Editorial
Published: 10 Jul, 2020

SF Journal of Agricultural and Crop Management

Next Generation Sequencing: Prospects in Plant Breeding and Crop Improvement

Aryadeep Roychoudhury*

Department of Biotechnology, St. Xavier's College, 30, Mother Teresa Sarani, Kolkata, West Bengal, India

The world population is expected to reach around nine billion by 2050 [1] so that increased yield and productivity of crops is of paramount importance to mitigate world hunger, malnourishment and also protect crops against abiotic and biotic stresses in the face of global climate change. For a long time, crop improvement programs mostly relied on breeding techniques to incorporate superior traits through hybridization, marker-assisted selection and genetic engineering to develop transgenic plants with one or more desirable genes through several plant transformation techniques. Identification of different molecular markers and Quantitative Trait Loci (QTL), in combination with conventional phenotype-based selection, formed the bases of plant improvement practices. With the passage of time, whole genome sequencing projects, initially that of *Arabidopsis thaliana* and rice, and later many other crop species including tomato, based on Sanger's sequencing method, have led to the identification of sequence-based Single Nucleotide Polymorphism (SNP), thereby increasing the number of suitable informative markers. The major problem with this approach was that it was extremely time-consuming, expensive with limited use in gene discovery. With the advent of Next Generation Sequencing (NGS) technology and sophisticated computer pipelines, the quality of whole genome sequencing of crops has largely improved because numerous DNA sequence polymorphism-based markers can be detected within a short time frame. The NGS platform has enabled investigation of functional and evolutionary divergence among plant species with proper understanding of genome complexity and relationships between genotypes and evolution [2]. Thus, domesticated and economically important plants can be sequenced to identify millions of new markers and agronomically important genes for their direct utilization in crop improvement.

Towards the beginning of the 21st century, the first commercially available next generation instrument, also called 454 sequencing, was launched by the Roche company. The basis of this technology, also called pyrosequencing, was sequencing by synthesis or chemi-luminiescent detection of pyrophosphate released during deoxynucleotide triphosphate (dNTP) incorporation. The single stranded DNA is fractionated into smaller fragments (300-1000 bp), made blunt- ended, followed by ligating short oligo adapters having 5' biotin tag. These adapters provide priming sequence for the attachment, amplification as well as sequencing the fragment. The yield of DNA sequence is measured in 'pyrogram' corresponding to the incorporated nucleotide order. This technology proved to be impressive for metagenomics and *de novo* sequencing assembly where longer read length is necessary, *viz*., almost 10 lakh individual reads are possible at the rate of 500-800 bases per 10 h run. However, homopolymeric or repetitive regions which cause dephasing due to asynchronous synthesis are difficult to be sequenced through pyrosequencing, increasing the sequence length and error rate [3]. This was followed by the launching of Illumina genome analyzer which utilizes reversible dye-terminator based sequencing by synthesis method *via* bridge amplification that enable identification of single bases as they are introduced into DNA strands. The automated nature of Illumina sequencing enables sequencing of multiple strands to obtain actual sequencing data very rapidly. Moreover, it requires only DNA polymerase, and not multiple costly enzymes as needed in pyrosequencing. This method can be successfully used for whole genome and transcriptome analysis, genome-wide protein-nucleic acid interaction analysis, small RNA (sRNA) discovery, methylation profiling and epigenomics. Illumina offers several sequencing platforms from the benchtop MiniSeq and MiSeq to large-scale HiSeqX and NovaSeq systems with extraordinarily high throughput and coverage (http://www.illumina.com/). The third next-generation sequencing, launched in 2007 was called SOLiD (Sequencing by Oligo Ligation and Detection), which employ the method of sequencing by ligation, instead of synthesis. A significantly smaller size of beads is adopted for DNA amplification with ordered array format. A mixture of different fluorescently labeled dinucleotide probes is pumped into the flow cell; when the correct dinucleotide probe incorporates the DNA template, it is ligated onto the pre-built primer on the solid phase. Each fluorescent wavelength corresponds to a particular dinucleotide combination [4].

OPEN ACCESS

Correspondence:
Aryadeep Roychoudhury, Department of Biotechnology, St. Xavier's College, 30, Mother Teresa Sarani, Kolkata, West Bengal, India.
E-mail: *aryadeep.rc@gmail.com*
Received Date: *28 Jun 2020*
Accepted Date: *06 Jul 2020*
Published Date: *10 Jul 2020*

Citation: *Aryadeep Roychoudhury. Next Generation Sequencing: Prospects in Plant Breeding and Crop Improvement. SF J Agri Crop Manag. 2020; 1(1): 1004.*

This technology generates a high data output (>60 Gb) from large numbers of short reads (~ 50 bp) with fairly high accuracy. However, short read lengths can cause an inadequate cover of repeat regions. The ligation platform, the Polonator G.007 produces 10-35 Gb of data per 2.5 day run. In the present times, the market is enriched with third generation technologies, which produce over 10,000 bp reads, allowing *de novo* assemblies and contiguous reconstruction of genomes rich in repetitive elements. The most advanced sequencing technologies nowadays include semiconductor sequencing by Ion Torrent and nano-pore sequencing. The Ion Torrent works on simple biochemistry independent of light. After nucleotide incorporation into the growing DNA strand by polymerase, H^+ ions are released as a byproduct causing a detectable local change in pH, which can be detected by ion sensor. The change in pH is directly proportional to the number of nucleotides added [5]. In case of nanopore sequencing, DNA is placed on thin membrane containing nano-pores of different materials (carbon nanotubes and thin films, α-hemolysin protein nano-pores, plastic materials) of 1.5-2.0 mm diameter. Upon applying current across the membrane, DNA molecules with different sequences can be discriminated based on alteration in electrical conductances due to channel blocking by the migrating DNA through the membrane. This method eliminates the use of fluorescent nucleotides, enzymes, cloning and amplification steps with easier way of sample preparation [6]. Single Molecule Real Time (SMRT) sequencing is another third-generation technique that can read the base sequence directly from individual strands of DNA present in a sample, without the need of an amplification step. Only one DNA template and one DNA polymerase can be immobilized in a hole, present in an aluminium cladding film called Zeromode Waveguide (ZMW) on a silica substrate (SMRT chip). When a nucleotide is incorporated during DNA synthesis, the laser beam illumination of small detection volume (20 zeptoliters=20×10^{-21}) causes fluorophores attached to the bases to light up, allowing the identification of each nucleotide incorporated, depending on the light pulse. This technique requires minimum amount of sample and reagent without any PCR amplification bias [7]. Several bioinformatic tools like GLIMMER, GenMark, GO, FGENES, etc. have also been developed, parallel with sophisticated NGS technologies, in order to analyze and manage a huge amount of sequence data per run in a high throughput manner. The steps in NGS data analysis include quality assessment, alignment, variant identification, variant annotation and data visualization.

Advances in NGS technologies have revolutionized *de novo* sequencing and re-sequencing of various crop species and also crop genomics in several areas like plant breeding, mutation mapping, developing Single Nucleotide Polymorphism (SNP)-based markers, transcript profiling of small RNAs, studying protein-DNA interaction, metagenomic studies, epigenetic modification and gene mining. Several hybrid methods based on pyrosequencing and Sanger long pairs could be used as strategy for developing *de novo* assembly algorithms like ABySS, ALLPATHS, Velvet, SOAPdenovo, CLC Bio's *de novo* assembler, etc., [8]. The *de novo* assembly of whole genome sequences has made reference genomes of many crop species available, so that agronomically important genes and SNPs can be easily identified *via* bioinformatic tools based on reference genomes and sequences from different cultivars. Studying genetic variations among populations, like insertions/deletions (indels), Structural Variation (SV) including translocations and chromosome fusions, and Copy Number Variations (CNVs), population structure and linkage disequilibrium are also possible which has proved beneficial in studying the evolutionary history of a crop species, their adaptability to environmental conditions and natural selection at the population level [9]. Metagenomics approaches and the sequencing of pooled amplicons generated for a large number of candidate genes across large populations offer possibilities in better understanding population biology and to study genome-wide association genetics. Marker Assisted Selection (MAS) technique has been improved since genome-wide genotyping method and Genotyping-By-Sequencing (GBS), using reference mapping can be applied because of the production of several molecular markers within a short period of time. GBS is considered as a robust approach in sequencing multiplexed samples, and has been used in development of high density map of 20000 SNPs in wheat and 34000 SNPs in barley [10]. A tool called Circos was used for comparing pigeon pea and soybean genomes, as well as visualizing the contigs of *Glycine max* var. Sinpaldalkong 2, mapped onto the reference genome (*G. max* var. Williams 82), suggesting that there was a recent duplication of the Sinpaldalkong 2 genome [11]. In the area of transcriptome analysis, RNA-Seq enables gene expression profiling under different environmental conditions or in different plant tissues, so that even rare and novel transcripts can be identified [12]. The assembly of transcriptomes can be mapped by using computer programs like SOAP, RMAP, Elnad, MAQ, SHRimp, SSAHA2, TopHat, Stampy, RNA-MATE, etc, whereas clustering analysis of transcriptomes are possible *via* MIRA, CLC Bio, BLAT, CAP3, TGICL, etc. Transcriptome of red clover (*Trifolium pratense* L.), for example, was sequenced using Illumina technology and genes responsible for drought tolerance were discovered. The concentration of the three metabolites (pinitol, proline, and malate) was increased in leaves as an impact of drought stress [13]. Combined techniques of bisulphite conversion and Illumina sequencing to analyze the methylated regions in tomato genome demonstrated that epigenetic regulation along with hormone application controls tomato fruit ripening [14]. Conventional ChIP (chromatin immunoprecipitation) followed by direct sequencing has also started its application in plant systems. The cytosine methylome (methylC-seq), transcriptome (mRNA-seq), and small RNA transcriptome (smRNAseq) were directly sequenced in Arabidopsis using Solexa technology, which led to the generation of highly integrated epigenome maps for wild-type Arabidopsis and for mutants defective in either DNA methyltransferase or demethylase activity [15]. Mining of molecular markers through NGS is possible via a new approach called Coverage-based Consensus Calling (CbCC) which has helped in searching SNPs in chickpea [16]. NGS technologies have also increased the availability of organellar (mitochondrial and chloroplast) genomes. The complete chloroplast and mitochondrial genomes of *Boea hygrometrica* was sequenced to have detailed insight into the evolution of plant organellar genomes [15]. It provided the information that the smaller chloroplast genome (150 kb) contains more coding genes (147 genes covering 72% genome size) and large mitochondrial (510.5 kb) genome contains less genes (65 genes covering 12% of genome). Sequencing of mitochondrial genome, containing male sterility genes, helps in hybrid crop production [17].

The use of third generation NGS technology has led to a quantum leap in deriving adequate genomic data for many crops like wheat, chickpea, common bean and pigeon pea, and also in orphan crops where resources were initially limited. The availability of large number of genetic markers has made linkage mapping and marker-assisted breeding easier in marker-deficient crops. Gene expression studies conducted through NGS technology provide insights into

the spatial and temporal control of expression because of the ability to identify all RNA transcripts produced at a specific time, which is not possible through microarray or real-time PCR analysis. NGS can also accelerate the development of transformation technologies for crops because it is easier to modify genes due to increasing availability of genomic data. However, limited bioinformatics tools and storage space for huge sequence data are still a challenge for NGS technology. Further reduction in cost for resequencing of genome will help in extending the facility to parental and progeny lines of mapping populations without being restricted only within model plants and major crop species, as well as will lead to multiple genome sequencing projects to provide information regarding numerous alleles of different genes. Moreover, the availability of genome sequences for many important crops will facilitate genome editing approaches, including CRISPR/Cas9 system, which largely depends on accurate sequence information for precise determination of the target position. To conclude, the continuing technological advancement and sophistication in NGS analysis is definitely going to set benchmark in further advancement of crop genomics as well as in other areas of omics.

References

1. Godfray HCJ, Beddubgton JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food security: the challenge of feeding 9 billion people. Science. 2010; 327: 812-818.
2. Barabaschi D, Guerra D, Lacrima K, Laino P, Michelotti V, Urso S, et al. Emerging knowledge from genome sequencing of crop species. Mol Biotechnol. 2012; 50: 250-266.
3. Ronaghi M. Pyrosequencing sheds light on DNA sequencing. Genome Research. 2001; 11: 3-11.
4. Shendure J, Porreca GJ, Reppas NB, Lin X, Mc-Cutcheon JP, Rosenbaum AM, et al. Accurate multiplex polony sequencing of an evolved bacterial genome. Science. 2005; 309: 1728-1732.
5. Niedringhaus TP, Milanova D, Kerby MB, Snyder MP, Barron AE. Landscape of next-generation sequencing technologies. Analytical Chemistry. 2011; 83: 4327-4341.
6. Healy K. Nanopore-based single-molecule DNA analysis. Future Medicines. 2007; 2: 459-481.
7. Flusberg BA, Webster DR, Lee JH, Travers KJ, Olivares EC, Clark TA, et al. Direct detection of DNA methylation during single-molecule, real-time sequencing. Nat Methods. 2010; 7: 461-465.
8. Lee HC, Lai K, Lorenc MT, Imelfort M, Duran C, Edwards D. Bioinformatics tools and databases for analysis of next-generation sequence data. Brief Func Genomics. 2012; 11: 12-24.
9. Varshney RK, Bansal KC, Aggarwal PK, Datta SK, Craufurd PQ. Agricultural biotechnology for crop improvement in a variable climate: hope or hype? Trends Plant Sci. 2011; 16: 363-371.
10. Elshire R.J, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE. 2011; 6: e19379.
11. Van K, Kang YJ, Shim SR, Lee SH. Genome-wide scan of the soybean genome using degenerate oligonucleotide primed PCR: an example for studying large complex genome structure. Genes & Genomics. 2012; 34: 467-474.
12. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57-63.
13. Yates SA, Swain MT, Hegarty MJ, Chernukin I, Lowe M, Allison GG, et al. De novo assembly of red clover transcriptome based on RNA-Seq data provides insight into drought response, gene discovery and marker identification. BMC Genomics. 2014; 15: 453.
14. Zhong S, Fei Z, Chen YR, Vrebalov J, Mcquinn R, Gapper N, et al. Single-base resolution methylomes of tomato fruit development reveal epig¬enome modifications associated with ripening. Nature Biotechnol. 2013; 31: 154-159.
15. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell. 2008; 133: 1-14.
16. Azam S, Thakur V, Ruperao P, Shah T, Balaji J, Amindala B, et al. Coverage-based consensus calling (CbCC) of short sequence reads and comparison of CbCC results to identify SNPs in chickpea (*Cicer arietinum*; Fabaceae), a crop species without a reference genome. Am J Bot. 2012; 99: 186-192.
17. Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 2009; 27: 522-530.