ADVANCED ANALYTICAL TECHNIQUES FOR NEXT GENERATION SEQUENCING: A NARRATIVE REVIEW

Tanveer Tara[1], Kamran[1], Zakir Ullah[1], Sajid Aziz[2], Asif Mehmood[3]

[1]Lecturer Institute of biological Sciences, Sarhad University of Science and Information Technology Peshawar

[2]Laboratory Technologist, Institute of Basic Medical Sciences, Khyber Medical University Peshawar

[3]Assistant Professor, Institute of biological Sciences , Sarhad University of Science and Information Technology Peshawar

Correspondence
Mr. Tanveer Tara
Lecturer Institute of biological Sciences, Sarhad University of Science and Information Technology Peshawar
Email: Tanveertara@gmail.com

ABSTRACT

The objective of current study was to provide historical perspective of DNA sequencing with focus on newer developments especially the sequencing techniques under the umbrella of next generation sequencing (NGS). Several methods of sequencing were compared based on time consuming, complexity of process, cost effectiveness, reading of base pairs per seconds/minutes, rate of errors, high quality and accuracy. Comparison of various sequencing techniques including sanger's sequencing, next generation sequencing and third generation sequencing were discussed in current review. Other sub-sequencing techniques of next generation sequencing were also described, in which Illumina sequencing was determined good range short read sequencing method with 300bp. 454 and ION torrent techniques was comparatively high read lengths. Gene Reader is fully automated from sample preparation to analysis of sample. On the side, PacBio was less time consuming in long read sequencing techniques. This paper provides enhanced understanding and judgmental comparison among different variants of sequencing technologies.

KEYWORDS: NGS, Sangers' sequencing, Illumina, Ion torrent, PacBio.
This article may be cited as: Tara T, Kamran, Ullah Z, Aziz S, Mehmood A. Advanced analytical techniques for next generation sequencing: A narrative review. Ann Allied Health Sci.2019; 5(2):3-6

INTRODUCTION

Since the discovery of the structure of Deoxyribonucleic acid (DNA), in 1953, great efforts have been made in understanding the complexity and diversity of genomes in health and disease.[1] DNA has four building blocks from which it is made of i.e. Adenine, Thymine, Cytosine and Guanine (A, T, C and G). The ceaseless repetitive order of these four building blocks (A, T, G and Cs) holds numerous secrets of life.[1,2] The accurate determination of the sequence of these building blocks is known as sequencing. The benefits of DNA sequencing spans over multiple disciplines including, medical, agriculture, social connections, evolution and human migration, and forensics. In medical arena, the genes associated with inherited diseases can be detected by DNA sequencing. Furthermore, doctors can identify and predict certain diseases from a particular order of DNA sequence. Consequently, disease oriented genes can also be replaced with healthy ones.[3] A lot of advancements have also been made in plant and animal breeding, production of crops and livestock through DNA sequencing. The productive and more immune genes towards pests and insects are detected and are exploited to gain more production and produce resistance towards different diseases. Social connections and hereditary can be found by analyzing DNA sequences.[4] In forensic area, skin, hair, blood or nail sample can be used to distinguish between innocents and criminals.[5]

Major benefits of DNA sequencing are obtained by comparisons. Exposure to diseases, inheritance role and environmental effects are yielded by information obtained through comparisons.[3] A program of National Institutes of Health (NIH) has examined gene activities that controlled different tissues and the role of gene instructions in disease. DNA sequencing has been extensively used to determine the growth of complex diseases such as cardiac arrest, paralysis, diabetes and other inherited diseases. Matching the genome sequences of unlike types of organisms, such as chimpanzees and yeast, can also deliver intuitions for biological developments.

SEQUENCING TECHNIQUES

Since 1977, DNA sequencing technologies have endured incredible developments. Though early sequencing techniques suffer from limited throughput, high cost and high time duration, but it laid foundation for latest sequencing techniques.[1,2]

Sequencing techniques can be classified as follows:

- Sanger sequencing
- Next Generation Sequencing
- Third Generation Sequencing

The sequencing can also be classified on the basis of their read length into two groups

- Short read length sequencing
- Long Read length sequencing

Long read sequencing are normally based on "single-molecule real-time sequencing approaches" e.g. PacBio and ONT. Or "a synthetic approach that depends upon shortread sequencing methods for constructing long reads in silico".

SANGER SEQUENCING

In 1970s, Dr. Frederick Sanger concocted a method for the DNA's sequencing known as "Sanger's sequencing" technique.[1,2] It was based on chain termination method. It was firstly used commercial and laboratory method for sequencing DNA. It required radio-active material and it was a laborious task. Sangers' sequencing was holding excellent accuracy as well as reasonable read length but its throughput was very low. It requires radio-active material and hence was a laborious task. In 1987 Applied Bio

System company introduced a capillary electrophoresis based first automatic sequencing technique namely AB370. Human Genome Project used all of the methods to complete and lead to the development of NGS (Next generation sequencing).[2]

NEXT GENERATION SEQUENCING (NGS)

Next generation sequencing (NGS) technologies are becoming popular in clinical sectors due to their high throughput, low cost and capacity to tackle the complexities of genomes. NGS are capable of extracting sequence data from single DNA molecules. They can provide physicians with the tools to translate genomic information onto clinical results. A common NGS strategy to read many different parallel DNA templates is to use either DNA synthesis or ligation process.[3] Next generation sequencing technologies provides better speed and throughput capacity of DNA.[4] NGS has made human genome sequencing an economical task and can be done in around $1000. This economical factor has brought the sequencing in clinics, where NGS can be used e.g. in broad detection of pathogens. NGS provides large data but with high error rate (0.1 to 15 %) and shorter read lengths (35 to 700bp) than traditional approaches.[5] The sample preparation method used by NGS for DNA or RNA differs from prior techniques.

There are multiple NGS platforms currently available:

- 454 pyro-sequencing (Roche)
- Illumina/Solexa Genome Analyzer
- ABI SOLiD analyzer
- GeneReader (Qiagen)
- Polonator G.007
- HelicosHeliScope
- Ion Torrent

Helicos and 454 pyro-sequencing are obsolete now and setback NGS.

NGS ROCHE 454

Roche 454 was launched by 454-Life-science in 2005. It was the first NGS sequencer which took revolution in DNA sequencing. It is based on the principle of Pyro-sequencing technology (Sequencing by synthesis) in contrast to Sanger's method which uses di-deoxy nucleotide for terminating chain amplification.[6] 454 provides average read length of 700bp which is superior to short range sequencers.(while ION torrent average read length is 400bp which is also better than short range sequencers). 454 platform depends upon SNA (Single nucleotide addition) technology. Roche 454 uses *Emulsion PCR* technique. This method uses enzymatic cascade from where signal will be generated. (Ion torrent does not use this method, rather it detects H+ ion which is released after incorporating dNTP).

METHODS

At first double stranded DNA helix is break down into single strand. These strands are captured by amplification beads which then go through the PCR (Polymerase chain reaction: numerous copies of DNA fragment on each bead, hence create millions of copies of DNA sequence) emulsion process, passing through a procedure it finally generates visible light of oxy-luciferin.[7] CCD (Charged coupled device camera) record this light. The intensity of light tells about one or more identical nucleotides (dNTPs).[7]

SOFTWARE

54 GS FLX Titanium Software is used for it. This software can take picture of background as well as it perform other tasks including to normalize background, to correct location of signal, make correction of cross-talk, can perform signal conversion, similarly the task of sequencing data generation.GS Run-Processor produces standard flow-gram format (SFF) which finally can be converted into FASTQ format for further data analyzing.

PROS

- In 2009 read length of Roche 454 GS Flix was 700bp with accuracy 99.9% and 14 G data per run within 24 hours.[8] Use of emulsion PCR instead of general cloning (used in sangers' sequencing) reduces time of sequencing from week to days or hours.[9]
- The Roche completes sequencing in just 10 hours from sequencing start till completion. The read length is also a distinguished character compared with other NGS systems.
- Used for structural variation studies such as indel (insertion, deletion), inversions.
- Useful for high resolution detection

CONS

- In case of similar bases (such as AAAA), it becomes difficult for Roche 454 to distinguish them.
- 454 can't compete with rapidly evolving NGS technologies in terms of yield and cost. This proved as a big limitation for 454 thus Roche had to discontinue from 2016 platform. (While Ion Torrent platform kept pace with rapidly evolving NGS.[10]
- 454 use high cost reagents, almost 12.56×10^{-6} per base.
- It could have relatively high error rate in terms of poly-bases which are longer than 6bp, but this drawback can be reduced with automated library and semi-automated polymerase chain reaction).

ILLUMINA/ HiSeq

Illumia is an amplified single molecule short read sequencing technique. It was introduced commercially, in 2007, after Roche 454 (2005). This method is based on sequencing by synthesis (SBS) approach (simultaneously add all four nucleotides along with polumerase). Its uses the principle of cyclic reversible termination (CRT) (in which terminator molecules are used to block ribose 3'-OH group) to prevent elongation. Illumina is an enhanced nanopore sequencing technique. It's read length ranges up to 300bp. Linearization enzyme or polymerase binding is used to convert DNA into single strand and then grafted to the flow cell, The bridge amplification process forms the clones of these DNA fragments in clusters. Each cluster is emit a color for specific base and after that cleavage is performed to repeat the process. Illumina identifies nucleotide dNTP using "total internal reflection fluorescence- TIRF" microscopy. In short read sequencing methods, Illumina sequencing is at edge from others due to its cross-platform compatibility and good range of platforms. The instruments that use illumine includes MiniSeq (low

throughput) and HiSeq X (for ultra high throughput).

PROS

- It is known as most easiest and adaptable sequencing platform.[4]
- Low error rate and high quality
- Lowest cost per base
- Tons of data can be processes quickly
- HiSeqX has ultra high throughput

CONS

- This technique is used very large scale
- It has short read length (50-300bp),
- Runs take multiple days,
- It has high startup costs,
- De Novo assembly difficult
- HiSeq X applications are very limited due to its optimization.
- HiSeqX is an all purpose instrument so users need to purchase 5 to 10 additional instruments to use it; which make it unreachable for some users.

ABI-SOLID (APPLIED BIO-SYSTEM - SEQUENCING BY OLIGO LIGATION DETECTION)

- ABI-SOLiD is a Next generation sequencing technique. It is an amplified single molecule sequencing technique. It is based on 'Sequence by Ligation' (SBL) and two-base coding approach to determine DNA sequence composition. In this sequencing technique SOLiDflowcell, ligation site (the first base), cleavage site (the fifth base), and 4 different fluorescent dyes (linked to the last base) are used for elongation.[11] It is one of the commercial high throughputs sequencing platform. In ABI SOLiD, amplification of DNA fragments is not performed using propagation of bacterial clone libraries as was in sangers' method.

METHODS

- DNA fragmented primed libraries are embedded on micro-beads using emPCR. These beads are adhered into a slide (made up of glass). This flow cell is now added with 1,2-probes (of different fluorophore). A matching 1,2-probe is ligated to the primer by DNA ligase. Fluorescence image determine that which probes is ligated, after that silver ions cleave the phosphorothiolate link, which re-generate 5′ phosphate group for next ligations.

PROS

- It has targeted re-sequencing
- It supports gene expression profiling.

CONS

- It is short read sequencing technique.
- ABI-SOLiD has high cost approximately $4 per 800bp.

GENEREADER (QIAGEN)

It is NGS technique based on SBS (Sequencing by synthesis) approach. Gene Reader was launched by Qiagen in 2015. Gene Reader is an updated form of Intelligent Biosystem CRT platform. This platform was intended to be automated NGS platform, with an ability to perform all steps i.e. from sample preparation to analysis in a single machine. Thus it has built in two components QIAcube system (for sample preparation) and the Qiagen Clinical Insight platform (for analysis).[12] Unlike Illumina, in GeneReader every template is not labeled with fluorophored NTP for identification. This technique is intended to be a used for cancer gene clinical device.

PROS

- It does not need separate or pre-sample preparation.
- It is also an economical NGS technique.

CONS

- It has limited applications in market due to its targeted cancer oriented approach.

POLONATOR G.007

This NGS technology was developed by Harvard University and Dover Systems and for the first time it was launched by Azco Biotech. Polonator uses emPCR for DNA template amplification (emPCR is shown in figure 4).[13] This technology is also based on SBL approach. It is a short read sequencer with read length 40bp with 80 hours run time.[14] Its protocol includes paired end tag library construction followed by template amplification and then DNA sequencing.

PROS

- It is an economical and accurate technique
- It provides bench-top open source platform.
- Currently the major system for the Personal Genome Project

CONS

- Although it collects a huge data in a specific time bit only 1 bit of information out of 10,000 bits is useful
- Non-uniformity of the relative amplification of individual targets lowers its efficiency.

HELICOS GENETIC ANALYSIS SYSTEM

This is the first commercially available NGS technology based on single nucleotide addition. It limits or control the incorporation of nucleotides using virtual terminators. Helicos has read 35bp length short read sequencing. However, it has been abandoned.

ION TORRENT

ION torrent is another NGS technology. It is based on Amplified single molecule sequencing technique with the read length of 100-400 bp. It has an accuracy of 99.5% and output data of approximately 1 Gb. It is an Emulsion PCR based technique under sequencing by synthesis.

ION Torrent don't use an optical sensing approach but it detect H^+ ion which are released on incorporation of dNTP nucleotide, the change in pH is recorded by Complementary Metal Oxide Semiconductor and Ion Sensitive Field Effect Transistor but it provide limited accuracy.[15]

PROS

- ION Torrent is fully scalable.
- Have high read-length then some other short-read NGS technologies.
- There is no pause after the detection of a base or series of bases
- The sequences are derived in real-time.
- Its platform without optical sensing
- Price = 25-2400$
- Supported sequencing types :

- Targeted Sequencing
- Exome Sequencing
- Transcriptome Sequencing
- Microbial Sequencing,
- Aneuploidy Sequencing,
- Genotyping by sequencing,
- Bacterial Typing,
- Viral Typing,
- Denovo Sequencing,
- Small RNA and miRNA sequencing

CONS:

- Rely on SNA

ION TORRENT

PacBio is single molecule real-time sequencing technique. It is longest read length with an ability to detect base modifications. PacBio is a technique with short run time.

CONS

- High cost per Mb
- High capital cost
- High error rates
- It has low number of reads per run

OXFORD NANOPORE

Oxford Nano-pore is third generation (next-next generation) single molecule sequencing platform. It's read length ranges from 230,000 bases to 200Kb. Its sequencing technique is Real-time sequencing /Nano-pore sequencing. Its accuracy is about 99.99%. It is fully scalable. This sequencing platform promises to decrease costs for reagents and instrumentation

Oxford Nanopore Technologies includes MiniON and GridION System.

CONCLUSION

This review describes about Next generation sequencing (NGS) technologies. NGS can be categorized in short read sequencing and long read sequencing. Illumina sequencing is at edge from others short read sequencing methods due to its cross-platform compatibility and good range of platforms having read length about 300bp. ION Torrent and 454 (Short read sequencing) technologies have relatively high read length 700 and 400 respectively as compared to other short read-length NGS technologies. ION Torrent detect H^+ ion for detection of bases while other techniques use optical sensing approach. GeneReader is an automated NGS platform, with an ability to perform sample preparation as well as analysis. It is a devise for cancer genes. Illumina (HiSeq X) provides ultra high throughput, low error rate and high quality as well as lowest cost per base but its applications are limited and it requires additional instruments for different purpose uses. In long read sequencing techniques, PacBoi has short run time.

REFERENCES

1. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. 1975 May 25;94(3):441-8.
2. Sanger F. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA. 1983;80:2432-6.
3. Fuller CW, Middendorf LR, Benner SA, Church GM, Harris T, Huang X, Jovanovich SB, Nelson JR, Schloss JA, Schwartz DC, Vezenov DV. The challenges of sequencing by synthesis. Nat. Biotechnol. 2009 Nov;27(11):1013.
4. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. J. of Genetics and Genomics. 2011 Mar 20;38(3):95-109.
5. Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. Comparison of next-generation sequencing systems. BioMed Res. Int. 2012 Jul 5;2012.
6. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB. Corrigendum: Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2006 May;441(7089):120.
7. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet. 2016 Jun;17(6):333.
8. Leamon JH, Lee WL, Tartaro KR, Lanza JR, Sarkis GJ, deWinter AD, Berka J, Lohman KL. A massively parallel PicoTiterPlate™ based platform for discrete picoliter-scale polymerase chain reactions. Electrophoresis. 2003 Nov;24(21):3769-77.
9. 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. 2003 Jul 22;100(15):8817-22.
10. Reporter AG. Roche shutting down 454 sequencing business. GenomeWeb Daily News. https://www. genomeweb. com/sequencing/roche-shutting-down-454-sequencing-business. 2013.
11. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 2008 Jul 1;18(7):1051-63.
12. Meldrum C, Doyle MA, Tothill RW. Next-generation sequencing for cancer diagnostics: a practical perspective. The Clinical Biochemist Reviews. 2011 Nov;32(4):177.
13. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005 Sep 9;309(5741):1728-32.
14. Deschamps S, Campbell MA. Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Molecular breeding. 2010 Apr 1;25(4):553-70.
1. Smith DR, McKernan K, inventors; Applied Biosystems LLC, assignee. Methods of producing and sequencing modified polynucleotides. U.S patent 8,058,030. 2011 Nov 15.