Next-Generation Sequencing: A Boon to Plant Science

Subaran Singh[1]* and Supriya Ambawat[2]

[1] College of Horticulture, Maharana Pratap Horticultural University, Karnal- 132001 (Haryana)

[2] ICAR-AICRP on Pearl Millet, Agriculture University, Jodhpur -342304 (Rajasthan)

Corresponding Author: subarans.apbiotech@mhu.ac.in

Abstract

Several scientific discoveries made by the use of next-generation DNA sequencing technologies emphasized the remarkable impact of these extremely useful platforms on genomics. These new methods have expanded previously focused readouts from a variety of DNA preparation protocols to a genome-wide scale and have fine tuned their resolution to single base precision. The novel sequencing techniques with increased accuracy and cost-effectiveness have been developed and now the sequencing cost of an entire genome has crossed the golden line of 1000-US$. There is an enormous augmentation in amount of sequencing data after next-generation sequencing (NGS) came into existence. There are different NGS techniques available and has different types of applications in plant science. Next-generation sequencing has also enabled novel applications such as the sequencing of ancient DNA samples and has substantially widened the scope of metagenomic analysis of environmentally derived samples. As a whole, an astonishing potential exists for these technologies to bring enormous change in genetic and biological research and to enhance our fundamental biological knowledge.

Introduction

Nucleic acid sequencing is a method for determining the exact order of nucleotides present in a given DNA or RNA molecule. The determination of base sequence of a DNA fragment is called DNA sequencing. DNA sequencing became feasible due to the availability of restriction enzymes, development of electrophoresis techniques capable of separating DNA fragments differing by a single nucleotide, and gene cloning and PCR techniques that make available very large number of copies of individual DNA fragments required for sequencing. Initially, two methods-chemical and enzymatic methods of DNA sequencing were developed. These methods are popularly termed as first-generation DNA sequencing procedures. Soon the second or next-generation DNA sequencing (NGS) methods were developed which use PCR for *in vitro* cloning in place of *in vivo* cloning and are much faster and cheaper. At present, the third-generation DNA sequencing (TGS) methods are becoming commercially available which sequence single DNA molecules without any cloning. During last 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. NGS techniques play a crucial role in understanding the biology of plants against various biotic and abiotic stresses through their role in genomics, epigenomics and transcriptomics studies.

The sequencing of the reference human genome was the capstone for many years of hard work spent developing high-throughput, high-capacity production DNA sequencing and associated sequence finishing pipelines. The approach used >20,000 large bacterial artificial chromosome (BAC) clones that each contained an approximately 100-kb fragment of the human genome, which together provided an over-lapping set or tiling path through each human chromosome as determined by physical mapping. In BAC-based sequencing, each BAC clone is amplified in bacterial culture, isolated in large quantities and sheared to produce size selected pieces of approximately 2-3 kb. These pieces are sub cloned into plasmid vectors, amplified in bacterial culture and the DNA is selectively extracted prior to sequencing. By generating approximately eightfold coverage of each BAC clone in plasmid sub-clone equivalents, computer-aided assembly can largely recreate the BAC insert sequence in contigs (contiguous stretches of assembled sequence reads). The substantive changes have occurred in the approach to genome sequencing that has moved away from BAC-based approaches and toward whole-genome sequencing (WGS), with changes in the accompanying assembly algorithms.

The NGS techniques are often classified as second and third generation sequencing technologies. But there is no consistent definition for classification of this terminology. In general, second-generation sequencing techniques refers to those methods which require a PCR step for signal intensification prior to sequencing and third generation sequencing

techniques are those which can do single molecule sequencing. Three platforms for massively parallel DNA sequencing read production are in reasonably widespread use at present: Roche/454 FLX, Illumina/ Solexa Genome Analyzer, Applied Biosystems SOLiD™ System. The Helicos Heliscope™, Pacific Biosciences SMRT and nanopore sequencing techniques are also included among the recent and more advanced techniques.

Different types of next-generation DNA sequencing techniques

The different types of DNA sequencing techniques can be categorized as first-generation sequencing techniques (FGSTs), second generation sequencing techniques (SGSTs), third generation sequencing techniques (SGSTs). They differ in their method of sequencing and read length (Table 1).

Table 1. Various types of sequencing methods and their read length

Name of Technique	Read length	Method
First generation sequencing techniques		
Maxam-Gillbert sequencing	400bp	Chemical method
Sanger sequencing	700 to 800bp	Enzymatic Chain termination method
Second generation sequencing techniques		
Roche-454 (Pyrosequencing)	400bp	Sequencing by synthesis
Illumina-Solexa	250bp	Sequencing by synthesis
Ion Torrent sequencing	100 to 200bp	Sequencing by synthesis
SOLiD sequencing	50bp	Sequencing by ligation
Polony sequencing	26bp	Sequencing by ligation
Third generation sequencing techniques		
Helicos (SMDS)	35bp	Sequencing by synthesis
Pacific Biosciences (SMRT)	1-10kb	Sequencing by synthesis
Nanopore sequencing	10-100kb	Oxford Nanopore

First generation sequencing techniques (FGSTs)

The Sanger's chain termination sequencing technique and the Maxam-Gillbert chemical-based sequencing technique are placed in this category. Due to use of less toxic chemicals and various other improvements, the Sanger's sequencing technique was used extensively in earlier times before discovery of different NGS techniques and regarded as classical DNA sequencing technique.

Second generation sequencing techniques (SGSTs)

The SGSTs are introduced as Next Generation Sequencing (NGS) techniques with lesser cost, high speed and high-throughput compared to FGSTs. SGSTs include- Roche 454 (pyrosequencing) (2005), Solexa/Illumina sequencing (2006), SOLiD sequencing (2007) and Ion Torrent (2008). These techniques are different from FGSTs and require simpler PCR based amplification steps for preparation of sequencing libraries. These PCRs are generally done in nanolitre volume with template DNA bound to a microbead or surface of the glass plate. The sequencing chemistry of various SGSTs is different and therefore their accuracy, speed, read length, throughput and cost are different. These SGSTs produce thousands to billions of 25-800 nucleotide long reads overnight with low cost compared to Sanger sequencing. These SGSTs can sequence entire human genome overnight following shotgun sequencing approach. The different second-generation sequencing techniques are described below:

Roche 454 (pyrosequencing)

Pyrosequencing technique is based on the detection of pyrophosphate released after the addition of each nucleotide to the growing DNA strand by DNA polymerase (sequencing by synthesis reaction). The inorganic PPi released is converted to ATP-by-ATP sulfurylase enzyme and finally the luciferase enzyme uses the ATP to generate light which is then detected by a Charged Couple Device (CCD) camera. First of all, the DNA to be sequenced is fractionated into small size (300-400 bp) and then linkers are ligated to both the ends. Single stranded DNA is isolated and captured on the microbeads. These microbeads containing DNA are then amplified in emulsified (water in oil) mixture with amplification reaction. Finally, microbeads are loaded onto a PicoTiter plate for sequencing.

Illumina-Solexa Sequencing

It uses solid phase bridge amplification for library construction for sequencing. (http://www.illumina.com). The sequencing chemistry of this technique is reversible cycling sequencing (sequencing by synthesis reaction). Elongation of DNA is done by cyclic process of adding 4nt (labeled with four different fluorescence dyes and modified 3′-O-azidomethyl reversible terminator), detecting the signal of nucleotide incorporated complementary to the natural DNA followed by removal of azidomethyl group from 3′-O and washing.

SOLiD (Support Oligonucleotide Ligation Detection)

It was developed by Life Technologies/Applied Biosystems. The preparation of DNA for sequencing (library preparation) is similar to pyrosequencing by emulsifying PCR but the sequencing chemistry is different. It exploits the mismatch sensitivity of DNA ligase. The fluorescent tagged oligonucleotide probes of varying length are hybridized with template single stranded DNA. DNA ligase is then added to the flow cell and signal is detected after washing step. This cycle is repeated and the sequence of nucleotide template DNA is determined.

Polony sequencing

Polony sequencing method was developed by George Church group at Harvard University. It uses polony amplification for library preparation for high throughput DNA sequencing. Polony amplification is a method in which DNA is amplified on a thin Polyacrilamide film so that the amplified DNA is localized in the gel and form polymerase colonies (or Polonies). The DNA is then sequenced by sequencing-by-synthesis technologies using fluorescent labeled nucleotides, sometimes referred as FISSEQ (fluorescent in situ sequencing). Scientists have developed a technology combining polony amplification and sequence-by-ligation method to sequence 14-base tags, named Polony Multiplexed Analysis of Gene Expression (PMAGE).

Ion torrent

It also uses clonal amplification of template DNA for preparing sequencing libraries. The multiplexed ion sensors used to detect hydrogen atoms released during the addition of nucleotide to the growing strand (sequencing by synthesis reaction). It is unique method as the detection is done by sensing the change in pH rather detecting fluorescent signals as used in other SGSTs. Ion Torrent may be considered as a TGST because it uses electronic signals for detecting addition of nucleotides, unlike other SGSTs which uses fluorescent dye based detection system.

The SGSTs are however, produce small reads imposing huge challenges for analysis and their accuracy also varies due to their dependency on multiple multiplication steps to make template DNA in a form that can be detected by electro-chemical signals by various sensing mechanisms. Each manipulation introduced to the natural DNA causes various artifacts in DNA measurements. Therefore novel technologies are being designed in such ways that involve minimum or no manipulation of the natural DNA molecule..

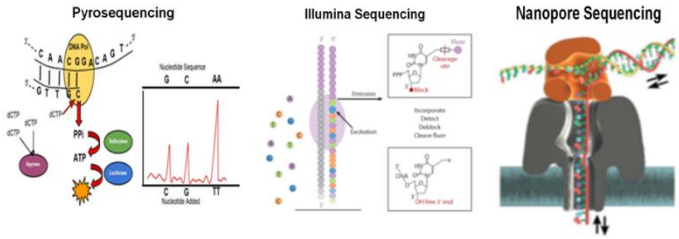

Fig 1. Different types of NGS sequencing methods

Third generation sequencing techniques (TGSTs)

The advancement in automated single molecule imaging and nanofluidics technologies pave the way of development of TGSTs, which can analyze natural DNA and RNA molecules without any manipulation in a high throughput manner. Direct sequencing of natural DNA with minimal or no manipulation prevents the inaccuracies caused by SGSTs and generate longer read length. The different third generation sequencing techniques are mentioned below:

Helicos Biosciences (SMDS: Single Molecule DNA Sequencing)

Helicos Biosciences was the first commercial TGST which can perform single molecule DNA sequencing (SMDS) and direct RNA sequencing (DRS). This technique was developed by Stephen R Quake laboratories based on sequencing-by-synthesis chemistry. DNA is isolated, fractionated and ligated with poly-A tail. The sequencing platform has a flow-cell coated with poly-(dT) oligonucleotide on which poly-A tail ligated DNA is fixed. After template

captured to flow-cells, the sequencing-by-synthesis is started with incubating the templates DNA molecule with one of the four Virtual Terminator™ (labeled with fluorescent dye and a chemically cleavable inhibitor) nucleotide and DNA polymerase in cyclic manner, with alternating the Virtual Terminator™ in subsequent cycles. The terminating dye moiety is removed after each cycle and the process repeated until read length is reached. The single molecular nature of sequencing avoid extensive manipulation steps of template DNA resulting no GC bias and better coverage compared to SGSTs. However, with the shorter read length (up to 55 nucleotides), it posses challenge for certain applications like detection of alternative splicing, sequencing of microbiome and *de nove* genome sequencing.

Pacific Biosciences (SMRT: Single Molecule Real Time Sequencing)

The SMRT technology was developed by Steve Turner and Jonas Korlach at Cornell University in mid 90s. The basic concept was to observe real time incorporation of fluorescently labeled nucleotides into growing strand by DNA polymerase. The advancement in biochemistry, photonics, nanofluidics , integration semiconductor processing and the zero-mode waveguide (ZMW) technique make it possible to observe fast incorporation kinetics of polymerase to a single molecule scale. ZMWs are microwells placed in a metal on a silica surface. Due to specific behavior of light travelled in small well, the illumination of bottom of the ZMW acquiesce detection of single without significant interference from upper portion of the wells. A single molecule of DNA polymerase molecule is immobilized to the bottom of the well and DNA sequence is loaded to it. Now the differently fluorescent labeled nucleotides are added to the growing strand by DNA polymerase and real time observation is done by ZMW imaging system.

The SMRT nucleotides contain fluorescently labeled group on its phosphate moiety, so it released with the pyrophosphate (PPi) release during nucleotide incorporation to the growing strand and through diffusion it goes outside the well and new nucleotide come to well. One of the limiting factors of SMRT sequencing technique is the polymerase damage by the use of high strength laser for imaging. The limitations of current optical imaging techniques, which cannot observe the incorporation of nucleotide by DNA polymerase at naturally speed (≈1000 nt/second) is another challenge to this sequencing technique.

Oxford Nanopore

The Nanopore sequencing refers to sequencing of a single DNA molecule by electrophoretically passing it through nano scale pore**.** Like other TGSTs, nanopore sequencing technique does not require an amplification step for the preparation of sequencing libraries. In this technique, a voltage bias is introduced across a nanopore which creates an electric field and it could drive ssDNA molecule through the nanopore. The passage of ssDNA is detected by measuring change in ionic current and duration across the nanopore. Nanopore sequencing has huge potential because it can use very low quantities of DNA and can sequence it directly with least or no manipulation to obtain very long reads. With this technique, original DNA/RNA molecules can be detected directly whereas in other sequencing-by-synthesis techniques, the copy of template synthesized during reaction is detected.

Applications of NGS

- Molecular marker development
- Phylogenetic and ecological studies
- Allele mining
- Elucidating DNA-Protein interactions
- Gene expression
- Genomes resurrection
- Epigenetics regulations
- Metagenomic analysis

Future prospects

The advantages of NGS techniques have demonstrated the application of diverse sequencing based novel approaches to study plant biology by various institutes and organizations. They provide opportunity to conduct small as well as large sequencing-based research at the same time. It has paved the way of improving applied science as well as deepening our understanding of basic functioning of plant genome. The TGSTs have opportunity to discover the unexplored area of genomics that cannot be analyzed by SGSTs. But, TGSTs are currently struggling to reduce their costs comparable to SGSTs and hopefully in future they will be more commercially available to provide their services.

Conclusion

The potential of various second and third generation sequencing techniques can be seen by its broad area of applications to study the plant biology. With the projected population of >9 billion people in 2050 and various climate change issues, it will be very challenging to produce sufficient amount of food, fiber, feed and fuel for human requirement. The increasing understanding of plant biology and their interaction with different biotic and abiotic factors has utmost importance to ensure our future needs. With decreasing costs and error rate and increasing throughput and sequencing read length the NGS techniques have even greater potential to be realized with ever increasing ability of data handling processing and analysis.

References

Braslavsky I, Hebert B, Kartalov E, Quake SR (2003) Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences* 100 (7): 3960-3964.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133-138.

Healy K (2007) Nanopore-based single-molecule DNA analysis.

Hutchison CA (2007) DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research* 35(18): 6227-6237.

Kim JB, Porreca GJ, Song L, Greenway SC, Gorham JM, Church GM, Seidman CE, Seidman J (2007) Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. *Science* 316:1481-1484.

Landegren U, Kaiser R, Sanders J, Hood L (1988) A ligase-mediated gene detection technique. *Science* 241:1077-1080.

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences* 74(2):560-564.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* 242 (1):84-89.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature* 475:348-352.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 74(12):5463-5467.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology* 26(10):1135-1145.

* * * * * * * * *