Agri Articles
(e-Magazine for Agricultural Articles)
Volume: 02, Issue: 06 (NOV-DEC, 2022)
Available online at http://www.agriarticles.com
©Agri Articles, ISSN: 2582-9882

NGS Techniques and their Applications in Crop Improvement

(*Jitendra Kumar Yadav[1], Seema Yadav[2], Ashish Yadav[3] and Suman Yadav[4])
[1]Indian Institute of Agriculture Research, New Delhi
[2]Shri Karan Narendra Agriculture University, Jobner, Jaipur, Rajasthan
[3]Rajmata Vijayaraje Scindia Krishi Vishwa Vidyalaya, Gwalior, M.P
[4]Agriculture University, Kota, Rajasthan
*Corresponding Author's email: jk180811@gmail.com

Abstract
Sequencing is the process of determining of base sequences of the DNA fragment. Initially, DNA sequencing was done by the Sanger method and Maxam gilbert method. Due to the rapid advancement in sequencing technology various new methods have been discovered which are known as next-generation sequencing techniques which increased data output, efficiencies, and applications. The next-generation sequencing techniques can be categorized based on reading length. The next-generation sequencing (NGS) technology refers to non-Sanger-based DNA sequencing methods that have replaced conventional sequencing methods. In the second-generation, technologies short-read sequencing approaches are represented. The most prevalent technologies are Illumina,454 (Roche), SOLiD, and Ion Torrent technologies. It is vividly used for analyses of the complete genome (whole genome sequencing), the coding exons within already reported genes (whole exome sequencing), and only coding regions of selected genes (targeted panel). In this presentation, we discuss an introduction to NGS technology as well as a gist of the different types and applications of NGS. First-generation sequencing is based on enzymatic digestion followed by electrophoresis and autoradiography which allows the deduction of the base sequence. Second-generation sequencing is based on PCR these are much faster and cheaper and third-generation sequencing is based on a single DNA molecule without any cloning. As advancements in NGS data analysis have opened up new therapeutic opportunities for disease diagnosis, the complementary approaches such as machine learning algorithms used in NGS are subtly dealt with at the end.

Dna Sequencing
DNA sequencing is the process of determining the exact order of nucleotides within a DNA molecule. This method is used to determine the order of the four bases—adenine (A), guanine (G), cytosine (CY), and thymine (T) in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research.

- **DNA sequencing has been feasible because of the following developments**

1. Availability of restriction endonucleases.
2. Development of a highly sensitive gel electrophoresis technique, which can separate DNA fragments, differing by only one nucleotide.
3. Availability of large quantities of individual DNA fragments due to the development of gene cloning and PCR techniques.
4. Development of reliable, easy, and rapid DNA techniques.

The appearance of sequencing technologies has played an important role in the analysis of genomic sequences of organisms. A DNA sequencer produces files containing DNA sequences. These sequences are strings called reads on an alphabet formed by five letters {A, T, C, G, N}. The symbol N is used to represent ambiguity. The first sequencing technologies were developed in 1977 by Sanger et al. from Cambridge University awarded a Nobel Prize in chemistry in 1980 and Maxam et al. from Harvard University. Their discovery opened the door to studying the genetic code of living beings and brought their inspiration to researchers to the development of faster and more efficient sequencing technology. Sanger sequencing has become the most applied technique of sequencing for its high efficiency and low radioactivity [6] and has been commercialized and automated as the "Sanger Sequencing Technology".

Sanger and Maxam-Gilbert sequencing technologies were the most common sequencing technologies used by biologists until the emergence of a new era of sequencing technologies opening new perspectives for genome exploration and analysis. These sequencing technologies first appeared through Roche's 454 technology in 2005 and were commercialized as technologies capable of producing sequences with very high throughput and at a much lower cost than the first sequencing technologies. These new sequencing technologies are generally known under the name of "Next Generation Sequencing (NGS) Technologies" or "High Throughput Sequencing Technologies".

NGS technologies produce a massively parallel analysis with high throughput from multiple samples at a much-reduced cost. NGS technologies can be sequenced in parallel millions to billions of reads in a single run and the time required to generate the Giga Base-sized reads is only a few days or hours making it best than first-generation sequencing such as Sanger sequencing. The human genome, for example, consists of 3 billion bps and is made up of DNA macromolecules of lengths varying from 33 to ~247 million bps, distributed in the 23 chromosomes located in each human cell nucleus, the sequencing of the human genome using the Sanger sequencing took almost 15 years, required the cooperation of many laboratories around the world and cost approximately 100 million US dollars, whereas the sequencing by NGS sequencers using the 454 Genome Sequencer FLX took two months and for approximate one-hundredth of the cost. Unfortunately, NGS is incapable to read the complete DNA sequence of the genome, they are limited to sequencing small DNA fragments and generating millions of reads. This limit remains a negative point especially for genome assembly projects because it requires high computing resources. NGS technologies continue to improve and the number of sequencers increases these last years. However, the literature divided NGS technologies into two types. We distinguish the second-generation sequencing technologies which refer to the newest sequencing technologies developed in the NGS environment after the first generation, they are characterized by the need to prepare amplified sequencing banks before starting the sequencing of amplified DNA clones and there are the third-generation sequencing technologies that are sequencing technologies recently appeared, in contrast to the second generation, these technologies are classified as Single Molecule Sequencing Technology because they can make sequencing a single molecule without the necessity to create the amplification libraries and that are capable of generating longer reads at much lower costs and in a shorter time.

❖ NGS Workflow

1. LIBRARY PREPARATION
2. AMPLIFICATION/ENRICHMENT
3. SEQUENCING
4. DATA PROCESSING

(Note- Methods involved in each step may differ by platform.)

Library preparation refers to the DNA pre-treatment process before sequencing and is typically combined with the amplification step. Actual sequencing is performed in the last step and the technology involved in this step is highly dependent on the preferred platform. Generated data during the sequencing step is then handled by bioinformatic tools and further analyzed as needed.

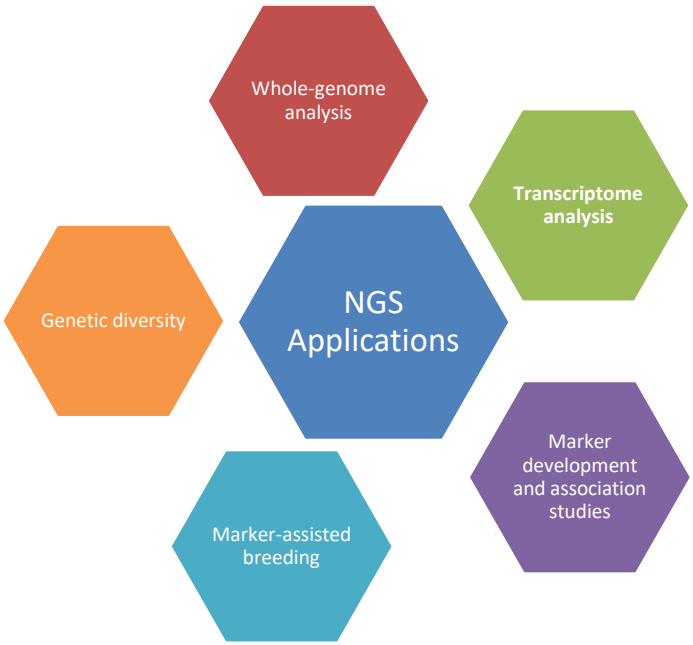
- **Next (2nd) Generation DNA Sequencing technology**

Platform	Capacity	Speed	Read length	Sequencing	Detection Based On	Cost/run	Amplification
454 Roche	35-700 Mb	10-23 Hrs	400-700 bp	Pyrosequencing (Sequencing By Synthesis)	Light emitted by luciferase	5,000 €	Emulsion PCR
SOLiD	90-180 Gb	7-12 Days	75 bp	Ligation (Sequencing By Ligation)	Fluorescence from fluorophore	5,000 €	Emulsion PCR
Illumina	6-600 Gb	2-14 Days	100-250 bp	Reversible terminator (Sequencing By Synthesis)	Fluorescence from fluorophore	10,000 - 20,000 €	Bridge PCR

- Note-: Although SBH (Sequencing By Hybridization) is cost-effective it is mostly used for genome-wide association studies and variant detection rather than de-nova sequencing.

Applications of NGS in Crop Improvement

CROP GENOME SEQUENCING STATUS- After the commonly-used for sequencing changed from the Sanger method to NGS, the number of plants with complete or draft genome sequences dramatically increased. Arabidopsis thaliana was the first plant to be completely sequenced, and sequencing was performed by the Arabidopsis Genome Initiative (AGI, 2000). Next, rice genome sequences became available (Yu et al., 2002; International Rice Genome Sequencing Project, 2005). Since then, the sequences of many important crop species, such as grape, sorghum, maize, and soybean, became available from studies that used the traditional Sanger method and NGS (Jallion et al., 2007; Paterson et al., 2009; Schnable et al., 2009; Schmutz et al., 2010). Genome sequencing projects involving the sequencing of many other important crop species (e.g., oil palm, banana, cotton, barley, and wheat) are still in progress (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). As of early July 2012, 42 genome sequences for 39 crop species were publicly available, and 12 crop sequencing projects were underway or not publicly available (Table 1). Of the 39 crop species with available genome sequences, most were sequenced after 2005, when NGS technology was developed. Recently, the tomato genome sequence has been published using information from studies that employed NGS as well as Sanger technology (The Tomato Genome Consortium, 2012). Crop genome sequencing is one of the beneficiaries of the rapid development of NGS technology. With lower costs and shorter time requirements, the quality of whole-genome sequencing of crops has been improved. Also, the whole-genome sequencing of many crop plants has enabled the progression of plant evolution studies from the gene to the nucleotide level. This will be helpful for understanding the complexity of existing genomes and the strong relationships between genotypes and evolution. Because whole-genome duplications and structural variations in chromosomes played a prominent role in plant evolution, the development of NGS technology may lead to the identification of new genes with new functions by investigating the functional and evolutionary divergence among plant species. Furthermore, a direct comparison between crop genome sequences has already led to the identification of conserved elements and species-specific differences that underlie unique traits (Barabaschi et al., 2012).



1) **Whole-genome analysis**- A comprehensive method for analysing the entire genome information which identifies inherited disorders, the mutations lead to various disease outbreaks and population genetics (Ng and Kirkness 2010). From the WGS results, we can analyse the read depth, gene density, insertion density, and SNP density and elucidate the unexplored genomic regions. The identified unique variants would reveal novel biological pathways that lead to complex disorders that provide high-resolution insights in the affected pathways (Sanders et al. 2017).
   NOTE- The whole exome sequencing (WES) is one of the broadly used NGS techniques where only protein-coding regions of the genome are sequenced. WES has been proven advantageous for identifying pathogenic variants in several Mendelian phenotypes, complex disorders as well as rare disorders (Jeste and Geschwind 2014; Mathur et al. 2018). The WES methods allow variant detection located in the coding exonic sites with an ability to extend the target regions to involve untranslated regions (UTRs), and in some cases microRNAs (de Carvalho et al. 2019) and even long non-coding RNAs to get a more detailed outlook of gene regulation in rare disorders (Gupta et al. 2018).
2) **Transcriptome analysis**- Transcriptome sequencing is the sequencing of mRNA isolated from different tissues of a plant at different time intervals, which focuses analysis of the transcribed portion of the genome. RNA-Seq is the popular choice for gene expression profiling via the sequencing of a whole-transcriptomes using NGS (Varshney et al., 2009; Jain, 2011; Strickler et al., 2012). Because the depth of sequence coverage is considered to be proportional to the expression level of the corresponding gene, even rare and novel transcripts can now be identified. Many researchers have therefore tried to elucidate a nearly complete picture of gene expression profiles under different environmental conditions using transcriptome analysis. Using a set of differentially expressed genes, transcriptome characterization and gene annotation were performed as transcriptomics downstream analysis. Also, SNP detection is a common application of RNA-Seq (Strickler et al., 2012).
3) **Marker development and association studies**- Compared to the traditional Sanger method, NGS helps discover and develop SSR or microsatellite loci efficiently. These markers are still commonly used for the construction of linkage maps, QTL mapping, MAS, cultivar fingerprinting, and studying gene flow. Zalapa et al. (2012) listed plant

SSR markers that were recently developed using the Sanger and NGS methods, but still, the majority of SSR markers were identified using Sanger technology. With the rapid development of NGS technologies, tremendous numbers of molecular markers like SNPs have been identified, and SNP-based resources are publically available for crop breeding programs (Kilian and Graner, 2012). Genome-wide marker discovery by NGS has become more feasible using new methods, such as reduced-representation libraries (Hyten et al., 2010), complexity reduction of polymorphic sequences (van Orsouw et al., 2007), restriction site-associated DNA sequencing (Baxter et al., 2011), and low coverage sequencing for genotyping (Huang et al., 2009a; Elshire et al., 2011). Since genome-wide markers were quickly developed in large quantities using NGS technologies, association mapping, patterns of natural population structure, and the decay of linkage disequilibrium (LD) can be studied more easily by whole-genome scanning using NGS (Varshney et al., 2009; Kilian and Graner, 2012). Also, whole-genome scanning has been performed using specially designed mapping populations.

4) **Marker-assisted breeding**- Different MAS strategies are used depending on the specific types of traits and breeding programs (Xu et al., 2012). Two major marker-assisted backcrossing (MABC) methods, marker-assisted foreground selection, and background selection are commonly used for breeding major gene-controlled traits. Marker-assisted foreground selection uses 2-10 markers for each target trait; both single and multiple traits are used for introgression with a population size of several hundred. By using advanced NGS technologies rather than earlier methods, the cost of genotyping with these markers and populations is dramatically decreased. But many breeding programs still demand much lower costs per sample. GBS is performed using libraries of reduced genome complexity that are created with the use of restriction enzymes. This simple, specific, and reproducible method has become a popular tool for population genotyping using NGS. High-throughput, large-scale genotyping methods using GBS have been introduced, and these strategies have already been applied to recombinant inbred lines (Huang et al., 2010; Elshire et al., 2011). After genotyping by NGS, high throughput and precise phenotyping are required for the genetic analysis of traits examined by MAS in crop breeding programs. Automated platforms in growth chambers or greenhouses are designed for phenotyping throughout the life cycle of the plant, and these plant materials are good resources for metabolomics and quantitative phenotyping (Bergelson and Roux, 2010; Massonnet et al., 2010).
5) **Genetic diversity**- To help counteract the loss of genetic diversity caused by agricultural practices, plant genetic resources (PGRs) including cultivars, landraces, wild species closely related to cultivated varieties, breeder's elite lines and mutants have been collected to increase the genetic variability of plants used in crop breeding programs. The collection of these PGRS was also performed to enhance future food security (Van et al., 2011; Barabaschi et al., 2012; Kilian and Graner, 2012). Since a barcoding system was developed for use with NGS technology, many individual plants could now be sequenced simultaneously at a lower cost. Sequencing at lower levels of coverage or sequencing only targeted regions of DNA are practical strategies for studying population genetics, conservation genetics, and molecular ecology. The genomics era provides a golden opportunity for categorizing PGRs by SNP marker instead of by phenotype. Therefore, the resequencing method using a wide range of PCR products is now affordable and enables genome-wide marker development, genotyping within populations, and the evaluation of genetic diversity (Barabaschi et al., 2012; Kilian and Graner, 2012).

Limitations

- Sequence properties and algorithmic challenges.

- Contamination or new insertions.
- Repeat content.
- Segmental duplications.
- Missing and fragmented genes.

Future of NGS in Crop Science

NGS technology provides a golden opportunity for understanding biological systems in crops. Compared to the traditional Sanger sequencing method, the cost of NGS is dramatically decreased, and employing advanced NGS technology is more feasible for many researchers who wish to sequence crop genomes. Some cash crops were considered to be less-studied/orphan crops due to a lack of sequence and marker information. But now, by employing de novo assembly strategies, whole-genome sequences of less-studied/orphan crops are becoming feasible for crop improvement. Also, more molecular markers like SNPs and indels have been rapidly developed at lower cost, and these markers are easily applicable to MAS in crop breeding programs. In this paper, several different applications for crop improvement were discussed. The identification of genes related to agronomic traits by crop breeding is important, but experiments for understanding the functions of these identified genes should be performed and could be applied to crop improvement in breeding programs. Although the development of bioinformatics tools and storage space for huge sequence data is still a challenge for NGS, the speed of crop improvement will be much faster than before because the third generation of sequencing platforms, such as HeliScope, Ion Torrent, single molecular real-time sequencing and Oxford Nanopore, have already been developed.

Conclusion

Although HT-NGS and NNGS have read lengths ranging from 25bp to 1000bp, but for genomic sequencing and for analysis of the ever more important structural genetic variation in genomes such as copy number variations, chromosomal translocations, inversions, large deletions, insertions, and duplications it would be a great advantage if sequence read length in the single DNA molecule could be increased to several 1000 bases. Plant breeding has a major role to play in increasing global food production while tackling the issues of limited land and water resources and changing climate. While the molecular era has laid the foundation for molecular breeding, the advent of genomic tools and technologies has been providing unprecedented capabilities for understanding the molecular basis of plant growth, development, and key traits towards improving crop productivity. The unending technological advancements in HT-NGS analysis are not only setting the benchmark in the advancement of crop genomics research, but also in the field of proteomics and other omics. It's not only the improvement of sequencing technologies producing thousands of terabytes of data that could solve the problem, but the technologies required for downstream processing of this huge data should also keep in pace with sequencing technologies. The paralleled development of these two would greatly help in genomics-assisted breeding for crop improvement and alleviating world hunger at large.

References

1. Berkman PJ, Lai K, Lorenc MT, Edwards D. Next-generation sequencing applications for wheat crop improvement. Am J Bot. 2012 Feb;99(2):365-71. doi: 10.3732/ajb.1100309. Epub 2012 Jan 20. PMID: 22268223.
2. https://www.researchgate.net/publication/346786600_Next_generation_sequencing_-Techniques_and_its_applications.
3. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect. 2018 Apr;24(4):335-341. doi:

10.1016/j.cmi.2017.10.013. Epub 2017 Oct 23. PMID: 29074157; PMCID: PMC5857210.

4. https://www.researchgate.net/publication/304018425_Principle_analysis_application_and_challenges_of_next-generation_sequencing_a_review.
5. https://www.researchgate.net/publication/301633710_Common_applications_of_next-generation_sequencing_technologies_in_genomic_research.
6. Yang, H., Tao, Y., Zheng, Z., Li, C., Sweetingham, M. W., & Howieson, J. G. (2012). Application of next-generation sequencing for rapid marker development in molecular plant breeding: a case study on anthracnose disease resistance in Lupinus angustifolius L. *BMC Genomics*, *13*(1), 318. https://doi.org/10.1186/1471-2164-13-318
7. Ahmad, R., Parfitt, D.E., Fass, J., Ogundiwin, E., Dhingra, A., Gradziel, T.M., Lin, D., Joshi, N.A., Martinez?Garcia, P.J. and Crisosto, C.H. (2011): Whole genome sequencing of peach (Prunus persica L.) for SNP identification and selection.
8. Acevedo-Garcia J, Spencer D, Thieron H, Reinstädler A, Hammond-Kosack K et al (2017) Mlo-based powdery mildew resistance in hexaploid bread wheat generated by a non-transgenic TILLING approach. Plant Biotechnol J 15:367
9. Alahmad S, Dinglasan E, Leung KM, Riaz A, Derbal N, Voss-Fels KP et al (2018) Speed breeding for multiple quantitative traits in durum wheat. Plant Methods 14:36.
10. [BOOK] Genotyping by Sequencing for Crop Improvement H Sonah, V Goyal, SM Shivaraj, RK Deshmukh - 2021 - Wiley Online Library.