Task: Hamza's research

Abstract:

- To process biological data we need a lot of computational resources and time, both of them are costly for us and we would like to process the data in the least time possible with the least resources available.
- To deal with the enormous and complex biological data, we need efficient ways to analyze that data.
- The target we are aiming for is human genome sequencing within a little unit of time.
- Achieving such a target is challenging but comparing with its results, we will be able to identify strange mutations and the causes of any genetic disease.

Introduction:

- The paper describes the advantages and limitations of current NGS platforms including those using sequencing by synthesis, sequencing by ligation, and real-time sequencing, as well as their significant impact on molecular oncology.
- The main purpose of sequencing the human genome is to obtain valuable information for future care. Genomic sequencing can provide information on genetic changes that can lead to disease or can increase the risk of disease development, even in people without symptoms.
- NGS(the next generation sequencers) enables a simultaneously and massively increased sequencing rate ranging from few gigabases per run to 6000 gigabases and therefore a possible human genome sequencing within 1 week

with only 999 US dollars according to Veritas genomic company (Müllauer, 2017; Goodwin et al, 2016).

- Current NGS is categorized into (1) systems that use sequencing by synthesis chemistry [Illumina® platforms (Illumina®, San Diego, CA, USA), Ion Torrent® platforms (Thermo Fisher Scientific, Waltham, MA, USA), QIAGEN GeneReader® (QIAGEN, Hilden, Germany), Roche® Sequencing platforms (Roche, Pleasanton, CA, USA)] and (2) systems that use sequencing by ligation [SOLiD® (Thermo Fisher, Waltham, MA, USA) and BGISEQ-500® (BGI (MGI) Tech, Shenzhen, China)] allowing short-read sequencing approaches (for review, see Goodwin et al. 2016).
- NGS ranges from the whole-genome sequencing analyzing the totality of the human genome to targeted exome sequencing and finally to focused single genetic alteration assays.
- In this paper, these methods are applied to the human genome noticing the efficiency of each one.

Related work:

- The availability of reference genomes was instrumental in the study of biology. Many competing technologies have been developed to improve the quality and robustness of genome assemblies during the past decade. The 2 widely used long-read sequencing provider Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) have recently updated their platforms: PacBio enables high-throughput HiFi reads with base-level resolution of >99%, and ONT generated reads each one equals 2 Mb. It shows that both PacBio HiFi reads and ONT ultralong reads had their own merits. Further genome reference constructions could leverage both techniques to lessen the impact of assembly errors and subsequent annotation mistakes rooted in each.

Sequencing by Synthesis Platforms:

Platforms Pyrosequencing Systems (Roche® and QIAGEN® PyroMark)

Pyro sequencing principle is based on single-nucleotide addition methods that quantify the liberated inorganic pyrophosphate (PPI) after incorporation of a nucleic base using a cascade of enzymatic reactions that produce detectable bioluminescence signals (Metzker 2010; Ronaghi et al. 1998). Instead of Sanger sequencing which needs the addition of complementary nucleotides all together at the same time into the reaction medium, pyrosequencing incorporates sequentially each known deoxyribonucleotide triphosphate (dNTP) in the elongation single-stranded amplicon by DNA polymerase. A PPI is therefore released and captured by an ATP sulfurylase to produce an ATP molecule which in turn is coupled to a luciferin to generate an oxyluciferin and light signals by luciferase-mediated conversion. An apyrase is added to the reaction wells to degrade the excess of dNTPs, and a camera called a charge-coupled device (CCD) enables high-resolution and sensitive detection of generated signals. Of note, recorded light peaks and intensity are proportional to the number of incorporated nucleotides and reveal DNA sequences using different programs (Fig. 3.1b). Before performing pyrosequencing using Roche® 454 platform, template preparation and amplification are required using a microfluidic emulsion PCR (EmPCR) technology that has the advantage to avoid loss of DNA sequences.

Illumina® Platforms

So far, Illumina is dominating the market of short-read NGS platforms as a result of its impressive high-throughput sequencing technology and low cost per base (van Dijk et al. 2014). The first NGS platform from Illumina (Genome Analyzer) was launched in 2006 by Solexa (acquired by Illumina one year later) allowing 1 gigabase/run (https://emea.illumina.com/science/technology/next-generation-sequencing/illumina-sequencing-history.html—accessed 18-05-2018). The foundation of Illumina instruments is based on sequencing by synthesis (base-by-base) technology using fluorescently labeled nucleotides. In the first step, DNA is fragmented and ligated to adapters and bound to a solid support (glass flow cell) that contains immobilized primers (two types of oligos, forward and reverse). The free end of DNA fragments interacts with close oligos, therefore creating bridges, and a clonal amplification PCR is used to generate the second strand. Finally, the bridge is denatured to form single-stranded DNA, the template is washed to remove reverse strands, and the process is repeated over again. In the second step, four differently labeled, fluorescent, and cleavable reversible terminator

dNTPs (blockade of their 3′-OH group to prevent elongation) and DNA polymerase are added to the reaction (Guo et al. 2008; Goodwin et al. 2016). Every nucleotide is incorporated one by one into the elongating strand, unbound dNTPs are washed away, and a CCD camera is used to scan and identify which nucleotide is added and another cycle is repeated (Goodwin et al. 2016).

Thermo Fisher Ion Torrent® Platforms

Ion Torrent® systems share sequencing by synthesis strategy used by other platforms such as pyrosequencing and employ a unique pH-mediated non-optical sequencing (Rothberg et al. 2011). Similar to pyrosequencing, Ion Torrent® uses EmPCR to prepare templates (Fig. 3.1a). DNA-amplified beads are incubated in microwells where sequencing takes place. Nucleotides are added into the reaction one species at a time, and if the dNTP incorporated in the elongation strand is complementary, hydrogen ions (H+) are released and induce pH changes which are detected by ion sensors [CMOS (complementary metal-oxide semiconductor) and ISFET (ion-sensitive field-effect transistor)] placed in the microwells and converted to voltage signals; the residual dNTPs are washed away and another cycle begins.

QIAGEN® GeneReader

QIAGEN® introduced its all-in-one NGS system named GeneReader in 2015 (Karow 2015). The GeneReader was developed to perform all the sequencing steps from nucleic acid extraction and clonal amplification using the QIAcube system until data analysis and interpretation workflow. Template enrichment during the preparation phase uses EmPCR as the one used by Roche® pyrosequencing, SOLiD®, and Ion Torrent® platforms. Typically, the GeneReader sequences incorporated fluorescent nucleotides by Illumina platforms and detect signals with imaging by TIRF (total internal reflection fluorescence) microscopy using laser channels (Goodwin et al. 2016). Sequencing of DNA from FFPE samples from CRC subjects using this NGS system was recently validated with reference to PCR, pyrosequencing, and Illumina MiSeq (Darwanto et al. 2017). Until this time, the GeneReader is intended for cancer clinical research use only.

Sequencing by Ligation Platforms:

Thermo Fisher SOLiD®

SOLiD NGS system was launched by Applied Biosystems Inc. in 2007 (purchased later by Thermo Fisher®) and is based on the use of two-base color encoding and sequencing by ligation strategies (Goodwin et al. 2016; Valouev et al. 2008).

BGI Complete Genomics Platforms (BGISEQ-500® and BGISEQ-50®)

BGISEQ sequencers are provided by the life sciences company “Complete Genomics” and use sequencing by ligation based on DNA nanoballs.

Real-Time Sequencing Platforms:

Pacific BioSciences® (PacBio)

Platforms in PacBio technology, template preparation avoids clonal amplification by using direct sequencing of modified DNA (Rhoads and Au 2015). DNA templates are ligated to two hairpin barcoded adapters followed by removal of templates with inadequate size using a selection process (Goodwin et al. 2016). Templates and fluorescently labeled dNTPs are then deposited in picoliter wells called zero-mode waveguide cells containing each single DNA polymerase immobilized at the bottom that can bind the hairpin adapters (Rhoads and Au 2015). The resulting light pulses corresponding to the colors emitted by the incorporated tagged nucleotides during amplification are detected and visualized using a camera and matched tags are cleaved off (Rhoads and Au 2015). With a great long read length estimated at ~20 Kb, the PacBio RS II platform is the most commonly used for this purpose, and it seems to be the gold standard for de novo assembly of genome projects (Giordano et al. 2017; Goodwin et al. 2016; Gordon et al. 2016).

Oxford Nanopore Technologies® Platforms

Oxford Nanopore Technologies® (ONT) is a rising star in real-time sequencing using pocket-sized devices. Compared to the other platforms that detect secondary signals (pH changes, light emission, or color) revealing the composition of DNA, the technology behind these long-read sequencers directly sequences

DNA fragments during their passage through a biological protein nanopore fixed on a microwell (Goodwin et al. 2016; Clarke et al. 2009).

Refrence:

https://link.springer.com/chapter/10.1007/978-3-030-53821-7_3