7

Genomics-Assisted Plant Breeding in the 21st Century: Technological Advances and Progress

Siva P. Kumpatla[1], Ramesh Buyyarapu[1],
Ibrokhim Y. Abdurakhmonov[2] and Jafar A. Mammadov[1]
[1]Department of Trait Genetics and Technologies, Dow AgroSciences LLC,
[2]Center of Genomic Technologies, Institute of Genetics and Plant Experimental Biology, Academy of Sciences of Uzbekistan,
[1]USA
[2]Uzbekistan

1. Introduction

One of the key global challenges of the 21st century is the production of enough food for the increasing world population. As per some recent reports the global population will continue to grow with some 9 billion people by the middle of the current century and the world will need 70 to 100% more food by that time (Godfray et al., 2010 and references therein). Agricultural productivity needs to be increased while addressing the issues of scarcity of arable land and water, impact of changing climate and preservation of natural resources. Improvement of crop yields on available agricultural land requires concerted efforts using modern scientific and technological advances in multiple disciplines (Hubert et al., 2010). Two such disciplines that have revolutionized crop improvement in the recent decades are molecular breeding and plant genomics. While the availability and application of molecular markers have accelerated the pace and precision of plant genetics and breeding, the introduction of a multitude of "omics" tools has provided unprecedented ability to dissect the molecular and genetic basis of traits as well as the characterization of whole genomes.

Molecular markers have occupied center stage in plant genetics since late 1980s. The advent of markers based on simple sequence repeats (SSRs) and single nucleotide polymorphisms (SNPs) and the availability of high-throughput (HTP) genotyping platforms have further accelerated the generation of dense genetic linkage maps and the routine use of the markers for marker-assisted breeding in several crops (Collard and Mackill, 2008). However, despite the routine use of markers for genome-wide profiling and trait-specific marker-assisted selection (MAS), breeding of crops with many traits of interest such as yield, improved nutritive value and resistance to several biotic and abiotic stresses is still a challenge due to complex inheritance of these traits. Therefore, there is a dire need for the molecular dissection of these traits in the context of the whole genome. This is where plant genomics plays a key role by providing the knowledge base required for the understanding and improvement of these traits. Genome is defined as a haploid (single set) content of all of the hereditary information of an organism, and genomics is the scientific discipline that studies the genome at the structural and functional levels towards understanding the genetic basis of inheritance, molecular basis of important intragenomic biological phenomena and the evolutionary history of genes. Plant genomics has enormous potential to revolutionize crop improvement by providing extensive knowledge from the analysis of genomes which in turn can be used for rapid and efficient plant breeding towards crop improvement. In the following sections we have reviewed prominent genomics tools and how technological advances in these as well as associated tools are contributing to the progress towards genomics-assisted plant breeding of 21st century.

2. Genomics tools and technologies and their applications

2.1 Structural genomics: random, targeted and whole genome approaches

Structural genomics is an approach in molecular genetics that enables researchers to detect segments of DNA with allelic variations, correlate those polymorphisms with phenotypic data and determine causative mutations underlying important traits. The scope of "structural genomics" discussed here needs to be distinguished from that coined by protein community where similarly-named approach has been used to investigate the comprehensive repertoire of protein folds to infer molecular functions of the proteins (Burley et al., 1999). Although the main goal of structural genomics is similar in both cases, i.e. from structure to function, researchers use different paths to achieve the final goal.

2.1.1 Molecular markers: development and applications

Structural allelic alterations, or polymorphisms, of a genome can be grouped into three major categories that include differences in the number of tandem repeats at a particular locus [microsatellites, or simple sequence repeats (SSRs)] (Weber and May, 1989), segmental insertions/deletions (InDels) (Ophir and Graur, 1997) and single nucleotide substitutions [single nucleotide polymorphisms (SNPs)] (Wang et al., 1998). In order to detect and track allelic variations in progeny, the scientific community has been developing genetic tools, called molecular markers, since the late 1980s (Botstein et al., 1980). Although SSRs, InDels and SNPs are the three major allelic variations discovered so far, a plethora of molecular markers have been developed to detect the above-mentioned polymorphisms (Bernardo, 2008; Gupta et al., 1999). The main drivers for the evolution of molecular markers have been throughput, level of reproducibility and cost reduction (Bernardo, 2008). Depending on the detection method and throughput, all molecular markers can be divided into three major groups: (1) low-throughput, hybridization-based markers such as restriction fragment length polymorphism [RFLP (Botstein et al., 1980)], (2) medium-throughput PCR-based markers, that include random amplification of polymorphic DNA (RAPD) (Welsh and McClelland, 1990; Williams et al., 1990), amplified fragment length polymorphism (AFLP) (Vos et al., 1995) and SSRs (Wang et al., 1998), and (3) HTP sequence-based markers: SNPs (Wang et al., 1998). In late eighties, RFLPs were the most popular molecular markers and were widely used in plant molecular genetics, because they were reproducible and co-dominant. However, the detection of RFLPs was very expensive, labor- and time-consuming process, which made these markers eventually obsolete. Additionally, RFLP markers were not amenable for automation. Invention of PCR technology and application of this method for the rapid detection of polymorphisms overthrew low-throughput RFLP markers, and new generation of PCR-based markers emerged in the beginning of 1990s. RAPD, AFLP and SSR markers are the major PCR-based markers that research community has been using in various plant systems. RAPDs were able to simultaneously detect polymorphic loci in various regions of a genome. However, they were anonymous and the level of their reproducibility was very low due to the non-specific binding of short, random primers. Although AFLPs were anonymous too, the level of their reproducibility and sensitivity was very high owing to the longer +1 and +3 selective primers and the presence of discriminatory nucleotides at 3′ end of each primer. That is why AFLP markers are still popular in molecular genetics research in crops with little to zero reference genome sequence available (Zhang et al., 2011). However, AFLP markers did not find widespread application in molecular genetics and molecular breeding applications, because the detection method was too long, laborious and not amenable to automation. Therefore, it was not surprising that in the beginning of 21st century SSR markers were declared as “markers of choice” (Powell et al., 1996). SSRs were no longer anonymous; they were highly-reproducible, highly-polymorphic, and amenable to automation. Despite the cost of detection remaining high, SSR markers pervaded all areas of plant molecular genetics and breeding. However, during the last five years, the hegemony of medium-throughput SSRs was eventually broken by SNP markers. First developed for human genome, SNPs have proven universal and are the most abundant forms of genetic variation among individuals within a species (Rafalski, 2002). Although SNPs are less polymorphic than SSR markers because of their bi-allelic nature, they easily compensate this drawback by being abundant, ubiquitous and amenable to high and ultra-high-throughput automation. Since SNPs are currently the most widely used markers in plant molecular genetics and breeding, they are discussed in great detail in the following sections.

2.1.2 SNP markers

Development of SNP markers usually consists of two parts: SNP discovery and SNP validation. SNP discovery in crops is not an easy task because of genome complexity and often the lack of reference genome sequences. Even in crops such as maize (*Zea mays*), where a reference genome sequence is available, large scale SNP discovery efforts are still impeded by the highly repetitive (Meyers et al., 2001) and duplicated (Gaut and Doebley, 1997) nature of the genome. In order to avoid repetitive sequences, maize researchers have focused on the discovery of SNPs within coding sequences by re-sequencing amplicons derived from unigenes (Wright et al., 2005) or by *in silico* mining of SNPs within ESTs (Batley et al., 2003). The advantage of these approaches is the detection of gene-based SNPs. However, both approaches have some drawbacks: they are low-throughput and are unable to detect SNPs located in low-copy non-coding regions and intergenic spaces. Additionally, amplicon re-sequencing is an expensive and labor intensive procedure (Ganal et al., 2009). The recent emergence of next generation sequencing (NGS) technologies such as 454 Life Sciences (Roche Applied Science, Indianapolis, IN), Hiseq (Illumina, San Diego, CA), SOLiD and Ion Torrent (Life Technologies Corporation, Carlsbad, CA) have elevated expectations towards the rapid genome-wide identification of a large number of SNPs at a much lower price tag (Mardis, 2008a). However, efficient application of these technologies for SNP discovery in a given crop depends on the availability of the reference genome sequence (Ganal et al., 2009) as well as the level of genome complexity. For instance, in maize, the availability of a reference sequence does not guarantee a painless SNP discovery using NGS technologies. The complexity and existence of re-arrangements in the maize genome complicate the assembly of short-read NGS sequences and their alignment to the reference genome (Morozova and Marra, 2008). Thus, the reduction of genome complexity becomes an important prerequisite for the genome-wide discovery of true SNPs in crops with and without reference genome using sequencing by synthesis (SBS) technologies. Several genome complexity reduction techniques have been developed, including High C_0t (DNA renaturation kinetics C_0t) selection (Yuan et al., 2003), methylation filtering (Emberton et al., 2005; Palmer et al., 2003), and microarray-based genomic selection (Okou et al., 2007). However, a majority of these techniques mainly reduce the number of repetitive sequences and are ineffective in the recognition and elimination of paralogues and homoeologues, which cause the detection of false-positive SNPs. Recently, computational SNP calling methods were developed that can drastically reduce the number of false SNPs resulting from the alignment of duplicated sequences and re-sequencing errors (Baird et al., 2008; Barbazuk et al., 2007; Gore et al., 2009; Van Orsouw et al., 2007). Hence, the availability of reference sequences, the application of genome complexity reduction techniques and NGS technologies coupled with post-re-sequencing computational treatment become important prerequisites for genome-wide detection of SNPs in complex genomes.

2.1.3 SNP validation and modern genotyping platforms and chemistries

The availability of reference sequence and sophisticated software do not always guarantee that the discovered SNP can be converted into a valid marker. In order to insure that the discovered SNP is a Mendelian locus, it has to be validated. The validation of a marker is a process of designing an assay based on the discovered polymorphism and genotyping a panel of diverse germplasm or segregating population. Segregating population is more informative as a validation panel than a collection of unrelated lines, because it not only allows inspection of the discriminatory ability of a marker but also its segregation patterns and ratios which helps researcher to understand whether it is a Mendelian locus or a duplicated/repetitive sequence that escaped software-filter (Mammadov et al., 2010).

In plants, SNPs can be validated using flexible and HTP assays, chemistries and genotyping platforms, including Illumina's BeadArray technology-based GoldenGate (GG) (Fan et al., 2003) and Infinium assays (Steemers and Gunderson, 2007), Life Technologies' TaqMan (Livak et al., 1995) assay coupled with OpenArray platform (TaqMan OpenArray Genotyping system, Product bulletin, 2010) and KBiosciences' Competitive Allele Specific PCR (KASPar) complemented with the SNP Line platform (SNP Line XL; http://www.kbioscience.co.uk). These modern genotyping assays and platforms differ from each other in chemistry, cost and throughput of samples to genotype and number of SNPs to validate. The choice of chemistry and genotyping platform depends on many factors that include the length of SNP fragment sequence, overall number of SNPs to genotype and finally the funds available to the research unit because most of these chemistries remain cost-intensive. Below is the summary of four SNP genotyping assays and platforms, which have been widely used in academia and industry.

2.1.3.1 Illumina's BeadArray platform

The Illumina's BeadArray platform (Fan et al., 2003) is capable of validation of a large number of SNPs in parallel by combining several technologies. The core of the technology is a collection of 3-micron silica beads that get self assembled in the wells, which are etched on the surface of a miniaturized matrix (either fiber optic bundles or planar silica slides) and evenly spaced at ~5.7 micron distance. Each bead is covered with hundreds of thousands of copies of a specific oligonucleotide that act as the capture sequences in one of Illumina's assays such as GG and Infinium. A high-resolution confocal scanner (iScan) is engineered to read arrays and generate intensity data, which is converted into genotypic data by reliable genotype-calling software, GenomeStudio. GG and Infinium are highly multiplexed chemistries and can genotype a maximum of 3,072 SNPs and ~1.1 million attempted bead types, respectively, in a single reaction without adverse effect on allele discrimination. All previous multiplexing efforts by various companies and academic labs had limited success mostly because of the interactions of primers and discrimination of alleles during amplification. In GG and Infinium, primers do not interact with each other. In GG assay all 3,072 loci are amplified with the same trio of universal primers (namely, P1, P2 and P3). In addition, allele discrimination occurs prior to PCR. Last but not least, small, newly synthesized DNA fragments, not entire genomic DNA, serve as templates for PCR amplification, which dramatically reduces the complexity of PCR reactions. Although the whole SNP genotyping process using the GG assay takes three days, the GG assay is a combination of simple molecular biology techniques, which is easy to follow and implement. In contrast to GG where all 3,072 assays [oligo pool assay (OPA)] are manufactured as a suspension in a single tube, in case of Infinium all assays are immobilized via beads on the surface of a chip. Depending on SNP type, two types of assay can be designed: Infinium I and Infinium II. Infinium I is designed for [A/G] and [T/C] SNPs and requires one bead type per allele (two bead types per SNP), while Infinium II is designed for all other SNPs and requires only one bead type for both alleles. That is why, the calculation of the price of Infinium assay is based on number of attempted bead types but not SNPs. The entire set of attempted bead types is called iSelect, which is an equivalent of OPA in GG assay. In both assays, the number of samples processed per day is restricted to three 96-well plates because of limited capacity of a liquid handler TECAN, which is a part of automation in BeadArray technology. Thus, BeadArray technology coupled with GG and Infinium assays is a robust and high-throughput platform designed to validate a large number of SNPs with relatively small number of samples: minimum 24 (one beadchip) and maximum 288 samples (12 beadchips).

2.1.3.2 The OpenArray technology

The chemistry of OpenArray technology (Brenan and Morrison, 2005) is based on Life Technologies' end-point TaqMan assay. The arrays require assays and samples on a small OpenArray plate. The physical size of an OpenArray plate is 1/8 of the size of 384-well plate. However, unique structural design of OpenArray plate allows accommodation of 3,072 assays within one plate, which is equal to the capacity of eight 384-well plates. The plate has 48 subarrays and each subarray has 64 holes, where nano-volumes of DNA get loaded. Another great feature of this platform is that it is very flexible and allows the user to array different SNP vs. sample combinations, including 64 x 48, 128 x 24, 256 x 12, 192 x 16, 32 x 96 and 16 x 144 formats. Another important feature of the OpenArray plate is that the hydrophobic and hydrophilic coatings of the surface and holes, respectively, enable reagents to stay in the bottomless through-holes via capillary action. OpenArray plates are preloaded with assay reagents by a vendor, and sent to the end-user to load DNA samples. The throughput of SNP genotyping using OpenArray technology can be greatly increased by attaching slide towers on top of the DNA engine, i.e. thermocycler. Each slide tower can harbor 32 slides. Throughput can be increased by using several two to three thermocyclers with slide-towers. For example, if 128 x 24 format is used, 128 SNPs can be validated with 2,048 (32 x 24 x 2) samples per day using two thermocyclers. In contrast to BeadArray platform, OpenArray technology can validate relatively small subset of SNPs with larger number of samples, which makes it very attractive for marker-assisted breeding projects, where gene or QTL region must be tracked by a few markers within a large number of samples.

2.1.3.3 KBiosciences' competitive allele specific PCR (KASPar)

In addition to above platforms, the Competitive Allele Specific PCR (KASPar) from KBiosciences (Hoddesdon, Hertfordshire, UK) (http://www.kbioscience.co.uk) is widely used in SNP validation, although it does not have any multiplexing capabilities. However, this chemistry is becoming widely used in SNP validation. KASPar assay uses a technique based on allele specific oligo extension and fluorescence resonance energy transfer (FRET) for signal generation. The fluorescent reporting system comprises of four single-labeled oligonucleotides that hybridize to one another in free solution to form a fluorescent quenched pair which upon introduction of complementary sequences generates a measurable signal. The kit requires two components, the assay mix (a mixture of three unlabelled primers: two allele specific oligos and one common reverse locus specific oligo) and the reaction mix (the other components required for PCR, including the universal fluorescent reporting system and Taq polymerase). KASPar is a very flexible assay, because SNP validation can be carried out in a variety of formats and the chemistry has been shown to function well in 96, 384 and 1536-well plates. One of the most attractive features of KASPar is the cost effectiveness and the duration of the synthesis of the assay. One KASPar assay will cost ~$15 and results can be delivered next day. Compared to KASPar assay, cost per one TaqMan and GG assays as well as Infinium bead type will be around $400, $42 and ~$9, respectively. Duration of synthesis of Taqman assay, GG OPA and Infinium iSelect is two, six and nine weeks, respectively. Also, depending on the size of the validation panel, GG and Infinium assays might not be suitable to validate SNPs, because Illumina imposes minimum sample order limitation per OPA and iSelect, which are 480 and 1152 samples, respectively. Finally, the choice of chemistry and genotyping platform for validation will also depend on the length of the context sequence based on which one can develop an assay. The length of context sequence is a crucial factor because most of the modern genotyping chemistries have a strict requirement for the length of the template strand. For example, 70 nucleotide (nt) short reads generated by Illumina's HiSeq NGS instrument can be suitable for validation using GG, Infinium and KASPar assays, which require minimum 50 nt template sequence from both sides of a SNP. At the same time, HiSeq output might not be suitable for TaqMan assay design which requires a longer input sequence (100 nt), because it needs enough space for a probe and two oligos flanking the SNP. Thus, there is no ideal genotyping platform and assay that a researcher can leverage for SNP validation. The choice of assay and genotyping system will depend on the number of SNPs, length of the template sequence, sample size, time-sensitivity of a project and the funds available to the researcher.

2.1.4 SNP Application

When SNP markers have passed the validation step, they are considered as viable markers and ready for use in various areas of molecular genetics and plant breeding, including gene/QTL mapping, linkage-disequilibrium-based association mapping, map-based gene/QTL cloning, germplasm characterization, genetic diagnostics, event characterization, marker-assisted trait introgression, and finally marker-assisted selection (MAS). In order to conduct most of the above-mentioned SNP applications, researcher must know the order of the markers on chromosomes, which can be obtained by constructing recombination-based genetic linkage maps. Genetic mapping is carried out using segregating populations, including F_2, backcross, recombinant inbred lines (RILs) or doubled haploids (DHs). Currently, most of the major crops possess genetic maps densely saturated with molecular markers. Publicly available genetic linkage maps that are constructed solely based on SNPs currently exist for rice (http://www.gramene.org) and maize (http://www.maizegdb.org) only. Remaining crops have genetic maps, which have been constructed by means of SNPs in combination with other markers such as SSRs and RFLPs that include barley, wheat, sorghum (http://wheat.pw.usda.gov/ggpages/map_shortlist.html) and soybean (http://soybase.org/).

2.2 Other key "Omics" tools needed for structural genomics work

Genotypic and the corresponding phenotypic data are the two major components required for understanding the genetic basis of traits through genetic linkage analysis. While advances in molecular marker fingerprinting and next generation sequencing are enabling economical and HTP genotyping of samples (Peleman and van der Voort, 2003), phenotyping of a large number of samples under field conditions is still a bottleneck. It is a laborious and expensive task and is a serious drawback for the dissection of complex and dynamic traits such as abiotic stress tolerance, yield and nutrient use efficiency where data needs to be collected form really large populations for efficient genetic analysis. Because of this drawback, many research efforts to date have treated dynamic traits as static traits and have relied on only one measurement for analysis, that too on small populations. Gathering multiple data points is even difficult for traits for which root measurements are needed. In order to address this situation, several HTP phenotyping techniques have been conceived and implemented. The discipline focused on developing such HTP phenotyping tools and platforms is termed as 'phenomics'. An example of HTP phenotyping is a near-infrared spectroscopy equipment, mounted on agricultural harvesters that can be used to collect spectral information about the plants during the harvesting of field trials (Montes et al., 2006). Spectral information thus collected can be condensed into a single near-infrared spectrum and analyzed using calibration models for the determination of information on several traits. Spectral reflectance of plant canopy using light curtains and spectral reflectance sensors mounted on a tractor is another phenotyping technique which non-invasively monitors several dynamic and complex traits such as biomass accumulation (Montes et al., 2011). The domain of phenomics concerned with the measurement of phenome (measurement of physical and biochemical attributes or phenotypes of traits of interest) has seen commendable efforts in the recent years in the automation of plant phenotyping. Several automated platforms and approaches such as Phenopsis (automated growth chambers for growing 504 pots of *Arabidopsis thaliana* at a time), Phenodyn (for simulating drought conditions and measuring transpiration and growth), Growscreen (digital imaging and processing for growth rate determination), Traitmill (fully automated growth facility) and LemnaTec (automated greenhouse) now exist for the HTP collection of plant phenotypic data on several traits of interest [reviewed in Kolukisaoglu and Thurow (2010)]. Availability of such automated phenotyping methods holds a great promise for the molecular and genetic dissection of complex traits by integrating this information with that of multiple datasets resulting from HTP genotyping as well as diverse 'omics' efforts.

2.3 Next generation sequencing

Improvements in crop productivity require adoption of new breeding technologies. Integration of genomic and transcriptomic data provides an opportunity to generate newer molecular resources for improved breeding technologies and crop improvement. Availability of DNA/RNA sequence information is highly critical to develop such resources. Until recently, sequencing efforts were dominated by Sanger sequencing method. Initial draft of human genome sequence was generated using BAC-by-BAC approach using Sanger's sequencing method by investing approximately three billion dollars into Human Genome Project (Venter et al., 2001). The availability of human genome reference sequence paved the way to a multitude of applications including detection of structural and copy number variations to understand the underlying genetic and epigenetic mechanisms. Though Sanger sequencing method dominated the industry for almost two decades and still considered the gold standard for sequencing, its limitations, especially with respect to throughput and cost, necessitated high demand for new and improved technologies for sequencing large and complex genomes. With advances made in the fields of microfluidics, microscale imaging, detection and computational tools, alternative sequencing technologies with increased throughput and lower sequencing cost are continuously emerging. These alternative technologies to Sanger's sequencing can be collectively termed as Next generation sequencing (NGS) technologies (Varshney et al., 2009). Since NGS technologies are impacting several 'omics' efforts, they are discussed in extended detail below.

2.3.1 NGS technologies

The advent of NGS technologies has changed the dynamics and the pace of genomic research in humans, plants, animals and microorganisms because of their rapid, inexpensive and highly accurate sequencing capabilities. Unlike Sanger sequencing method which depends upon capillary electrophoresis, these NGS technologies are highly dependent on massive parallel sequencing, high resolution imaging, and complex algorithms to deconvolute the signal data to generate sequence data. NGS technologies offer a wide variety of applications such as whole genome *de novo* and re-sequencing, transcriptome sequencing (RNA-seq), microRNA sequencing, amplicon sequencing, targeted sequencing, chromatin immunoprecipitated DNA sequencing (ChIP-seq), methylome sequencing and many others. Before dwelling into the use of this wide variety of NGS applications for crop improvement, various NGS technologies and their capabilities are briefly reviewed first.

Current NGS technologies can be broadly grouped into long and short read length technologies based on the number of bases they can sequence in a single sequencing reaction. Long read length technologies are preferred for applications involving *de novo* sequencing while short read length technologies are relatively inexpensive and mostly used for re-sequencing applications. Most of the NGS technologies monitor millions of sequencing reactions in parallel and thus result in a massive amount of sequencing data. The output capacities of these instruments outpaced the development of computational tools and hardware for data processing needs. Sophisticated computer programs are created to handle and process large amounts of sequencing data before final data analysis. Several bioinformatics tools were designed for diverse purposes such as *de novo* sequence assembly, mapping sequences to an existing reference genome sequence, mutation detection and annotation. Long read technologies include Roche/454 GS FLX and Pacific Biosciences RS systems while short read technologies include Illumina Genome Analyzer IIx, HiSeq 2000, MiSeq, Life Technologies' SOLiD™ system, Helicos Genetic Analysis system and Life technologies/Ion Torrent Personal Genome Machine (PGM). Mardis (2008b) and Metzker (2009) provided detailed reviews of these NGS technologies. NGS technologies that are widely used at present are briefly reviewed below and sequencing capabilities of instruments are summarized in Table 1.

2.3.1.1 Roche/454 GS FLX – pyrosequencing

This is the first NGS technology commercially introduced and is based on pyrosequencing method (Margulies et al., 2005). This technology is relatively rapid and inexpensive as it omits the expensive *in vivo* sub-cloning of sheared fragments for template amplification. Instead of cloning, sheared fragments are attached to microbeads and amplified in an emulsion-based PCR. These microbeads are further distributed to a fiber optic slide (PicoTiterPlate™), where the four dNTPs are added in turns. In pyrosequencing, the DNA sequence is determined by analyzing the fluorescence emitted by the activity of luciferase during the process of template extension by a single nucleotide addition. The fluorescence emitted is captured by a high resolution CCD camera for each type of nucleotide passed in a flow cycle. The intensity of the fluorescence is proportional to the number of nucleotides integrated in each step. The first commercial 454 instrument was able to generate >25 milion bases in short reads of 100 bp or more per 4 hr run. With the improvements in sequencing chemistry, PicoTiterPlate (PTP), reagent volumes and the number of nucleotide flow cycles in the instrument, the current GS FLX plus instrument was able to achieve an average read length of ~750 bp across 1 – 1.5 million sequences in ~20 hr runtime. Long read length capabilities of this instrument enable *de novo* sequencing of genomes and transcriptomes with ease compared to short read technologies. However, this technology is prone to sequencing errors in the homopolymer regions. Since the advent of 454 sequencing technology, there are ~1331 peer reviewed publications as of July, 2011 (http://454.com/publications/all-publications. asp) across a wide range of topics.

2.3.1.2 Illumina Genome Analyzer/HiSeq/MiSeq – sequencing-by-synthesis

Illumina sequencing method utilizes clonal array formation and proprietary reversible terminator reaction chemistry for rapid, accurate and large scale sequencing. DNA template fragments were immobilized in an 8-channel microfabricated flow cell where they were amplified up to 1000 copies in close proximity by bridge amplification method. Sequencing-by-Synthesis uses all four fluorescently labeled nucleotides to sequence millions of clusters on the flow cell surface. The fluorescent label in each nucleotide blocks the 3'–OH group and thus acts as a terminator for polymerase extension. At the incorporation of each nucleotide,

Platform	Template preparation	Method	No. of reads	Read length (bases)	Run time	Throughput (Gb)	Advantag
Roche/454's GS FLX +	emPCR	Pyrosequencing	1 - 1.5 million	750	18-20 hrs	0.75	Longer readlengths, time, highly useful fo sequencing applicati
Illumina HiSeq 2000	bridge amplification	Sequencing-by-synthesis using reversible terminator nucleotides	3^1 or 6^2 billion	100	6 or 11 days	150^1 or 300^2 per flow cell	Very high throughpu base sequecing cost a resequencing applica
Life technologies / ABI's SOLiD 5500	emPCR	Sequencing-by-Ligation using di-base probes	1.4 billion	35-75	7 or 14 days	90	HTP, lower per base cost and useful for re applications
Polonator G.007	emPCR	Sequencing-by-Ligation using non-cleavable probes	7-12 million	26	5 days	12	Open source and less NGS platform
Helicos Biosciences Heliscope	single molecule	True single molecule sequencing	0.6 - 1 billion	35	8 days	37	Non-biased represen templates and high t
Pacific Bioscience's RS	single molecule	Single molecule real time sequencing	30,000-50,000	1000 - 10000	30 min	0.03 - 0.05 per SMRT cell	Extra longer read len run times, highly use *novo* sequencing
Ion Torrent PGM	bead based	Semiconductor sequencing	1 - 12 millionᶲ	200	2 hrs	0.01 - 1	Shorter run times, re cheap NGS technolog Useful for amplicon

* Based on the information provided by Metzker (2009) and other company web resources

[1] Single end read chemistry, [2]paired end read chemistry

ᶲ Sequence capacity change with the type of chip used for sequencing.

Table 1. Comparison of NGS technologies and capabilities *

fluorescent dye is imaged to identify the dye and then the label is enzymatically cleaved to allow the incorporation of next base (Bentley et al., 2008; Ju et al., 2006). As each nucleotide base incorporation is a unique event, the error rate in homopolymer regions is minimal compared to 454 pyrosequencing method (http://www.illumina.com/technology/sequencing_technology.ilmn). Illumina has a range of sequencing instruments that can generate from ~1 Gibabase (Gb) from ~3-6 million sequences (MiSeq) and up to 600 Gb from 6 billion paired end reads per two flow cells (HiSeq 2000) in a single sequencing run. Though the output capabilities of Illumina sequencing instruments are large, they also take longer sequencing time from 3 – 11 days depending on the machine, single end or paired end protocol and number of flow cycles. This technology has revolutionized the pace of re-sequencing efforts in human and other genomes besides bringing down per base cost to a bare minimum. As of July, 2011, there are ~1746 peer reviewed publications that have used this technology.

2.3.1.3 Life technologies SOLiD™ – Sequencing-by-Ligation

Life technologies, previously Applied Biosystems, developed another short read sequencing technology which utilizes sequencing-by-ligation method. Template DNA fragments are clonally amplified in an emulsion PCR reaction similar to that of 454 sequencing and the clonal bead populations are covalently bound to a slide by 3′ modification of the beads. During the sequencing reaction, a fluorescently labeled di-base probe hybridizes to the complementary sequence adjacent to primed template and DNA ligase enzyme joins the dye-labeled probe to the primer. After the non-ligated probes are washed off, fluorescence is imaged to identify the nucleotides incorporated at first and second base (http:// www.appliedbiosystems.com/absite/us/en/home/applications-technologies/solid-nextgeneration-sequencing/next-generation-systems/solid-sequencing-chemistry.html). The cycle can be repeated either by using cleavable probes to remove the fluorescent dye and regenerate a 5′ -PO_4 group for subsequent ligation cycles or by removing and hybridizing a new primer to the template (Metzker, 2009; Valouev et al., 2008). SOLiD 5500, the recent version of this technology can generate up to 90 Gb of sequence data from ~1.4 billion reads of 35-75 bases in length over ~7 days of time. Due to its massive outputs and short read length capabilities, this system is heavily used for re-sequencing and RNA-Seq applications.

2.3.1.4 Pacific Biosciences *RS* – SMRT™ (single molecule real time) sequencing

Pacific Biosciences developed SMRT technology which implements detection of fluorescently labeled nucleotides as they are incorporated over a single DNA molecule in real time. A single Φ-29 DNA polymerase enzyme molecule, a highly processive and strand displacing enzyme, is immobilized in a small hole called zero-mode wave guide (ZMW) to process the extension of a single molecule of primed DNA template (Eid et al., 2009). Four color phospholinked dye labeled nucleotides are used in this process and their fluorescence is quenched until they are incorporated during the sequencing reaction (Korlach et al., 2008). In the ZMW, as each nucleotide is incorporated by the anchored DNA polymerase, the phospholinked dye label is cleaved and its fluorescence light pulses are captured by four single photon sensitive cameras in the sequencing instrument (Lundquist et al., 2008). The real time light pulse information coming from 75000 ZMWs in a SMRT cell is converted to A, C, G, or T based on quality metrics to provide the sequencing information. The biggest advantage of this technology is longer read lengths of ~1000 – 10000 bases which facilitates easy sequence assemblies especially for *de novo* sequencing applications. As the sequencing reaction in a SMRT cell is monitored in real time, each typical sequencing run requires as little as 30 minutes compared to other technologies which can take up to 11 days. Strobe sequencing was used to achieve higher read lengths with higher accuracy (Lo et al., 2011). Though the cost of sequencing is relatively cheap, observed sequencing error rates are higher compared to other NGS technologies.

2.3.1.5 Life technologies/Ion Torrent PGM – Semi conductor sequencing

Ion Torrent PGM machine uses semi conductor technology with simple, non-fluorescent sequencing chemistry to generate the sequencing information. It is based on the detection of H^+ ions released (pH change) during a natural polymerase reaction using an ion sensor underneath the micro machined wells in a semiconductor chip, each containing a different DNA template. As each nucleotide flows in one at a time during the sequencing reaction, pH change is observed in all wells where the complementary nucleotide is incorporated (Pennisi, 2010). Change in pH is relative to the number of bases added to the template strand, and thus can sequence the homopolymer regions. As there is no involvement of fluorescent labeled nucleotides or imaging, incorporation of each nucleotide is recorded in seconds and the cost of sequencing is relatively cheap compared to other NGS technologies (http://www.iontorrent.com/technology-scalability-simplicity-speed/). Current read lengths are ~200 bp and each run takes about 2 hrs.

Existing and emerging NGS technologies are helping to bring down the sequencing costs towards making personalized genome services, personalized medicine and other applications possible in the near future. Third generation sequencing technologies such as Oxford's nanopore sequencing and VisiGen's nano sequencing technologies are currently being developed and would help the genome research more affordable than any time before.

2.3.2 NGS Applications for crop research

Widely available and cost-effective NGS technologies enabled many exciting opportunities for crop research in plants with or without a reference genome. Availability of reference genome/transcriptome sequence greatly enhances our ability to decipher the underlying molecular mechanisms of a trait, understand the gene regulatory mechanisms, determine gene expression differences and variations in expressed gene sequences, and other structural variations such as copy number variations (CNV) and presence-absence variations (PAV). NGS technologies can be applied to answer a wide variety of biological questions such as sequencing of complete genomes and transcriptomes and genome wide analysis of DNA-protein interactions (Bräutigam and Gowik, 2010). To facilitate crop improvement, NGS and other accessory technologies can be used for whole genome sequencing, transcriptome sequencing, genome wide and candidate gene marker development, targeted enrichment and sequencing and other applications. These NGS technologies even hold promise for a methodological leap towards genotyping–by-sequencing (GBS) and genetic mapping applications. Analysis of NGS data from genome wide association studies, transcriptomics and epigenomics in combination with data from proteomics, metabolomics and other 'omics' can provide an integrative systems biology approach to understand the regulation of complex traits.

2.3.2.1 Whole genome de novo / re-sequencing

Recently, whole genome sequencing efforts of many plant species including model and non-model crop species gained momentum due to lower sequencing costs and turnaround time. Genomes of model plant species such as Arabidopsis, rice and maize have been sequenced using Sanger sequencing method. *De novo* and re-sequencing of the genomes of several crop species with and without reference genome sequence are currently being accomplished by various NGS technologies. NGS technologies typically employ either multiplexed BAC pool sequencing or shotgun sequencing approaches to derive the whole genome sequence. Despite the availability of sophisticated assembly software, *de novo* assembly of large, complex and highly repetitive genomes poses enormous challenges to generate genomic reference sequence. To overcome the challenges of *de novo* sequence assembly of genomes, it is ideal to build a genome scaffold using long read length technologies and then use the short read length technologies to support the consensus sequence and thus minimize sequencing errors. Gaps generated during sequence assembly can be mitigated by using paired end and mate pair library sequencing in both long and short read length technologies. Information from the recent plant genome sequencing efforts using NGS technologies is summarized in Table 2. Genome re-sequencing efforts are currently in progress for crop species with reference genome such as corn and rice to understand important agronomic traits. These efforts are using structural, CNVs and PAVs and comparative genomics approaches for understanding variations, especially in closely related cultivars. Short read length technologies are routinely used for re-sequencing applications and re-sequence data is analyzed by mapping the reads back to genome scaffold to identify different kinds of variations in genic and non-genic regions.

Plant Species	Ploidy	Genome size (Mb)	Sequencing Technology	Reference
Theobroma cacao (Cocoa)	diploid	430	Roche 454	Argout et al. (2011)
Malus × domestica (domesticated apple)	diploid	743	Sanger paired end / Roche 454	Velasco et al. (2010)
Fragaria vesca (Woodland Strawberry)	diploid	240	Rohe 454 /Illumina GA/ABI SOLiD	Shulaev et al. (2011)
Vitis vinifera ssp. sativa (Grapevine)	diploid	504	Sanger paired end / Illumina GA	Velasco et al. (2007)
Jatropha curcas	diploid	410	Sanger WGS/Roche 454/Illumina GA	Zieler et al. (2010)
Elaeis guineensis (Oil Palm)	diploid	1700	Roche 454	Zieler et al. (2010)
Gossypium raimondii (cotton)	diploid	880	Roche 454 / Illumina GA	http://www.jgi.doe.gov/sequencing/why/ gossypium.html
Triticum aestivum ('Chinese Spring' wheat)	Hexaploid	16000	Roche 454	http://www.wheatgenome.org
Musa spp (Banana)	diploid	550-650	Sanger/Roche 454/Illumina GA	Hribova et al. (2009)

Table 2. Examples of whole genome sequencing in crop species.

2.3.2.2 Transcriptome / siRNA / miRNA sequencing

Transcriptome sequencing provides information about functional genes in an organism and helps in gene discovery. Collection of mRNA from different tissues and different stages of plant growth provides a comprehensive set of expressed genes in the template libraries for transcriptional profiling of even non-model organisms. Such libraries can be *de novo* or re-sequenced using NGS technologies more efficiently compared to earlier gene cloning and sequencing methodologies. Although Sanger sequencing method provides longer EST contigs, detection of allelic variations in the gene sequences is an expensive effort. The assembly becomes increasingly difficult when the read length gets shorter and shorter, which is the most compelling reason for choosing a long read technology for *de novo* sequencing. Many protocols are currently available to prepare the template libraries in normalized and non-normalized fashion either from cDNA or directly from mRNA/total RNA for diverse applications. Due to its longer read length capabilities and improved protocols to overcome the inherent problems of homopolymer region sequencing, 454 sequencing is well suited for *de novo* sequencing of EST libraries. Gene regulatory networks and pathways could be easily developed using transcriptome profiling experiments. Transcriptome data is highly useful not only to know the gene content and transcriptional status in various tissues but also helps in identifying SSRs and SNPs in the genic regions, which can be converted to gene-based markers (Narina et al., 2011).

Comparative transcriptomic approaches using NGS technologies can be applied to find functional gene homologs and orthologs in non-model organisms to support the gene discovery efforts. In several model plant species, EST libraries have been deeply sequenced and annotated to provide information on reference transcriptome and led to the creation of transcriptome databases (Morozova and Marra, 2008). For crop species with such annotated reference transcriptome dataset, RNA-seq and serial analysis of gene expression (SAGE) experiments using short read length NGS technologies could provide the expression levels of various genes very cost effectively. Illumina and Life technologies can be used for efficient sequencing of novel variants. They can also facilitate detection of homolog and paralogs of functional genes due to their high data throughput. Apart from expressed genes, other RNA molecules such as microRNA (miRNA), short interfering RNA (siRNA) are also present in the cell and are involved in the regulation of gene expression. Novel miRNA molecules can be easily sequenced from different tissues using NGS technologies to understand the mechanisms of gene regulation. For example, miRNA molecules thus detected during a biotic or abiotic stress condition were utilized to develop improved cultivars through transgenic approaches (Sindhu et al., 2009).

2.3.2.3 Molecular marker development

Genetic variation is the key for implementing molecular breeding approaches in any crop improvement project. Genetic variation is usually detected by identifying the polymorphisms exhibited at restriction site, as fragment lengths, or at single nucleotide levels either in genic or intergenic regions of the genome. Traditionally, the development of markers such as microsatellites, RFLPs and AFLPs was a costly iterative process that involved time-consuming cloning and primer design steps that could not easily be parallelized. In recent years, SNPs have been the markers of choice for the researchers due to their high abundance and amenability for automation and HTP genotyping capabilities.

However, prior availability of sequencing information is absolutely necessary to identify and design assays using SNPs. Genomic and transcriptomic resources can be easily generated by NGS technologies to rapidly and cost-effectively develop molecular markers such as SNPs and SSRs. NGS technologies have been recently used for whole genome and re-sequencing projects where the genomes of several specimens were sequenced to discover large numbers of SNPs for exploring within-species diversity, constructing haplotype maps and performing genome-wide association studies (GWAS) (Elshire et al., 2011).

Genome wide marker development is often achieved by comparing sequences from either whole genome re-sequencing efforts or genome complexity reduction approaches. SNP marker development in plant species with reference genome is relatively easier where the NGS data from different genotypes is mapped against the reference genome to identify SNPs. However, overall size and structure of plant genomes constitute a major hurdle for non-model crop species. Sequencing the whole genome of every individual in a population is costly and often unnecessary, as many biological questions can be answered using polymorphisms that are measured in a subset of genomic regions. Sequencing of libraries generated from reduced representation or target enrichment techniques as well as the DNA fragments resulting from the application of restriction-site associated DNA (RAD) or Complexity Reduction of Polymorphic Sequences (CRoPS) approaches are some of the methods for sampling and sequencing a small set of genome-wide regions without sequencing the entire genome and all these processes are often coupled with NGS technologies. Genome complexity reduction approaches greatly help marker development and have been used for SNP marker discovery in multiple crop and animal species (Davey et al., 2011). Complexity reduction methods include restriction digestion of genomic DNA using methylation sensitive and other restriction enzymes to exclude the repetitive regions and retroposon/transposon sequences during sequencing; and target enrichment for selective sequencing of regions of interest (Deschamps and Campbell, 2010).

Transcriptome data could also be used as a resource for detecting genetic variants in many crop species (Hamilton et al., 2011; Oliver et al., 2011). NGS technologies have been routinely used to generate huge EST datasets by RNA-seq experiments, which can be used for identifying SNPs in functional genes. Comparison of transcriptome datasets from parental genotypes could derive polymorphic SNPs. SNPs derived from gene sequences have higher significance compared to those of non-genic regions, as they can be directly associated with the gene function. Libraries enriched with PCR amplicons from target gene regions were sequenced by Roche 454 technology to use the NGS data as a resource to detect SNP markers in sugarcane, a complex polyploid crop species (Bundock et al., 2009). Bioinformatics tools such as AutoSNP (Wang and Liu, 2011), HaploSNPer (Tang et al., 2008) have been designed to detect the variations in NGS data by mathematical calculations of minor allele frequency or haplotype information. Minor allele frequency can be used as a measure to identify candidate SNPs in simple diploid species while calculation of haplotype information improves the SNP confidence in polyploid crop species such as potato and cotton.

2.3.2.4 Genotyping-by-sequencing and genetic mapping

The development of PCR based markers has revolutionized marker development and genotyping procedures to identify QTL regions associated with important traits. However, these markers, although still widely used, have shown growing limitations in chromosomal coverage, time, and cost effectiveness. The development of genomics concepts and tools and genome-based HTP strategies has provided an alternative approach to marker based mapping approaches. The NGS technologies coupled with the growing number of genome sequences opens the opportunity to redesign genotyping strategies for more effective genetic mapping and genome analysis. Although array-based genotyping methods such as Illumina Infinium iSelect assays provide HTP genotyping ability, it is laborious, time-consuming, and expensive to design, produce, and process arrays suited for specific mapping populations (Huang et al., 2009). Current SNP genotyping technologies often interrogate two alleles at a polymorphic site and this could limit genotyping of other alleles, especially in diverse natural populations, for association mapping.

Advances in next generation technologies have driven the costs of DNA sequencing down to the point that genotyping-by-sequencing (GBS) is now feasible for high diversity, large genome species. The new sequencing techniques not only increase sequencing throughput by several orders of magnitude but also allow simultaneous sequencing of a large number of samples using a multiplexed sequencing. These recent technical advances have paved the way for the development of a sequencing-based HTP genotyping method that combines the advantages of time and cost effectiveness, dense marker coverage, high mapping accuracy and resolution, and more comparable genome and genetic maps among mapping populations and organisms (Elshire et al., 2011)**.**

GBS strategies often depend upon the resources available and type of mapping population used in the study. Availability of reference genome is always encouraged, but not an absolute requirement to implement GBS approaches. GBS experiments are often complemented with sequencing of samples in a mapping population in a multiplexed format and HTP manner to derive the genotyping information. This type of SNP data differ from that of traditional genetic markers primarily in two aspects. First, it is often not the case that all members of a recombinant population can be scored at a given SNP site. Second, an individual SNP site is no longer a reliable marker or locus for genotyping due to several potential sources of sequence errors. To overcome these difficulties, bioinformatics and statistical tools are used to validate the SNPs in a genotype at a specific locus and such tools included sliding window approach where the SNP information is confirmed not only at a single polymorphic site, but also in the flanking regions by verifying the haplotype information.

In rice, Huang et al. (2009) first demonstrated a HTP GBS method by whole genome re-sequencing in a 150 recombinant inbred line (RIL) population, where 16 samples were multiplexed per lane, and a total of 112 samples per flow cell were sequenced using Illumina GA. In this case, the mapping population was derived from only a set of two parents vs. *indica* and *japonica* cultivars. Sequence alignment of the population data and validation of the SNPs using sliding window approach provided the genotype calls for the population. Another parent independent GBS approach was also implemented in rice by Xie et al. (2010), where an ultra-high density linkage map was constructed by low coverage sequencing of mapping population. In this study, genotype calls in the population were derived by maximum parsimonious inference of recombination assisted by Hidden Markov Model (HMM) and were validated with Bayesian inference. This approach can be implemented in crops with no reference genome.

Whole genome re-sequencing is not always an option to implement GBS especially in crop species with large, complex, and repetitive genomes. Although GBS is fairly straightforward for small genomes, target enrichment or reduction of genome complexity must be employed to ensure sufficient overlap in sequence coverage for species with large genomes. Reducing genome complexity with restriction enzymes is relatively easy and reproducible compared to other target enrichment methods such as use of long range PCR, molecular inversion and capture probes etc. Elshire et al. (2011) applied complexity reduction approaches using *Ape*KI, a type II restriction endonuclease, to generate reduced representation libraries and then generated sequence data across these libraries from RIL mapping populations in both maize and barley. They analyzed the data to demonstrate GBS as a proof of concept for routine mapping and QTL identification studies. These studies illustrate and promise eventual application of GBS in introgression programs for traits of interest.

In human research, multiple cancer and other disease traits were investigated by GWAS using NGS technologies and that was possible due to the existence of narrow genetic variation in humans. Though GBS is an attractive option for populations derived from a set of parents, its application is very challenging for association mapping studies in plant species due to huge variations existing in natural populations. In GBS, variations are typically detected by aligning to the reference genome, but in natural populations, the variations are not limited to SNPs but also have PAVs. Detection of PAVs becomes exceptionally difficult unless comparative genome hybridization (CGH) approaches are applied along with NGS. Complex computational tools and deep sequencing data would help to overcome these problems.

2.3.2.5 Targeted sequencing, Methylation profiling and DNA-protein interactions

NGS technologies are often paired up with multiple accessory molecular biology methods to achieve the project-specific goals more efficiently and cost-effectively. Target enrichment is one of those accessory molecular techniques often used in NGS applications. Target enrichment methods mainly help to derive the NGS data from targeted regions such as candidate genes/exome regions and QTL regions and reduce the noise from unwanted regions in an experiment. Several commercial technologies are available for target enrichment and they include Roche/Nimblegen sequence capture arrays, Agilent SureSelect™ platform, RainDance technologies' RainStorm™ microdroplet-based PCR technology and Fluidigm Access Array technologies. For a review of these technologies and applications in NGS refer to Mamanova et al. (2010). The NGS data derived from targeted regions could be used for variant detection, identification of gene analogs and paralogs, SNP discovery in QTL/exome regions (Nijman et al., 2010) and also for fine mapping efforts. Gene sequences are usually conserved and exome enrichment methods help to enrich the gene regions in libraries from genomic DNA by not only capturing exon regions but also the intron and other flanking regions next to the gene sequence. This helps to detect the mutations in the genes and intron-exon junctions to identify the splice variants (Ng et al., 2009).

Genome-wide sequence data should greatly facilitate our understanding of complex phenomena, such as heterosis and epigenetics, which have implications for crop genetics and breeding (Varshney et al., 2009). Expression of the genome is influenced by chromatin structure, which is governed by processes often associated with epigenetic regulation, namely histone variants, histone post-translational modifications, and DNA methylation. Developmental and environmental signals can induce epigenetic modifications in the genome, and thus, the single genome in a plant cell gives rise to multiple epigenomes in response to developmental and environmental cues. N-terminal regions of nucleosome core complex histones undergo various post-translational modifications, namely acetylation, phosphorylation and ubiquitination that enhance transcription, while biotinylation and sumoylation repress genes (Chinnusamy and Zhu, 2009). Such modifications can be easily detected by combining the chromatin immunoprecipitation (ChiP) procedures with NGS technologies to analyze genome-wide histone modifications.

Methylation of cytosine bases in DNA provides a layer of epigenetic control in many eukaryotes that has important implications for normal biology and disease. Therefore, profiling DNA methylation across the genome is vital to understand the influence of epigenetics (Laird, 2010). Determination of DNA methylation patterns usually requires the use of methylation-sensitive restriction endonucleases or affinity chromatography with methyl-binding proteins, or anti-mC antibodies. Reinders et al. (2008) have used bisulfite conversion for genome-wide DNA methylation profiling. In Arabidopsis, DNA methylation patterns and effects of methylation mutants were studied using Illumina GA sequencing technology by Cokus et al. (2008).

NGS technologies have been leading genome sequencing initiatives across many non-model and orphan crops in recent years to answer the complex biological questions. Newer NGS technologies are being developed and implemented to meet the ever increasing needs of research community and to solve the complex puzzles of nature. NGS methods are being extended to study population genetics, evolutionary biology, molecular ecology, host-pathogen interactions, organellar development, genotype-phenotype interactions and many others. NGS can also accelerate the development of better transformation technologies to modify genes and transform plants easily. In a nut shell, NGS technologies have already demonstrated significant impact on crop breeding and would certainly help to transform the practices and pace of molecular breeding of crops.

2.4 Functional genomics

Functional genomics is the field of molecular biology that utilizes the vast wealth of data produced by genome sequencing projects to understand the gene functions, and their interactions. It is often referred to the study of the genes, their functions, interactions, and regulation to provide a biological function in an organism. Functional genomics mainly focuses on dynamic aspects such as gene transcription, translation, their interaction with other genes and proteins to define the gene function and their regulation. Functional genomics helps to understand the mechanism of a biological function and usually involves combination of both transcriptomics and proteomics.

Genome-wide expression analysis is rapidly becoming an essential tool for identifying and analyzing genes involved in, or controlling, various biological processes ranging from development to responses to environmental cues (Breyne and Zabeau, 2001). Transcriptional profiling studies routinely generate huge EST datasets using sequencing technologies to understand the biological significance of the genes. Availability of gene function information enables various applications of functional genomics. Assignment of gene function (annotation) in most instances is facilitated by comparing them with the genes of known function. Gene prediction modeling tools such as FGENESH, GENESCAN, GLIMMER, SNAP are used to predict the coding regions (exons) in the genome and compare the translated protein to existing protein database for finding the gene function (Korf, 2004). Gene annotation tools such as BLAST2GO use the BLAST algorithm to find the similarity with the existing gene information to derive the gene annotation information (Conesa et al., 2005). Analysis of transcriptomic data using these bioinformatics tools helps gene discovery and associated pathways.

While structural genomics uses genetic variations to understand the phenotypic changes, functional genomics often uses gene expression differences to understand the same. Gene expression differences are usually measured by estimating mRNA expression either by relative or absolute quantification methods. These methods frequently involve PCR-based, hybridization-based or sequencing-based approaches. Differential gene expression of known genes are usually characterized using quantitative PCR (qPCR), microarray technologies, serial analysis of gene expression (SAGE) and RNA-seq methods. Variations in gene expression data could be used to generate expression QTL (eQTL) information, similar to that of genetic markers. Different techniques and methods that are routinely used for gene expression studies are briefly reviewed below.

2.4.1 Quantitative PCR

Measurement of gene expression (RNA) has been used extensively in monitoring biological responses to various environmental conditions. Quantitative gene analysis has been used for detecting the CNVs of a particular gene in the genome or in the transcriptome. PCR method has revolutionized many aspects of molecular biology including gene quantification. PCR protocols are modified and optimized by using either fluorescent probes or dyes in the reaction mixture to obtain accurate quantification of genes in the input DNA or RNA, and these procedures can be collectively called as quantitative PCR (qPCR). The qPCR approaches frequently use either fluorescent probes such as TaqMan® probes (Applied Biosystems) which detect a specific PCR product as it accumulates during PCR cycles, or fluorescent dyes such as SYBR Green which detects all double stranded DNA in a PCR reaction.

In TaqMan® assays, probes are designed specific to a target gene along with a pair of primer sequences in flanking regions of a probe. TaqMan® probes while hybridized to the target gene during PCR do not emit the fluorescence, but as the DNA extension is continued by the DNA polymerase using flanking primers in the PCR reaction, these hybridized probes are displaced and emit the fluorescence thereby facilitating the quantification of the target gene. TaqMan® probes have been routinely used for quantification of genes in the genome for allelic discrimination and zygosity studies and also in reverse transcribed mRNA/cDNA to study the gene expression. Using different reporter dyes, one can quantify two or more genes in the same PCR reaction. However, poor probe design could result in false positive signals. SYBR Green assays exploit the double stranded binding ability of the dye molecules during a PCR reaction. As the target gene product is accumulated in the PCR reaction, the fluorescence emitted by the dye increases and thus can quantify the gene. Though this procedure is simple to set up, there is no specificity to the target gene as it quantifies the entire double stranded DNA in the PCR reaction. There are several publications in the literature that employed these techniques.

2.4.2 Microarray technology

Parallel quantification of large numbers of mRNA transcripts for studying the regulation of gene expression was made possible by microarray technologies. The use of microarrays to analyze gene expression on a global level has recently received a great deal of attention. This should allow new understanding of gene signaling and regulatory networks that operate in various cell processes. The principle of a microarray experiment, as opposed to the classical northern-blotting analysis, is that mRNA from a given cell line or tissue is used to generate a labeled sample, sometimes termed the 'target', which is hybridized in parallel to a large number of DNA sequences, immobilized on a solid surface in an ordered array. Tens of thousands of transcript species can be detected and quantified simultaneously (Schulze and Downward, 2001).

The probes used in this technology could be cDNA fragments generated by PCR or synthetic oligonucleotides and these probes vary in their length based upon the instrumentation and technology available for synthesizing these arrays. Also, probes can be designed to represent the most unique part of a given transcript, making the detection of closely related genes or splice variants possible. To understand complex traits in crops, microarrays can be designed from existing sources of EST/functional gene information within the same crop or could leverage others with better resources. For example, Yang et al. (2007) utilized the microarrays designed from the model plant *Arabidopsis thaliana* to conduct transcriptional profiling experiments in canola for a disease condition.

Generally, in these experiments, mRNA from cells or tissue is extracted, converted to cDNA and labeled, hybridized to the DNA elements on the surface of the array, and detected by phospho-imaging or fluorescence scanning. The use of different fluorescent dyes (such as Cy3 and Cy5) allows mRNAs from two different cell populations or tissues to be labeled in different colors, mixed and hybridized to the same array, which results in competitive binding of the target to the arrayed sequences. After hybridization and washing, the slide is scanned at two different wavelengths corresponding to the dyes used, and the intensity of the same spot in both channels is compared. This results in the measurement of the ratio of transcript levels for each gene represented on the array. Microarrays can be used to investigate the changes in expression at single gene or across the whole genes to infer the changes in the phenotype. Analysis of expression differences at multigene level would facilitate understanding of gene regulatory pathways and gene-to-gene interactions besides providing the information of up or down regulation to a particular experimental condition. Several statistical and data analysis tools are available to interpret the microarray data (Schulze and Downward, 2001).

In traditional QTL analyses, linkage mapping leads to the detection of genomic regions which are associated with phenotypic variations within a population. Genetical genomics employs this same approach, except that the phenotypes are levels in gene expression resulting in the detection of expression QTL (eQTL). The eQTL do not necessarily result from sequence polymorphisms proximal to the gene being measured (cis-acting) but could result from differences in genes unlinked to the target. In these cases, the eQTL function in a trans-acting manner (Holloway et al., 2011). When the gene expression data is collected from a specialized tissue to understand the phenotype of the trait, we can use that data as a marker system to derive the eQTL information. In a recent genome wide eQTL study, Holloway et al. (2011) utilized expression data from 50,000 maize genes to identify both cis-acting and trans-acting genetic elements that cooperate to regulate gene expression in maize crown roots and described the pitfalls of detecting false cis- or trans-acting eQTL in the absence of perfect genomic sequences from both parents. Multiple examples of eQTL analysis were reported in many plants and crop species (Druka et al., 2010). Despite many advantages, microarray technologies have their limitations, because expression profiling is conducted for only a limited set of genes that are currently known and we cannot detect the influence of unknown genes for the phenotype.

2.4.3 Serial analysis of gene expression (SAGE)

The SAGE technique is based on counting sequence tags of 14–15 bases from cDNA libraries. These tags are generally derived from either 5′ or 3′ end of the expressed genes by restriction enzyme and can be up to 22 bases. Earlier, these tags used to be ligated to each other to create longer stretches of tags (Super-SAGE) and then sequenced by Sanger sequencing. But now, with the use of NGS short read technologies, these tags can be individually sequenced to generate the expression information. Each unique tag can represent a copy of the gene in the cDNA and thus by counting these tags, the gene expression can be quantified. The principal advantage of SAGE is that it gives an absolute measure of gene expression instead of measuring relative expression levels. Indeed, by counting the number of tags from each cDNA, one obtains an accurate measure of the number of transcripts present in the mRNA sample. This technology has been widely used to monitor gene expression in human cell cultures and tissue samples, but was used sporadically in plants. The principal limitation of SAGE is the need to sequence large numbers of tags in order to monitor scarcely expressed genes. Another drawback of SAGE is that the obtained tags are very short and hence not always unambiguous. Gene identification on the basis of short sequence tags relies on the availability of large databases of well-characterized ESTs (Breyne and Zabeau, 2001).

2.4.4 RNA-Seq/digital gene expression

Recent developments in NGS technologies have transformed the way through which quantitative transcriptomics can be done. Due to their massively parallel sequencing abilities, these NGS technologies have been driving down the sequencing costs and time required to generate large amounts of sequencing data. Using these NGS technologies, RNA content of the cells can be directly sequenced without requiring any of the traditional cloning associated with EST sequencing. This approach, called ‘‘RNA-seq”, can generate quantitative expression scores that are comparable to microarrays, with the added benefit that the entire transcriptome is surveyed without the requirement of *a priori* knowledge of transcribed regions. One key advantage of this technique is that not only quantitative expression measures can be made, but transcript structures including alternatively spliced transcript isoforms, can also be identified (Wilhelm and Landry, 2009).

RNA-seq procedures usually involve generation of multiplexed sequencing libraries from the mRNA/total RNA followed by high throughput sequencing. Short read sequencing instruments are more cost-effective for RNA-seq studies. Using the efficient bioinformatics tools, the sequence data is mapped to the reference genome to provide the multiple sequence alignment. Removal of repetitive sequences from the dataset would improve the mapping process. Sequence data can be converted into expression data in different ways: i) by simply adding the number of reads which fall within the co-ordinates of each element (either exon or gene), and then normalizing the data for the length of the element; ii) by calculating a sequence score for each nucleotide in the genome based on the number of reads which cover each base position, and again normalizing for element lengths (Wilhelm and Landry, 2009); iii) by calculating RPKM (reads mapping to the genome per kilobase of transcript per million reads sequenced) values (Mortazavi et al., 2008) using a mathematical formula and use them as a measure of gene expression. These processes are often referred to as 'Digital Gene Expression'.

Compared to microarrays, the limits of the dynamic range measured in RNA-seq experiments are only determined by the amount of sequencing obtained. This means that through the continued sequencing of a given library, it should be possible to eventually measure the expression of every transcript present and so the ''dynamic range" only represents the actual biological diversity of the transcriptomes. To implement RNA-seq experiments effectively, availability of accurate annotation of the reference genome is necessary. This is particularly challenging for higher eukaryotes especially plant species with large genomes. However, RNA-Seq promises the gene quantification for studying complex phenotypic traits in cost-effective way in the near future for many crop species.

2.5 Comparative genomics

Comparative genomics relates the structure and function of genomes of evolutionarily close species. It is a tool that helps researchers to study complex genomes of plants by leveraging sequence information of related species with smaller and less complicated genomes. However, in 1990s and at the beginning of 21st century when no reference sequences existed, comparative genomics was limited to comparative mapping. Both in dicot and monocot plants, collinearity at micro level, i.e. the same order of the molecular markers in genetic maps, were observed. In dicots, comparative sequencing approach revealed collinearity in gene order within several chromosomal segments of Arabidopsis, Capsella and Brassica genomes (Rossberg et al., 2001). Later Schranz et al. (2006) integrated all comparative genomics data in Brassicaceae and constructed a set of 24 genomics blocks, which represented the conserved segments of ancestral karyotype, *Arabidopsis thaliana*, and *Brassica rapa.* In monocots, comparative genetic mapping in oats, maize, rice, barley, wheat, sorghum, sugarcane and fox millet resulted in the construction of the "Crop Circle", which placed the small genome of rice in the center of the circle and aligned with maps of crops with larger genomes (Gale and Devos, 1998). Comparative mapping revealed 30 blocks of rice genome that could be found within genomes of other crops (Devos, 2005).

Sequencing of several model plants, including Arabidopsis, rice, *Medicago truncatula*, *Lotus japonicum* and Brachypodium, as well as crops such as maize, rice (both a model and a crop plant), sorghum and soybean confirmed, in general, the existence of synteny between genomes of related species at DNA level, which was, however, reported to be less obvious because of the unique patterns of distribution of repetitive sequences, duplications, insertions and deletion of genes (Dubcovsky et al., 2001). Nevertheless, comparative genomics has been a valuable tool for the development of molecular resources for crops and the identification of key genes for crop improvement.

Owing to the collinearity between rice and other cereals and *Arabidopsis thaliana* and Brassicas, genomic resources of these model plants were leveraged to boost map-based cloning of genes from the genomes of other cereals with larger and complex genomes (Salse et al., 2008). Although the model plant, *A. thaliana*, and the field crop, rice, are the only plants with completely sequenced genomes, availability of high quality draft genome sequences of Brachypodium, sorghum and maize are believed to provide even more opportunities in gene and QTL discovery in orphan crops with zero to little genomic resources (Mayer et al., 2011).

Molecular markers and EST collection of the above-mentioned crops, have been widely used to saturate genomic regions of other crops to narrow down the location of economically important QTL and genes. Using rice and Brachypodium ESTs, a powdery mildew resistance gene Ml3D232 was isolated from bread wheat (*Triticum aestivum*) (Zhang et al., 2010). Wheat tiller inhibition gene, tin3, was mapped using molecular markers developed based on ESTs from syntenic region of rice. Comparative analysis revealed collinear regions between perennial ryegrass (*Lolium perenne L.)*, *B. distachyon* and sorghum, which facilitated cloning of the self-incompatibility genes in the former (Shinozuka et al., 2010). Using synteny between rice chromosome 9 and Italian ryegrass LG5 and rice EST-based molecular markers, the location of LMPi1 gene conferring resistance to grey leaf spot was delimited to short chromosomal segment of the latter (Takahashi et al., 2010). Using homology between *B. rapa* and *A. thaliana*, the TRANSPARENT TESTA GLABRA 1 (TTG1) gene controlling both hairiness and seed coat color traits in Brassica species was isolated (Zhang et al., 2009). Sequences from Arabidopsis chromosome1 were used as RFLP probes to genetically map fertility restorer gene, *Rfp*, in *B. napus* genome.

Comparative genomics has been an indispensable tool to study evolution of genomes and gene families. *Brachypodium distachyon* has been widely used to study the evolution of important traits in barley and wheat, including flowering time pathways and (1,3;1,4)-b-D-glucans in plant cell walls (Higgins et al., 2010). Recently, the genomes of Brachypodium, rice and sorghum were used to assign 32,000 barley genes to the corresponding individual chromosomes (Mayer et al., 2011). Genome of Arabidopsis expanded its value to study xylem genomics of conifers (Xinguo et al., 2010). Two legume species, *Medicago truncatula* and *Lotus japonica,* are being sequenced to study genetic background of legume-specific phenomena such as symbiotic nitrogen fixation (Sato et al., 2008). Genome of *M. truncatula* should also facilitate the assembly of next generation sequence data in closely related taxa such as alfalfa (Young and Udvardi, 2009). Tremendous progress in sequencing technologies is opening new avenues for comparative genomics and enabling researchers to do genome-wide comparisons. Current trends indicate that NGS technologies might change the focus of comparative studies by shifting them more towards evolutionary genomics rather than synteny-based gene-cloning or marker development. No matter what direction comparative genomics will take in the future, it is certain that it has the potential to broaden our current understanding of complex biological processes. An understanding of complex traits and processes such as adaptation of plants to biotic and abiotic stresses, yield, gene regulation, polyploidy and the influence of natural selection on gene and protein function can be translated into development of new strategies for crop improvement.

2.6 Genetical genomics

The concept and strategy of genetical genomics, outlined by Jansen and Nap (2001), aims to rapidly identify key gene targets by super-imposing gene expression data on that of genetic mapping. Although molecular markers linked to quantitative trait loci (QTL), the genomic regions genetically determined to be associated with the observed phenotypic variation, can be used in marker-assisted breeding programs, in order to identify the gene(s) underlying the QTLs, comparison of gene expression differences between contrasting lines is very important. Genetical genomics approach involves expression profiling and molecular marker based genotyping of all individuals in a segregating population. It is followed by a comprehensive analysis using all statistical tools that are normally used in the analysis of quantitative trait loci. These analyses result in the identification of expression QTL (eQTL). Based on the nature of gene expression variation, there are two types of eQTL, cis-eQTL and trans-eQTL. If the variation or polymorphism is located near the gene, it is classified as a cis-eQTL, whereas if the source of variation is located at a distant location in the genome then it is a trans-eQTL. Genetical genomics has been applied to a broad range of organisms and all studies have demonstrated the power of combining gene expression data with that of genetic analysis to fine tune pathways involved in complex phenotypes thereby enabling the identification of key genes (Joosen et al., 2009). Genetical genomics studies benefit greatly with the availability of reference genome sequence as well as with the use of large populations. In a genetical genomics study of the model plant Arabidopsis where a 162-line RIL population was used and data analyzed in conjunction with the genome sequence, researchers successfully predicted key regulators of flowering time and circadian rhythms through the construction of genetic regulatory network by combining eQTL mapping and regulator candidate gene selection (Keurentjes et al., 2007). Examples also exist where synteny of the target genome with other species can be used in genetical genomics. For example, in a wheat study, synteny with the sequenced rice genome was used to map the eQTL for seed quality parameters (Jordan et al., 2007). The use of genetical genomics is on the rise. Several technological advances such as expression profiling using next generation sequencing and the availability of several types of 'omics' data sets are enabling rapid construction of biological networks for not only identifying key genes for phenotypic variations but also for understanding the pathways and systems.

2.7 Systems biology

Systems biology is the study of interactions among biological components using models and/or networks to integrate genes, metabolites, proteins, regulatory elements and other biological components (Yuan et al., 2008). Although integration of data from multiple areas of research is not a new concept in biological research, the availability of huge amounts of diverse sets of data obtained from modern, HTP 'omics' technologies has renewed interest in systems biology. In particular, next generation genome sequencing as well as HTP profiling of transcripts, proteins, metabolites, phenotypes etc. are providing the necessary raw material that can enable us to construct interaction networks of genes, their products and many associated players. It must be emphasized that the goal of systems biology exceeds that of the omics components in that it looks holistically at the biological systems with respect to the structure and dynamics.

Considering the complexity and diversity of datasets analyzed and integrated in systems biology, strategy for these projects can be broadly divided into 4 steps. Step 1 is the development of a sound experimental strategy and collection of reliable and reproducible data. Step 2 is the annotation of the data components, for example, genes and proteins investigated in the study. While gene ontology system and other molecular function information resources are used for the functional classification of genes/proteins, pathway databases (Bauer-Mehren et al., 2009) such as KEGG (Kyoto Encyclopedia of Genes and Genomes; http://www.genome.jp/kegg/), Reactome (http://www.reactome.org), PANTHER (Protein ANalysis THrough Evolutionary Relationships; http://www. pantherdb.org/) and plant metabolic pathway databases (http://plantcyc. org/) can be used for obtaining pathway information of the candidates. In step 3, the annotation information from step 2 is used for the generation of mathematical models and networks based on the associations and interactions observed in the datasets in comparison to known pathway information from the databases. Step 4 is the model development and validation where generated models are validated, and then variations introduced and models revised and validated again. Several software tools are used for this iterative cycle of model development – validation – perturbation – validation (Endler et al., 2009). In addition to the tools needed in these steps, availability of efficient platforms and infrastructure that can archive and support analysis of data from diverse sources and labs is a critical requirement for any systems biology project.

There can be multiple types of output expected from systems biology depending on the datasets used for integration. The output can either be a snap shot of a system with respect to a key gene or protein such as gene regulatory or biochemical networks or it can be based on the quantitative or qualitative dynamics of the system based on multiple perturbations. The most popular outcomes to date have been the construction of biological networks that include gene and transcriptional regulatory networks and interactome networks. A major promise of systems biology for crop improvement lies in the area of understanding quantitative traits and abiotic and biotic stresses and examples currently exist where systems biology approach has been explored in these areas. Cooper et al. (2003) developed a network of genes associated with developmental and stress responses in rice by measuring the gene expression changes with respect to environmental, biological and stress treatments related to interaction domains for 200 proteins from stressed and developing tissues. Data obtained from this study resulted in the identification of several stress response genes and were also found to be useful for the prediction of gene function in monocots and dicots. Use of systems biology in unraveling plant defense response has to deal with several dynamics related layers before key resistance gene candidates can be identified. In a study where mutant analysis and whole genome gene expression profiling was used for determining the regulatory roles of WRKY genes in systemic acquired resistance (SAR), a 'regulatory node' was identified in the transcriptional regulatory network controlling SAR (Wang et al., 2006). Although reports that utilized input from several omics for a detailed analysis of quantitative traits are yet to appear, approaches that utilized genome wide eQTL or metabolite QTL (mQTL) do exist that led to the construction of networks and identification of key metabolic QTLs that could be leveraged for crop improvement (Schauer et al., 2006). Current efforts in this area also include extending the impact of systems biology in understanding plant communities using a holistic approach by merging systems biology and systems ecology in order to improve agricultural productivity(Keurentjes et al., 2011).

2.8 Bioinformatics

The field of genomics has grown leaps and bounds during the recent decades. While the development and implementation of HTP genome sequencing and more than a dozen omics technologies revolutionized our ability to study whole genomes and biological systems, one important capability that has also played a major role in this unprecedented growth is bioinformatics. The explosive growth of information from the biological and genomics research has accentuated the need for computational requirements. The development of computational biology tools enabled rapid and HTP analysis of large amounts of data generated by research community while the creation of comprehensive databases made the information available globally, further enhancing the pace of research. A recent review by Mochida and Shinozaki (2010) on various genomics and bioinformatics resources available and how they are helping in biological research provides a greater appreciation for the impactful contributions of bioinformatics over the years. Systems biology is one area where we need additional technological advances as well as improvements in our computational tools in order to efficiently analyze, integrate and interpret huge data sets resulting from dynamic biological systems and currently efforts are underway to this end.

3. Molecular breeding approaches

3.1 Mapping and map-based cloning of QTL

The availability of dense genetic maps with informative markers makes it possible to map genes and QTL. Numerous studies have been dedicated to mapping QTL governing major agronomically important traits. As of July 06, 2011, there were about 24,000 research articles in this direction in Google Scholar. However, majority of those studies ended at the mapping step and did not pursue the ultimate goal of cloning the gene(s) underlying the QTL due to many reasons including lack of funding, insufficient genomic resources, inadequate phenotypic data collection methodology, availability of experimental designs with limited detection power and finally complexity of the trait and the genome. Due to these factors, QTL cloning in crops remains a very challenging process. Researchers have to undergo multiple labor-, time- and cost-intensive steps prior to answering the question "to clone or not to clone plant QTL" (Salvi and Tuberosa, 2005, 2007). What are those steps and challenges?

The first step in every map-based cloning (MBC) project is genetic mapping of QTL using a small mapping population (200-300 individuals) and identifying flanking markers. Provided that phenotypic data collection methodology and experimental design are at adequate level, the main constraint that researchers face at this initial step is very large confidence interval (CI) of QTL, which can span 10-30 cM (Kearsey and Farquhar, 1998). However, even in cases where the CI is limited to a few centimorgans, the interpretation of this genetic distance is not a straight forward process. The reason is that the genetic distance largely depends on the rate of recombination frequency in the region and the size of the mapping population used to construct the linkage map. In many cases, markers that flank QTL are physically far from the target, and the interval between markers contains a large number of genes. Depending on the length of DNA segment spanning QTL, strategies to reduce the distance between markers and QTL have to be designed. One of the strategies to narrow down the CI is the right choice of molecular markers for QTL mapping. Majority of QTL mapping studies have been using a pool of publicly available markers, which are not always informative for a particular bi-parental cross under investigation. Another most important factor that can condition the success of QTL mapping and subsequent cloning is knowledge of all possible allelic variations existing between parents. For instance, most of the public SNP markers in maize are developed from B73 and Mo17 cultivars. Taking into account massive intraspecific variations among maize inbred lines (Eichten et al., 2011; Fu and Dooner, 2002; Springer et al., 2009), SNPs developed from a few lines will capture only a small portion of all allelic variations happening between parents of a cross designed for a QTL study. Consequently, there is a big chance that majority of allelic variations, including the causative mutation between parents of this cross will be missing. In order to avoid this situation, re-sequencing of genomes of both parents and discovery of allelic variations in low and single copy regions could be implemented using NGS technologies coupled with genome complexity reduction techniques. Discovered cross-specific polymorphisms can later be converted into any modern SNP genotyping assay (Mammadov et al., 2010; Trebbi et al., 2011). For instance, technologies such as complexity reduction of polymorphic sequences (CRoPS™) (Van Orsouw et al., 2007) or restriction site associated DNA (RAD) (Baird et al., 2008) can be successfully applied to generate cross-specific SNPs (Mammadov et al., 2010). Depending on the organism, this approach may result in the validation of about 1000 robust cross-specific markers, which can be later combined with public SNPs and used for mapping. NimbleGen Sequence Capture technology (Roche Applied Science) has also been used for the detection of cross-specific polymorphisms within low and single copy sequences (Springer et al., 2009). However, this technology can be applied only to crops for which reference genome sequences are available because of the necessity to design capture probes. The approach of development of cross-specific markers increases the precision of QTL mapping and consequently narrows down CI and can lead to the detection of causative mutation(s).

If the heritability and the effect of QTL governing the trait are high and CI is narrow enough, then the development of large mapping populations and the creation of high-resolution or fine mapping can be sufficient (Jiao et al., 2010). However, in many cases there is a risk that the detected major QTL is in reality represented by several co-segregating loci with minor effects. In order to eliminate the effects of other co-segregating alleles affecting the phenotype, target locus is recommended to be Mendelized through the painfully long and expensive process of developing near isogenic lines (NILs) (Kearsey and Farquhar, 1998). In QTL-NIL the target QTL will behave as a single Mendelian gene. Development of NILs begins with the same population where all QTL were mapped, and can be carried out by marker-assisted backcross introgression. QTL Mendelization is believed to give an answer on the feasibility of cloning the target locus. In some cases, after Mendelization, the effect of target QTL may become so negligible that further cloning activities will not make any sense. However, if the high-resolution reveals that target QTL still explains the large portion of phenotypic variation, the physical mapping of the locus must be implemented (Saito et al., 2004).

In order to physically anchor the QTL, bacterial artificial chromosome (BAC) library has to be developed, which is a very important prerequisite for map-based cloning. Researchers use markers flanking the gene of interest as probes to implement chromosome walking (Tanksley et al., 1995). When the price of sequencing was high, chromosome walking was a very tedious process, which included several rounds of marker-library hybridization, identification of new BAC clones spanning the region between the flanking markers, BAC fingerprinting to construct contigs, identification of minimum tiling path (MTP) and development of new markers from the extreme left and right BAC clones representing the

MTP. Chromosome walking had not been a straight forward process either. It was especially complicated in crops with complex genomes. In many cases, the isolation of BAC ends resulted in the dissection of repetitive sequences which were of no use in designing a new probe for the subsequent library hybridization. Currently, instead of doing BAC fingerprinting, NGS technology allows direct sequencing of all identified BACs and construction of BAC contigs based on sequence similarity (Schneeberger and Weigel, 2011). Availability of a reference sequence simplifies the assembly of BAC clone sequences and contig construction. Lack of reference sequence will force researchers to construct contigs *de novo*, which is not an easy task taking into account the complexity of a plant genome. From this point of view, MBC in crops with available reference genome sequence is supposed to be easier than in crops with no genome sequence available. However, this is not always the case because success of MBC depends not only on available structural genomics resources but also the complexity of a trait and availability of adequate phenotypic data collection methodology. Because QTL mapping results from the comparison of marker and phenotypic data, the accuracy of the latter is of great importance. Although accurate phenotypic data is equally important both for QTL and single gene cloning, the former is more sensitive to the robustness of phenotypic data due to their dependence on environmental conditions. Nowadays, precision phenotypic data collection remains the major bottleneck for the successful QTL mapping and cloning.

If the position of a QTL is delimited to one BAC clone, this large insert clone must be sequenced to identify the candidate genes. Ideally, homologous BACs from both parents should be sequenced, because this will facilitate the identification of gene candidates. Sequencing of a BAC clone using NGS technology is a very straightforward process if the target crop has reference genome sequence. For example, sequencing can be done using fairly cost-effective HiSeq instrument (Illumina, San Diego, CA), which generates small 70-100 bp fragments which can later be mapped back to reference genome to facilitate the assembly. In orphan crops sequencing efforts are impeded due to the absence of reference sequence. Sequencing using traditional Sanger could be the only choice. However, it is very expensive and a long process. Another alternative is sequencing the same BAC clone with two NGS instruments in parallel such as GS20 Sequencer (454/Roche) and Illumina's HiSeq, which generate long (1 kb) and short (100 bp) sequences, respectively. A combination of short and long sequences may improve the assembly of a BAC clone from a crop with no reference sequence.

When a BAC clone is sequenced and assembled, next step is sequence analysis of the large insert clone. Sequence analysis of the BAC clone is necessary to reveal gene candidates. Normally, the entire BAC clone first gets scanned for the repetitive sequences using publicly available Repeat_Masker software (http://www.repeatmasker.org/cgi-bin/WEBRepeat Masker). The Repeat_Masker output can be used as an input template for any protein prediction software, including FGENESH (http://linux1.softberry.com/ berry.phtml?topic =fgenesh&group=programs&subgroup=gfind) and GenScan (http:// genes.mit.edu/ GENSCAN.html) and GenMark (http://exon.gatech.edu/). Finally, all predicted open reading frames (ORFs) must be BLASTed against protein database using 'blastp' algorithm to reveal their biological functions. Depending on the genome size and repetitive sequence content one BAC clone may have dozen or so genes. The number of candidate genes can be reduced by aligning homologous BAC clones from both parents. Further, any allelic variations between two parents within target BAC could be converted into KASPar assay and bulk segregant analysis (BSA) can be performed. If BSA does not reduce the number of candidate genes to one, then causative mutation can be identified through functional prediction. For example, if the target gene confers resistance to a disease and among the remaining candidate genes there is an NBS-LRR gene, which belongs to one of the classes of defense-related genes, then it might be considered as a gene of interest. However, if the researcher wants to confirm the gene candidate using classical methods, then either further increase in the resolution within a locus or functional characterization of each predicted protein is recommended. Both methods are very expensive and labor- and time-intensive. Functional characterization of gene candidates can be done using several methods including genetic complementation through transformation technique, down-regulation of a gene via RNAi, complementation of a known mutant or by marker-assisted trait introgression (Borevitz and Chory 2004).

3.2 QTL cloning through association mapping

One of the major limitations of QTL cloning using bi-parental mapping approach is insufficient amount of meiotic recombination in mapping populations such as F2, DH and RIL, which lead to a strong statistical association of QTL with the block of markers that physically span large chromosomal segments. QTL-mapping approach requires that specific mapping populations, usually consisting of several hundred F2 or RIL progeny, be developed from each germplasm accession to be examined for important genes effecting traits of interest. Each population must be genotyped using hundreds, perhaps thousands of molecular markers. This population development and marker screening is extremely time-consuming, high-risk and expensive work - prohibitively expensive if dozens, let alone hundreds or thousands, of germplasm accessions are to be examined (Abdurakhmonov et al., 2008; Abdurakhmonov and Abdukarimov, 2008). However, geneticists mapping complex traits in the human genome have circumvented the need for large F2 or RI mapping populations (which are not available in humans) by making use of information contained within the genetic recombinations that have occurred in typical human populations during the course of recent evolution. Genetic loci linked to a specific disease will show historically reduced level of recombination (non equilibrium) with specific alleles at different loci controlling particular genetic variation in a population (Abdurakhmonov et al., 2008). This "linkage disequilibrium (LD)" can be detected statistically, and has been used to map and eventually clone a number of genes underlying complex genetic traits in humans (Schulze and McMahon, 2002; Weiss and Clark, 2002). LD-Mapping, referred also as Association mapping (AM), is an alternative approach to a now classical QTL-mapping in bi-parental mapping populations because it overcomes the problem related to the lack of recombination events. AM population is composed of genetically un-related individuals with unknown pedigrees and accumulates larger number of historical recombination events that occurred in the past (Nordborg and Tavaré, 2002). Multiple increases in number of recombination events will break the block of markers that is associated with QTL and increase the resolution of QTL region. Association Mapping is linkage disequilibrium (LD)-based association while bi-parental approach is a genetic recombination-based mapping. AM attempts to reveal a significant association between a trait and a gene, or a molecular marker, or block of molecular markers, which are at LD. Marker and trait are in disequilibrium if they are truly linked to each other and historically have been passing from generation to generation together. The theory of AM is based on the idea that LD tends to be preserved over many generations between loci, which are linked to one another and form haplotypes. The higher the LD, the tighter is the linkage between markers. However, in bi-parental mapping, because of lack of recombination event, several haplotypes could be grouped into one linkage and be statistically associated with QTL, which decreases the resolution of the map. To summarize, AM theoretically has several advantages over classical bi-parental approach: (1) increased mapping resolution; (2) availability of more genetic variability for marker-trait correlations and detection of multiple-alleles simultaneously; (3) elimination of the necessity to develop large populations for fine mapping which saves expenses and time; (4) one AM population can be leveraged for dissection of many traits, while bi-parental crosses are dead-end and in many cases can be used to study one trait only and (5) feasibility of employing historical phenotypic data collected over many years (Zhu et al., 2008).

However, not everything is that smooth and painless in the implementation of AM. This approach is often criticized for (1) detecting a large number of false-positive QTL due to population confounding effects and 2) influence of allele frequency distributions (rare and minor allele frequencies) of functional polymorphisms to the power of the detected associations (Abdurakhmonov and Abdukarimov, 2008; Aranzana et al., 2005; Stich and Melchinger, 2010). In order to avoid false positives, several factors have to be taken into consideration. Structure of the population must be carefully analyzed using various computational methods. Too structured population with too many sub-groups will detect pseudo LDs between loci that in reality are not linked, and cause false-positive association between a marker and a trait. In order to avoid this, prior to AM implementation, the population must be analyzed for the presence of hidden sub-structures. One of the popular software programs that researchers use to resolve this issue is publicly available software called STRUCTURE. Removal of rare alleles is a choice in AM to reduce false-positives (Abdurakhmonov and Abdukarimov, 2008), but studies showed that most phenotypic variations are due to rare alleles (Stich and Melchinger, 2010), suggesting importance of these rare alleles in tagging biologically meaningful associations. Both structured population and rare allele frequency issues can be greatly minimized by creating segregating populations and performing genetic crosses between several reference populations with known allele frequencies for functional polymorphisms. Such approach is referred to as nested association mapping (NAM) and NAM populations greatly enhance the power of association mapping in plants (Stich and Melchinger, 2010).

Because AM is LD-based, another consideration is the rate of LD decay in a crop under the study. The pattern of LD throughout the genome will determine the appropriate marker density for whole genome scanning (Yu et al., 2008). The longer the haplotype, the lesser is the marker density needed because all markers within a haplotype will behave similarly. In contrary, if the segment of a chromosome is characterized by the presence of short haplotypes then the density of markers has to be increased correspondingly. Additionally, the rate of LD decay with physical and genetic distances is important to determine the maximum resolution that can be achieved for association mapping. Length of haplotypes depends on a genome, the number of loci investigated, and the reproductive history of a population. It was previously reported that in maize, LD decay distance was on an average less that 2000bp (Remington et al., 2001). Later studies suggested that in commercial inbred lines, LD decay may span more than 100-500 Kb (Jung et al., 2004; Tian et al., 2011). Recently developed first-generation haplotype map of maize presented the evidence for much longer haplotypes spanning several million bases (Gore et al., 2009), which was later supported by Mammadov et al. (2010). In cultivated barley, LD has been reported to span from 1 cM to 10 cM (Rostoks et al., 2006). Use of different molecular markers, can also significantly change the length of LD. In rice, it was indicated that LD decays within 1 cM or less using SNP markers (Agrama and Eizenga, 2008), whereas others reported 20-30 cM length of LD while using simple sequence repeat markers (SSRs) (Jin et al., 2010). Differences in the rate of LD decay within a crop could be explained by the nature of unit it was represented. Centimorgan is a unit of recombination, and the rate of recombination has proven to be not uniform across the genome. Consequently, 30 cM LD span in one region of rice genome physically may carry the same value as 1 cM LD span in another region of the genome. Physical distance seems to be more realistic way to designate LD decay. However, absence of reference sequence in some crops limits researcher to centimorgans only. Now-a-days researchers successfully use AM to study genetics of complex traits in all major crops, including rice, maize (Poland et al., 2011), barley (Massman et al., 2011), soybean (Wang et al., 2008) and canola (Honsdorf et al., 2010) and other crops (Abdurakhmonov and Abdukarimov, 2008; Stich and Melchinger, 2010). The common feature of these studies is the detection of large number of QTL with small effects. Whether or not the information on these QTL can be translated into real use in crop improvement is unclear and the answer is yet to come.

To date there are only a few successful QTL cloning studies. Most of them have been reported in rice, which is not surprising because rice genome has been sequenced since 2004 (Table 3). Additionally rice genome is smaller and less complex. QTL cloning is in progress in maize and soybean, two crops that have reference genome sequence available. Undoubtedly, the progress in sequencing technologies, availability of reference genome in major crops, rapid evolution in high-throughput polymorphism detection platforms and bioinformatics tools and last, but not least, the development of accurate phenotypic data collection methodology will increase the precision of QTL mapping in major and orphan crops bringing us closer to the discovery of the Holy Grail of molecular geneticists, causative mutations, that will be subsequently translated into robust diagnostic tools to implement marker-assisted selection.

3.3 Marker-assisted selection

Marker-assisted selection (MAS) as a process refers to the selection of superior genotypes using molecular markers. MAS is thought to have substantial advantages over conventional phenotypic selection because the latter could be (1) unreliable when the expression of the trait is environmentally dependent, (2) biologically deadline-sensitive, (3) expensive and difficult to screen and (4) subject to the mercy of weather. In contrast to phenotypic selection, MAS (1) does not rely on environmental conditions because it detects the structural polymorphisms at molecular level, (2) requires leaf tissue collected at seedling stage, which is very useful for traits that are expressed at later stages of development and which also helps to avoid adverse weather conditions that could kill the plant at adult stage, (3) could be cheaper and less labor intensive, (4) allows selection in off-season nurseries and has a potential to accelerate breeding process.

Crop	Gene	Trait	Function	Reference
Rice	*Ctb1*	Cold tolerance	F-box protein	Saito et al 2010
	OsSPL14	Panicle number, grain productivity	Transcription factor, a protein similar to Squamosa promoter binding protein	Jiao et al., 2010; Miura et al., 2010
	ERECT PANICLE 3	Erect Panicle architecture	F-box protein	Piao et al 2009
	DEP1	Panicle number, grain number	Gain-of-function mutation causing truncation of a phosphatidylethanolamine-binding protein-like domain protein	Huang et al., 2009
	SK1/SK2	Deepwater tolerance	Ethilene responsive factors	Hattori et al., 2009
	Mt1	Rice tillering	carotenoid cleavage dioxygenase 8 (CCD8)	Zhoua et al 2009
	qSW5(GW5)	Grain width and weight	novel nuclear protein likely acts in the ubiquitin-proteasome pathway to regulate cell division during seed development	Shomura et al., 2008; Weng, et al., 2008
	Ghd7	Grain number, plant height, heading date	CCT (CO, CO-LIKE and TIMING OF CAB1) domain protein	Xue et al., 2008
	GW2	Grain width and weight	RING-type protein with E3 ubiquitin ligase activity	Song et al., 2007
	TAC1	Tiller angle	Unknown	Yu et al., 2007
	GS3	Grain weight and length	VWFC membrane protein	Fan et al., 2006
	sh4	Shattering	Transcription factor	Li et al., 2006a
	qSH1	Shattering	BEL1-homeobox	Konishi et al., 2006
	Sub1A	Submergence tolerance	Transcription factor	Xu et al., 2006
	Gn1a	Grain productivity	Cytokinin oxidase	Ashikari et al., 2005
	PSR1	Regenerability	Nitrite reductase	Nishimura et al. 2005
	qUVR-10	UV resistance	CDP photlyase	Ueda et al. 2005
	SKC1	Salt tolerance	HKT transporter	Ren et al., 2005
	Ehd1	Heading date	B-type response regulator	Doi et al., 2004

	Hd3a	Heading date	Unknown protein	Kojima et al., 2002
	Hd6	Heading date	Protein kinase	Takahashi et al., 2001
	Hd1	Heading date	Transcription factor	Yano et al., 2000
Maize	Vgt1	Flowering time	Non-coding sequence	Salvi et al. 2007
	Tga1	Glume architecture	Transcription factor	Wang et al 2005
	Tb1	Plant architecture	Transcription factor	Doebley et al 1995, 1997
	DGAT	High oil content	acyl-CoA:diacylglycerol acyltransferase	Zheng et al 2008
Soybean	*E3 locus*	Flowering time	Phytochrome A	Watanabe et al 2009

Table 3. QTL cloning studies in the literature

In a review article, Xu and Crouch (2008) demonstrated an interesting chronology on the evolution of MAS as a technology. According to them, the term was coined in the mid – eighties by Beckmann and Soller (1986) as a technology that might have a potential use in plant breeding. Ten years later, MAS was already considered as a possible technology to tag genes (Concibido et al. 1996). Although as of June 27, 2011, according to Google Scholar, there were about 32,300 articles containing the keyword "marker-assisted selection", most of them still have been referring to potential application of MAS in plant breeding. A vast majority of those publications were from academia. Although private sector does not normally release the details of their breeding methodologies to public domain, several articles on successful application of MAS in the development of varieties of maize (Ragot et al., 2007) and soybean (Cahill and Schmidt, 2004; Crosbie et al., 2003) came mainly from industry. Fairly low impact of academic research in developing varieties using MAS can be explained by the lack of funding to complete the entire marker development pipeline (MDP), which can be long-term and cost-intensive task. MDP includes several steps such as (1) population development, (2) initial QTL mapping, (3) QTL validation (testing in several locations and years and implementing fine mapping) and (4) marker validation (development of inexpensive but high-throughput assays that are amenable to automation) (Collard and Mackill, 2008). Every step of the development of markers linked to QTL are associated with numerous constraints which may take several years and substantial funding to resolve. In the 1990s it was believed that molecular markers identified at step 2 were enough for successful MAS. In the 1990s it was observed that markers that were previously declared as tightly linked were failing to confirm the phenotype at advanced stages of MAS. One of the main reasons for the failure of a marker in MAS, which was identified at pre-fine mapping step, was the inconsistency in QTL mapping. Detection of QTL within one year and in one location was proved to be not enough to claim the robust QTL location because the expression of latter has been environmentally dependent. Thus, QTL validation and confirmation was required, which foresaw QTL mapping based on data collected within several years and multiple locations. Molecular markers that were tightly linked to QTL and were consistent across several years and locations did have a potential in MAS. However, even after QTL validation, so called "tightly linked marker" hardly met the expectations because the confidence interval (CI) of QTL peak is so large that it is very difficult to predict the real distance between marker and QTL. Moreover, there are several hundreds of candidate genes within CI, and it is impossible to predict which gene explains the phenotypic variation. Genetic proximity of a marker to QTL depends on two factors such as the size of the population and the region of a genome where the marker and QTL are residing. If the size of the mapping population is small (100-200 individuals), then the claims of having a marker closely-linked to a gene are barely valid, because fine mapping will identify many crossing-over events happening within marker-gene complex. Occurrence of the recombination events between marker and QTL makes the marker unable to track a target. This type of situation is especially true for the regions of the genome with high rate of recombination frequency. However, certain regions of genome exhibit very low rate of recombination frequency. If QTL was mapped to low recombination frequency region, then it will be very difficult to prove that a marker tagging QTL is indeed physically close to the locus. In some cases, even fine mapping will not help to break the linkage between marker and QTL. Availability of reference genome sequence is helpful to define the physical proximity of the marker to a gene. However, even physical proximity may not insure successful MAS. There are examples showing that out of several mutations, which occurred within a target gene, only one of them, the causative, can be converted into viable assay to leverage in MAS (Zheng et al., 2008). In human molecular genetics, mostly causative mutations have been used to develop diagnostic tools to detect diseases. However, in plants development of gene-based diagnostic tools for MAS has been limited to major crops with available reference genome sequence, e.g. rice and maize (Chen et al., 2010; Zheng et al., 2008). With respect to orphan crops, including wheat and barley, it may require several years before gene-based molecular markers derived from QTL cloning can be used in MAS.

Tracking QTL using molecular markers is just one of the MAS applications in plant breeding. This application uses mostly one or a couple of markers ideally developed based on causative mutations. Applications of MAS in plant breeding were grouped into five broad categories (Collard and Mackill, 2008): (1) marker-assisted germplasm evaluation including pedigree verification, purity assessment, evaluation of genetic diversity, identification of heterotic patterns and event characterization; (2) marker-assisted trait introgression, (3) marker-assisted pyramiding of genes and (4) genomic selection (GS). The nature of the MAS-based molecular breeding projects determines the marker and sample throughput and consequently requires specific marker genotyping technologies. Most of the contemporary marker genotyping technologies are oriented towards SNP detection, because SNPs are amenable for high-throughput automation, and are preferred type of polymorphism in molecular genetics research projects because of their abundance and resolution (Chagné et al., 2007). Majority of SNP genotyping technologies that have been described in this chapter were originally developed for SNP detection in human genetics research. However, the rapid growth and expansion of agribusiness, challenges to 'increase the slope' and 'stay competitive' forced major seed companies to adapt those technologies to plant genome and use in high-throughput SNP genotyping for MAS projects. Although MAS projects are diverse, in terms of sample and marker throughput, they all can be divided into two major groups with opposite tasks: (1) projects that deal with large sample volume (>10,000 plants) to be genotyped with a few markers (1-96 SNPs) and (2) projects that require genotyping a fewer samples (1-300) with large number of SNPs (384 to several millions of SNPs). The projects that fall into the first group are related to categories such as marker-assisted germplasm evaluation, marker-assisted trait introgression and marker-assisted gene pyramiding. The categories of MAS-based projects such as genome wide selection (GWS) for complex traits that fall into the opposite category will require genotyping of several millions of SNPs in fairly small subset of samples. SNP genotyping platforms that would match the requirements of the project from group 1 could include OpenArray platform coupled with TaqMan chemistry and KBiosciences' Competitive Allele Specific PCR (KASPar) complemented with the SNP Line platform (SNP Line XL, Kbiosciences, Hoddesdon, England). The latter has proven to be more cost effective and flexible compared to TaqMan assay (Chen et al., 2010). The second group of projects can be implemented using Illumina's BeadArray technology coupled with GG and Infinium assays. Current throughput of Infinium assay is ~1.1 MM SNPs per iSelect. However, GWS might require several millions of SNPs depending on the complexity of genome and its LD decay rate. If this is the case, then genotyping-by-sequencing (Elshire et al., 2011) could be another alternative.

3.4 Genomic selection approach towards breeding complex traits

Current MAS strategies fit the breeding programs for traits with high heritability and are governed by a single gene or one major QTL that explains large portion of the phenotypic variability. However, the application of MAS for breeding traits with complex genetics based on the interaction of multiple QTL with minor effects has been inefficient. Examples of complex traits are yield, drought tolerance, and nitrogen and water use efficiency. In classical MAS projects researchers use molecular markers that show statistically significant association with a phenotype and are linked to major QTL. Because minor QTL have small effects on phenotype, they have not been applicable in MAS. Meuwissen et al. (2001) described a new methodology in plant breeding, called genomic selection (GS) that was believed to solve problems related to MAS of complex traits. This methodology also applies to molecular markers but in different fashion. Unlike MAS, in GS markers are not used for tracking a trait. In GS high density marker coverage is needed to potentially have all QTL in LD with at least one marker. Then the comprehensive information on all possible loci, haplotypes and marker effects across the entire genome is used to calculate genomic estimated breeding value (GEBV) of a particular line in the breeding population.

Genomic selection of superior lines can be carried out within any breeding population. In order to enable successful GS, the experimental population must be identified. The population should not be necessarily derived from bi-parental cross but must be representative of selection candidates in the breeding program to which GS will be applied (Heffner et al., 2009). Experimental population must be genotyped with large number of markers. Taking into account the low cost of sequencing, the best choice is the implementation of genotyping-by-sequencing which will yield maximum number of polymorphisms. The sequence of the two events, i.e. phenotypic and genotypic data collection, is arbitrary and can be done in parallel. When both phenotypic and genotypic data are ready, one can start "training" molecular markers (Zhong S. 2009). In order to train

GS model, the effect of each marker is calculated computationally. The effect of a marker is represented by a number with a positive or negative sign that indicates the positive or negative effect, respectively, of a particular locus to phenotype. When the effects of all markers are known, they are considered "trained" and ready to assess any breeding population different from the experimental one for the same trait. Availability of trained GS model does not require the collection of phenotypic data from new breeding populations. The same set of "trained" markers will be used to genotype a new breeding population. Based on genotypic data, the known effects of each marker will be summed and GEBV of each line will be calculated. The higher the GEBV value of an individual line, the more likely that this line will be selected and advanced in the breeding cycle. Thus, GS using high-density marker coverage enables to capture QTL with major and minor effects and eliminates the need to collect phenotypic data in all breeding cycles. Also, the application of GS was demonstrated to reduce the number of breeding cycles and increases the annual gain (Heffner et al., 2009). One of the problems of GS is the level of GEBV accuracy. Simulation studies based on simulated and empirical data demonstrated that GEBV accuracy could be within 0.62-0.85. Heffner et al. (2009) used previously reported GEBV accuracy of 0.53 and reported three- and two-fold annual gain in maize and winter barley, respectively.

The obvious advantages of GS over traditional MAS have been successfully proven in animal breeding (Hayes and Goddard, 2010). Rapid evolution of sequencing technologies and high-throughput SNP genotyping systems are enabling generation and validation of millions of markers, giving a "cautious optimism" for successful application of GS in plant breeding of complex traits. Thus, considering current application level and success in various crops, MAS technology still remains in its development stages but attractive for 21st century breeding. Successful efforts, as a wake-up call, further require incorporation of abovementioned advances in large-scale modern genotyping, precise phenotyping, statistically improved genetic mapping and data analysis as well as genome characterization of the crop species.

4. Genomics efforts in understanding molecular basis of plant growth, development and traits of interest towards crop improvement

4.1 Genomics tools for understanding natural variation

The availability of Next generation sequencing (NGS) technologies has paved the way for discovery of genetic variation at whole genome level of multiple genotypes. Ossowski et al. (2008) have demonstrated that even the short reads derived from Illumina Genome Analyzer could reveal most of the sequence variations in *A. thaliana* strains/accessions. The 1001 Arabidopsis Genomes project that is currently ongoing aims to discover the whole genome sequence variation in 1001 distinct accessions of Arabidopsis. The wealth of information from this project enables large scale genetic and functional analyses to address key biological phenomena and leverage that information for improvement in cultivated crop species. NGS technologies have also enabled other whole genome exploratory research on variation detection such as genome-wide DNA methylation detection (Lister et al., 2008), mutation mapping (Ossowski et al., 2008) and DNA-protein interactions (Bernatavichute et al., 2008). Section 2.3.2.3 provides additional details on how NGS technologies are being leveraged for variation detection and HTP molecular marker development. Together these efforts are instrumental in developing molecular markers and other diagnostic tools thereby enabling or accelerating molecular breeding.

4.2 Genomics efforts in understanding root growth, development and architecture

Water and nutrient uptake by roots plays a significant role in the growth of plants. In addition to providing anchorage in the soil, roots can adapt developmentally and physiologically to environmental changes. Efforts in the past to understand the molecular basis of root development have focused on single mutant analysis. While these approaches shed light on the cell type patterning in root, they have revealed the complex interactions underlying root growth and development accentuating the need for the use of exhaustive global "omics" analyses. Recent availability of a root expression map (Brady et al., 2007), root proteome (Baerenfaller et al., 2008) and environment-specific expression data are revealing complex transcriptional and pot-transcriptional pathways in root development (Iyer-Pascuzzi et al., 2009). These efforts and the initiatives to integrate the data from multiple "omics" studies are paving the way for the understanding of root biology across plant species.

In a recent review Hochholdinger and Tuberosa (2009) summarized the latest results on the genetic and genomic dissection of maize root development and on the cloning of underlying genes using root architecture mutants. Maize root system has complex architecture and is controlled by many genes. Characterization of *rtcs* mutant (rootless concerning crown and seminal roots) and the map-based cloning of underlying *RTCS* gene revealed that this codes for a transcription factor involved in early events responsible for root initiation. Similarly analysis of *rth1* and *rth3* mutants (roothairless 1 and 3) demonstrated their involvement in root hair elongation and the corresponding genes were found to be parts of machinery responsible for tethering exocytotic vesicles (rth1) (Wen et al., 2005) and cell expansion and cell wall biosynthesis related processes. While QTL mapping has identified some regions that influence root features and thereby yield, the use of 'omics' technologies provided unprecedented capability in obtaining significant insights into maize root development. For example, comparative laser capture microdissection (LCM) gene expression profiles of primary root meristem and root cap cells identified gene clusters linked to transport, environmental interactions and hormonal and carbohydrate signaling (Jiang et al., 2006). Similarly, comparison of LCM microarray profiles between pericle cells of wild-type and mutant rum1 (rootless with undetectable meristems 1) seedlings revealed a set of genes related to signal transduction, cell cycle, transcription and translation that are probably linked to lateral root initiation (Woll et al., 2005). In another transcriptome study that analyzed maize root responses towards environmental stimuli, highly differentially expressed transcripts were found to be those arising from reactive oxygen species (ROS) and carbon metabolism in root tips and elongation zone (Spollen et al., 2008). Comparative proteome analysis of maize roots from mutants and wild type as well as before and after a given treatment led to the identification of proteins that are likely to be associated with influence of lateral roots on the proteome composition of the primary root, phosphorus depletion and water deficit (reviewed in Hochholdinger & Tuberosa, 2009).

4.3 Genomics-based dissection of molecular basis of biomass production and Cell wall composition

Yield is the most important yet one of the most intriguing traits in agriculture. Despite its economic importance very little is known about the mechanisms underlying yield. One approach to identify the candidate genes responsible for yield is the use of information from model plants such as Arabidopsis and leverage this information in cultivated crop plants. Using genetics and genomics approaches candidate genes were identified in Arabidopsis, many of which had a significant effect on the biomass production through an increase in the size of leaves or roots (Gonzalez et al., 2009). Genes thus identified belonged to different functional classes that include transcriptional factors, translational regulators (protein synthesis and modification), signaling pathways, hormonal regulation, cell division and expansion. Examples now exist where some of the candidate genes belonging to these categories have been demonstrated to positively influence yield based on transgenic studies (Wu et al., 2008). In order to get a better handle on yield, it is proposed to employ a systems biology approach for obtaining additional 'omics' data and generating an integrated network of pathways in which the candidate genes are key players.

One of the key distinguishing features of grasses is the presence of (1,3;1,4)-β-D-glucans in their cell walls. These (1,3;1,4)-β-D-glucans are almost exclusively distributed within Poaceae where they are present in both primary and secondary walls. Considering the undesirable characteristics of barley (1,3;1,4)-β-D-glucans in malting and brewing industries as well as in animal feeds, many researchers set out to investigate the molecular mechanism underlying the accumulation of this class of polysaccharides in barley by developing molecular markers and mapping QTL. Mapping efforts led to the identification of a region on barley chromosome 2 that controls the production of (1,3;1,4)-β-D-glucans (Han et al., 1995). Although many biochemical efforts during 1980s and 1990s focused on the isolation of (1,3;1,4)-β-D-glucan synthase, the genes coding for these synthases could not be identified due to issues in the purification of these enzymes. In the recent years, the beneficial effects of (1,3;1,4)-β-D-glucans in human health (Wood, 2007) as well as their ability to positively influence biofuel industry through increased biomass production led to the implementation of molecular and genomics approaches for rapid identification of genes underlying these glucans (Fincher, 2009). In order to identify the genes and proteins responsible for the biosynthesis of β-glucans, Burton et al. (2006) have used the mapping information of Han et al. (1995) along with the information on conserved genome structure, gene collinearity or synteny between barley and rice, for which complete genome sequence is available. Using this approach they have successfully demonstrated the presence of a rice locus corresponding to mapped barley region and showed that the rice genome contains six cellulose synthase like (CslF) genes, thus identifying strong candidate genes for β-glucan biosynthesis.

4.4 Dissecting leaf architecture using genomics

Photosynthesis, the process through which plants harvest light energy and convert it into the building blocks of life is an extremely important biological phenomenon. While details on the process of photosynthesis have been worked out over the years, we are at the beginning in understanding the genetic and molecular basis of this complex process. Many efforts are currently underway in dissecting different components of photosynthesis. Using a high-throughput Illumina sequencing approach, Li et al. (2010) have analyzed maize leaf transcriptome and identified differential mRNA processing events for most maize genes. Their data revealed maize transcriptome to be a dynamic one with transcripts for primary cell wall and basic cellular metabolism at the leaf base transitioning to transcripts for secondary cell wall biosynthesis and C_4 photosynthetic development toward the tip. They found that as the leaf develops, large numbers of genes are turned on and off. Such information could not be obtained prior to the availability of massively parallel techniques such as the next generation sequencing. The comprehensive information from this and other studies will serve as the foundation for a systems biology approach for the understanding of photosynthetic development of maize.

4.5 Genomics for improving abiotic stress tolerance of crops

Abiotic stresses have become major concerns for global crop production and conventional approaches for developing tolerant cultivars have been difficult due to the complex inheritance of stress tolerance traits. For example, drought is the most recalcitrant abiotic stress trait for crop improvement due to its quantitative genetic inheritance and the involvement of multiple physiological effects on the ultimate yield (Passioura, 2002). Recent years have seen a renewed interested in understanding the molecular basis of drought tolerance with many studies reporting the mapping of QTL underlying this trait and the availability of high-throughput sequencing and associated computational tools are providing new avenues for the characterization of this complex trait (Tuberosa and Salvi, 2006). Genomics efforts to date in drought tolerance research could be broadly divided into two approaches. In one approach, QTL maps were combined with maps containing genic information or annotated genome sequence (Varshney et al., 2005) for the rapid identification of candidate genes. Use of this approach for the analysis of root trait QTL along with EST and cDNA screening has identified OsEXP2 and EGase genes in rice that were found to be involved in cell expansion (Zheng et al., 2003). The second approach is the employment of high-throughput transcriptomic profiling to investigate the changes in gene expression in response to drought. A recurring theme based on the comparison of multiple transcriptomics studies is the central role of transcription factors (TFs) in drought as well as the complex hierarchy of regulatory networks that modulate the tissue-specific differential expression of candidate genes (Yamaguchi-Shinozaki and Shinozaki, 2005). Proteomic profiling has also revealed several lead candidates for drought resistance in rice and maize. The actin depolymerizing factor (ADF) in rice has displayed most significant drought-induced fluctuations with its concentration increasing in leaves (especially in leaf blades and sheath) and roots after exposure to dehydration in drought-tolerant cultivars (Ali and Komatsu, 2006). In another approach, proteomic profiling was carried out on a mapping population to identify protein quantity loci (PQLs) i.e., QTLs influencing quantity of protein. Such an analysis in maize led to the identification of a putative transcription factor gene (*Asr1)* that co-localized with a PQL for the ASR1 protein and a QTL for ASI and leaf senescence (Jeanneau et al., 2002).

Salt tolerance is another complex trait threatening crop production in many countries worldwide and genomics efforts are underway to dissect this trait. Sanchez et al. (2011) have used comparative functional genomics techniques such as ionomics, transcriptomics and metabolomics to distinguish genotype-specific transcriptional and metabolic changes from those of true salinity responses leading to the identification of conserved and tolerance-specific responses towards achieving salinity tolerance across species.

5. Conclusion

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 during the last quarter of twentieth century, the advent of genomics tools and technologies has been providing unprecedented capabilities for understanding the molecular basis of plant growth, development and key traits towards improving crop productivity in the 21st century. A multitude of omics and associated HTP technologies are enabling systematic dissection and understanding of plants that was not possible previously. The knowledge derived from such efforts will certainly be useful in developing "designer plants" that can yield better through improved growth and ability to withstand biotic and abiotic stresses. In addition to the insights, derived from individual or a combination of omics technologies applied to specific traits of interest, the renewal of 'holistic' systems biology concept and genome-wide measurements of components of interest certainly has the potential in dissecting the molecular, biochemical, physiological and evolutionary basis of traits and biological phenomena. This holds a great promise for crop improvement. Continued development of 'omics' technologies and computational tools for accurate analysis, and integration and interpretation of massive amounts of data are key challenges that need to be addressed to reap the full potential of genomics and systems biology approaches. The progress made so far through marker-assisted breeding and genomics and the promising technological breakthroughs are certainly paving the way for "Genomics-assisted breeding" in the 21st century!

6. References

Abdurakhmonov, I., Kohel, R., Yu, J., Pepper, A., Abdullaev, A., Kushanov, F., Salakhutdinov, I., Buriev, Z., Saha, S., and Scheffler, B. (2008). Molecular diversity and association mapping of fiber quality traits in exotic G. hirsutum L. germplasm. *Genomics* Vol. 92, No. 6, pp. 478-487, ISSN 0888-7543.

Abdurakhmonov, I. Y., and Abdukarimov, A. (2008). Application of association mapping to understanding the genetic diversity of plant germplasm resources. *Int J Plant Genomics* Vol. 574927.

Agrama, H., and Eizenga, G. (2008). Molecular diversity and genome-wide linkage disequilibrium patterns in a worldwide collection of Oryza sativa and its wild relatives. *Euphytica* Vol. 160, No. 3, pp. 339-355, ISSN 0014-2336.

Ali, G. M., and Komatsu, S. (2006). Proteomic analysis of rice leaf sheath during drought stress. *Journal of proteome research* Vol. 5, No. 2, pp. 396-403, ISSN 1535-3893.

Aranzana, M. J., Kim, S., Zhao, K., Bakker, E., Horton, M., Jakob, K., Lister, C., Molitor, J., Shindo, C., and Tang, C. (2005). Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. *PLoS genetics* Vol. 1, No. 5, pp. e60, ISSN 1553-7404.

Argout, X., Salse, J., Aury, J.-M., Guiltinan, M. J., Droc, G., Gouzy, J., Allegre, M., Chaparro, C., Legavre, T., Maximova, S. N., Abrouk, M., Murat, F., Fouet, O., Poulain, J., Ruiz, M., Roguet, Y., Rodier-Goud, M., Barbosa-Neto, J. F., Sabot, F., Kudrna, D., Ammiraju, J. S. S., Schuster, S. C., Carlson, J. E., Sallet, E., Schiex, T., Dievart, A., Kramer, M., Gelley, L., Shi, Z., Berard, A., Viot, C., Boccara, M., Risterucci, A. M., Guignon, V., Sabau, X., Axtell, M. J., Ma, Z., Zhang, Y., Brown, S., Bourge, M., Golser, W., Song, X., Clement, D., Rivallan, R., Tahi, M., Akaza, J. M., Pitollat, B., Gramacho, K., D'Hont, A., Brunel, D., Infante, D., Kebe, I., Costet, P., Wing, R., McCombie, W. R., Guiderdoni, E., Quetier, F., Panaud, O., Wincker, P., Bocs, S., and Lanaud, C. (2011). The genome of Theobroma cacao. *Nat Genet* Vol. 43, No. 2, pp. 101-108, ISSN 1061-4036.

Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R., Hirsch-Hoffmann, M., Yalovsky, S., Zimmermann, P., Grossniklaus, U., Gruissem, W., and Baginsky, S. (2008). Genome-scale proteomics reveals Arabidopsis thaliana gene models and proteome dynamics. *Science* Vol. 320, No. 5878, pp. 938, ISSN 0036-8075.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., and Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* Vol. 3, No. 10, pp. e3376, ISSN 1932-6203.

Barbazuk, W. B., Emrich, S. J., Chen, H. D., Li, L., and Schnable, P. S. (2007). SNP discovery via 454 transcriptome sequencing. *The Plant Journal* Vol. 51, No. 5, pp. 910-918, ISSN 1365-313X.

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 physiology* Vol. 132, No. 1, pp. 84, ISSN 0032-0889.

Bauer-Mehren, A., Furlong, L. I., and Sanz, F. (2009). Pathway databases and tools for their exploitation: benefits, current limitations and challenges. *Molecular systems biology* Vol. 5, No. 1.

Beckmann, J., and Soller, M. (1986). Restriction fragment length polymorphisms in plant genetic improvement. *Oxford surveys of plant molecular and cell biology.* Vol. 3, No., pp. 196-250.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., Hall, K. P., Evers, D. J., Barnes, C. L., and Bignell, H. R. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* Vol. 456, No. 7218, pp. 53-59, ISSN 0028-0836.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Science* Vol. 48, No. 5, pp. 1649-1664.

Bernatavichute, Y. V., Zhang, X., Cokus, S., Pellegrini, M., and Jacobsen, S. E. (2008). Genome-wide association of histone H3 lysine nine methylation with CHG DNA methylation in Arabidopsis thaliana. *PLoS ONE* Vol. 3, No. 9, pp. e3156, ISSN 1932-6203.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American Journal of Human Genetics* Vol. 32, No. 3, pp. 314.

Brady, S. M., Orlando, D. A., Lee, J. Y., Wang, J. Y., Koch, J., Dinneny, J. R., Mace, D., Ohler, U., and Benfey, P. N. (2007). A high-resolution root spatiotemporal map reveals dominant expression patterns. *Science* Vol. 318, No. 5851, pp. 801, ISSN 0036-8075.

Bräutigam, A., and Gowik, U. (2010). What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. *Plant Biology* Vol. 12, No. 6, pp. 831-841, ISSN 1438-8677.

Brenan, C., and Morrison, T. (2005). High throughput, nanoliter quantitative PCR. *Drug Discovery Today: Technologies* Vol. 2, No. 3, pp. 247-253, ISSN 1740-6749.

Breyne, P., and Zabeau, M. (2001). Genome-wide expression analysis of plant cell cycle modulated genes. *Current opinion in plant biology* Vol. 4, No. 2, pp. 136-142, ISSN 1369-5266.

Bundock, P. C., Eliott, F. G., Ablett, G., Benson, A. D., Casu, R. E., Aitken, K. S., and Henry, R. J. (2009). Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. *Plant biotechnology journal* Vol. 7, No. 4, pp. 347-354, ISSN 1467-7652.

Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W., and Swaminathan, S. (1999). Structural genomics: beyond the human genome project. *Nat Genet* Vol. 23, No. 2, pp. 151-7,(Oct), ISSN 1061-4036.

Burton, R. A., Wilson, S. M., Hrmova, M., Harvey, A. J., Shirley, N. J., Medhurst, A., Stone, B. A., Newbigin, E. J., Bacic, A., and Fincher, G. B. (2006). Cellulose synthase-like CslF genes mediate the synthesis of cell wall (1, 3; 1, 4)-ß-D-glucans. *Science* Vol. 311, No. 5769, pp. 1940, ISSN 0036-8075.

Cahill, D. J., and Schmidt, D. H. (2004). Use of marker assisted selection in a product development breeding program. *Proc. 4th Int. Crop. Sci. Cong* Vol. 26.

Chagné, D., Batley, J., Edwards, D., and Forster, J. W. (2007). Single Nucleotide Polymorphism Genotyping in Plants. *In* "Association Mapping in Plants" (N. C. Oraguzie, E. H. A. Rikkerink, S. E. Gardiner and H. N. Silva, eds.), pp. 77-94. Springer New York.

Chen, W., Mingus, J., Mammadov, J., Backlund, J. E., Greene, T., Thompson , S., and Kumpatla, S. (2010). *In* "International Plant & Animal Genomes XVIII Conference", San Diego.

Chinnusamy, V., and Zhu, J. K. (2009). Epigenetic regulation of stress responses in plants. *Current opinion in plant biology* Vol. 12, No. 2, pp. 133-139, ISSN 1369-5266.

Cokus, S. J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C. D., Pradhan, S., Nelson, S. F., Pellegrini, M., and Jacobsen, S. E. (2008). Shotgun bisulfite sequencing of the Arabidopsis genome reveals DNA methylation patterning. *Nature* Vol. 452, No. 7184, pp. 215.

Collard, B. C., and Mackill, D. J. (2008). Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci* Vol. 363, No. 1491, pp. 557-72,(Feb 12), ISSN 0962-8436.

Concibido, V., Denny, R., Lange, D., Orf, J., and Young, N. (1996). RFLP mapping and marker-assisted selection of soybean cyst nematode resistance in PI 209332. *Crop Science* Vol. 36, No. 6, pp. 1643-1650, ISSN 0011-183X.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* Vol. 21, No. 18, pp. 3674, ISSN 1367-4803.

Cooper, B., Clarke, J. D., Budworth, P., Kreps, J., Hutchison, D., Park, S., Guimil, S., Dunn, M., Luginbühl, P., and Ellero, C. (2003). A network of rice genes associated with stress response and seed development. *Proceedings of the National Academy of Sciences* Vol. 100, No. 8, pp. 4945, ISSN 0027-8424.

Crosbie, T. M., Eathington, S. R., Johnson Sr, G. R., Edwards, M., Reiter, R., Stark, S., Mohanty, R. G., Oyervides, M., Buehler, R. E., and Walker, A. K. (2003). Plant breeding: Past, present, and future. pp. 3-50. Wiley Online Library.

Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., and Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat Rev Genet* Vol. 12, No. 7, pp. 499-510, ISSN 1471-0056.

Deschamps, S., and Campbell, M. (2010). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. *Molecular Breeding* Vol. 25, No. 4, pp. 553-570, ISSN 1380-3743.

Devos, K. M. (2005). Updating the 'Crop Circle'. *Current opinion in plant biology* Vol. 8, No. 2, pp. 155-162, ISSN 1369-5266.

Druka, A., Potokina, E., Luo, Z., Jiang, N., Chen, X., Kearsey, M., and Waugh, R. (2010). Expression quantitative trait loci analysis in plants. *Plant biotechnology journal* Vol. 8, No. 1, pp. 10-27, ISSN 1467-7652.

Dubcovsky, J., Ramakrishna, W., SanMiguel, P. J., Busso, C. S., Yan, L., Shiloff, B. A., and Bennetzen, J. L. (2001). Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant physiology* Vol. 125, No. 3, pp. 1342, ISSN 0032-0889.

Eichten, S. R., Foerster, J., de Leon, N., Ying, K., Yeh, C. T., Liu, S., Jeddeloh, J., Schnable, P., Kaeppler, S. M., and Springer, N. M. (2011). B73-Mo17 near isogenic lines (NILs) demonstrate dispersed structural variation in maize. *Plant physiology* Vol., No., ISSN 0032-0889.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., and Bettman, B. (2009). Real-time DNA sequencing from single polymerase molecules. *Science* Vol. 323, No. 5910, pp. 133, ISSN 0036-8075.

Elshire, R. J., Glaubitz, J. C., Sun, Q., Poland, J. A., Kawamoto, K., Buckler, E. S., and Mitchell, S. E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* Vol. 6, No. 5, pp. e19379.

Emberton, J., Ma, J., Yuan, Y., SanMiguel, P., and Bennetzen, J. L. (2005). Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Res* Vol. 15, No. 10, pp. 1441, ISSN 1088-9051.

Endler, L., Rodriguez, N., Juty, N., Chelliah, V., Laibe, C., Li, C., and Le Novère, N. (2009). Designing and encoding models for synthetic biology. *Journal of The Royal Society Interface* Vol. 6, No. Suppl 4, pp. S405, ISSN 1742-5689.

Fan, J. B., Oliphant, A., Shen, R., Kermani, B., Garcia, F., Gunderson, K., Hansen, M., Steemers, F., Butler, S., and Deloukas, P. (2003). Highly parallel SNP genotyping. Vol. 68, pp. 69. Cold Spring Harbor Laboratory Press.

Fincher, G. B. (2009). Exploring the evolution of (1, 3; 1, 4)-[beta]-d-glucans in plant cell walls: comparative genomics can help! *Current opinion in plant biology* Vol. 12, No. 2, pp. 140-147, ISSN 1369-5266.

Fu, H., and Dooner, H. K. (2002). Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences* Vol. 99, No. 14, pp. 9573, ISSN 0027-8424.

Gale, M. D., and Devos, K. M. (1998). Comparative genetics in the grasses. *Proceedings of the National Academy of Sciences* Vol. 95, No. 5, pp. 1971, ISSN 0027-8424.

Ganal, M. W., Altmann, T., and Röder, M. S. (2009). SNP identification in crop plants. *Current opinion in plant biology* Vol. 12, No. 2, pp. 211-217, ISSN 1369-5266.

Gaut, B. S., and Doebley, J. F. (1997). DNA sequence evidence for the segmental allotetraploid origin of maize. *Proceedings of the National Academy of Sciences* Vol. 94, No. 13, pp. 6809, ISSN 0027-8424.

Godfray, H. C., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., Pretty, J., Robinson, S., Thomas, S. M., and Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. *Science* Vol. 327, No. 5967, pp. 812-8,(Feb 12), ISSN 1095-9203.

Gonzalez, N., Beemster, G. T. S., and Inzé, D. (2009). David and Goliath: what can the tiny weed Arabidopsis teach us to improve biomass production in crops? *Current opinion in plant biology* Vol. 12, No. 2, pp. 157-164, ISSN 1369-5266.

Gore, M. A. W., Ersoz, M. H., Bouffard, E. S., Szekeres, P., Jarvie, E. S., Hurwitz, T. P., Narechania, B. L., Harkins, A., Grills, T. T., and Ware, G. S. (2009). Large-scale discovery of gene-enriched SNPs. *The Plant Genome* Vol. 2, No. 2, pp. 121, ISSN 1940-3372.

Gupta, P., Varshney, R., Sharma, P., and Ramesh, B. (1999). Molecular markers and their applications in wheat breeding. *Plant breeding* Vol. 118, No. 5, pp. 369-390, ISSN 1439-0523.

Hamilton, J., Hansey, C., Whitty, B., Stoffel, K., Massa, A., Van Deynze, A., De Jong, W., Douches, D., and Buell, C. R. (2011). Single nucleotide polymorphism discovery in elite north american potato germplasm. *BMC Genomics* Vol. 12, No. 1, pp. 302, ISSN 1471-2164.

Han, F., Ullrich, S., Chirat, S., Menteur, S., Jestin, L., Sarrafi, A., Hayes, P., Jones, B., Blake, T., and Wesenberg, D. (1995). Mapping of b-glucan content and b-glucanase activity loci in barley grain and malt. *Theoretical and Applied Genetics* Vol. 91, No. 6, pp. 921-927, ISSN 0040-5752.

Hayes, B., and Goddard, M. (2010). Genome-wide association and genomic selection in animal breeding. *Genome* Vol. 53, No. 11, pp. 876-883, ISSN 0831-2796.

Heffner, E. L., Sorrells, M. E., and Jannink, J. L. (2009). Genomic selection for crop improvement. *Crop Science*, Vol. 49, No. 1, pp 1-12.

Higgins, J. A., Bailey, P. C., and Laurie, D. A. (2010). Comparative genomics of flowering time pathways using Brachypodium distachyon as a model for the temperate grasses. *PLoS ONE* Vol. 5, No. 4, pp. e10065, ISSN 1932-6203.

Hochholdinger, F., and Tuberosa, R. (2009). Genetic and genomic dissection of maize root development and architecture. *Current opinion in plant biology* Vol. 12, No. 2, pp. 172-177, ISSN 1369-5266.

Holloway, B., Luck, S., Beatty, M., Rafalski, J. A., and Li, B. (2011). Genome-wide expression quantitative trait loci (eQTL) analysis in maize. *BMC Genomics* Vol. 12, No. 1, pp. 336, ISSN 1471-2164.

Honsdorf, N., Becker, H. C., and Ecke, W. (2010). Association mapping for phenological, morphological, and quality traits in canola quality winter rapeseed (Brassica napus L.). *Genome* Vol. 53, No. 11, pp. 899-907, ISSN 0831-2796.

Hribova, E., Neumann, P., Macas, J., and Dolezel, J. (2009). Analysis of genome structure and organization in banana (Musa acuminata) using 454 sequencing. *Plant and Animal Genomes XVII. San Diego, CA*.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., and Huang, T. (2009). High-throughput genotyping by whole-genome resequencing. *Genome Res* Vol. 19, No. 6, pp. 1068, ISSN 1088-9051.

Hubert, B., Rosegrant, M., van Boekel, M., and Ortiz, R. (2010). The future of food: scenarios for 2050. *Crop Science* Vol. 50.

Iyer-Pascuzzi, A., Simpson, J., Herrera-Estrella, L., and Benfey, P. N. (2009). Functional genomics of root growth and development in Arabidopsis. *Current opinion in plant biology* Vol. 12, No. 2, pp. 165-171, ISSN 1369-5266.

Jansen, R. C., and Nap, J. P. (2001). Genetical genomics: the added value from segregation. *Trends in Genetics* Vol. 17, No. 7, pp. 388-391, ISSN 0168-9525.

Jeanneau, M., Gerentes, D., Foueillassar, X., Zivy, M., Vidal, J., Toppan, A., and Perez, P. (2002). Improvement of drought tolerance in maize: towards the functional validation of the Zm-Asr1 gene and increase of water use efficiency by over-expressing C4-PEPC. *Biochimie* Vol. 84, No. 11, pp. 1127-1135, ISSN 0300-9084.

Jiang, K., Zhang, S., Lee, S., Tsai, G., Kim, K., Huang, H., Chilcott, C., Zhu, T., and Feldman, L. J. (2006). Transcription profile analyses identify genes and pathways central to root cap functions in maize. *Plant molecular biology* Vol. 60, No. 3, pp. 343-363, ISSN 0167-4412.

Jiao, Y., Wang, Y., Xue, D., Wang, J., Yan, M., Liu, G., Dong, G., Zeng, D., Lu, Z., and Zhu, X. (2010). Regulation of OsSPL14 by OsmiR156 defines ideal plant architecture in rice. *Nature genetics* Vol. 42, No. 6, pp. 541-544, ISSN 1061-4036.

Jin, L., Lu, Y., Xiao, P., Sun, M., Corke, H., and Bao, J. (2010). Genetic diversity and population structure of a diverse set of rice germplasm for association mapping. *TAG Theoretical and Applied Genetics* Vol. 121, No. 3, pp. 475-487, ISSN 0040-5752.

Joosen, R., Ligterink, W., Hilhorst, H., and Keurentjes, J. (2009). Advances in genetical genomics of plants. *Current Genomics* Vol. 10, No. 8, pp. 540.

Jordan, M. C., Somers, D. J., and Banks, T. W. (2007). Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci†. *Plant biotechnology journal* Vol. 5, No. 3, pp. 442-453, ISSN 1467-7652.

Ju, J., Kim, D. H., Bi, L., Meng, Q., Bai, X., Li, Z., Li, X., Marma, M. S., Shi, S., and Wu, J. (2006). Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences* Vol. 103, No. 52, pp. 19635, ISSN 0027-8424.

Jung, M., Ching, A., Bhattramakki, D., Dolan, M., Tingey, S., Morgante, M., and Rafalski, A. (2004). Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. *TAG Theoretical and Applied Genetics* Vol. 109, No. 4, pp. 681-689, ISSN 0040-5752.

Kearsey, M., and Farquhar, A. (1998). QTL analysis in plants; where are we now? *Heredity* Vol. 80, No. 2, pp. 137-142, ISSN 0018-067X.

Keurentjes, J. J. B., Angenent, G. C., Dicke, M., Santos, V. A. P., Molenaar, J., and van der Putten, W. H. (2011). Redefining plant systems biology: from cell to ecosystem. *Trends in Plant Science* Vol., No., ISSN 1360-1385.

Keurentjes, J. J. B., Fu, J., Terpstra, I. R., Garcia, J. M., Van Den Ackerveken, G., Snoek, L. B., Peeters, A. J. M., Vreugdenhil, D., Koornneef, M., and Jansen, R. C. (2007). Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences* Vol. 104, No. 5, pp. 1708, ISSN 0027-8424.

Kolukisaoglu, Ü., and Thurow, K. (2010). Future and frontiers of automated screening in plant sciences. *Plant Science* Vol. 178, No. 6, pp. 476-484, ISSN 0168-9452.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* Vol. 5, No. 1, pp. 59, ISSN 1471-2105.

Korlach, J., Bibillo, A., Wegener, J., Peluso, P., Pham, T. T., Park, I., Clark, S., Otto, G. A., and Turner, S. W. (2008). Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides. *Nucleosides, Nucleotides and Nucleic Acids* Vol. 27, No. 9, pp. 1072-1082, ISSN 1525-7770.

Laird, P. W. (2010). Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics* Vol. 11, No. 3, pp. 191-203, ISSN 1471-0056.

Li, P., Ponnala, L., Gandotra, N., Wang, L., Si, Y., Tausta, S. L., Kebrom, T. H., Provart, N., Patel, R., and Myers, C. R. (2010). The developmental dynamics of the maize leaf transcriptome. *Nature genetics* Vol., No., ISSN 1061-4036.

Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* Vol. 133, No. 3, pp. 523-536, ISSN 0092-8674.

Livak, K. J., Flood, S., Marmaro, J., Giusti, W., and Deetz, K. (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. *Genome Res* Vol. 4, No. 6, pp. 357, ISSN 1088-9051.

Lo, C., Bashir, A., Bansal, V., and Bafna, V. (2011). Strobe sequence design for Haplotype assembly. *BMC Bioinformatics* Vol. 12, No. Suppl 1, pp. S24, ISSN 1471-2105.

Lundquist, P. M., Zhong, C. F., Zhao, P., Tomaney, A. B., Peluso, P. S., Dixon, J., Bettman, B., Lacroix, Y., Kwo, D. P., and McCullough, E. (2008). Parallel confocal detection of single molecules in real time. *Optics letters* Vol. 33, No. 9, pp. 1026-1028, ISSN 1539-4794.

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., Howard, E., Shendure, J., and Turner, D. J. (2010). Target-enrichment strategies for next-generation sequencing. *Nature methods* Vol. 7, No. 2, pp. 111-118, ISSN 1548-7091.

Mammadov, J. A., Chen, W., Ren, R., Pai, R., Marchione, W., Yalçin, F., Witsenboer, H., Greene, T. W., Thompson, S. A., and Kumpatla, S. P. (2010). Development of highly polymorphic SNP markers from the complexity reduced portion of maize [Zea mays L.] genome for use in marker-assisted breeding. *TAG Theoretical and Applied Genetics* Vol. 121, No. 3, pp. 577-588, ISSN 0040-5752.

Mardis, E. R. (2008a). The impact of next-generation sequencing technology on genetics. *Trends in Genetics* Vol. 24, No. 3, pp. 133-141, ISSN 0168-9525.

Mardis, E. R. (2008b). Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.* Vol. 9, No., pp. 387-402, ISSN 1527-8204.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* Vol. 437, No. 7057, pp. 376-80,(Sep 15), ISSN 1476-4687.

Massman, J., Cooper, B., Horsley, R., Neate, S., Dill-Macky, R., Chao, S., Dong, Y., Schwarz, P., Muehlbauer, G., and Smith, K. (2011). Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. *Molecular Breeding* Vol. 27, No. 4, pp. 439-454, ISSN 1380-3743.

Mayer, K. F. X., Martis, M., Hedley, P. E., Šimková, H., Liu, H., Morris, J. A., Steuernagel, B., Taudien, S., Roessner, S., and Gundlach, H. (2011). Unlocking the barley genome by chromosomal and comparative genomics. *The Plant Cell Online* Vol. 23, No. 4, pp. 1249, ISSN 1040-4651.

Metzker, M. L. (2009). Sequencing technologies—the next generation. *Nature Reviews Genetics* Vol. 11, No. 1, pp. 31-46, ISSN 1471-0056.

Meuwissen, T., Hayes, B., and Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* Vol. 157, No. 4, pp. 1819.

Meyers, B. C., Tingey, S. V., and Morgante, M. (2001). Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. *Genome Res* Vol. 11, No. 10, pp. 1660, ISSN 1088-9051.

Mochida, K., and Shinozaki, K. (2010). Genomics and bioinformatics resources for crop improvement. *Plant and Cell Physiology* Vol. 51, No. 4, pp. 497, ISSN 0032-0781.

Montes, J., Paul, C., Kusterer, B., and Melchinger, A. (2006). Near infrared spectroscopy to measure maize grain composition on plot combine harvesters: evaluation of calibration techniques, mathematical transformations and scatter corrections. *Journal of near infrared spectroscopy* Vol. 14, No. 6, pp. 387-394, ISSN 0967-0335.

Montes, J., Technow, F., Dhillon, B., Mauch, F., and Melchinger, A. (2011). High-throughput non-destructive biomass determination during early plant development in maize under field conditions. *Field Crops Research* Vol., No., ISSN 0378-4290.

Morozova, O., and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics* Vol. 92, No. 5, pp. 255-264, ISSN 0888-7543.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* Vol. 5, No. 7, pp. 621-628.

Narina, S., Buyyarapu, R., Kottapalli, K., Sartie, A., Ali, M., Robert, A., Hodeba, M., Sayre, B., and Scheffler, B. (2011). Generation and analysis of Expressed Sequence Tags (ESTs) for marker development in yam (Dioscorea alata L.). *BMC Genomics* Vol. 12, No. 1, pp. 100, ISSN 1471-2164.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., and Nickerson, D. A. (2009). Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* Vol. 42, No. 1, pp. 30-35, ISSN 1061-4036.

Nijman, I. J., Mokry, M., van Boxtel, R., Toonen, P., de Bruijn, E., and Cuppen, E. (2010). Mutation discovery by targeted genomic enrichment of multiplexed barcoded samples. *Nature methods* Vol. 7, No. 11, pp. 913-915, ISSN 1548-7091.

Nordborg, M., and Tavaré, S. (2002). Linkage disequilibrium: what history has to tell us. *Trends in Genetics* Vol. 18, No. 2, pp. 83-90, ISSN 0168-9525.

Okou, D. T., Steinberg, K. M., Middle, C., Cutler, D. J., Albert, T. J., and Zwick, M. E. (2007). Microarray-based genomic selection for high-throughput resequencing. *Nature methods* Vol. 4, No. 11, pp. 907-909, ISSN 1548-7091.

Oliver, R., Lazo, G., Lutz, J., Rubenfield, M., Tinker, N., Anderson, J., Wisniewski Morehead, N., Adhikary, D., Jellen, E., Maughan, P. J., Brown Guedira, G., Chao, S., Beattie, A., Carson, M., Rines, H., Obert, D., Bonman, J. M., and Jackson, E. (2011). Model SNP development for complex genomes based on hexaploid oat using high-throughput 454 sequencing technology. *BMC Genomics* Vol. 12, No. 1, pp. 77, ISSN 1471-2164.

Ophir, R., and Graur, D. (1997). Patterns and rates of indel evolution in processed pseudogenes from humans and murids. *Gene* Vol. 205, No. 1-2, pp. 191-202, ISSN 0378-1119.

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res* Vol. 18, No. 12, pp. 2024, ISSN 1088-9051.

Palmer, L. E., Rabinowicz, P. D., O'Shaughnessy, A. L., Balija, V. S., Nascimento, L. U., Dike, S., de la Bastide, M., Martienssen, R. A., and McCombie, W. R. (2003). Maize genome sequencing by methylation filtration. *Science* Vol. 302, No. 5653, pp. 2115, ISSN 0036-8075.

Passioura, J. B. (2002). Review: Environmental biology and crop improvement. *Functional Plant Biology* Vol. 29, No. 5, pp. 537-546, ISSN 1445-4416.

Peleman, J. D., and van der Voort, J. R. (2003). Breeding by design. *Trends in Plant Science* Vol. 8, No. 7, pp. 330-334, ISSN 1360-1385.

Pennisi, E. (2010). Semiconductors inspire new sequencing technologies. *Science* Vol. 327, No. 5970, pp. 1190, ISSN 0036-8075.

Poland, J. A., Bradbury, P. J., Buckler, E. S., and Nelson, R. J. (2011). Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences* Vol. 108, No. 17, pp. 6893, ISSN 0027-8424.

Powell, W., Machray, G. C., and Provan, J. (1996). Polymorphism revealed by simple sequence repeats. *Trends in Plant Science* Vol. 1, No. 7, pp. 215-222, ISSN 1360-1385.

Rafalski, J. A. (2002). Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Science* Vol. 162, No. 3, pp. 329-333, ISSN 0168-9452.

Ragot, M., Lee, M., Guimarães, E., Ruane, J., Scherf, B., Sonnino, A., and Dargie, J. (2007). Marker-assisted selection in maize: current status, potential, limitations and perspertives from the private and public sectors. *Marker-Assisted Selection, Current Status and Future Perspectives in Crops, Livestock, Forestry and Fish* Vol., No., pp. 117–150.

Reinders, J., Delucinge Vivier, C., Theiler, G., Chollet, D., Descombes, P., and Paszkowski, J. (2008). Genome-wide, high-resolution DNA methylation profiling using bisulfite-mediated cytosine conversion. *Genome Res* Vol. 18, No. 3, pp. 469, ISSN 1088-9051.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences* Vol. 98, No. 20, pp. 11479, ISSN 0027-8424.

Rossberg, M., Theres, K., Acarkan, A., Herrero, R., Schmitt, T., Schumacher, K., Schmitz, G., and Schmidt, R. (2001). Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, Arabidopsis, and Capsella genomes. *The Plant Cell Online* Vol. 13, No. 4, pp. 979, ISSN 1040-4651.

Rostoks, N., Ramsay, L., MacKenzie, K., Cardle, L., Bhat, P. R., Roose, M. L., Svensson, J. T., Stein, N., Varshney, R. K., and Marshall, D. F. (2006). Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proceedings of the National Academy of Sciences* Vol. 103, No. 49, pp. 18656, ISSN 0027-8424.

Saito, K., Hayano-Saito, Y., Maruyama-Funatsuki, W., Sato, Y., and Kato, A. (2004). Physical mapping and putative candidate gene identification of a quantitative trait locus Ctb1 for cold tolerance at the booting stage of rice. *TAG Theoretical and Applied Genetics* Vol. 109, No. 3, pp. 515-522, ISSN 0040-5752.

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U. M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C. (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant Cell Online* Vol. 20, No. 1, pp. 11, ISSN 1040-4651.

Salvi, S., and Tuberosa, R. (2005). To clone or not to clone plant QTLs: present and future challenges. *Trends in Plant Science* Vol. 10, No. 6, pp. 297-304, ISSN 1360-1385.

Salvi, S., and Tuberosa, R. (2007). Cloning QTLs in plants. *Genomics-assisted crop improvement* Vol., No., pp. 207-225.

Sanchez, D. H., Pieckenstain, F. L., Szymanski, J., Erban, A., Bromke, M., Hannah, M. A., Kraemer, U., Kopka, J., and Udvardi, M. K. (2011). Comparative Functional Genomics of Salt Stress in Related Model and Cultivated Plants Identifies and Overcomes Limitations to Translational Genomics. *PLoS ONE* Vol. 6, No. 2, pp. e17094, ISSN 1932-6203.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E., Kato, T., Nakao, M., Sasamoto, S., Watanabe, A., Ono, A., and Kawashima, K. (2008). Genome structure of the legume, Lotus japonicus. *DNA research* Vol. 15, No. 4, pp. 227, ISSN 1340-2838.

Schauer, N., Semel, Y., Roessner, U., Gur, A., Balbo, I., Carrari, F., Pleban, T., Perez-Melis, A., Bruedigam, C., and Kopka, J. (2006). Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature biotechnology* Vol. 24, No. 4, pp. 447-454, ISSN 1087-0156.

Schneeberger, K., and Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. *Trends in Plant Science* Vol., No., ISSN 1360-1385.

Schranz, M. E., Lysak, M. A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. *Trends in Plant Science* Vol. 11, No. 11, pp. 535-542, ISSN 1360-1385.

Schulze, A., and Downward, J. (2001). Navigating gene expression using microarrays-a technology review. *Nature Cell Biology* Vol. 3, No. 8, pp. 190-195, ISSN 1465-7392.

Schulze, T. G., and McMahon, F. J. (2002). Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines. *American journal of medical genetics* Vol. 114, No. 1, pp. 1-11, ISSN 1096-8628.

Shinozuka, H., Cogan, N. O. I., Smith, K. F., Spangenberg, G. C., and Forster, J. W. (2010). Fine-scale comparative genetic and physical mapping supports map-based cloning strategies for the self-incompatibility loci of perennial ryegrass (Lolium perenne L.). *Plant molecular biology* Vol. 72, No. 3, pp. 343-355, ISSN 0167-4412.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., Jaiswal, P., Mockaitis, K., Liston, A., Mane, S. P., Burns, P., Davis, T. M., Slovin, J. P., Bassil, N., Hellens, R. P., Evans, C., Harkins, T., Kodira, C., Desany, B., Crasta, O. R., Jensen, R. V., Allan, A. C., Michael, T. P., Setubal, J. C., Celton, J.-M., Rees, D. J. G., Williams, K. P., Holt, S. H., Rojas, J. J. R., Chatterjee, M., Liu, B., Silva, H., Meisel, L., Adato, A., Filichkin, S. A., Troggio, M., Viola, R., Ashman, T.-L., Wang, H., Dharmawardhana, P., Elser, J., Raja, R., Priest, H. D., Bryant, D. W., Fox, S. E., Givan, S. A., Wilhelm, L. J., Naithani, S., Christoffels, A., Salama, D. Y., Carter, J., Girona, E. L., Zdepski, A., Wang, W., Kerstetter, R. A., Schwab, W., Korban, S. S., Davik, J., Monfort, A., Denoyes-Rothan, B., Arus, P., Mittler, R., Flinn, B., Aharoni, A., Bennetzen, J. L., Salzberg, S. L., Dickerman, A. W., Velasco, R., Borodovsky, M., Veilleux, R. E., and Folta, K. M. (2011). The genome of woodland strawberry (Fragaria vesca). *Nat Genet* Vol. 43, No. 2, pp. 109-116, ISSN 1061-4036.

Sindhu, A. S., Maier, T. R., Mitchum, M. G., Hussey, R. S., Davis, E. L., and Baum, T. J. (2009). Effective and specific in planta RNAi in cyst nematodes: expression interference of four parasitism genes reduces parasitic success. *Journal of experimental botany* Vol. 60, No. 1, pp. 315, ISSN 0022-0957.

Spollen, W., Tao, W., Valliyodan, B., Chen, K., Hejlek, L., Kim, J. J., LeNoble, M., Zhu, J., Bohnert, H., and Henderson, D. (2008). Spatial distribution of transcript changes in the maize primary root elongation zone at low water potential. *BMC plant biology* Vol. 8, No. 1, pp. 32, ISSN 1471-2229.

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C. T., Jia, Y., Wu, W., Richmond, T., Kitzman, J., and Rosenbaum, H. (2009). Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS genetics* Vol. 5, No. 11, pp. e1000734, ISSN 1553-7404.

Steemers, F. J., and Gunderson, K. L. (2007). Whole genome genotyping technologies on the BeadArray™ platform. *Biotechnology journal* Vol. 2, No. 1, pp. 41-49, ISSN 1860-7314.

Stich, B., and Melchinger, A. (2010). An introduction to association mapping in plants. *CAB Reviews* Vol..

Takahashi, W., Miura, Y., Sasaki, T., and Takamizo, T. (2010). Targeted mapping of rice ESTs to the LmPi1 locus for grey leaf spot resistance in Italian ryegrass. *European journal of plant pathology* Vol. 126, No. 3, pp. 333-342, ISSN 0929-1873.

Tang, J., Leunissen, J. A. M., Voorrips, R. E., Van Der Linden, C. G., and Vosman, B. (2008). HaploSNPer: a web-based allele and SNP detection tool. *BMC genetics* Vol. 9, No. 1, pp. 23, ISSN 1471-2156.

Tanksley, S. D., Ganal, M. W., and Martin, G. B. (1995). Chromosome landing: a paradigm for map-based gene cloning in plants with large genomes. *Trends in Genetics* Vol. 11, No. 2, pp. 63-68, ISSN 0168-9525.

Tian, F., Bradbury, P. J., Brown, P. J., Hung, H., Sun, Q., Flint-Garcia, S., Rocheford, T. R., McMullen, M. D., Holland, J. B., and Buckler, E. S. (2011). Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nature genetics* Vol., No., ISSN 1061-4036.

Trebbi, D., Maccaferri, M., de Heer, P., Sørensen, A., Giuliani, S., Salvi, S., Sanguineti, M. C., Massi, A., van der Vossen, E. A. G., and Tuberosa, R. (2011). High-throughput SNP discovery and genotyping in durum wheat (Triticum durum Desf.). *TAG Theoretical and Applied Genetics* Vol., No., pp. 1-15, ISSN 0040-5752.

Tuberosa, R., and Salvi, S. (2006). Genomics-based approaches to improve drought tolerance of crops. *Trends in Plant Science* Vol. 11, No. 8, pp. 405-412, ISSN 1360-1385.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., Zeng, K., Malek, J. A., Costa, G., and McKernan, K. (2008). A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res* Vol. 18, No. 7, pp. 1051, ISSN 1088-9051.

Van Orsouw, N. J., Hogers, R. C. J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., Van Der Poel, H., Van Oeveren, J., and Verstegen, H. (2007). Complexity reduction of polymorphic sequences (CRoPS™): a novel approach for large-scale polymorphism discovery in complex genomes. *PLoS ONE* Vol. 2, No. 11, pp. e1172, ISSN 1932-6203.

Varshney, R. K., Graner, A., and Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends in Plant Science* Vol. 10, No. 12, pp. 621-630, ISSN 1360-1385.

Varshney, R. K., Nayak, S. N., May, G. D., and Jackson, S. A. (2009). Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends in Biotechnology* Vol. 27, No. 9, pp. 522-530, ISSN 0167-7799.

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., Fontana, P., Bhatnagar, S. K., Troggio, M., and Pruss, D. (2010). The genome of the domesticated apple (Malus [times] domestica Borkh.). *Nature genetics* Vol. 42, No. 10, pp. 833-839, ISSN 1061-4036.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D. A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L. M., Vezzulli, S., and Reid, J. (2007). A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* Vol. 2, No. 12, pp. e1326, ISSN 1932-6203.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Francesco, V. D., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R.-R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z. Y., Wang, A., Wang, X., Wang, J., Wei, M.-H., Wides, R., Xiao, C., Yan, C., et al. (2001). The Sequence of the Human Genome. *Science* Vol. 291, No. 5507, pp. 1304-1351.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., and Kuiper, M. (1995). AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* Vol. 23, No. 21, pp. 4407, ISSN 0305-1048.

Wang, D., Amornsiripanitch, N., and Dong, X. (2006). A genomic approach to identify regulatory nodes in the transcriptional network of systemic acquired resistance in plants. *PLoS Pathogens* Vol. 2, No. 11, pp. e123, ISSN 1553-7374.

Wang, D. G., Fan, J. B., Siao, C. J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., and Spencer, J. (1998). Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* Vol. 280, No. 5366, pp. 1077, ISSN 0036-8075.

Wang, J., McClean, P. E., Lee, R., Goos, R. J., and Helms, T. (2008). Association mapping of iron deficiency chlorosis loci in soybean (Glycine max L. Merr.) advanced breeding lines. *TAG Theoretical and Applied Genetics* Vol. 116, No. 6, pp. 777-787, ISSN 0040-5752.

Wang, S., and Liu, Z. (2011). SNP Discovery through EST Data Mining. *In* "Next Generation Sequencing and Whole Genome Selection in Aquaculture", pp. 91-108. Wiley-Blackwell.

Weber, J. L., and May, P. E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *American Journal of Human Genetics* Vol. 44, No. 3, pp. 388.

Weiss, K. M., and Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* Vol. 18, No. 1, pp. 19-24, ISSN 0168-9525.

Welsh, J., and McClelland, M. (1990). Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res* Vol. 18, No. 24, pp. 7213, ISSN 0305-1048.

Wen, T. J., Hochholdinger, F., Sauer, M., Bruce, W., and Schnable, P. S. (2005). The roothairless1 gene of maize encodes a homolog of sec3, which is involved in polar exocytosis. *Plant physiology* Vol. 138, No. 3, pp. 1637, ISSN 0032-0889.

Wilhelm, B. T., and Landry, J. R. (2009). RNA-Seq--quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* Vol. 48, No. 3, pp. 249-257, ISSN 1046-2023.

Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A., and Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res* Vol. 18, No. 22, pp. 6531, ISSN 0305-1048.

Woll, K., Borsuk, L. A., Stransky, H., Nettleton, D., Schnable, P. S., and Hochholdinger, F. (2005). Isolation, characterization, and pericycle-specific transcriptome analyses of the novel maize lateral and seminal root initiation mutant rum1. *Plant physiology* Vol. 139, No. 3, pp. 1255, ISSN 0032-0889.

Wood, P. J. (2007). Cereal [beta]-glucans in diet and health. *Journal of Cereal Science* Vol. 46, No. 3, pp. 230-238, ISSN 0733-5210.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., and Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science* Vol. 308, No. 5726, pp. 1310, ISSN 0036-8075.

Wu, C., Trieu, A., Radhakrishnan, P., Kwok, S. F., Harris, S., Zhang, K., Wang, J., Wan, J., Zhai, H., and Takatsuto, S. (2008). Brassinosteroids regulate grain filling in rice. *The Plant Cell Online* Vol. 20, No. 8, pp. 2130, ISSN 1040-4651.

Xie, W., Feng, Q., Yu, H., Huang, X., Zhao, Q., Xing, Y., Yu, S., Han, B., and Zhang, Q. (2010). Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. *Proceedings of the National Academy of Sciences* Vol. 107, No. 23, pp. 10578.

Xinguo, L., Harry, W., and Simon, S. (2010). Comparative genomics reveals conservative evolution of the xylem transcriptome in vascular plants. *BMC Evolutionary Biology* Vol. 10, No., ISSN 1471-2148.

Xu, Y., and Crouch, J. H. (2008). Marker-assisted selection in plant breeding: from publications to practice. *Crop Science* Vol. 48, No. 2, pp. 391–407.

Yamaguchi-Shinozaki, K., and Shinozaki, K. (2005). Organization of cis-acting regulatory elements in osmotic-and cold-stress-responsive promoters. *Trends in Plant Science* Vol. 10, No. 2, pp. 88-94, ISSN 1360-1385.

Yang, B., Srivastava, S., Deyholos, M. K., and Kav, N. N. V. (2007). Transcriptional profiling of canola (Brassica napus L.) responses to the fungal pathogen Sclerotinia sclerotiorum. *Plant Science* Vol. 173, No. 2, pp. 156-171, ISSN 0168-9452.

Young, N. D., and Udvardi, M. (2009). Translating Medicago truncatula genomics to crop legumes. *Current opinion in plant biology* Vol. 12, No. 2, pp. 193-201, ISSN 1369-5266.

Yu, J., Holland, J. B., McMullen, M. D., and Buckler, E. S. (2008). Genetic design and statistical power of nested association mapping in maize. *Genetics* Vol. 178, No. 1, pp. 539, ISSN 0016-6731.

Yuan, J. S., Galbraith, D. W., Dai, S. Y., Griffin, P., and Stewart Jr, C. N. (2008). Plant systems biology comes of age. *Trends in Plant Science* Vol. 13, No. 4, pp. 165-171, ISSN 1360-1385.

Yuan, Y., SanMiguel, P. J., and Bennetzen, J. L. (2003). High Cot sequence analysis of the maize genome. *The Plant Journal* Vol. 34, No. 2, pp. 249-255, ISSN 1365-313X.

Zhang, H., Guan, H., Li, J., Zhu, J., Xie, C., Zhou, Y., Duan, X., Yang, T., Sun, Q., and Liu, Z. (2010). Genetic and comparative genomics mapping reveals that a powdery mildew resistance gene Ml3D232 originating from wild emmer co-segregates with an NBS-LRR analog in common wheat (Triticum aestivum L.). *TAG Theoretical and Applied Genetics* Vol., No., pp. 1-9, ISSN 0040-5752.

Zhang, J., Lu, Y., Yuan, Y., Zhang, X., Geng, J., Chen, Y., Cloutier, S., McVetty, P. B. E., and Li, G. (2009). Map-based cloning and characterization of a gene controlling hairiness and seed coat color traits in Brassica rapa. *Plant molecular biology* Vol. 69, No. 5, pp. 553-563, ISSN 0167-4412.

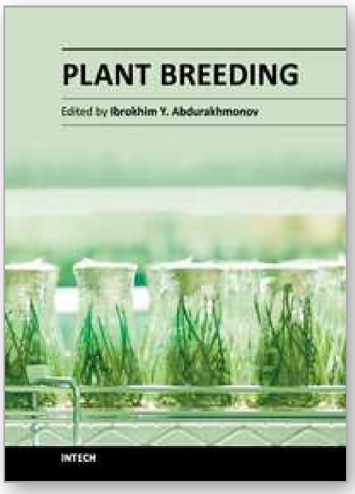
Zhang, Z., Guo, X., Liu, B., Tang, L., and Chen, F. (2011). Genetic diversity and genetic relationship of Jatropha curcas between China and Southeast Asian revealed by amplified fragment length polymorphisms. *African Journal of Biotechnology* Vol. 10, No. 15, pp. 2825-2832, ISSN 1684-5315.

Zheng, B., Yang, L., Zhang, W., Mao, C., Wu, Y., Yi, K., Liu, F., and Wu, P. (2003). Mapping QTLs and candidate genes for rice root traits under different water-supply conditions and comparative analysis across three populations. *TAG Theoretical and Applied Genetics* Vol. 107, No. 8, pp. 1505-1515, ISSN 0040-5752.

Zheng, P., Allen, W. B., Roesler, K., Williams, M. E., Zhang, S., Li, J., Glassman, K., Ranch, J., Nubel, D., and Solawetz, W. (2008). A phenylalanine in DGAT is a key determinant of oil content and composition in maize. *Nature genetics* Vol. 40, No. 3, pp. 367-372, ISSN 1061-4036.

Zhu, C. G., Buckler, M., and Yu, E. S. (2008). Status and prospects of association mapping in plants. *The Plant Genome* Vol. 1, No. 1, pp. 5, ISSN 1940-3372.

Zieler, H., Richardson, T., Schwartz, A., Herrgard, M., Lomelin, D., Mathur, E., Cheah, S., Tee, T., Lee, W., and Chua, K. (2010). Whole-Genome shotgun sequencing of the oil palm and Jatropha genomes. pp. 9-13.



Plant Breeding

Edited by Dr. Ibrokhim Abdurakhmonov

ISBN 978-953-307-932-5
Hard cover, 352 pages
Publisher InTech
Published online 11, January, 2012
Published in print edition January, 2012

Modern plant breeding is considered a discipline originating from the science of genetics. It is a complex subject, involving the use of many interdisciplinary modern sciences and technologies that became art, science and business. Revolutionary developments in plant genetics and genomics and coupling plant "omics" achievements with advances on computer science and informatics, as well as laboratory robotics further resulted in unprecedented developments in modern plant breeding, enriching the traditional breeding practices with precise, fast, efficient and cost-effective breeding tools and approaches. The objective of this Plant Breeding book is to present some of the recent advances of 21st century plant breeding, exemplifying novel views, approaches, research efforts, achievements, challenges and perspectives in breeding of some crop species. The book chapters have presented the latest advances and comprehensive information on selected topics that will enhance the reader's knowledge of contemporary plant breeding.