Crop Genomics for Food Security

Priti Saini
CCS Haryana Agricultural University
4 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Sumit Jangra
Indian Agricultural Research Institute
38 PUBLICATIONS **91** CITATIONS
SEE PROFILE

Disha Kamboj
CCS Haryana Agricultural University
7 PUBLICATIONS **18** CITATIONS
SEE PROFILE

Neelam Rani Yadav
CCS Haryana Agricultural University
69 PUBLICATIONS **841** CITATIONS
SEE PROFILE

20

Crop Genomics for Food Security

**Priti[1], Sumit Jangra[1*], Disha Kamboj[1],
Neelam R. Yadav[1] and Ram C. Yadav[2]**

[1]Department of Molecular Biology, Biotechnology & Bioinformatics,
CCS Haryana Agricultural University, Hisar 125004, India
[2]Centre for Plant Biotechnology, CCS Haryana Agricultural University Campus, Hisar-125004, India
*Corresponding author: sumit.jangra712@gmail.com

Abstract

Making the world free of hunger and malnutrition is the ultimate goal of agricultural scientists. Decrease in the arable land and increasing global population has made it much more difficult. So, there is a need of advance techniques like genomics which will facilitate science based agricultural innovations such as development of nutrient rich and stress tolerant crop plants to eradicate hunger and malnutrition. Genomics or decoding the plant genome sequence using high-throughput techniques will allow scientific community to access agronomically important genes and will speed up the breeding programs for the development of superior varieties with higher yield and stress tolerance. Next generation genomics will help in the development of improved varieties which will be able to grow in harsh environmental and soil conditions without any compromise in the yield, which will lead to increase in farmers income and will make the world food secure.

Keywords: Genomics, food security, DNA sequencing, crop improvement and population.

Introduction

World population is increasing with high rate and is going to touch a mark of 9 billion by 2050 but the agricultural productivity is not able to cope with the increasing population. Although many countries have made significant progress in last few decades, but preponderance of famishment and malnutrition within the world population at dismaying rates challenges the global food security. According to FAO Hunger Report 2015, about 10.9% of the global population (one in nine people) is starving. Approximately 3.1 million children die every year due to hunger (http://www.worldhunger.org/world-child-hunger-facts/). Thus, seeing food security is the contiguous requisite for saving the global human potential, where the role of crop genomics is significant. Increasing agricultural productivity can increase access to food for people. So, the most populous countries like India has taken a step to increase land productivity. Green revolution, revolutionised the production of the major cereal crops with the introduction of semi-dwarf varieties of wheat and rice. But green revolution is not able to meet the demand of today's world. So, there is a need to focus on the development of high-yielding varieties. Agricultural productivity can be increased with the help of mechanisation, fertilisers, liming of soil acids to increase pH and to provide calcium and magnesium, herbicides, pesticides and development of high-yield varieties with the help of conventional breeding and genetic engineering. Conventional breeding has been contributing from a long time to crop improvement and development of high yielding varieties to meet the demand of increasing population. However, due to tapering of germplasm, identification of variability for incorporation into new cultivars is becoming more difficult. Therefore, there has been recourse to alternative approaches including mutagenesis, tissue culture and genetic transformation to aid breeding programs. Furthermore, close relatives of domesticated plants that is crop wild relatives represent gene pool that can be used for crop improvement by plant breeders. Since plant breeding is time-consuming and laborious, molecular breeding can speed up conventional breeding but still, there are some tasks which are unprocurable which can be achieved through genetic engineering. Genomics evolved with the advent of the invention of techniques of genetic engineering can generate data that can help to identify which region of genome needs to be targeted for crop improvement. Data generated from genomics will help in better understanding of gene expression, molecular and biochemical pathways which will aid plant breeders to entertain a different selection approach based on expression quantitative traits to maximise combinations of genes capable of conferring high performance. The advancements in high-throughput sequencing technologies have speed up the crop genomic research, role of genomic in ensuring food security is high lightened in this chapter.

Genomics

The term Genomics coined by Tom Roderick in 1986, is an interdisciplinary field of science which focuses on the study of genome of any organism. More specifically it refers to the analysis of the genome of organisms, both anatomically (sequences and organisation) and physiologically (expression and regulation) which further helps in deducing structure, function and mapping of genome (haploid set of chromosomes in a gamete or in each cell of multicellular organisms).

Research Areas

Structural Genomics

Branch of genomics that determines the three-dimensional structure of proteins encoded by a genome. This information can be helpful to find out where the desired gene resides in the genome.

Functional Genomics

Branch of genomics that determines the biological function of the genes and their products. The ability to find out gene and its function lay the basis of functional genomics. Various processes such as transcription, translation and epigenetics are investigated by functional genomics and answer when, where and how genes are expressed.

Epigenomics

It is the study of total gene expression that has not undergone any mutation in the DNA sequence. Various types of epigenetic changes like histone modification, DNA methylation, Nucleosome position occur in a cell. Epigenetic changes are reversible and heritable and change the overall chromatin structure.

Metagenomics

All the genetic material present in the environmental sample of the community of any organism is called metagenome. Metagenomics is the study of metagenome, genetic material recovered directly from the environment. Also called as environmental genomics, ecogenomics and community genomics. Metagenomics has made possible the study of non-culturable microbes.The phrase metagenome of soil first used by Handelsman *et al.*, (1998) to describe collective genome of soil microflora.

Genomics aid conventional breeding by gaining deeper insights into the biological mechanisms and can led to the development of new or improved screening methods for selecting superior genotypes more efficiently. These advances and development will provide an opportunity for efficient transfer of information systems from model species and major crops to orphan crops and will help in the development of crop plants with enhanced crop and productivity which will help in attaining food security.

Genome Sequencing Techniques

Background

DNA sequencing is the method of determining the actual arrangement of nucleotides within a DNA molecule. DNA sequencing includes any process or engineering which can be used to find out the order of the four bases: adenine, guanine, cytosine, and thymine within the DNA strand. Data generated from sequencing plays very crucial role in research perspectives as well as in a number of biological and its applied fields viz. forensics, biotechnology, virus studies, classification, medical science and many more. Initially, sequencing strategies were developed to sequence proteins but these were not applicable to nucleic acids. Because DNA molecule is larger and made up of fewer same units which make it difficult to sequence. Newer techniques were needed to sequence DNA. RNA was the first to be sequenced among nucleic acids, alanine tRNA from *Saccharomyces cerevisiae* was sequenced by Robert Holley and colleagues in 1965. The major milestone in RNA sequencing was the complete genome sequence of Bacteriophage in 1976 by Walter Fiers and coworkers. A huge number of efforts has been made to develop efficient DNA sequencing technology (fig. 1). Gilbert and Maxam in 1973 reported the sequence of Lac operator (24 bp) using wandering spot analysis. Frederick Sanger also followed this extension strategy and produced speedy method for sequencing DNA during his work at the MRC Centre, UK and released DNA sequencing with chain-terminating inhibitors in 1977 (Sanger *et al.*, 1977). Hood advanced the existing Sanger method of DNA sequencing which was becoming the common laboratory method and is the first semi-automated DNA sequencing machine in 1986. The first fully automated DNA sequencing machine (ABI 370) was marketed by Applied Biosystems in 1987. Revolution in DNA sequencing technology has brought down the cost of DNA sequencing and made the sequencing of an increased number of genomes both feasible and cost effective. The first plant genome Arabidopsis was completely sequenced in December 2000, and it was the third complete genome of a higher eukaryote. Subsequently, after Arabidopsis, several other crop plants have been sequenced (Sequence history of few important plants is represented in fig. 2).These genomes reveal numerous species-specific details, including genome size, gene number, patterns of sequence duplication, a catalogue of transposable elements, and syntenic relationships. To understand the complex instructions contained in all these raw sequence information of the plant genome, large-scale functional genomics projects are required. Progress towards a complete understanding of gene regulatory networks shared among many crop plants is important for improving cultivated species and for understanding different stress responsive mechanisms for crop improvement to increase productivity so that ample amount of food can be made available for growing population.

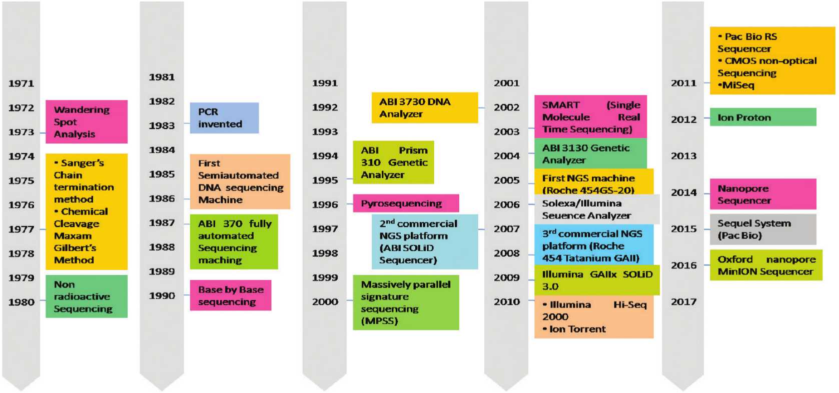
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
Wandering Spot Analysis
• Sanger's Chain termination method
• Chemical Cleavage Maxam Gilbert's Method
Non radioactive Sequencing
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
PCR invented
First Semiautomated DNA sequencing Machine
ABI 370 fully automated Sequencing maching
Base by Base sequencing
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
ABI 3730 DNA Analyzer
ABI Prism 310 Genetic Analyzer
Pyrosequencing
2nd commercial NGS platform (ABI SOLiD Sequencer)
Massively parallel signature sequencing (MPSS)
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
SMART (Single Molecule Real Time Sequencing)
ABI 3130 Genetic Analyzer
First NGS machine (Roche 454GS-20)
Solexa/Illumina Seuence Analyzer
3rd commercial NGS platform (Roche 454 Tatanium GAII)
Illumina GAIIx SOLiD 3.0
• Illumina Hi-Seq 2000
• Ion Torrent
2011
2012
2013
2014
2015
2016
2017
• Pac Bio RS Sequencer
• CMOS non-optical Sequencing
•MiSeq
Ion Proton
Nanopore Sequencer
Sequel System (Pac Bio)
Oxford nanopore MinION Sequencer

Fig. 1: Timeline representing the major breakthrough in DNA sequencing

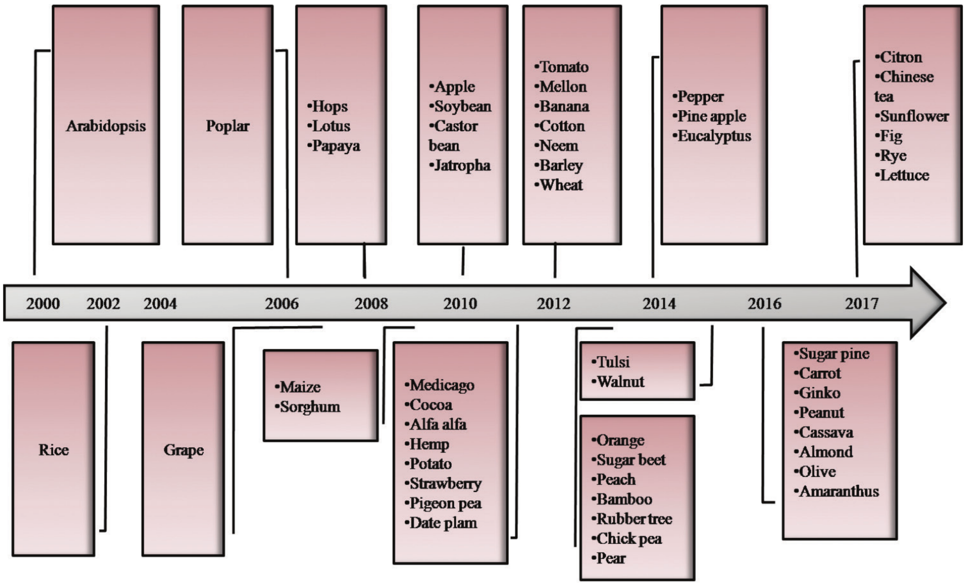

Fig. 2: Sequence history of some important plants

Types of DNA Sequencing Techniques

1. Sanger's method

The first DNA sequencing method devised by Sanger and Coulson in 1975 was called plus and minus sequencing that utilised *E. coli* DNA pol I and DNA polymerase from bacteriophage T4 with different limiting triphosphates. This method is also known as "Dideoxy Method". It prevailed from the 1980s until the mid-2000s. In this method, the synthetic nucleotides that lack the OH at the 3′ carbon atom are used for preparing DNA to be sequenced. When dideoxynucleotide added to the growing DNA, it stops chain elongation because there is no 3′ OH for the next nucleotide to be attached. The target DNA is assembled with the deoxynucleotides, tagged dideoxynucleotides and DNA polymerase I. These dideoxynucleotides are tagged such that each one fluorescence dissimilar colour. The elongation process goes on until it finds a dideoxynucleotide which leads to arrest the process. At the end, the fragments are separated according to length and the resolution is so good that a slight difference of single base pair is enough to separate that strand from the next shorter and next longer strand. When illuminated by laser beam, each of the four dideoxynucleotides fluorescence a give different colour and an automatic scanner provides a printout of the sequence.

2. Maxam-Gilbert method

This method required radioactive labelling at one 5′ end of the DNA (by kinase reaction using gamma-^{32}P ATP) and purification of the DNA fragment to be sequenced. Chemical treatment modifies the nucleotide base and then generates breaks at a small proportion of one or two of the four nucleotide bases in each of four reactions (G, A+G, C, C+T). The chemicals used for modification used in varied concentration such that they bring out only single change per molecule of DNA. As a result, a chain of labelled fragments is developed. Then, electrophoresis is performed for detachment according to size. Then the gel is disclosed under UV for visualisation. The sequence can be determined by the series of the band formation.

3. Pyrosequencing/Pal Nyren's Method

In 1996, Pal Nyren's group reported that natural nucleotide can be used to obtain efficient incorporation during a sequencing by synthesis protocol. Sequence detection is based on the release of pyrophosphate during the DNA polymerase activity, the pyrophosphate so generated is converted to ATP by sulfurylase and firefly luciferase converts this ATP to visible light which is recorded. Inclusion of dATPaS instead of dATP in the polymerization reaction was the first major improvement, which enables pyrosequencing to be done in real time. The second improvement was the introduction of apyrase to the reaction to make a four enzyme system. Apyrase allows nucleotides to be added sequentially without any intermediate washing step. In a cascade of enzymatic reaction, visible light is generated that is proportional to the number of incorporated nucleotides. The cascade starts with a nucleic acid polymerization reaction in which inorganic biphosphate (PPi) is released as a result of nucleotide incorporation by polymerase. The released PPi is subsequently converted to ATP by ATP sulfurylase, which provides the energy to luciferase to oxidise luciferin and generate light. The light so generated is captured by a CCD camera and recorded in the form of peaks known as pyrogram. Because the added nucleotide is known the sequence of the template can be determined.

3. Roche/454 FLX Pyrosequencer

The technology was commercialized in 2004 and is based on pyrosequencing. In this approach, the library fragments are mixed with agarose beads carrying oligonucleotides on the surface complementary to the 454-specific adapter sequences in the fragment library, so each bead is combined with a single fragment. The fragment:bead complexes isolated from oil:water-micelle comprises of the reacting agents of PCR. After thermal cycling, amplification of each DNA occurs on the bead surface. These single amplified DNA molecules are then sequenced. The process involves arrangement of beads into picotiter plate in such a way that each well has one bead in contact. Then, enzymes catalysing the reaction are added to each bead. The plate acts as a flow cell, which note downs every nucleotide's addition because of emission of light which is recorded by the CCD camera.

4. The Illumina genome analyzer

In this technique sequencing is performed by an automated device (Cluster Station), in this single molecule amplification step starts with an Illumina-specific adapter library. Sequencing takes place on the oligo-derivatized surface of a flow cell. The flow cell is an eight-channel sealed microfabricated device which let the bridge amplification of fragments to take place on its surface and employs DNA polymerase to make multiple copies or clusters of the single molecule that initiated the amplification. Same, separate or combination of both libraries can be added to each of the eight channels. Approximately one million copies of each fragment are present in a single cluster, which is enough for reporting bases at the required signal intensity for detection while sequencing. Sequencing-by-synthesis approach is utilised by Illumina system, in which all the four nucleotides are simultaneously added to the flow cell along with DNA polymerase. The nucleotides are labelled with a base-specific fluorescent label at 3-OH group. These steps are repeated for a specific number of cycles, as determined by user-defined instrument settings, which permits discrete read lengths of 25–35 bases. A base-calling algorithm assigns sequences and associated quality values to each read and a quality checking pipeline evaluates the Illumina data from each run, removing poor-quality sequences.

5. Applied Biosystems SOLiDTM system

The SOLiD (sequencing by oligonucleotide ligation and detection) platform uses an adapter-ligated fragment library other next-generation platforms, and uses an emulsion PCR approach with small magnetic beads to amplify the fragments for sequencing. This technology was developed by George Church in 2005 and was further improved and distributed by Applied Biosytem in 2007. The principle of this approach relies on DNA ligase to detect and incorporate bases in specific manner. Two flow cells are processed per instrument run, each of which can be divided to contain different libraries in up to four quadrants. Each Solid run requires 5 days and yield 2-4Gb DNA sequence data with average read length of between 25–35 bps. Once the reads are base called, have quality values, and low-quality sequences are removed, the reads are aligned to a reference genome to enable the second tier of quality evaluation called two-base encoding.

6. Helicos Heliscope TM

The Helicos sequencer, is based on work by Quake's group, which relies on cyclic interrogation of a dense array of sequencing features. No clonal amplification is needed in this method. A very sensitive fluorescence detection system is applied to directly find the DNA molecules through sequencing-by-synthesis method. The target DNA molecules libraries are devised by fragmenting randomly and addition of poly A tails. These are then hybridised with surface-attached poly T oligomers and are thus captured producing an irregular chain of single molecule sequencing templates. At each cycle, DNA polymerase and a single species of fluorescently labelled nucleotide are added, resulting in a template-dependent extension of the surface-immobilized, primer-template duplexes. Extension and imaging is permitted by the chemical cleavage and release of fluorescent labels after acquisition of images. An average read length of 25 bp or greater is generated by single base extension (i.e. A, G, C, T, A, G, C, T….) in several hundred cycles.

7. Pacific Biosciences SMRT

The Single Molecule Real Time Sequencing was developed by Pacific Biosciences. This approach involves the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can potentially be detected through FRET (fluorescence resonance energy transfer) interactions between a fluorophore-bearing polymerase and gamma phosphate labelled nucleotides. This technique is ramped on two conceptions: zero-mode waveguides (ZMWs) and phosphor linked nucleotides. ZMWs helps in illuminating light at the bottom of the well where DNA polymerase-template complex is trapped. Phospho linked nucleotides grants recording of the trapped complex as DNA polymerase develops a DNA strand.

8. Nanopore-based Fourth Generation DNA Sequencing

This technology uses single-molecule technique which enables us to further examine the interaction between DNA and protein, as well as protein-protein interaction. Nanopore analysis opens a new door to molecular biology investigation at the single-molecule scale. The advantages of nanopores include label-free, ultra-longreads (104–106 bases), low cost, low material requirement and display results in real time, which simplifies theprocess and thus can be used for various DNA sequencing applications. Its principle is based on the fact that each base could produce different ionic current when DNA passed through nanopore, so it would be possible to distinguish different nucleotides. The nanometre-sized pores are in a biological membrane or formed in the solid-state film, which then separates two compartments containing conductive electrolytes and electrodes are emerged in each compartment. Upon applying voltage, electrolyte ions in solution move electrophoretically through the pore, thereby generating an ionic current signal. When the pore is blocked due to passage of biomolecule, current flowing through the nanopore would be blocked, interrupting the current signal. The physical and chemical properties of the target molecules can be determined by analyzing the amplitude and duration of current blockades from translocation events. The advantages and disadvantages of various DNA sequencing techniques is drafted in table 1.

Table 1: Comparison of various DNA sequencing techniques

Techniques	Read length	Accuracy	Reads per run	Time per run	Advantages	Disadvantages
Sanger Method	400-900bp	99.9%	NA	25 minutes to 3 hours	Long individual reads	Expensive and takes more time for plasmid cloning

Techniques	Read length	Accuracy	Reads per run	Time per run	Advantages	Disadvantages
Maxam and Gilbert Method	100 bp				DNA can be read directly, used to analyze DNA-protein interactions, nucleic acid structure and epigenetic modifications in DNA	extensive use of hazardous chemicals, cannot be used to analyze more than 500 base pairs, the read-length decreases due to incomplete cleavage reactions
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	Fast and long read size	High cost, low throughput, homo-polymer errors
Illumina	50, 150, 250 or 300 bp	99.9%	1-25 million, 300 million -2 billion, 3 billion	6 hrs - 3 days	High sequence yield	Expensive, needs concentrated DNA
SOLiD	600 bases	99.9%	1.2-1.4 billion	1-2weeks	Low cost	Slower, error during palindromic sequences
Helicose	32 nucleotides	99.5%	32 bases	8 days	Simplifies data analysis, avoids PCR bias	High error rate, time -consuming
SMRT	10,000 bp -15,000bp	87%	50,000 per cell	30 min - 4 hrs	Longest read length, fast, reliable	Costly, throughput is moderate
Nanopore-based sequencing	Depends on library size	92-97%	Selectedby user	1 min – 48 hrs	Portable, easy, very long read	Throughput is low, single read takes times

Genomics-Assisted Crop Improvement

Initially genomics research was focused on understanding the fundamentals of biology. To meet the demand of increasing population growth and to fight the climate changes, there has been growth in modernization of agricultural practices. Genomics is a rapidly expanding field of research fueled by reducing cost of DNA sequencing and genotyping (Edwards *et al.,* 2013). The complete and annotated DNA sequence of a plant genome canlocalise important traits that will be a valuable source for the plant breeders. The annotated DNA sequence of a plant genome informs about gene location, gene structure, distance between genes and locations of repetitive DNA sequence such as microsatellite and transposons. This information can further assist plant breeders to select effective breeding strategies. High throughput whole genome genotyping platforms such as VeraCode, Illumina Golden Gate, Infinium or DArT could be employed in breeding programs. High throughput screening methods could be developed based on genome sequence and by using genomics technologies, to identify desired progeny rapid and more efficiently than traditional methods. Accessibility of DNA sequence of the plant genome and transcriptome allows the development of DNA markers for marker-assisted breeding, allele mining, germplasm characterization, interpretation of gene function, analysis of genome evolution and study population genetics. Through these methods novel genes governing important traits can be found and could become new breeding targets for development of superior varieties to meet the global food requirement.

Genomics of Model Plants

Arabidopsis thaliana was the first plant and third multicellular organism whose complete genome was sequenced at the end of the year 2000. At that time, it was claimed that genome of *Arabidopsis thaliana* will help in the deeper understanding of plant development and environmental response and permits the structure and dynamics of plant genome to be accessed. The genome sequencing project began in 1996 with the formation of the Arabidopsis Genome Initiative, a consortium of international laboratories that utilised a BAC by BAC approach with Sanger sequencing to complete the genome sequence. The sequenced genome (135 Mb) is considered the gold standard for plant genomes due to the high quality and finished nature of the sequence.

Using the approach taken by Arabidopsis Genome Initiative, the genome of rice (*Oryza sativa*) was also sequenced by the International Rice Genome Sequencing Project (IRGSP). IRGSP was one of the few truly multinational plant genome projects with flags planted in chromosomes. The IRGSP was established in 1998 and taken up by ten nations [Chromosome 1 (Japan, Korea), Chromosome 2 (UK), Chromosome 3 (USA), Chromosome 4 (China), Chromosome 5 (Taiwan), Chromosome 6,7,8 (Japan), Chromosome 9 (Thailand& Canada), Chromosome 10 (USA), Chromosome 11 (USA & India) and Chromosome 12 (France &Brazil)] to obtain a complete finished quality sequence of rice genome (*Oryza sativa* L. sp.Japonica cv. Nipponbare). In December 2002 IRGSP released a high-quality map based draft sequence. This has permitted rice geneticists to identify several genes underlying traits and revealed very large and previously unknown segmental duplications that comprise 60% of the genome (Peterson *et al.,* 2014). The public sequence has also revealed new details about the syntenic relationships and gene mobility between rice, maize and sorghum(Paterson *et al.,* Salse *et al.,* and Lai *et* *al.,* 2004). Map based, finished quality sequence covers 95% of the 389 Mb genome, including all of the chromatin and two complete centromeres.

Nicotiana benthamiana a close relative of tobacco, widely used model for plant-microbe biology and other research applications. It is particularly useful because it is related to tomato, potato and will help in designing constructs for virus-induced gene silencing in order to reduce the possibility of 'off-target' gene silencing (Bombarley *et al.,* 2012). A new draft sequence of *N. benthamiana* genome has been released by researchers from the Boyce Thompson Institute for Plant Research (BTI) in 2012 and its estimated size was found to be 3Gb. Genome sequence of *Nicotiana slvestris* and *Nicotiana tomentosiformis* were released in 2013. Draft genomes of *N. sylvestris* and *N. tomentosiformis* were assembled to 82.9% and 71.6% of their expected size respectively.

Cajanus cajan, the first seed legume plant to have its complete genome sequenced in 2012. The sequencing efforts led by group of 31 Indian scientists from the Indian Council of Agricultural Research, it was then pursued by global research partnership, the International Initiative for Pigeon pea Genomics (IIPG), led by ICRISAT with partners such as BGI– Shenzhen (China), US research laboratories like University of Georgia, University of California-Davis, Cold Spring Harbor Laboratory and National Centre for Genome Resources, European research institutes like National University of Ireland, Galway, CGIAR Challenge Generation Programme and US National Science Foundation. Pigeon pea was considered an orphan crop but now significant amount of data has been generated, mainly because of the efforts by Indo-US Agricultural Knowledge Initiative (AKI), NSF and GCP-funded projects, (Varshney *et al.,* 2009, 2010a; Dutta *et al.,* 2011; Bohra *et al.,* 2011).

Medicago truncatula (~384 Mb) is a close relative of Alfalfa and is a preeminent model for the study of the processes of nitrogen fixation, symbiosis and legume genomics. The Medicago sequencing project began in 2003 with aim to elucidate sequences originates from the euchromatin portion of the genome. Among the eight chromosomes of Medicago, six were sequenced in US, chromosome no 5 was sequenced in France and chromosome no 3 was sequenced in the United Kingdom. In 2011 draft sequence was published based on BAC tilling path, supplemented with Illumina shotgun

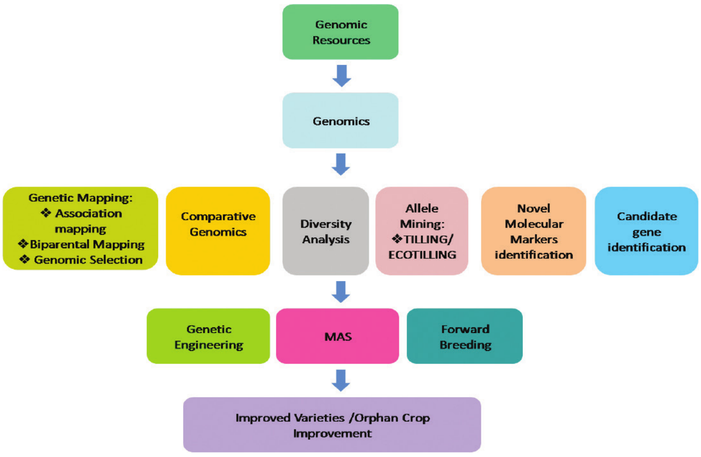

Fig. 3: Workflow of crop improvement through genomics

sequence (Mt3.5), together capturing 94% of all *M. truncatula* genes. In 2014 improved and more refined version of sequence was published based on *de novo* whole genome shotgun assembly of a majority of Illumina and 454 reads using ALLPATHS-LG.

Sequencing of model crop plants is not the end by itself but is just the beginning of a new venture to unravel genetic information and to gain better insights into the genetics of other species under investigation. The workflow of improvement of plant through genomics is depicted in fig. 3.

Mining of Genes for Agronomic Importance

With the progress in genome sequencing efforts for reducing cost, error rate and time have enabled to develop millions of novel markers, in non-model crop species, as well as identification of genes which are agronomically important. Identification of all genes within a species helps in understanding of how important agronomic traits are being governed, this information can further help in crop improvement. Biochemical function of these encoded proteins may be to gain greater understanding of the mechanism underlying the trait, whether change in gene structure or expression may further improve the trait. Knowledge of the gene responsible for trait may be further transferred to different cultivar by marker assisted selection or to species through genetic modification. While gene underlying the simple traits are simple to characterise at genome level, but there are many complex traits which are controlled by interacting gene networks.

Like related SNP-based functional molecular marker have been developed for SIDREB-2 gene involved in dehydration response in foxtail millet (Lata & Prasad, 2013). The EST-based SSR markers were developed for NBS-LRR R-gene in finger millet (Panwar *et al.,* 2011) to screen blast resistance. An allele-specific marker was developed based on the SNP detected in the eight exons of TaGW2 gene in wheat, to screen genotypes for increased kernel width and high thousand-kernel weight (Yang *et al.,* 2012)

Novel Molecular Markers

Various important and diverse agronomic traits coupled with genes and molecular markers which can be localized with availability of DNA sequence information creating new opportunities for crop improvement. Since 1980, with the discovery of very first molecular marker i.e. Restriction Fragment Length Polymorphism (RFLP), breeders employed various molecular markers in breeding programs like such as random amplified polymorphic DNA(RAPD), Amplified Fragment Length Polymorphism (AFLP), Sequence Tagged Sites (STS) and Simple Sequence Repeats (SSR). The major drawback with these random markers is that their prediction mainly depends on linkage between marker and target locus and for non-model crops it is difficult to develop sequence based markers due to lack of genomic information. Devlopment of next generation sequencing in transcrptomics creating coding sequence collection, which is much more cost effective has facilitated development of molecular markers from genic regions. As a result, functional molecular markers like genic SSRs, Single nucleotide polymorphism and other SSRs have been revealed and developed for non-model plant species. For example, genic SSRs are the functional molecular markers, derived from the primers which have been designed from the most conserved region of the gene (Varhsney *et al.*, 2005b).

Genic SSRs are electronically mined from publically available ESTs or gene sequence information using different bioinformatics softwares. The average frequency of SSRs in expressed regions of barley, sorghum, maize, rye, rice and wheat was reported to be 1 in every 6 kb (Varshney *et al.*, 2002, Varshney *et al.*, 2005c). The genic SSRs can be used in genetic mapping, functional diversity studies and these can be transferred among distantly related species (Varshney *et al.*, 2005c, Yu *et al.*, 2004). Like cssu45 a genic-SSR marker can distinguish the presence of Sr45 (stem rust resistance loci 45) across different genetic backgrounds in wheat (Periyannan *et al.*, 2014). Functional SNPs or insertions or deletions can cause direct phenotypic effect and such polymorphism are indispensable for the development of functional markers which can be further employed for crop improvement (Anderson & Lubberstedt, 2003; Bagge *et al.*, 2007; Domon *et al.*, 2004). Discovery of SNPs can further assist in GWAS using SNP array because we cannot afford whole genome sequencing in all individual plants. Traditionally SNP discovery involved PCR amplification of genes/genomic regions of interest from multiple individuals selected to represent diversity in the species or population of interest, followed by either direct sequencing of these amplicons, or the more expensive method of sequencing or cloning. Sequences are then aligned followed by screening for polymorphism. This approach is very expensive and time taking as large number of SNPs are required for most applications such as genetic mapping and association studies. SNP and SSR discovery through *in silico* methods are now being popular because of providing cheap and efficient methods for marker identification (Batley *et al.*, 2003; Robinson *etal.*, 2004; Jewell *et al.*, 2006; Duran *et al.*, 2009a,b). Large amount of sequencing data being produced in less time along more accuracy with latest sequencing technologies which provide a valuable resource for the mining of molecular markers (Imelfort *et al.*, 2009).

While next generation sequencing data produces a sequence data of quality, large amount of genome sequencing data makes it feasible to differentiate between true SNP and sequencing error. To identify genetic diversity in population and to study relationship between inherited genome and inherited traits, whole genome sequencing is the most potent method.

In plants one of the first application of next generation sequencing was identification of 36,000 maize SNPs using 260,000 and 280,000 EST, sequenced using Roche GS20. These SNP were identified between B73 and Mo17 inbred maize lines (Barbazuk *et al.*, 2007). Strict post processingof data reduced this number to >7000 putative SNP and over 85 % (94/110) of a sample of these SNPs were successfully validated by Sanger sequencing. Based on this validation rate, this pilot experiment conservatively identified >4900 valid SNP within >2400 maize genes, demonstrating the suitability and potential of the approach.

Genetic Mapping

Genetic mapping identify order and relative positions of molecular markers in linkage groups based on their pattern of inheritance. Genetic maps are prepared by analyzing populations derived from crosses of genetically diverse parents or from diverse line from natural environment. These markers can be further used to transfer gene of interest from donor to target genotypes.

Biparental QTL Mapping

Biparental mapping creates genetic maps on the basis of segregating population derived from biparental cross. The type of marker used for map construction should be present in enough density to increase resolution of the map. Genetic mapping now become robust method with the SNPs identified within whole genome sequence. This becomes possible to map a specific gene of interest and assist in the identification of linked or perfect markers for traits as well as increasing the density of markers on genetic maps (Rafalski, 2002).

The development of these markers also allows the integration of genetic and physical maps. By using molecular markers from related species allows the comparison of linkage maps. This allows the translation of information between model species with sequenced genomes and non-model species (Moore *et al.*, 1995). Furthermore, the integration of molecular marker data with phenomics, genomics and proteomics data allows researchers to link sequenced genome data with observed traits, bridging the genome to phenome divide. These markers can then be further employed in crop breeding programs.

Association Mapping

Association mapping (AM) identifies quantitative trait loci (QTL) by examining the marker-trait associations that can be attributed to the strength of linkage disequilibrium (LD) between markers and functional polymorphisms across a set of diverse germplasm. First reports of AM in plants emerged in 1996 in rice and 1997 in oat by Virk*et al.,* (1996) and Beer *et al.,* (1997) respectively. Association mapping is better than QTL mapping because of availability of broader genetic variations with wider background for marker trait correlation, high resolution, exploitation of historically measured trait data for association, time-saving and cost effective. In association mapping, unstructured populations represent many recombination events and are often many generations from common ancestor, providing the potential of great resolution for a particular set of population size. Sequencing led to development of large amount of molecular marker genotypic data which favours association studies over QTL mapping.

Conclusion

Recent advances in DNA sequencing technology has revolutionised crop improvement programs. Genomic tools reveal the information underlying plants responses to environmental stress and large inputs are needed to translate this information to climate resilient crops. Increased knowledge in crop genomics, along with reduced cost is expected to ease the sped-up betterment of orphan crops. Omics data generated from crop plants will allow us to address key agronomic traits that were difficult or impossible to address previously. Work on model plants has led to the refinement of technologies for later application to crop improvement. Genomics will help in identification of molecular markers linked to important agronomic traits which will help in the development of improved varieties with improved yield, quality, stress tolerance and disease resistance. All these will aid in making the world food secure.

References

Andersen, J. R. and Lübberstedt, T. 2003. Functional markers in plant. *Trends Plant Sci*, 8: 554-560.

Bagge, M., Xia, X. and Lubberstedt, T. 2007. Functional markers in wheat.*Curr Opin Plant Biol,* 10: 211–216.

Barbazuk, W.B., Emrich, S.J., Chen, H.D., Li, L. and Schnable, P.S. 2007. SNP discovery via 454 transcriptome sequencing. *Plant J,* 51: 910–918.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J. and Edwards, D. 2003.Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol,* 132: 84–91.

Beer, S.C., Siripoonwiwat, W., O'Donoughue, L.S., Souza, E., Matthews, D. and Sorrells, M.E. 1997. Associations between molecular markers and quantitative traits in an oat germplasm pool: Can we infer linkages? *J. Agric. Genom,* 3: 1-16.

Bohra, A., Dubey, A., Saxena, R. K., Varma, P. R., Poornima, K. N., Kumar, N., Farmer, A. D., Srivani, G., Upadhyaya, H. D., Gothalwal, R., Ramesh, S., Singh, D., Saxena, K. B., Kavikishor, P. B., Singh, N. K., Town, C. D., May, G. D., Cook, D. R. and Varshney, R. K. 2011. Analysis of BAC-end sequences (BESs) and development of BES-SSR markers for genetic mapping and hybrid purity assessment in pigeonpea (*Cajanus* spp.). *BMC Plant Biol*, 11(56): 1-15.

Bombarely, A., Rosli, H.G., Vrebalov, J., Moffett, P., Mueller, L. A. and Martin, G.B. 2012. A draft genome sequence of *Nicotiana benthamiana* to enhance molecular plant- microbe biology research. *Mol Plant Microbe Interact,* 25(12): 1523-1530.

Domon, E., Yanagisawa, T., Saito, A. and Takeda, K. 2004. Single nucleotide polymorphism genotyping of the barley waxy gene by polymerase chain reaction with confronting two-pair primers. *Plant Breed,*123: 225–228.

Duran, C., Appleby, N., Clark, T., Wood, D., Imelfort, M., Batley, J. and Edwards, D. 2009a. Auto SNPdb: an annotated single nucleotide polymorphism database for crop plants. *Nucleic Acids Res,* 37: 951–953.

Duran, C., Appleby, N., Vardy, M., Imelfort, M., Edwards, D. and Batley, J. 2009b. Single nucleotide polymorphism discovery in barley using autoSNPdb. *Plant Biotechnol. J,* 7: 326–333.

Dutta, S., Kumawat, G., Singh, B. P., Gupta, D. K., Singh, S., Dogra, V., Gaikwad, K., Sharma, T. R., Raje, R. S., Bandhopadhya, T. K., Datta, S., Singh, M. N., Fakrudin, B., Kulwal, P., Wanjari, K. B., Varshney, R. K., Cook, D. R., Singh, N. K. 2011. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. *BMC Plant Biol,* 11(17): 1-13.

Edwards, D., Batley, J. and Snowdon, R. J. 2013. Accessing complex crop genomes with next-generation sequencing. *Theor Appl Genet,*126: 1-11.

Gilbert, W. and Maxam, A. 1973. The Nucleotide Sequence of the lac Operator. *Proc. Natl Acad. Sci,*70(12): 3581–3584.

Guyot, R. and Keller, B. 2004. Ancestral genome duplication in rice. *Genome,* 47: 610−614.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M. 1998. "Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products". *Chem. Biol,*5 (10): R245–R249.

Imelfort, M., Batley, J., Grimmond, S., and Edwards, D. 2009. Genome sequencing approaches and successes. *Methods Mol. Biol,* 513: 345–358.

Jewell, E., Robinson, A., Savage, D., Erwin, T., Love, C.G., Lim, G.A.C., Li, X., Batley, J., Spangenberg, G.C. and Edwards, D. 2006.SSR Primer and SSR taxonomy tree: Biome SSR discovery. *Nucleic Acids Res,* 34: 656–659.

Lai, J., Ma, J., Swigonova, Z., Ramakrishna, W., Linton, E., Llaca, V., Tanyolac, B., Park, Y.J., Jeong, O.Y., Bennetzen, J.L. and Messing, J. 2004. Gene loss and movement in the maize genome. *Genome Res,*14: 1924–1931.

Lata, C. and Prasad, M. 2013. Validation of an allele-specific marker associated with dehydration stress tolerance in a core set of foxtail millet accessions. *Plant Breed,* 132: 496–469.

Moore, G., Devos, K. M., Wang, Z. and Gale, M. D. 1995.Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol,*5: 737–739.

Panwar, P., Jha, A., Pandey P. K. and Kumar, A. 2011. Functional markers based molecular characterization and cloning of resistance gene analogs encoding NBS-LRR disease resistance proteins in finger millet (*Eleusine coracana*). *Mol Biol Rep,* 38: 3427–3436.

Paterson, A.H., Bowers, J.E. and Chapman, B.A. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc. Natl Acad. Sci,* 101: 9903–9908.

Periyannan, S., Bansal, U., Bariana, H., Deal, K., luo, M. C., Dvorak, J. and Lagudah, E. (2014). Identification of a robust molecular marker for the detection of the stem rust resistance gene Sr45 in common wheat. *Theor. Appl. Genet,* 127 (4): 947-955.

Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin.Plant Biol,*5: 94–100.

Robinson, A.J., Love, C.G., Batley, J., Barker, G. and Edwards, D. 2004.Simple sequence repeat marker loci discovery using SSR primer. *Bioinformatics,* 20: 1475–1476.

Salse, J., Piegu, B., Cooke, R. and Delseny, M. 2004. New *in silico* insight in to the synteny between rice (*Oryza sativa* L.) and maize (*Zea mays* L.) highlights reshuffling and identifies new duplications in the rice genome. *Plant J,* 38(3): 396-409.

Sanger, F., Nicklen, S. and Coulson, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci,* 74(12): 5463–5467.

The 3,000 Rice Genome Project: The Rice 3,000 Genome Project. GigaScience Database. 2014. *http://dx.doi.org/10.5524/200001*

Varshney, R. K., Close, T. J. Singh, N. K., Hoisington, D. A. Cook, D. R. 2009. Orphan legume crops enter the genomics era. Curr opin Plant Biol, 12:202-210.

Varshney, R. K., Graner, A. and Sorrells, M. E. 2005c. Genic microsatellitemarkers in plants: features and applications. *Trends Biotechnol,* 23:48–55.

Varshney, R. K., Sigmund, R., Borner, A., Korzunb, V., Steina, N., Sorrellsc, M. E., Langridged, P. and Graner, A. 2005b. Inter-specific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci,* 168: 195–202.

Varshney, R. K., Thiel, T., Stein, N., Langridge, P. and Graner, A. 2002. In silico analysis onfrequency and distribution of microsatellites in ESTs of some cereal species. *Cell Mol Biol Lett,*7:537–546.

Virk, P. S., B. V. Ford-Lloyd, M. T. Jackson, H. S. Pooni, T. P. Clemeno *et al.*, 1996. Predicting quantitative variation within rice germplasm using molecular markers. *Heredity,*76: 296–304.

Yang, Z., Bai, Z., Li, X., Wang, P., Wu, Q., Yang, L., Li, L. and Li, X. 2012. SNP identification and allelic-specific PCR markers development for TaGW2, a gene linked to wheat kernel weight. *Theor Appl Genet,* 125: 1057–68.

Yu, J., Fleming, S.L., Williams, B., Williams, E.V., Li, Z., Somma, P., Rieder, C.L., Goldberg, M.L. 2004. Greatwall kinase: a nuclear protein required for proper chromosome condensation and mitotic progression in Drosophila. *J. Cell Biol,* 164(4): 487-492.