Biochimica et Biophysica Acta 1842 (2014) 1932–1941

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

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

Review

Next generation sequencing technology: Advances and applications ☆

H.P.J. Buermans [1], J.T. den Dunnen *

Leiden Genome Technology Center, Leiden University Medical Center, Postbus 9600, 2300 RC Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 22 November 2013
Received in revised form 5 June 2014
Accepted 15 June 2014
Available online 1 July 2014

Keywords:
Next generation sequencing
Sequence by synthesis
Nanopore
Single molecule sequencing
Basic technology
Applications

ABSTRACT

Impressive progress has been made in the field of Next Generation Sequencing (NGS). Through advancements in the fields of molecular biology and technical engineering, parallelization of the sequencing reaction has profoundly increased the total number of produced sequence reads per run. Current sequencing platforms allow for a previously unprecedented view into complex mixtures of RNA and DNA samples. NGS is currently evolving into a molecular microscope finding its way into virtually every fields of biomedical research. In this chapter we review the technical background of the different commercially available NGS platforms with respect to template generation and the sequencing reaction and take a small step towards what the upcoming NGS technologies will bring. We close with an overview of different implementations of NGS into biomedical research. This article is part of a Special Issue entitled: From Genome to Function.

1. Introduction

The growing power and reducing cost sparked an enormous range of applications of Next generation sequencing (NGS) technology. Gradually, sequencing is starting to become the standard technology to apply, certainly at the first step where the main question is "what's all involved", "what's the basis". It should be realized that for many applications sequencing would always have been the method of choice, yet it was science-fiction, technically unthinkable and later possible but far too costly. We perform genome-wide association studies (GWAS) using SNP-arrays simply because we cannot afford to perform whole-genome sequencing in ten-thousands of individuals. This is changing rapidly and sequencing will become our molecular microscope, the tool to get a first look. Although replication, transcription, translation, methylation and nuclear DNA folding are completely different processes, they can all be studied using sequencing.

An important advantage of sequence data is its quality, robustness and low noise. It should be noted that a successful NGS project requires expertise both at the wet lab as well as the bioinformatics side in order to warrant high quality data and data interpretation. The sequence itself is hard evidence of its correctness. A sequencing system will not produce "random" sequences and when it does this becomes evident immediately from QC calls obtained from spike-in controls. Furthermore random sequences will have no match and can be easily discarded during data analysis and when their number exceeds a certain threshold it is evident that there is a serious problem somewhere in the study.

2. Sequence library preparation

All currently available sequencing platforms require some level of DNA pre-processing into a library suitable for sequencing. In general, these steps involve shearing of high molecular weight DNA into an appropriate platform-specific size range, followed by an end polishing step to generate blunt ended DNA fragments. Specific adapters are ligated to these fragments by either A/T overhang or direct blunt ligation. A functional library requires having specific adapter sequences to be added to the 3′ and 5′ ends. Each of the sequence platforms uses a different set of unique adapter sequences to be compatible with the further steps of the process (Fig. 1).

Following adapter ligation Life Technologies (Solid, PGM, Proton) libraries require a nick translation step to get functional molecules while for the other technologies the sample is in principle ready for loading immediately after ligation. One may then choose to sequence these libraries directly as amplification free libraries or introduce a pre-amplification step prior to sequencing. It is important to realize that any step during pre-processing which involves amplification of the molecules [1] or which has been shown to be sequence biased, like ligations [2], will impose a selection on molecules that end up in the sequenceable libraries.

3. Current sequencing technology

The different sequence platform vendors have devised different strategies to prepare the sequence libraries into suitable templates as

☆ This article is part of a Special Issue entitled: From Genome to Function.

* Corresponding author: Tel.: + 31 715269400; fax: + 31 71 5268285.
E-mail addresses: h.buermans@lumc.nl (H.P.J. Buermans), ddunnen@humgen.nl (J.T. den Dunnen).
[1] Tel.: +31 715269400; fax: +31 71 5268285.

http://dx.doi.org/10.1016/j.bbadis.2014.06.015
0925-4439/

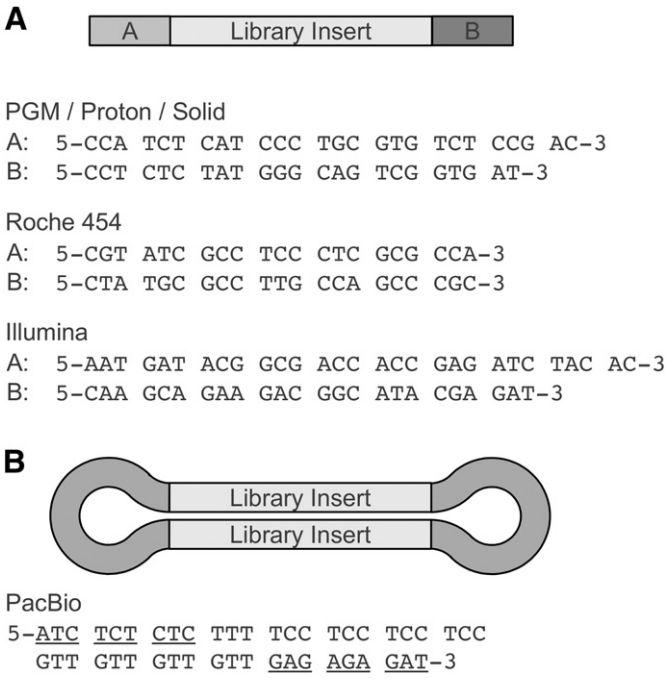

Fig. 1. Structure of sequence library molecules for the different technologies. Linear library molecules (Panel A) contain different adapter sequences at the 5′ [A] and 3′ [B] ends of the library inserts. Circular library molecules (Panel B) contain identical adapter molecules at both ends of the insert.

well as to detect the signal and ultimately read the DNA sequence. For the Illumina, Solid, PGM and 454 systems a local clonal amplification of the initial template molecules into polonies [3] is required to increase the signal-to-noise ratio because the systems are not sensitive enough to detect the extension of one base at the individual DNA template molecule level. On the other hand, the Heliscope and PacBio SMRT systems do not need any pre-amplification steps as these systems are sensitive enough to detect individual single molecule template extensions. The different strategies to generate the sequence reads also lead to differences in the output capacity for the different platforms (Table 1). Below we will focus on the newer sequencing platforms, being the Illumina, LifeTechnologies Semiconductor sequencing and PacBio. Other older platforms will briefly be discussed in Online Supplement 1.

3.1. Illumina technology

All of the enzymatic processes and imaging steps of the Illumina technology take place in a flow cell. Depending on the specific Illumina platform it may be partitioned into 1 (miSeq), 2 (HiSeq2500) or 8 (HiSeq2000, HiSeq2500) separate lanes. The Illumina platform uses bridge amplification for polony generation and a sequencing by synthesis (SBS) approach (Fig. 2A). Forward and reverse oligos for amplification (one with a cleavable site), complementary to the adapter sequences introduced during the library preparation steps, are attached to the entire inside surface of the flow cell lanes. The first step for loading the library onto the flow-cell is denaturation of the dsDNA fragments into individual ssDNA molecules. When on the flow-cell, these hybridize to the oligo nucleotides on the surface (Fig. 1A; step 1) which are used as primers to form an initial copy of the individual sequencing template molecule (Fig. 1A; step 2). The initial library molecules are removed and the copied, flow cell-attached fragments are used to generate a cluster of identical template molecules using isothermal amplification. This is done through cyclic alternations of three specific buffers that mediate the denaturation, annealing and extension steps at 60 °C. During these steps the 3′ end of the copied library molecules can hybridize to the complementary oligos on the flow cell, thus forming a bridge structure (Fig. 1A; steps 3–5). The final step is to remove one strand of the dsDNA fragments using the cleavable site in the surface oligo (Fig. 1A; step 6) and to block all 3′ ends with ddNTP to prevent the otherwise open 3′ ends to act as sequencing primer sites on adjacent library molecules [4].

With optimal loading of library molecules one flow-cell lane will yield approximately 800–1000 K clusters per mm^2. Optimal amounts depend not only on the concentration of the library, but also on the length of the molecules. Short molecules yield clusters with a small area that are denser and therefore generate more intense signals. Loading a wide fragment size distribution will generate clusters varying widely in size and signal strength which may impair the number of passing filter reads.

Bridge amplification is not a very efficient method for clonal amplification, i.e., the 35 cycles of isothermal amplification yield a mere ~1000 copies of the initial molecule. Moreover, there will be predominantly outward growth of the clusters, there is a high probability of the template strands to re-hybridize instead of annealing to a new primer site on the glass surface and there is both an upper and a lower limit to the length of the template molecules that can be reliably amplified. In addition, DNA polymerases, which are known to have biases towards specific DNA templates are used during the amplification processes. The bridge amplification scheme that Illumina exploits yields a high number of clusters, i.e., with good loading of the flow cell, the total number of reads generated per Hiseq2000 lane may reach ~180 million. With a paired-end 2 × 100 bp read format the total output of one flow-cell lane is up to ~36 Gb. A full run of 2 flow cells sequencing in parallel may yield ~600 Gb of data.

During sequencing, the polonies on the flow cell are read one nucleotide at a time in repetitive cycles. During these cycles, fluorescently labeled dNTPs are incorporated into the growing DNA chain. Each of the four dNTP species (A, C, T, G) has a single different fluorescent label which serves to identify the base and act as a reversible terminator to prevent multiple extension events. After imaging the fluorescent group is cleaved off, the reversible terminator is de-activated and the template strands are ready for the next incorporation cycle. The sequence is read by following the fluorescent signal per extension step for each cluster. Under ideal circumstances, all bases within a cluster will be extended in phase. However, a small portion of the molecules do not extend properly and fall either behind (phasing) or advance a base (pre-phasing). Over many cycles, these errors will accumulate and decrease the signal to noise ratio per cluster, causing a decrease in quality towards the ends of the reads.

The cycle time for the HiSeq2000 is approximately 1 h. The major contributor is the imaging of the flow-cell. The enzymatic reactions take very little time at all. By reducing the imaging time, the whole sequencing process can be sped up considerably. This is implemented in the miSeq and Hiseq2500 platforms by providing the option to decrease the total surface area to be imaged. In rapid mode, cycle time can thereby be reduced to 5 and 10 min for the miSeq and Hiseq 2500, respectively. Furthermore, with optimized reagent kits for these short cycle times it is possible to achieve a 2 × 300 bp paired end run on the miSeq, with 85% of data points above Q30 and run times of ~65 h. However, the increased sequencing speed does come at a price. With the decreased surface area, the total number of data points that can be generated per run will reduce, increasing sequencing cost per nucleotide significantly.

Early 2014, Illumina has announced the release of two new sequencer models, i.e., the NextSeq 500 and the HiSeq X Ten. The former system was designed to be a highly flexible, smaller version of the HiSeq2500, providing both medium (40 Gb) and a high output (120 Gb) modes both with run times under 30 h. The HiSeq X Ten was designed for one main purpose: enabling whole human genome sequencing and reaching the $1000 genome in run costs. The main advancement enabling this is the introduction of the patterned flowcells. In contrast to the spatial random cluster generation of the HiSeq and MiSeq flowcells, the X Ten flowcells contain a pre-formatted grid of nano-wells, which each can produce