Journal of Biotechnology and Crop Science
4(4): 4-18, 2015

Genomics and gene based markers for crop improvement

A Selvi, GP Mishra, RK Singh, Rajesh Singh

Received: 23 February 2015; Revised Accepted: 28 April 2015

ABSTRACT

Since the development of first molecular markers in 1980, a diverse array of molecular marker technologies have come into being revolutionizing conventional plant breeding efforts for crop improvement. Significant progress has been made in crop improvement through these classical markers or conventional random molecular markers (RDMs). Still, the biological function of most of the markers is unknown. Besides throwing light on organization, conservation and evolution of plant genomes, these markers have also aided geneticists and plant breeders to map QTLs for the traits of economic importance and to identify genes. Further advancements in genomics with high throughput sequencing methods and bioinformatics aided in the characterization of these genes and to date sequences of several genes are available in databases. The markers derived from the genes or ESTs are commonly called as functional markers or genic molecular markers (GMM). The availability of technologies for precise manipulation of these genes and their deployment is helping plant breeders in a way as never before for evolving better crop varieties. The following write-up focuses on the advancement of genomic tools and approaches that are available, strategies for tagging genes, candidate genes for traits of interest, and their applications for improving crop plants.

Keywords: Functional Markers, Gene based markers, Random Markers

Introduction

In the past fifty years, plant breeders have exerted tremendous effort to create new cultivars of crop species that have complement of genes, for better performance that have often been sourced from different wild populations. Plant breeding has seen a major transition in the past decade as advances in plant sciences helped in evolving tools that can be applied to commonly accepted field techniques. In the history of plant sciences and biotechnology, the recent developments in the area of genomics and gene technology deserve emphasis. Impressive advances in molecular genetics over the last two decades have provided a range of tools and techniques for analyzing genomes e.g. isolation, characterization, and functional determination of several genes have become possible. Many plant genes were rather rapidly identified, and the number is increasing at enormous speed. The discoveries and concepts are being fully exploited for crop improvement by manipulating the genome of crop species with the emerging genomic tools. The major achievements in this area include the sequencing of the several plant genomes like *Arabidopsis*, rice, sorghum and poplar, the generation of expressed sequence tag (EST) databases, the development of microarray technologies, molecular markers, gene expression methodologies etc. Genomic approaches are beginning to impact the conventional breeding processes. With the advent of genomic tools like DNA marker technology, several types of DNA markers both random and functional are now available to plant

A Selvi (✉)
Division of Biotech., Sugarcane Breeding Coimbatore
e-mail: a.selvi@icar.gov.in

GP Mishra
Defence Institute of High Altitude Research, DRDO, C/o 56 APO, Leh, 901205

RK Singh
Visiting Scientist, Plant Genome Mapping Laboratory, 111 River Bend Road, University of Georgia, Athens, GA (USA)-30603

Rajesh Singh
Genetics and Plant Breeding, IAS, BHU, Varanasi-221005

breeders and geneticists, helping them to overcome many of the problems faced in conventional breeding. Advances in automated sequencing, methodologies for gene expression studies, development of computational abilities and algorithms have enabled the structural and functional characterization of several genes governing economically important traits and their further use in enhancing the breeding efficiency. It is now perceivable that this scientific development is heading towards sequence-based knowledge that would highly impact the reliability and precision with which plant breeding will progress.

Molecular markers as genomic tools

Of the several genomic tools that were developed in the past decade molecular markers deserve special mention. The DNA based markers were the first genomic tools that were developed and used for mapping several genes in a variety of crop species. Several molecular marker techniques were developed from the genome of the crop species as well as from random amplification of the genome. These techniques include the first generation restriction based markers like RFLP followed by the second generation amplification based markers like RAPD, AFLP, SSR, ISSR and the third generation sequence based markers like SNPs. The RFLP markers are relatively though highly polymorphic, co-dominantly inherited and highly reproducible have been used less frequently owing to the laborious procedure involved. RAPDs and ISSRs had been the marker of choice in the nineties and soon interest had shifted to more reliable and reproducible marker systems like AFLPs and SSRs. An increasing number of agronomically important genes have been correlated using DNA based molecular markers. Molecular markers offered numerous advantages over conventional phenotype based alternatives, as they are stable and detectable in all tissues regardless of growth, differentiation, development, or defense status of the cell.

They are not confounded by the environment, pleiotropic and epistatic effects. They have been used for several purposes including screening programs for selecting desirable clones, introgression breeding, gene pyramiding etc. Marker assisted breeding has helped in the indirect selection of difficult traits at the seedling stage and has helped in the speeding up of conventional breeding in several crop plants.

Random DNA markers (RDMs)

Most of the above mentioned markers are developed from genomic DNA, and therefore they may arise from both the transcribed regions and the non-transcribed regions of the genome. These DNA-based markers derived from any region of the genome have also been described as RDMs. These markers when used for indirect selection are completely independent on any functional knowledge about the underlying DNA sequences. Thus even for tightly linked markers, the effectiveness of marker aided selection is greatly diminished by the occasional uncoupling of the marker from the trait during many cycles of meiosis in the breeding program. Also the application of such random markers for selections across populations has been limited. Recently, interests have shifted towards the development of molecular markers from the genes that are responsible for the expression of phenotypic trait variations. These markers from transcribed region of the genome, target the functional polymorphism in the gene sequences and allows selection in different genetic backgrounds which is not always possible with random markers.

Genomics resources

Genome and Gene space sequencing

The emphasis laid on genomics in the recent past had led to several gene discovery projects by genome sequencing, analysis of transcriptome, gene expression studies etc. Genomic approaches have led to the study of many genes that occur throughout the genome in a simultaneous manner. This has resulted in the expanded databanks that are available for crop improvement. An increasingly large number of genes have been identified in both wet lab and by *in-silico*

analysis of available databases. Gene sequences have been stored in public databases both in the form of genomic sequences and EST sequences, or as BAC clones as well as full length cDNA clones and genes.

The whole genome sequencing projects of *Arabidopsis*, rice and poplar has made available complete genome sequences of these species. Gene space sequencing that targets sequencing of long gene-rich regions containing many genes that are separated by long gene-poor regions has been carried out for plant species like maize (http://www.maizegenome.org/), sorghum, wheat (http://www.wheatgenome.org/), tomato (http://sgn.cornell.edu/help/about/tomato_sequencing.html), tobacco (http://www.intl-pag.org/13/abstracts/PAG13_P027.html), poplar (http://genome.jgi-psf.org/Poptr1/), Medicago (http://www.medicago.org/genome/) and lotus (http://www.kazusa.or.jp/lotus/). Several genomics based databases are now available to identify gene sequences that would help in molecular marker development (Table 1).

EST programs

Apart from whole genome sequencing projects and gene space sequencing, strategies to study the transcriptome of several important crop species have lead to the establishment of expressed sequence tag (EST) sequencing projects for gene discovery. Generation of ESTs is a quick and simple strategy involving partial sequencing of 5' or 3' end of the cDNAs. A wealth of DNA sequence information has been generated from these projects and deposited in online databases like http://www.ncbi.nlm.nih.go (National Centre for Biotechnology Information), http://www.tigr.org (The Institute of Genome Research), http://www.ebi.ac.uk (European Bioinformatics Institute). The plant EST database at EMBL has exceeded over five million EST sequences. ESTs have been developed from more than 50 plant species and in each species more than 5000 ESTs have been made available. These species represent important crops and their sequences are a potential source from which several valuable molecular markers that are of interest to plant breeders can be generated. Several crop specific EST databases are available for crops like wheat (http://genome.arizona.edu, http://wheat.pw.usda.gov/genome, http://wheat.pw.usda.gov/NSF/),barley (http://hordeum.ipk-gatersleben.de/est/est.html), sugarcane (http://sucest.lad.ic.unicamp.br/en/) etc.

In species that lack EST resources, comparative mapping strategies have facilitated the isolation of genes from these species. GRAMENE (http://www.gramene.org) is a curated open source database that is available for comparative genome analysis for grasses. HarVEST (http://harvest.ucr.edu) is asoftware developed to enable searching for differentially expressed ESTs among cDNA libraries and is oriented towards comparative genomics. However since the EST sequences are generated through partial sequencing of the 3' or 5' ends of cDNAs there is redundancy in the genes sequences that are obtained from the databases. The EST data has to be clustered to identify unigenes from random EST sequences using bioinformatics tools. The NCBI UniGene Resources is an excellent system for automatically partitioning GenBank sequences into a non-redundant set of gene oriented clusters. UniGene Sets are available for wheat, barley and many other crop plants in the NCBI database. The TIGR Gene Indices includes ESTs clustered into Tentative Consensi (TC), with top 5 peptide hits, and alignment to rice BACs and *Arabidopsis* chromosomes. These unigenes are then utilized to develop molecular markers.

Bioinformatics

The advances made in bioinformatics have made it possible to acquire and organize large amounts of information and also allows the visualization of information from heterogeneous datasets. It facilitates both the analysis of genomic and post-genomic data, and the integration of data from the related fields of transcriptomics, proteomics, metabolomics and

phenomics. Improved algorithms and increased computing power has helped to analyze DNA sequences, marker discovery and analyzing the information generated. It provides tools to integrate phenotypic and genotypic data and to derive meaningful conclusions for important agronomic traits.

Development of bioinformatic tools, databases and integration of information from different fields enable the identification of genes and gene products, and can elucidate the functional relationships between genotype and observed phenotype. Bioinformatic tools that are commonly used for datamining, gene identification, analysis and molecular marker development include MISA helps to identify SSRs, AutoSNP used for SNP identification, SNP2CAPS-used foe developing CAPs from SNP markers, TASSEL which is a tool for microarray data analysis, datamining and data visualization etc. Databases like NCBI, EMBL, Swissprot, AceDB, Plantmarkers, HarvEST, PEDANT, tools for datamining, Bioinformatic.net etc also provides multiple bioinformatics tools.

Gene based markers

With the advent of high throughput sequencing and ever enlarging sequence databases gene based markers came into existence and more recently with the development of gene expression studies several functional markers are being identified and are proving to be efficient. In contrast to random markers that are developed from any part of the genome, gene based markers are derived from polymorphic sites within gene sequences. These are also called 'Genic' markers. The gene based markers are further classed as Gene Targeted Markers (GTMs) and Functional Markers (FMs) depending on the functional characterization of the polymorphisms that are generated by these markers (Anderson and Luebberstedt 2003).

Gene Targeted Markers (GTMs)

It generates polymorphisms from gene sequences which are obtained from EST databases or genome sequences or cDNA sequences and the functions of these genes may be already known or unknown. Some common examples of GTMs are cDNA-RFLP, EST

Table 1 Databases related to crop genomics.

Crops	Details	Websites
Barley Genomics	Databases on molecular markers, QTLs ESTs, maps, mutants barley BACs	http://barleygenomics.wsu.edu http://uscrop.net/perl/ace/search/BarleyDB
MaizeDB	Comprehensive information source on the genetics and molecular biology, analysis tool for sequence, expression and phenotype data, online ordering for	http://www.zmdb.iastate.edu http://www.agron.missouri.edu
MilletGenes	ESTs, seed and microarrays	http://www.zmdb.iastate.edu
SorghumDB	AceDB database with molecular markers, ESTs, QTL, maps	http://uscrop.net/perl/ace/search/MilletGenes
ZmDB	AceDB database with molecular markers, ESTs, QTL, maps	http://algodon.tamu.edu/sorghumdb.html
(RyeGI), Triticum aestivum	Gene Index	http://www.tigr.org
Wheat	compile and distribute SSR-containing ESTs SNP Development	http://wheat.pw.usda.gov/ITMI/EST-SSR/ http://wheat.pw.usda.gov/ITMI/WheatSNP

-simple sequence repeats (EST-SSRs) (Varshney *et al* 2005), EST-Single nucleotide polymorphisms (EST-SNPs) (Rafalski 2002) and conserved orthologous sets of markers (COSs).

cDNA-RFLP

The first genic markers that were developed were in the form of cDNA-RFLP (Graner et al. 1991, Causse et al. 1994). Any transcript derived polymorphic fragment can be cloned and used as a probe for developing gene based markers for the trait of interest. In the past several random markers have been developed from mRNA or cDNA from a number of plant species that are subjected to stress. The polymorphic fragments obtained from these experiments like differential display, cDNA-AFLP, or PCR products obtained by amplifying gene specific fragments from cDNA or genomic DNA can be cloned and used as probes. These probes have the potential of being used as heterologous probes across species and genus.

EST-SSRs

The availability of sequences from many genomes has led to the mining of these sources using computational approaches and has permitted rapid and economical marker development programs. ESTs are ideal candidates for mining SSRs not only because of their availability in large numbers but also due to the fact that they represent expressed genes. Recent studies have observed that the frequency of microsatellites was significantly higher in ESTs than in genomic DNA in several plant species investigated (Morgante *et al.* 2002; Toth *et al.* 2000).

The generation of SSRs from EST sequences allows the identification of polymorphic loci directly from sequence data, if the sequence information for the same gene is available from more than one genotype of the same species. Most of the efforts till date for finding SSRs in EST sequences use several bioinformatics tools that have been developed for mining SSRs from EST databases (Table-2). Some important ones are:

Sputnik which is a simple program written in C programming language that searches DNA sequence files in FASTA format for microsatellite repeats (Abajian, 1994).

Find Patterns is one of the programs available in the Genetics Computer Group (GCG), now Accelrys, package (www.accelrys.com). It looks through large data sets and identifies short nucleotide or amino acid patterns specified by the user.

Repeat Finder is a web-based program specifically developed for the identification of SSRs (http://www.genet.sickkids.on.ca/~ali/repeatfinder.html). This program was originally developed for identifying repeats in a single input sequence, however, later upgraded to handle batch files containing multiple sequences.

Simple Sequence Repeats Identification Tool

(SSRIT) is a simple program available through Gramene / Genome databases portal at Cornell Universityhttp://brie2.cshl.org:8082/gramene/searches/ssrtool). The program helps in the identification of “perfect” simple sequence repeats and can handle moderate-sized datasets. Recently websites have also been created for documentation, curation and transaction of EST-SSRs data.

Plant SSR database is a major source of information on plant EST-SSRs, which was established at Clemson University Genomics Institute (CUGI). The fact that the EST-SSRs are derived from transcripts, they have been found useful for assaying the functional diversity in natural populations and germplasm collections. These markers are highly transferable to related species, and are useful for comparative mapping and evolutionary studies. When the EST-SSRs are generated from genes responsible for a phenotypic trait they are more effectively used for marker assisted selections.

Table 2 Tools for database mining for SSRs.

Script or program	References
MIcroSAtellite (MISA)	http://pgrc.ipk-gatersleben.de/misa/; Thiel T et al 2003
SSRFinder	Gao et al 2003
BuildSSR	Rungis *et al* 2004
SSR Identification Tool (SSRIT)	Kantaty et al 2002
Tandem Repeat Finder (TRF)	Benson 1999
Tandem Repeat Occurrence Locator (TROLL)	Castelo 2002
CUGIssr	http://www.genome.clemson.edu/projects/ssr/
Sputnik C. Abajian;	http://abajian.net/sputnik/index.html
Modified Sputnik	Morgante *et al* 2002
Modified Sputnik II	http://wheat.pw.usda.gov/ITMI/EST-SSR/LaRota/
SSRSEARCH	

EST-SNPs

Recent developments in sequencing techniques had made the discovery of single nucleotide polymorphisms (SNPs) and insertions/deletions, which are the basis of most differences between alleles more effecient. SNPs are generated by two methods. Direct sequencing of DNA segments that are amplified by PCR from several individuals is used for identification of SNP polymorphisms (Gaut and Clegg 1993, Shattuck-Eidens et al 1990). PCR primers are designed from genes of interest, to amplify 400–700 bp segments of DNA from a diverse set of individuals that represent a population. The resulting sequences are aligned and polymorphisms are identified.

The second method involves *in-silico* methods to identify SNPs by aligning EST sequences derived from different genotypes and available in public databases. A large number of SNP mining tools to automate the process of SNP discovery are available (Table-3) and SNPs have been generated in a number of species. ESTs were generated by sequencing shoot apical meristem (SAM) cDNA from maize inbred lines. The computational tool PolyBayes was used to identify single-nucleotide polymorphisms in 454 EST sequences of maize (Brad Barbazuk et al 2007). A comparative study of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barley cultivars revealed that EST-SNPs are the best class of markers for characterizing and conserving the gene bank materials (Varshney *et al* 2007).

Conserved Orthologous Set of markers (COS)

The increasing information on genomics and functional genomics from model plants like *Arabidopsis* and the evolution of new tools in functional genomics provides opportunities to develop gene based markers in crops where information is unavailable. COS markers represent orthologous genes from known sequences of a given species that can be used for related species through comparative genomics. It helps in identifying a set of genes conserved throughout evolution in both sequence and copy number in members of related species.

The EST database of tomato was computationally compared with the *Arabidopsis* genomic sequence and a set of conserved genes were identified as COS markers for tomato. These COS markers, 1,025 in numbers, represented functional genes and have shown to be conserved over a wide range of dicotyledonous plants. These genes were annotated, and most of them were identified to have putative functions that are associated with basic metabolic processes, such as energy-generating processes and the biosynthesis and degradation of cellular building blocks. Similarly 1130 potential COS markers for Lettuce, 426 for Sunflower, 1860 for Tomato and 1413 for Corn were identified by

screening 2185 sequences from *Arabidopsis* (Alexander Kozik and Richard Michelmore 2002, http://cgpdb.ucdavis.edu/COS_Markers/COS_Markers.html). EST sequences of three drought tolerance related genes like chalcone synthase (CHS), dihydroflavanol-4-reductase (DHRF-1) and drought responsive element binding factor (DREB-1) from *Musa* were used to identify Cassava homologs that were screened against *Arabidopsis* genome database to identify COS markers (Castelblanco and Fregene 2006). These markers have proved useful in comparative mapping among divergent genomes, and are useful for taxonomic studies and in deducing phylogenetic relationships between different genera and species.

Table 3 Softwares used for *in-silico* mining of SNPs.

Software tools for detecting SNPs automatically	Reference
PolyPhred	Nickerson et al. (1997)
TRACE_DIFF	Bonfield et al. (1998)
PolyBayes	Marth et al. (1999)
AutoSNP	Barker et al. (2003)
SNP locator (SNPL)	In house VB program
PARSESNP	Taylor and Greene (2003)
SNiPpER	Kota et al. (2003)
Single Nucleotide Polymorphism Finder (SNPF)	http://jic-bioinfo.bbsrc.ac.uk/cereals/webstart/snpf/snpf.html
Quality SNP	Tang et al 2006
SEAN	Huntley et al 2006

SEAN: SNP prediction and display program utilizing EST sequence clusters

Resistance Gene Analogues (RGAs)

Another important class of GTMs is resistance gene analogues. Several disease resistant genes from diverse sources are available now and are being used as markers thereby increasing gene based selections of superior genotypes. Genes conferring resistance to major classes of plant pathogens, including bacteria, virus, fungi and nematodes have been isolated from different plant species. Numerous genes involved in pathogen recognition, signal transduction and defense have been isolated. Nearly 40 resistance genes have been cloned during the past ten years. Many of these cloned genes are related in sequence and encode a limited number of functional classes.

Based on common molecular features, the R-genes are classified into several classes. The Class 1 encodes cytoplasmic receptor like proteins that contain a leucine rich repeat (LRR) domain and a nucleotide binding site (NBS). The genes that fall in this category are the *Arabidopsis RPS2, RPM1*, tomato *Prf* genes conferring resistance to *Pseudomonas syringae*, the *12-C* gene conferring resistance to fungus *Fusarium oxysporum*, the tobacco *N* gene conferring resistance to Tobacco mosaic virus, the rust resistance gene *L6* of flax and the *Arabidopsis RPP5* gene conferring resistance to downy mildew. Another class of resistance genes include the *Pto* which confers resistance to bacterial pathogen *Pseudomonas syringae* pv. *Tomato*. The Pto does not possess any LRR or NBS region and is dependent on the NBS-LRR containing protein *Prf* for its function. The third category includes the tomato Cf-2 and Cf-9 genes conferring resistance to the fungus *Cladosporium fulvum*. Yet another class of resistant genes consists of trans-membrane receptor with an extracellular LRR domain and an intracellular serine threonine kinase like the rice Xa 21 gene.

Resistance Gene Homologue Polymorphism (RGHPs)

It is based on the availability of candidate resistant gene sequences. RGHPs target groups of resistance genes by PCR, using primers for conserved domains of resistance genes, such as the Leucine Rich Repeat (LRR) or the Nucleotide Binding Site (NBS), both involved in resistance mechanisms. These RGHPs are then used to identify linkage with known disease resistant loci for use in marker assisted selections as well as to clone the resistant genes. Many RHGPs have been located to chromosome regions containing major R genes as well as QTLs. The cosegregation of

Table 4 Resistance gene homologue polymorphisms (RGHPs) in plants and their co- segregations with major genes and/or QTLs involved in disease resistance.

Plant species	Type of RGA	Resistance locus	Disease or pathogen	References
Arabidopsis thaliana	NBS	major R-genes and gene cluster	Pseudomonas syringae Peronospora parasitica Turnip crinckle virus Albugo candida Erisyphe cichoracearum Caulimovirus Turnip crinckle virus Tobacco ring spot virus	Aarts *et al* 1998 Speulman *et al* 1998
soybean	NBS	major R-genes	Phytophthora sojae Microspora diffusa Bradyrhizobia japonicum Potyvirus	Kanazin *et al* 1996 Yu *et al* 1996
		QTL	Cyst nematode	Kanazin *et al* 1996
Common bean	NBS	R gene cluster and QTL	Colletotrichum lindemuthianum Xanthomonas	Geffroy *et al* 1998
	Kinase	Major R gene	Uromyces appendiculatus	Nodari *et al* 1993 Rivkin *et al* 1999
Sunflower	NBS	R gene cluster	Plasmopara halstedii	Gentebittel *et al* 1998
	Kinase	QTL	Sclerotinia sclerotiorum	
Lettuce	NBS	Major R genes	Bremia lactucae	Woo *et al* 1998
Potato	NBS	Major R genes	Globodera rostochiensis Phytophthora infestans Potato virus Y Potato virus A Potato virus A	Leister *et al* 1996 Hamalainen *et al* 1998
Pepper	NBS	QTL	Cucumber mosaic virus	Pflieger *et al* 1999
	Kinase	QTL	Potyvirus	
Rapeseed	NBS	QTL	Leptosphaeria maculans	Pilet *et al* 1999
	Kinase	QTL	Pyrenopeziza brassicae	
		Major R genes	Leptosphaeria maculans	
Poncirus	NBS/LRR	Major R genes	Citrus tristeza virusCitrus nematode	Pilet 1999
Sugarbeet	NBS/LRR	Major R gene	Cercospora rhizomania	Weiland and Koch 2004
Pepper	NBS/kinase	QTLs	Phytopthora capsici	Donnelly *et al* 2005
Maize	NBS/LRR	Major genes	Sugarcane mosaic virus	Quint *et al* 2002
Sugarcane	NBS/LRR	Major R genes and QTLs	Puccinia melanocephala Ustilago scitaminia SCMV	Rossi *et al* 2003 McIntyre *et al* 2005
Wheat	RGAs	Major gene	Puccinia striiformis	Chen *et al* 1998
Rice	NBS	Major R gene Gene cluster QTLs	Xanthomonas oryzae Pyricularia oryzae Magnaparthae grisea	Leister *et al* 1998
Barley	NBS Kinase/LRR	Major R Genes Leaf rust R gene cluster	Erysiphe graminis	Leister *et al* 1998 Chen *et al* 1998

RHGPs with major disease resistant genes and quantitative trait loci (QTL) have been reported in several crops species (Table-4) (P flieger *et al* 2001). The disease R-gene database (available on line from the National Center for Genome Resources web site: http://www.ncgr.org/research/rgenes) facilitates access to R-gene and R-gene-like sequence data collected from public sequence and protein databases. The second database contains information about genes for both pathogen recognition (resistance genes and homologs) and plant defense responses (defense genes) (Chittoor *et al.* 1999).

Functional markers (FM)

The development of functional markers requires functionally characterized genes, allele sequences from such genes, the identification of polymorphic, functional motifs that affect plant phenotype within these genes, and the validation of associations between DNA polymorphisms and trait variation. Functional markers are further classed as direct functional markers and indirect functional markers.

Direct Functional Markers

These markers are developed from gene sequences with known expression and hence it is of prime importance to establish proof of gene function affecting a particular phenotype. The most direct means of obtaining proof of sequence motif function is by comparing isogenic genotypes differing in single sequence motifs. At current, the most appropriate approach for generating isogenic lines in crops is by Targeting Induced Local Lesions In Genomes (TILLING). TILLING provides point mutant alleles that are usually induced by chemical mutagens like ethyl methane sulphonate and these mutations are used in functional genomics and gene characterization.

The ability to detect point-mutations in specific genes within a large population of mutagenized plants was first demonstrated by Claire McCallum from the Henik off and Comai laboratories at the Fred Hutchinson Cancer Center and the University of Washington in Seattle and coined the word TILLING. Since then TILLING has been developed in several species including *Arabidopsis*, rice, maize, sorghum, wheat, barley, tomato, soybean, rapeseed etc. TILLING helps in generating a series of missense mutations in the alleles of a target gene. The polymorphisms that are generated by these mutations are compared with phenotypic variation to provide direct functional markers. Thus TILLING represents a viable method by which spontaneous and induced mutants help in the direct identification of beneficial nucleotide and amino acid changes in genes with known functions that can further be used in the development of diagnostic functional markers for selection.

Indirect Functional Markers

Association studies have the potential to identify sequence motifs affecting trait expression. They provide indirect evidence for the function of a sequence motif. In association studies the polymorphisms that exist within the gene such as a few nucleotides differences or insertions/deletions (indels), are correlated to the phenotype of interest. With the advent sequencing methods it is possible to find the SNP variations that occur through out the gene and the linkage disequilibrium (LD) that exists between these SNPs. Based on the LD of the SNPs they can be grouped as haplotype SNPs and those haplotype SNPs that are involved in the functional variation of the gene can be identified. This strategy would considerably reduce the number of SNPs that has to be genotyped. The selected SNPs which are functional variants can be genotyped on a set of accessions to find associations with the phenotype of interest.

In one of the pioneering studies, nine sequence motifs in the *dwarf8* gene of maize were shown to be associated with variation for flowering time. A set of 92 inbred lines were genotyped and associations between the polymorphisms in the *dwarf8* gene and flowering time was established. These associations of the polymorphisms in the *dwarf8* gene aided in the selection of lines that exhibited early flowering by 7-

11 days (Thornsberry *et al* 2001). Similarly associations of three haplotypes of the *StVe1* locus of potato that confers resistance to *V. alboatrum* were used to genotype 30 potato cultivars and one haplotype showed significant associations with *V. alboatrum* resistance (Simko *et al* 2004). A recent study in grapes showed that allelic variation in the gene *VvmybA1* that is responsible for transcriptional regulation of anthocyanin biosynthesis, was associated with multiple classes of fruit color in the 200 accessions of cultivated grapevine that were studied. One SNP and three indels were found to be significantly associated with fruit skin color in the structured association analysis of pigmented accessions. All four polymorphisms were associated with genetic differences separating black or gray-skinned accessions from red and pink skinned accessions (This *et al* 2007). The use of functional motifs to correlate with phenotyopes requires comprehensive allele sequencing, a relatively low LD between haplotypes and a phenotypically well characterized population. Several studies have been carried out on LD decay and haplotype diversity (Dvornyk *et al* 2002, van der Voort *et al* 2004, Oleson *et al* 2004, Neale and Savolainen 2004). In crops with low LD and high resolution of intragenic polymorphisms, association studies have the potential to identify sequence motifs that are correlated with trait variation.

Application of genetic molecular markers

As we know, molecular markers have already shown their applications in a variety of ways in several plant species. Hence, now with the development of GMMS, it is possible to have a targeted approach for detection of nucleotide diversity in genes which control agronomic traits in plant populations. The GMMS implementation will prove quite useful and can be utilized in three main areas of plant breeding and genetics, which are outlined below:

Functional genetic diversity

Identification of diverse genotypes is the prerequisite for improvement of any trait in the crop plants. Furthermore, monitoring the genetic variability within gene pool of elite breeding material could make crop improvement more efficient by the direct accumulation of favored alleles. DNA markers are being increasingly utilized in cultivar development, quality control of seed production, measurement of genetic diversity for conservation and management, varietal identification and intellectual property protection (IPP). Recent studies have used molecular markers to help in identification of genetically diverse genotypes to use in crosses in cultivar improvement programme. These studies have more success than conventional selection programme in producing productive lines from plant introduction/exotic lines crosses with elite lines (Thompson *et al.* 1998 a,b). Molecular markers have proven useful for assessment of genetic variation in germplasm collections (Hausmann *et al.* 2004; Maccaferri *et al.*, 2006). Evaluation of germplasm with GMMs might enhance the role of genetic markers by assaying the variation in transcribed and known function gene, although there may be higher probability of bias owing to selection.

According to Ayers *et al.* (1997) the expansion and contraction of SSR repeats in genes of known function can be tested for association with phenotypic variation or, more desirably, biological function by using genic SSR markers for diversity studies. The SSRs may have role in gene expression or function as suggested by the presence of SSRs in transcripts of genes. However, it is yet to be determined whether any unusual phenotypic variation is associated with the length of SSRs in coding regions as was reported for several diseases in human (Cummings and Zoghbi 2000).

Similarly, the use of SNP markers for diversity studies may correlate the SNPs of coding *vs* non coding regions of the gene with trait variation. The variation associated with deleterious characters, however, is less likely to be represented in the germplasm collections of crop species than among natural populations because undesirable mutations are commonly culled from breeding populations (Cho *et al.* 2000). Several studies involving GMMs, especially genic SSRs, have

been found useful for estimation of genetic relationship (Gupta and Rustgi 2004, Varshney *et al.* 2005a) and opportunities to examine functional diversity in relation to adaptive variation (Russel *et al.* 2004) can be seen in several studies using GMMs. Very soon with the development of more GMMs in major crop species, genetic diversity studies will become more meaningful if functional genetic diversity were to be given more importance than the evaluation of anonymous diversity. But use of the neutral traditional molecular markers will remain useful in situations where: (a) GMMS would not be available, and (b) to address some specific objectives e.g. neutral grouping of germplasm.

Cross transferability among the species or genera

The genic markers provide high degree of transferability among distantly related species which is one of the most important features of these markers while among the RDMs only RFLPs shows transferability. Transferability of GMM markers to related species or genera has now been demonstrated in several studies. A study based on analysis of ˜ 1000 barely GMMS suggested a theoretical transferability of barley markers to wheat (95.2%), maize (69.3%), sorghum (65.9%), rye (38.1%) and even to dicot species (16.0%). In fact, *in-silico* analysis of GMMs of wheat, maize and sorghum with complete rice genome sequence data have provided a larger number of anchoring points among different cereal genomes as well as provided insight into cereal genome evolution (Salse *et al.* 2004).

Genic markers are now used to enrich the genetic maps of related crop species (Varshney *et al.* 2004, 2005, 2007). Furthermore, genic markers from the related plant species offers the possibility to develop anchor or conserved orthologous sets (COS) for genetic analysis and breeding in different species. Based on these information workers identified a large repository of such COS markers and developed a database called "Plant Markers" (Rudd *et al.* 2005).

Tagging and mapping of traits/QTLs

One of the most important applications of molecular markers in plant breeding is their use as diagnostic markers for the trait in the selection. However, if random markers (RDMs) are used there is a risk of losing the linkage through genetic recombination. Such type of situation is also happened in case of GMMs, when the polymorphism for the gene-targeted markers (GTMs) was discovered through one allele analysis without any further specifications of the polymorphic sequence motif are threatened by the same way (Rafalski and Tingey, 1993). While the functional markers (FMs) /DFMs or IFMs allow reliable application of markers in populations without prior mapping and the use of markers in mapped populations without risk of information loss owing to recombination, in comparison to random markers. GMM have been developed and mapped in several plant species. Since the development of FMs is expensive it can not be undertaken for all the traits and in all crop species. The "transcript" or "gene" maps are the genetic maps, developed after mapping/integration of genic markers. Rostocks *et al.* (2005) have developed a "gene map" using a comprehensive set of >200 gene-based markers developed from candidate genes for drought tolerance in barely. Later, a "transcript map" of barley after integrating more than 1000 gene-based markers (GTMs) has been also developed (Stein *et al.* 2007). Such molecular maps can not only be compared with those of other related plant species in an efficient manner but also provide gene based molecular markers associated with trait of interest after the QTL analysis.

Conclusion

New genomic technologies are expensive to develop, and returns from the initial research can take time. However, once the knowledge reaches a critical level gains accelerate enormously. All the information and accumulation of knowledge gained since 1990 will allow breeders, for future decades of plant breeding, to incorporate and stack useful genes into several crop species. Products of breeding supplemented with MAS

using gene based markers are just now beginning to become available and work is continuing to maximize the utility of the sequence databases to integrate desirable genes in crop plants.

References

Aarts M, Hekkert B, Holub E, Beynon J, Stiekema W, Pereira A (1998) Identification of R-gene homologous DNA fragments genetically linked to disease resistance loci in *Arabidopsis thaliana*. Mol. Plant-Microbe Interact. 11: 251–258.

Abajian C (1994) Sputnik. (http://espressosoftware.com/pages/sputnik.jsp).

Andersen JR, Lubberstedt T (2003) Functional markers in plants. Trends in Plant Sci., 8: 554–560.

Ayers NM, McClung AM, Larkin PD, Bligh HFJ, Jones CA, Park WD (1997) Microsatellite and single nucleotide polymorphism differentiate apparent amylase classes in an extended pedigree of US rice germplasm. Theor. Appl. Genet., 94: 773-781.

Barbazuk WB, Emrich S, Schnable PS (2007) SNP Mining from Maize 454 EST Sequences Cold Spring Harb. Protoc., doi:10.1101/pdb.prot4786.

Barker G, Batley JO, Sullivan H, Edwards KJ, Edwards D (2003) Redundancy based detection of sequence polymorphism in expressed sequence tag data using auto SNP. Bioinformatics, 19: 421-422.

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res., 27: 573–580.

Bonfield JK, Rada C, Staden R (1998) Automated detection of point mutations using florescent sequence trace subtraction. Nucl. Acids Res., 26: 3404-3409.

Castelblanco W, Fregene M (2006) SSCP-SNP based conserved ortholog set (COS) markers for comparative genomics in Cassava (*Manihot esculenta* Crantz). Plant Molecular Biology Reporter, 24: 229-236.

Castelo AT, Wellington M, Gao GR (2002) Troll-Tandem Repeat Occurrence Locator. Bioinformatics, 18: 634-636.

Causse MA, Fulton TM, Cho YG, Ahn SN, Chunwongse J, Wu K, Xiao J, Yu Z, Ronald PC, Harrington SE, Second G, McCouch SR, Tanksley SD (1994) Saturated molecular map of the rice genome based on an inter-specific backcross population. Genetics, 138:1251-1274.

Chen X, Line RF, Leung H (1998) Resistance gene analogs associated with a barley locus for resistance to stripe rust. In: Heller SR (ed.), International Conference on the Status of Plant and Animal Genome Research VI. Abstracts. San Diego, CA.

Chittoor J, Kukreja K, Leung H, Nelson R, Hulbert S, Leach J (1999) A database of candidate genes for utilization QTL analyses of disease resistance in plants. In: Heller S.R. (ed.), International Conference on the Status of Plant and Animal Genome Research VII. Abstracts. San Diego, CA.

Cho YG, Ishii T, Temnykh S, Chen X, Lipovich L, McCouch SR, Parl WD, Ayers N, Cartinhour S (2000) Diversity of microsatellites derived from genomic liberaries and GenBank sequences in rice (*Oryza sativa* L.). Theor. Appl. Genet., 100: 713-722.

Cummings CJ, Zoghbi HY (2000) Fourteen and counting: unraveling trinucleotide repeat diseases. Hum. Mol. Genet., 9: 909-916.

Donnelly LM, Gomes VM, Ogundiwin EA, Glosier BR, Sidhu GS, Prince JP (2005) Pepper (*Capsicum sp.*) and *Phytophthora capsici*: Molecular genetic analysis of a host/pathogen system. Plant & Animal Genomes XIII Conference.

Dvornyk V, Sirvio A, Mikkonen M, Savolainen O (2002) Low nucleotide diversity at the pal1 locus in the widely distributed Pinus sylvestris. Mol. Biol. Evol., 19: 179–188.

Gao L, Tang J, Li H, Jia J (2003) Analysis of microsatellites in major crops assessed by computational and experimental approaches. Mol. Breed., 12: 245-261.

Gaut BS, Clegg MT (1993) Nucleotide polymorphism in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). Genetics, 135: 1091-1097.

Geffroy V, Sicard D, de Oliveira J, Sévignac M, Cohen S, Gepts P, Neema C, Langin T, Dron M (1999) Identification of an ancestral gene cluster involved in the coevolution process between *Phaseolus vulgaris* and its fungal pathogen *Colletotrichum lindemuthianum*. Mol. Plant-Microbe Interact., 12: 774–784.

Gentzbittel L, Mouzeyar S, Badaoui S, Mestries E, Vear F, Tourvieille de Labrouhe D, Nicolas P (1998) Cloning of markers for disease resistance in sunflower, *Helianthus annuus* L. Theor. Appl. Genet., 96: 519–525.

Graner A, Jahoor A, Schondelmaier J, Siedler H, Pillen K, Fischbeck G, Wenzel G, Herrmann RG (1991) Construction of an RFLP map of barley. Theor. Appl. Genet., 83: 250-256.

Gupta PK, Rustgi S (2004) Molecular markers from the transcribed/expressed region of the genome in higher plants. Funct. Integr. Genomics, 4: 139-162.

Hämäläinen JH, Sorri VA, Watanabe KN, Gebhardt C, Valkonen JPT (1998) Molecular examination of a chromosome region that controls resistance to potato Y and A potyviruses in potato. Theor. Appl. Genet., 96: 1036–1043.

Hausmann BI, Hess DE, Omanya GO, Folkertsma RT, Reddy BV, Kayento M, Welz HG, Geiger HH (2004) Genomic regions influencing resistance to the parasitic weed *Striga hermonthica* in two recombinant inbred populations of sorghum. Theor. Appl. Genet., 109: 1005-1016.

Huntley D, Baldo A, Johri S, Sergot M (2006) SEAN: SNP prediction and display program utilizing EST sequence clusters. Bioinformatics, 22: 495-496.

Kanazin V, Marek L, Shoemaker RC (1996) Resistance gene analogs are conserved and clustered in soybean. Proc. Natl. Acad. Sci. USA, 93: 11746–11750.

Kantety RV, La Rota M, Matthews DE, Sorrells ME (2002) Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol. Biol., 48: 501-510.

Kota R, Rudd S, Facius A, Kolesov G, Thiel T, Zhang H, Stein N, Mayer K, Graner A (2003) Snipping polymorphism from large EST collections in barley (*Hordeum vulgare* L.). Mol. Genet. Genomics, 270: 24-33.

Kozik A, Michelmore R (2002) Compositae Genome Project Database. http://cgpdb.ucdavis.edu/COS_Arabidopsis/

Leister D, Ballvora A, Salamini F, Gebhardt C (1996) A PCR-based approach for isolating pathogen resistance genes from potato with potential for wide application in plants. Nature Genet., 14: 421–429.

Leister D, Kurth J, Laurie DA, Yano M, Sasaki T, Devos K, Graner A, Schulze-Lefert P (1998) Rapid reorganization of resistance gene homologues in cereal genomes. Proc. Natl. Acad. Sci. USA, 95: 370–375.

Maccaferri M, Sanguineti MC, Natoli E, Arous-Ortega JL, Ben Salem M, Bort J, Chenenaoui S, Deambrogio E, Garcia DML, De Montis A (2006) A panel of elite accessions of durum wheat (*Triticum durum* Defs) suitable for association mapping studies. Plant Genet Res., 4: 79-85.

Marth GT, Korf I, Yandell MD, Yeh RT, Zhijie G, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR (1999) A general approach to single-nucleotide polymorphism discovery. Nat. Genet., 23: 452-456.

McIntyre CL, Casu RE, Drenth J, Knight D, Whan VA, Croft BJ, Jordan DR, Manners JM (2005) Resistance gene analogues in sugarcane and sorghum and their association

with quantitative trait loci for rust resistance. Genome, 48: 391-400.

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat. Genet., 30: 194–195.

Neale DB, Savolainen O (2004) Association genetics ofcomplex traits in conifers. Trends Plant Sci., 9: 325–330.

Nickerson DA, Tobe VO, Taylor SL (1997) Polyphred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucl. Acids Res., 25: 2745-2751.

Nodari RO, Tsai SM, Guzman P, Gilbertson RL, Gepts P (1993) Toward an integrated linkage map of common bean. III. Mapping genetic factors controlling host-bacteria interactions. Genetics, 134: 341–350.

Olsen KM, Halldorsdottir SS, Stinchcombe JR, Weinig C, Schmitt J, Purugganan MD (2004) Linkage disequilibrium mapping of *Arabidopsis* CRY2 flowering time alleles. Genetics, 167: 1361–1369.

Pflieger S, Lefebvre V, Causse M (2001) The candidate gene approach in plant genetics: a review**.** Molecular Breeding, 7: 275–291.

Pflieger S, Lefebvre V, Caranta C, Blattes A, Goffinet B, Palloix A (1999) Disease resistance gene analogs as candidates for QTLs involved in pepper/pathogen interactions. Genome 42: 1100–1110.

Pilet ML (1999) Analyse génétique de la résistance du colza à la nécrose du collet et à la cylindrosporiose à l'aide des marqueurs moléculaires. Ph-D Thesis of Institut National Agronomique Paris-Grignon, France, 182 pp.

Quint M, Mihaljevic R, Dussle C, Xu M, Melchinger A, Lübberstedt T (2002) Development of RGA-CAPS markers and genetic mapping of candidate genes for sugarcane mosaic virus resistance in maize. Theor. Appl. Genet., 105: 355-363.

Rafalski JA, Tingey SV (1993) Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. Trends Genet., 9: 275-280.

Rivkin MI, Vallejos CE, McClean PE (1999) Disease resistance related sequences in common bean. Genome, 42: 41–47.

Rossi M, Araujo PG, Florence P, Grasmeur O, Dias VM, Hui C, Sluys MAV, D'Hont A (2003) Genome distribution and characterization of EST derived sugarcane resistance gene analogs. Mol. Genet. Genomics, 269: 406-419.

Rudd S, Schoof H, Klaus M (2005) PlantMarkers- A database of predicted molecular markers from plants. Nucleic Acids Res., 33: D628-D632.

Rungis D, Berube Y, Zhang J, Ralph S, Ritland CE, Ellis BE, Douglas C, Bohlmann J, Ritland K (2004) Robust simple sequence repeat markers for spruce *(Picea* spp.) from expressed sequence tags. Theor. Appl. Genet., 109: 1283-1294.

Russel J, Booth A, Fuller J, Harrower B, Hedley P (2004) A comparison sequence based polymorphism and haplotype content in transcrbed and anonymous regions of the barley genome. Genome, 47: 389-398.

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

Shattuck-Eidens DM, Bell RN, Neuhausen SL, Helentjaris T (1990) DNA sequence variation within maize and melon: observations from polymerase chain reaction amplification and direct sequencing. Genetics, 126: 207-217.

Simko I, Haynes KG, Ewing EE, Costanzo S, Christ BJ, Jones RW (2004) Mapping genes for resistance to Verticillium albo-atrum in tetraploid and diploid potato populations using haplotype association tests and genetic analysis. Mol. Genet. Genomics, 271: 522–531.

Speulman E, Bouchez D, Holub EB, Beynon JL (1998) Disease resistance gene homologs correlate with disease resistance loci of *Arabidopsis thaliana*. Plant J., 14: 467–474.

Tang J, Gao L, Cao Y, Jia J (2006) Homologous analysis of SSR-ESTs and transferability of wheat SSR-EST markers across barley, rice and maize. Euphytica, 151: 87-93.

Taylor NE, Greene EA (2003) PARSESNP: a tool for the analysis of nucleotide polymorphisms. Nucl. Acids Res., 31: 3808-3811.

Thiel T, Michalek W, Varshney RK, Graner A (2003) Exploiting EST database for the development and characterization of gene derived SSR-markers in barley (*Hordeum vulgare* L.). Theor. Appl. Genet., 106: 411-422.

This P, Lacombe T, Cadle Davidson M, Owens CL (2007) Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene Vvmyb A 1. Theor. Appl. Genet., 114: 1432-2242.

Thompson JA, Nelson RL (1998b) Utilization of diverse germplasm for soybean yield improvement. Crop Sci., 38: 1362-1368.

Thompson JA, Nelson RL, Vodkin LD, (1998a) Identification of diverse soybean germplasm using RAPD markers. Crop Sci., 38: 1348-1355.

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES IV (2001) Dwarf8 polymorphism associate with variation in flowering time. Nat Genet., 28: 286–289.

Toth G, Gaspari Z, Zurka J (2000) Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res., 10: 967-981.

Van Der Voort JR, Sorensen A, Lensink D, Van der Meulen M, Michelmore R, Peleman J (2004) Decay of linkage disequilibrium in the dm3 resistance-gene cluster of lettuce. In: Plant & Animal Genomes XII Conference, 10–14 January, Town & Country Convention Center, San Diego, CA, P748. Pp

Varshney RK, Korzun V, Borner A (2004) Molecular maps in cereals: Methodology and preogress. In: Gupta PK, Varshney RK (eds.) Cereal genomics. Kluwer Academic Publishers, The Netherlands, pp 35-60.

Varshney RK, Mahendar T, Aggarwal RK, Borner A (2007) Genic molecular markers in plants: development and applications Varshney R.K and R. Tuberosa (eds.), Genomics-Assisted Crop Improvement: Vol. 1: Genomics Approaches and Platforms, 13-29.

Varshney RK, Chabane K, Hendre PS, Aggarwal RK, Graner A (2007) Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Science 173: 638-649.

Varshney RK, Graner A, Sorrells ME (2005) Genomic-assisted breeding for crop improvement. Trends Plant Sci., 10: 621-630.

Weiland J, Koch G (2004) Sugarbeet leaf spot disease (*Cercospora beticola* Sacc.) Mol. Plant Path., 5: 157-166.

Woo SS, Sicard D, Arroyo-Garcia R, Ochoa O, Nevo E, Korol A, Fahima T, Michelmore RW (1998) Many diverged resistance genes of ancient origin exist in lettuce. In: Heller S.R. (ed.), International Conference on the Status of Plant and Animal Genome Research VI. Abstracts. San Diego, CA.

Yu YG, Buss GR, Saghai Maroof MA (1996) Isolation of a superfamily of candidate disease-resistance genes in soybean based on a conserved nucleotide-binding site. Proc. Natl. Acad. Sci. USA, 93: 11751–11756.