulsion PCR as in the 454 protocol. Sequencing is performed through successive cycles of ligation, in which each sequencing primer is ligated to a specific fluorescence-labeled octamer probe according to the complementarity between the di-bases of the probe and the template. Since each four di-bases (e.g., AG, GA, TC, CT) are tagged with one of four fluorescent dyes, the di-nucleotides at the same positions of each template are associated with a unique fluorescent color. Across ligation cycles, di-nucleotides are read at intervals of five bases, that is, di-nucleotides at position 4–5, 9–10, 14–15, 19–20 and so forth. After five ligation rounds, each nucleotide in the template is read twice by two fluorescent signals, greatly improving base-calling accuracy. Among the current NGS platforms, the SOLiD system presents the lowest error rate. Its most common error type is substitution. In addition, an underrepresentation of AT-rich regions has also been shown in the SOLiD data [10].

2.4. Emerging technologies

The emergence of single-molecule sequencing has provided a technological leap forward in the evolution of next generation sequencing. The promise of this technology lies in its ability to directly sequence single DNA or RNA molecules in biological samples without amplification. The single-molecule sequencing strategy promises significant advantages over current NGS technologies in that it minimizes sample handling, reduces sample input requirements, avoids amplification-induced bias and errors, increases read length flexibility and enables accurate quantitation of nucleic acid molecules. The simplicity, sensitivity and quantitative capabilities of single-molecule sequencing make it highly promising for molecular diagnostics [11].

The Helicos Genetic Analysis System is the first commercially available single-molecule sequencing platform [12]. In this system, poly(A)-tailed single-stranded DNA templates are captured by poly(T) oligonucleotide primers tethered to the surface of a flow cell. Sequencing is performed through iterative cycles of DNA polymerase-mediated single-base primer extension using a series of four fluorescent Virtual Terminator nucleotides, each of which represents a 3′-unblocked reversible terminator with a fluorophore-labeled inhibitory moiety [13]. In a standard run, the sequencer with two 25-channel flow cells is capable of capturing billions of single DNA molecules and generating over 21–35 Gb of sequence data with an average read length of 35 bp. Although the sequencing process is asynchronous in the Helicos system, dephasing effects that commonly exist in amplification-based sequencing platforms are not present. Moreover, there is no GC-content bias in read coverage. However, the current error rate in Helicos reads is relatively high (~3–5%), and the dominant error type is deletion, which presumably results from incorporation of unlabeled nucleotides and/or detection errors. The use of Virtual Terminator chemistry solves the homopolymer sequencing problem, and the base-by-base incorporation manner results in very low substitution error rates (typically 0.2%). When a two-pass strategy is applied, in which individual template molecules are sequenced twice, the error rates can be further reduced.

Other single-molecule sequencing technologies with longer read lengths, higher sequencing speed or lower overall cost are also emerging. One example is the PacBio *RS*, a single-molecule real-time sequencing system developed by Pacific Biosciences [14]. In this system, a single template-bound DNA polymerase molecule is immobilized to the bottom of a zero-mode waveguide, which functions as a nanophotonic visualization chamber for monitoring the polymerization reaction in a detection volume on the order of zeptoliters (10^{-21} l). During sequencing, template-directed incorporation of four fluorescent phospholinked nucleotides into the growing complementary strand is optically recorded in real time. The fluorescent dye attached to the terminal phosphate moiety of each phospholinked nucleotide is naturally removed by enzymatic cleavage upon incorporation. This allows rapid and processive DNA synthesis by the polymerase, yielding sequence reads of thousands of bases. Nonetheless, the PacBio system presently offers a throughput of approximately 50–100 Mb per run, which is much lower than current NGS platforms. Moreover, the single-read error rate is typically 15%, exceeding the error tolerance of many applications. A second example is nanopore sequencing technologies, in which single-stranded nucleic acid molecules are electrophoretically driven through a nanometer-sized pore and detected by their effect on an ionic current or optical signal [15]. Nanopore sequencing potentially offers long read lengths of up to tens of kilobases, minimal requirements of reagent and sample preparation, and high sequencing pace at low cost. However, several problems remain to be solved before the application of nanopore sequencing. The high speed of DNA translocation through nanopores makes it challenging to distinguish base signals from background noises by an electronic sensor. The random motion of molecules during translocation also adds to the difficulty in reaching single-base resolution. As a solution, IBM is developing a DNA