Chapter 5
Gene-Based Marker Systems in Plants: High Throughput Approaches for Marker Discovery and Genotyping

Rajeev K Varshney

Abstract Development and application of molecular markers derived from genes, commonly called genic markers or sometimes functional markers, is gaining momentum in plant genetics and breeding. Availability of large amount of sequence data coming from genome/transcriptome sequencing projects as well as advent of next generation sequencing technologies together with advances in bioinformatics tools, marker discovery is becoming cheaper and faster. The availability of inexpensive high-density SNP-genotyping arrays is encouraging the plant genetics and breeding community to undertake genome-wide marker genotyping for a variety of applications. For instance, high-throughput and low cost genotyping assays for gene-based markers offers the possibility to accelerate the trait mapping based on high-density linkage mapping as well as genome-scanning based association mapping approaches in addition to facilitate physical mapping, comparative mapping, phylogenetic studies and understanding genome organization in crop plant species. Marker discovery, genotyping and molecular breeding practices would be routine in near future for crop improvement in many crop species. Advances in the area of marker discovery and genotyping using highly parallel genomics assays and also a few applications have been discussed in this chapter.

5.1 Introduction

Molecular markers are important genetic tools for plant breeders to detect the genetic variation available in the germplasm collection. During last two decades, varieties of molecular markers and in large numbers have been developed for almost all major crop species. Genetic variation detected by molecular markers has

R.K. Varshney
International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru (AP), India; and
Generation Challenge Programme (GCP) CIMMYT, Mexico D.F, Mexico
e-mail: r.k.varshney@cgiar.org

S.M. Jain and D.S. Brar (eds.), *Molecular Techniques in Crop Improvement*,
DOI 10.1007/978-90-481-2967-6_5,

been useful for understanding the genome dynamics as well improving the breeding efficiency. For instance, these markers have been utilized extensively for the preparation of saturated molecular maps (genetic and physical) and their association with genes/QTLs controlling the traits of economic importance has been utilized in several cases for marker assisted selection (MAS) (Varshney et al. 2005b, 2006). As a result of extensive efforts undertaken at international level to identify molecular markers tightly linked with a large number of agronomic traits as well as tolerance/resistance to abiotic and biotic stresses in major crop species, it has been possible to realize the potential of molecular markers to track loci and genome regions in several crop-breeding programmes (Gupta and Varshney 2004; Varshney et al. 2006, 2007b). Other important uses of molecular markers include germplasm characterization, genetic diagnostics, genome organization studies and phylogenetic analysis (see Jain et al. 2002; Varshney and Tuberosa 2007).

Classically, the molecular markers can be grouped in three main categories (Gupta et al. 2002): (1) hybridization-based markers: restriction fragment length polymorphism (RFLP), (2) PCR-based markers: random amplification of polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP) and microsatellite or simple sequence repeat (SSR), and (3) sequence or chip-based markers: single nucleotide polymorphism (SNP), diversity array technology (DArTs) and single feature polymorphism (SFP). Indeed, till recent past, molecular markers from the first two categories have been developed in majority of the crop species, which belonged to genomic DNA, and therefore could belong to either the transcribed or the non-transcribed part of the genome without any information available on their functions. Nevertheless, during last few years, it has been possible to develop the markers from transcribed region of the genome or genes. Such markers have been popularly referred as functional markers/FMs (Anderson and Lübberstedt 2003; Gupta and Rustgi 2004), genic molecular markers/ GMMs (Varshney et al. 2007c) and gene expression markers/ GEMs (West et al. 2006b). Although development of gene-based markers is currently restricted to only limited crop species, the next generation sequencing technologies available very recently are enabling development of gene-based markers even in "orphan" crop species that are deficient in genomic resources (see Varshney et al. 2009). The present chapter deals with the advances made recently in the area of development of gene-based markers and methods of genotyping in crop species.

5.2 Gene-Based Marker System: Moving from Genes to Genome

As a result of establishment of several large scale genome/transcriptome sequencing and gene discovery projects in several plant species, a large number of genes have been identified through *wet lab* as well as *in silico* studies and a wealth of sequence data have been accumulated in public databases (e.g. http://www.ncbi.nlm.nih.gov; http://www.ebi.ac.uk) in the form of BACs (bacterial artificial chromosomes), ESTs (expressed sequence tags), GSSs (genome survey sequences), full length cDNA clones and genes. Furthermore, excellent progress has been made in

the area of development of bioinformatics tools and databases (see the companion chapter in this volume by Jayashree and Hoisington). Because of these two main factors, the era of development of gene-based markers has taken off in plant systems and is in advanced stage at present (Gupta and Rustgi 2004; Varshney et al. 2005a, 2007c, 2007d).

Based on the origin (polymorphic or non-polymorphic site in the gene), genic markers have been classified into two groups (Anderson and Lübberstedt 2003): (a) gene targeted markers (GTMs), that are derived from polymorphisms within genes, however not necessarily involved in phenotypic trait variation, e.g. EST-based molecular markers (Schmitt et al. 2006); (b) functional markers (FMs) are derived from polymorphic sites within genes causally involved in phenotypic trait variation, e.g. candidate gene-based molecular markers. Depending on the involvement in the phenotypic trait variation, the FMs can be classified further into two subgroups: (a) direct functional markers (DFMs), for which the role for the phenotypic trait variation is well proven, and (b) indirect functional markers (IFMs), for which the role for phenotypic trait variation is indirectly known (Anderson and Lübberstedt 2003).

5.3 Marker Discovery

Several types of molecular markers can be developed from genes and therefore can be grouped under genic marker category. Since several review articles (e.g. Gupta and Rustgi 2004; Varshney et al. 2005a, 2007c) have discussed at length the development of cDNA or EST-based RFLP, SSR and SNP markers, this article would mainly discuss the development of genic markers from sequence and expression data.

5.3.1 Sanger Sequencing-Based Marker Development

As a result of sequencing the genomes and transcriptomes followed by annotation provided the entire/partial gene repository of several model plant species (e.g. *Arabidopsis*, *Medicago*, *Lotus*, *Poplar*) and major crop species (rice, wheat, maize, sorghum, soybean, grape). In such plant species, based on functional annotation, gene expression and physiological studies, candidate genes can be identified for marker development. However in several crop species e.g. rye, sweet potato, pigeonpea and millets that have less genomic resources, identification of candidate genes is not easily feasible. However, because of the availability of next generation sequencing technologies, it has been shown very recently that the genic markers can be developed in so-called orphan crop species also (Varshney et al. 2009).

5.3.1.1 Candidate Gene-Based Marker Discovery

Based on positional cloning and/or knowledge of plant metabolic cycles, information on candidate gene(s) involved in the expression of a particular phenotype, is

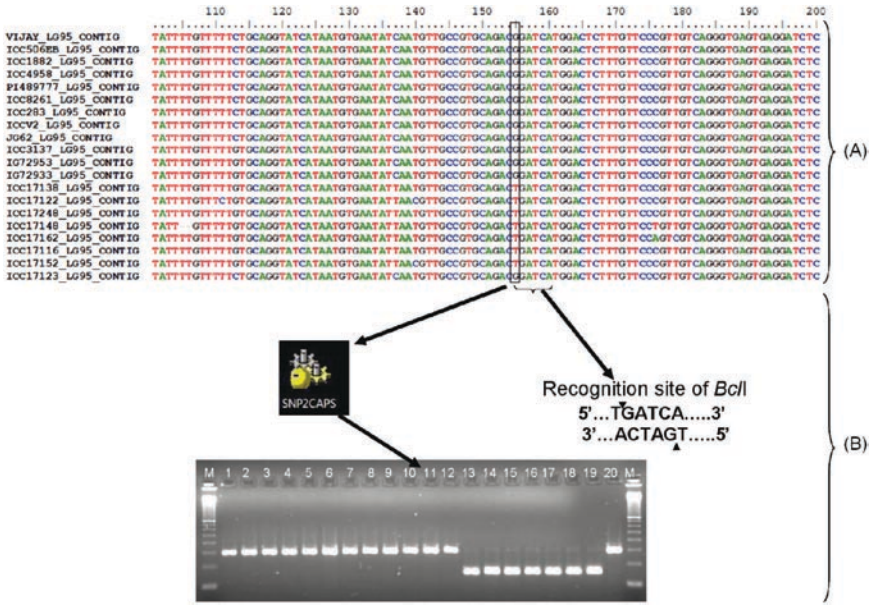

Fig. 5.1 An example of SNP discovery and conversion of SNP into CAPS assay. Multiple sequence alignment for sequence data generated for 20 genotypes using LG95 marker has been shown in panel A. This alignment reveals 7 SNPs at positions 111 (T/C), 113 (C/G), 137 (C/T), 140 (T/C), 155 (G/T), 176 (C/T/A) and 179 (T/C) in a window of 100 bp. SNP2CAPS analysis of these sequence data using SNP Pipeline developed at ICRISAT predicted the recognition site for restriction enzyme *Bcl*I at position 155. To verify the in silico identified restriction site involving SNP at 155 position, the PCR products of 20 genotypes generated by using LG95 marker were digested with restriction enzyme *Bcl*I. As a result, the PCR products, containing T nucleotide at 155 position in 7 genotypes were digested, while the amplicons, containing G nucleotide at the same position in remaining 13 genotypes remained intact (panel B). Thus the SNP at 155 position could be easily genotyped on agarose gel using CAPS assay (unpublished results).

available in many cases in model or major crop species. The primer pairs designed for the candidate gene(s) can be used to generate the amplicon in several genotypes of a species. Sanger dideoxy-nucleotide sequencing of these amplicons provides the allele specific sequencing data which can be subjected for multiple sequence alignment (MSA) by using bioinformatic tools and analyzed for identification of SNPs in the candidate gene (Fig. 5.1).

5.3.1.2 EST-Based Marker Discovery

For majority of main crop species, a large number of ESTs generated from several genotypes of a species are available in the sequence databases. The redundant set of ESTs for a given species can be used for *in silico* SNP discovery by employing bioinformatic scripts and tools for SNP discovery. A large number of bioinformatics tools or pipelines are available in public domain for identification of SNPs

Table 5.1 Some important SNP related tools and databases

Tool name	Web address	Description	Reference
SNP discovery tools			
SNP-VISTA	http://genome.lbl.gov/vista/snpvista/	A programme for visualization of mutations/SNPs in genes and discovery of recombination points	Shah et al. (2005)
SNPselector	http://primer.duhs.duke.edu/	A web tool for selecting SNPs for genetic association studies and it works on data available from Ensembl	Xu et al. (2005)
Forage	http://biobase.biotech.kth.se/forage/	A software application for SNP discovery which is based on two neural networks that have been trained on a set of validated SNPs	Unneberg et al. (2005)
SNPHunter	http://www.hsph.harvard.edu/ppg/software.htm	A software programme that allows for both ad hoc-mode and batch-mode SNP search, automatic SNP filtering, and retrieval of SNP data, including physical position, function class, flanking sequences at user-defined lengths, and heterozygosity from NCBI dbSNP	Wang et al. (2005)
SNPServer	http://hornbill.cspp.latrobe.edu.au/snpdiscovery.html	A real-time flexible tool for the discovery of SNPs and insertions/deletions	Savage et al. (2005)
SEAN	http//zebrafish.doc.ic.ac.uk/Sean	An application that predicts SNPs using multiple sequence alignments produced from EST clusters; the algorithm uses rules of sequence identity and SNP abundance to determine the quality of the prediction	Huntley et al. (2006)
QualitySNP	http://www.bioinformatics.nl/tools/snpweb/	A pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species	Tang et al. (2006)
MAVIANT (Multipurpose Alignment VIewing and Annotation Tool)	http://snp.agrsci.dk/maviant/	A platform-independent tool that provides DNA chromatogram and alignment views and facilitates evaluation of predictions	Panitz et al. (2007)

(continued)

Table 5.1 (continued)

Tool name	Web address	Description	Reference
MSQT (Multiple SNP Query Tool)	http://msqt.weigelworld.org/	MSQT extracts SNP information from multiple sequence alignments, stores it in a database, provides a web interface to query the database and outputs SNP information in a format directly applicable for SNP-assay design	Warthmann et al. (2007)
QuickSNP	http://bioinformoodics.jhmi.edu/quickSNP.pl	An automated web server for selection of tagSNPs	Grover et al. (2007)
WASP	http://bioinfo.biotec.or.th/WASP	A tool for designing allele specific primers for both SNPs and mutations. By integrating the database for known SNPs, this tool facilitates the awkward process of getting flanking sequences and other related information from public SNP databases	Wangkumhang et al. (2007)
Haplo*SNP*er	http://www.bioinformatics.nl/tools/haplosnper/	A flexible web-based tool for detecting SNPs and alleles in user-specified input sequences from both diploid and polyploid species	Tang et al. (2008)
SNP-PHAGE	http://bfgl.anri.barc.usda.gov/ML/snp-phage/	A SNP discovery pipeline with additional features for identification of common haplotypes within a sequence tagged site (Haplotype Analysis) and GenBank (-dbSNP) submissions	Matukumalli et al. (2006)
Tools to convert SNPs to CAPS or dCAPS			
SNP2CAPS	http://pgrc.ipk-gatersleben.de/snp2caps/	A simple algorithm that involves the screening of multiply-aligned sequences for restriction sites followed by a selection pipeline that allows the deduction of CAPS candidates by the identification of putative alternative restriction sites	Thiel et al. (2004)

SNP Cutter	http://bioinfo.bsd.uchicago.edu/SNP_cutter.htm	The program is capable of designing primers for either natural PCR-RFLP or mismatch PCR-RFLP, depending on the SNP sequence data; SNP Cutter generates the information needed to evaluate and perform genotyping experiments, including a PCR primers list, sizes of original amplicons and different allelic fragment after enzyme digestion	Zhang et al. (2005)
SNP pipeline	http://www.icrisat.org/gt-bt/download_SNP.htm	A pipeline of open source to mine large EST datasets for SNPs and identify restriction sites for assaying the SNPs so that cost-effective CAPS assays can be developed for SNP genotyping.	Jayashree et al. (2007)
dCAPS Finder	http://helix.wustl.edu/dcaps/dcaps.html	The tool generates mismatches in a PCR primer that are used to create a polymorphism based on the target SNP and the tool identifies the restriction enzymes to genotype the SNP using CAPS assay	Neff et al. (2002)

(Table 5.1). Basically, all these tools/pipelines perform clustering on redundant set of ESTs and the ESTs representing one gene are grouped under one cluster, which can be visualized for the occurrence of SNPs. The SNP(s) identified by using this approach, however, should be verified either by checking the sequence chromatogram of ESTs or validated by *wet lab* experiments, as *in silico* SNPs may have been observed as a result of sequencing error.

5.3.1.3 Intron Targeted Marker Development by Using Comparative Genomics Approach

Sequence diversity projects undertaken in several crop species using large number of genes (and different parts of gene) suggested higher frequency of SNP in intronic regions of the gene (Ching et al. 2002; Rajesh and Muehlbauer 2008). Therefore the markers developed from intronic region in general show high level of polymorphism (Bertin et al. 2005). In case of well-characterized species, exon-intron boundaries can be defined using tools like FEGNESH, etc. and subsequently primer pairs can be designed using the flanking exonic sequence of an intron to amplify intronic region (Feltus et al. 2006).

By using the genome sequence data of *Medicago* and *Lotus* and EST data of soybean and *Medicago*, >3,000 primer pairs targeting intronic regions have been developed in the laboratory of Doug Cook at UC-Davis, USA (DR Cook, personal communication) to develop gene-based conserved orthologous sequence (COS) markers in legume species. These primer pairs have been used to amplify the intronic regions in the parental genotypes of mapping populations of seven legume species including chickpea, common bean, groundnut, cowpea and pigeonpea which are being sequenced at present to identify the SNPs between parental genotypes of the mapping populations of these species. This study is expected to develop the extended version of comparative legume genetic maps developed earlier by Choi et al. (2004). Development of intron targeted markers has been successful even in under-resourced crop species. In such cases, the ESTs of the targeted species are BLASTed against the genome sequence of closest model genome sequence data to identify the exon–intron boundaries. In this way, intronic sequence can be identified and exonic sequence of the ESTs are used to design the primer pairs to amplify the intronic region. Based on this concept, after aligning the ESTs of sorghum, pearl millet, *Allium* and *Musa* with rice genome, >3,600 primer pairs, called conserved intron spanning primers (CISPs) have been developed for monocot species (Feltus et al. 2006; Lohithaswa et al. 2007; http://www.plantgenome.uga.edu/CISP/). Following the similar approach, a larger number of gene-based markers have been developed and used for diversity and mapping studies in pearl millet (Bertin et al. 2005), lupin (Phan et al. 2007), etc. Recently developed "cisprimerTOOL" at ICRISAT (http://www.icrisat.org/gt-bt/CISPTool.htm) for the identification of conserved intron scanning regions using EST alignments to a completely sequenced model crop genome and designing conserved intron scanning primers will greatly facilitate development of CISPs in several orphan crop species (Jayashree et al. 2008).

5.3.2 *Expression Polymorphism-Based Markers*

Comprehensive gene expression platforms are available in several crop species that has made it possible to undertake transcriptome profile of different tissues of the same genotype or same tissue of different genotypes (Sreenivasulu et al. 2004; Krist and Yu 2007). Several recent studies have demonstrated the use of transcript abundance data from genomic DNA/reduced complexity genomic DNA or cRNA hybridizations to microarrays (Affymetrix) to reveal genetic polymorphisms, if the transcript profiling has been done on different genotypes of the species. This polymorphism, also called expression level polymorphism (ELP), has been used as marker to genotype individuals in mapping populations (West et al. 2006a, 2006b).

Initially, Affymetrix GeneChips were used for identifying ELPs (Winzeler et al. 1998). Affymetrix GeneChips basically contain 11 different 25 bp-oligos covering features of the transcribed regions of each of several thousand genes. Each of these features for every gene on the GeneChip is present as a so-called perfect match (PM) and mismatch (MM) oligonucleotide. While the PM exactly matches the sequence of a particular standard genotype (e.g. one parent of a mapping population), the MM differs from this in a single substitution in the central 13th base. Therefore, if the parental genotypes of a mapping population, used for expression study, differ in the amount of mRNA produced by the particular tissue, this should result in a relatively uniform difference in their hybridizations across the 11 features. Furthermore, if the parental genotypes produce the same amount of mRNA but contain a genetic polymorphism within their DNA which coincides with one particular feature (or overlapping features), this would also result in differential hybridizations, however confined to that feature alone. Such polymorphism observed has been termed as single feature polymorphism (SFP) (Borevitz et al. 2003). Majority of studies dealing with discovery and genotyping of SFP have been conducted in sequenced and well-characterized model species such as yeast, mouse and *Arabidopsis* (Brem et al. 2002; Borevitz et al. 2003; Bing and Hoeschele 2005; Bystrykh et al. 2005; Ronald et al. 2005; Kumar et al. 2007). Nevertheless, recently in large and complex genome species such as barley, Cui et al. (2005) and Rostoks et al. (2005) hybridized barley expression microarrays with cRNA, to reduce target complexity, and detected thousands of SFPs. To establish the sensitivity and specificity of SFP prediction in species with complex and unsequenced genomes, Luo et al. (2007) explored four methods for identifying SFPs from microarray experiments involving two tissues in two barley genotypes and their doubled haploid (DH) progeny. They identified >4,000 separate SFPs that accurately predicted the SNP genotypes of >98% of DH lines. Very recently, the use of cross species platform has been demonstrated for identification of SFPs. For instance, Das et al. (2008) hybridized soybean microarray with cRNA of cowpea and identified >1,000 SFPs in parental genotypes of mapping populations of cowpea.

5.3.3 Next Generation Sequencing Technologies for Genome-Wide Marker Discovery

Several crop genomes have already been sequenced or in advanced stage of sequencing, that has enhanced our understanding of genome architecture. However, such data will have limited relevance to many other important species which are generally distantly related to model organisms. Genome sequencing in non-model organisms has the potential to also be greatly enhanced by developments in sequencing technologies. Recent developments, due to growing interest in human genome re-sequencing, nucleic acid chemistry, nanotechnology and microscopy have led to a new generation of sequencing and genotyping technologies. These new technologies sequence DNA very fast and cheap, however in short fragments. These new methods are currently driving down sequencing costs and increasing capacity at an unprecedented rate that makes the whole genome resequencing possible (Hudson 2008; Mardis 2008; Gupta 2008).

At present, three main sequencing methods of next generation sequencing technologies are commercially available: (a) 454/FLX sequencing, (b) Solexa/Illumina 1 GB SBS (sequencing-by-synthesis) technology, and (c) AB SOLiD (Sequencing by Oligonucleotide Ligation and Detection) technology (Varshney et al. 2009). For all these DNA sequencing methods, genomic DNA is randomly sheared and individual DNA molecules are then immobilized on a solid support, which can be a microscopic bead (in this case one molecule is affixed to each bead) or a macroscopic support such as a flow cell or slide (in this case many molecules are arrayed randomly on the support) (Fan et al. 2006). Subsequently these individual DNA molecules are then amplified using the PCR. In case of bead-based methods, amplification is done in an emulsion phase where the beads are protected from cross-contamination by the barrier of an immiscible solvent. The polymerase colonies, often called 'polonies' or 'clusters', which are clonally identical DNA molecules either attached to a single bead or attached to a localized region on a solid support. While in bead-based methods, the beads are then either themselves immobilized on a planar support, or placed in individual microscopic wells, in case of non-bead-based methods the polonies are generated *in situ*. After producing a planar array of polonies, the respective sequencing chemistry is applied directly to the molecules on the support. Instead of separating elongation products, the sequence is interrogated at every base, by the use of either fluorescence or chemiluminescence to directly detect the incorporation of a base-specific chemical probe.

In addition to above mentioned three technologies, one new method called Single Molecule Sequencing, also known as third generation technology has received a great deal of attention and has potential to further increase throughput. Infact, various single-molecule and other sequencing and resequencing methods are under development in academic laboratories, and at several companies such as Biotage, Helicos, Li-Cor, Microchip Biotechnologies, Nanofluidics, Nanogen, Network Biosystems and Visigen (Hudson 2008). These next generation sequencing

technologies can be used for genome-wide marker discovery in both model/major and under-resourced crop species. Several bioinformatics tools and pipelines have been developed recently for analyzing the next generation sequence data for SNP discovery (Table 5.2). Indeed the bioinformatics community across the world is actively engaged in improving the tools for analyzing the next generation sequence data with higher accuracy and efficiency (Hillier et al. 2008; Smith et al. 2008; Varshney et al. 2009).

5.3.3.1 Re-Sequencing in Well-Characterized Species

In the species, which have genome or EST sequence data, genotypes of interest e.g. parental genotypes of mapping populations can be subjected for next generation sequencing technologies for genome-wide marker discovery. Re-sequencing can be done on cDNA population as well as genomic DNA (reduced representation genome) of different genotypes. Sequencing data generated using the next generation sequencing technologies can be aligned with the reference genome (genome/ transcriptome assembly). In such a way, genome-wide variants can be identified between the genotypes or compared to the reference genotype. For instance, based on 454 sequencing the transcriptomes of shoot apical meristems from two maize inbred lines namely B73 (260,000 reads) and Mo17 (280,000 ESTs), >36,000 SNPs were detected within 9,980 unique B73 genomic anchor sequences (Maize Assembled Genomic Islands, called MAGIs). Stringent post-processing reduced this number to >7,000 putative SNPs; over 85% (94/110) of a sample of these putative SNPs were successfully validated by Sanger sequencing (Barbazuk et al. 2007). Similar kind of SNP discovery projects using 454 and/or Solexa sequencing are underway in several other crop species like soybean (Hyten et al. 2008), chickpea (May et al. 2008), pigeonpea (ongoing studies at ICRISAT and NRCPB, India).

In case of rice, where the genome sequence is available, International Rice Research Institute (IRRI), in collaboration with its partners, employed array-based resequencing technology using very high-density oligomer arrays for genome-wide SNP discovery in 20 diverse varieties at Perlegen Biosciences (Kenneth McNally, personal communication). Under this project, 100 Mb of the Nipponbare rice genome (IRGSP release 4) corresponding to the fraction with little or no repetitiveness was chosen for SNP discovery. Perlegen designed 25-mer oligos with single base offsets tiled across the 100 Mb fraction of the genome for both strands with the 13th base in full degeneracy with each target position in the reference sequence interrogated by eight oligos. Subsequently, independent long-range PCR amplicons were produced for target pools across the regions arrayed on a particular wafer or chip. These LR-PCR amplicons were pooled, labeled and hybridized to the wafers. Based on Perlegen's model-based algorithms, 259,721 non-redundant SNPs have been predicted among 20 varieties. The estimated SNP frequency (2.6 SNPs per kb) is comparable to the figures obtained from pairwise comparisons of *indica* and *japonica* rice varieties.

Table 5.2 Important bioinformatic tools for analyzing next generation sequence data for marker discovery and gene expression studies

Tool	Web address	Description	Reference
Atlas-SNP	http://code.google.com/p/atlas-snp/	A tool for SNP/indel discovery from genome re-sequencing using next-generation sequencing technologies	Wheeler et al. (2008)
Next*GENe*	http://www.softgenetics.com/NextGENe.html	A software to analyze the next generation sequence data for de novo assembly, SNP/indel detection and transcriptome analysis	SoftGenetics, USA
PanGEA	http://www.kofler.or.at/Bioinformatics/PanGEA/index.html	The programme allows to map ESTs/ sequence tags to genes or whole genomes using algorithm which are especially adapted to next-generation sequencing technologies and mapping results may be displayed as a gene-expression-profile or used for SNP identification	Kofler et al. unpublished
Alpheus	http://alpheus.ncgr.org/	A web-based cyberinfrastructure platform for pipelining, visualization and analysis of GigaBase-scale sequence data generated from Sanger, Roche-454, Illumnia-Solexa, ABI SOLiD methodologies. Alpheus provides data management services, an analysis pipeline, and internet-accessible software for variant discovery and isoform identification.	National Centre for Genome Resource, USA
SNPsniffer	http://bioinformatics.bc.edu/marthlab/Polymorphism_Discovery_in_Next-Generation_Sequence_Data	A SNP discovery tool specifically designed for 454 sequences (currently being developed)	Aaron Quinlan and Weichun Huang, Boston College, USA

5.3.3.2 *De novo* Sequencing of Under-Resourced Crop Species

Although next generation sequencing technologies are ideally meant for re-sequencing, *de novo* sequencing can also be undertaken with these sequencing technologies. Alignment of the smaller fragments without the availability of the reference genome, however, becomes quite tedious if not impossible. In such case more than one genotype can be used for generating the sequence data using 454/Solexa/AB SOLiD technologies. Alignment of these sequence data can be facilitated by: (1) genome or transcriptome sequence data of model/major crop species closely related with the species; (2) whole transcriptome or reduced representation genome sequence data of the species, generated using 454 sequence technology. Aligning of sequence data of more than two genotypes of the species by using one of the above approaches provides the confidence in aligning the short sequence and detecting the sequence variants. Although several bioinformatic tools and algorithms are currently available (Table 5.2), efforts are continuously underway at several places to improve the accuracy of alignment of next generation sequence data (Smith et al. 2008). A web-based cyber infrastructure platform, called Alpheus (http://alpheus.ncgr.org/), is very useful for pipelining, visualization and analysis of GigaBase-scale sequence data for identification of SNPs.

A preliminary study dealing with Solexa sequencing of drought challenged root trasncriptomes of two genotypes of chickpea, ICC 4958 and ICC 1882 carried out at ICRISAT in collaboration with National Centre for Genome Resources (NCGR), USA (Greg May and Andrew Farmer) and University of California, Davis, USA (Doug Cook), has demonstrated the utility of next generation sequencing technology for SNP discovery in a species without the reference genome (May et al. 2008). Half run of Solexa sequencing on the pooled RNA samples from ICC 4958 and ICC 1882 yielded 5.2×10^6 and 3.6×10^6 sequence reads respectively. In order to analyze the generated Solexa datasets, the following three set of sequence resources were used in Alpheus pipeline: (1) *Medicago truncatula* (Mt) IMGAG (International *Medicago* Genome Annotation Group) gene assembly representing 29.5 Mb sequence data, (2) *Cicer arietinum* transcript assembly (Ca TA) of JCVI (The James Craig Venter Institute) representing 681 kb sequence data and (3) *Cicer arietinum* (Ca) BAC-end sequence (Ca BES) data representing 16.4 Mb sequence data. Bioinformatic analysis revealed matches of Solexa tags with 5,886 genes in cases of ICC 4958 and 7,338 genes in ICC 1882, respectively. Although detailed analysis for SNP discovery is underway, the preliminary analysis suggested the occurrence of SNPs at least in 500 cases.

5.4 Genotyping Assays

After identifying the SNPs in genes, optimizing or developing the appropriate platform for SNP genotyping is another important task. At present more than 30 SNP genotyping assays are available and each of them is having some merits as

well as constraints (Gupta et al. 2001, 2008). A critical comparison of a selected SNP genotyping assays has been made recently by Bagge and Lübberstedt (2008). Instead of discussing different kinds of assays in this article, some important and/or most commonly used SNP genotyping assays are discussed. These assays can be selected and optimized based on costs available and intended objective.

5.4.1 Low-Throughput and Inexpensive Genotyping Assay

When only few SNP-based genic markers need to be genotyped or limited financial resources are available, inexpensive SNP genotyping assays can be used. Different kinds of inexpensive SNP genotyping assays are currently available, only two assays have been given here.

5.4.1.1 Cleaved Amplified Polymorphic Sequences (CAPS)

Under this approach, the sequence alignment for more than two genotypes that contained SNPs is subjected to identify the restriction sites for restriction enzymes. This procedure can be facilitated by using bioinformatics tools, available in public domain (Table 5.1) that use multiple sequence alignments for several genes/markers in a batch file. The principle of these programmes is identification of recognition site and their corresponding restriction enzyme if the SNP present in multiple sequence alignment creates some recognition site for a restriction enzyme. Subsequently, the gene sequence can be amplified in germplasm through PCR and amplicon can be digested with the restriction enzyme identified by the programme and visualized on agarose gel. By using such methodology more than 80 EST-based SNP markers were converted into cost-effective CAPS markers (Kota et al. 2007). The approach has been used for assaying SNPs in many crop species, e.g. chickpea (Varshney et al. 2007d; Rajesh and Muehlbauer 2008), rye (Varshney et al. 2007a), rice (Komuri and Nitta 2005). Development of SNP markers and their optimization into CAPS assay is underway at ICRISAT (Reddy et al. unpublished). An example of such a CAPS-based genic SNP marker has been shown in Fig. 5.1.

In case, the SNP present in the genotypes of interest does not provide the recognition site for a restriction enzyme for CAPS assay, a modified technique called dCAPS (derived cleaved amplified polymorphic sequences) assay can be developed by creating a mismatch in a PCR primer to create a polymorphism based on the target mutation (Neff et al. 1998, 2002). In case of chickpea, genic SNP markers, which could not be converted to CAPS markers, are being assayed as dCAPS markers at ICRISAT (Reddy et al. unpublished).

5.4.1.2 Single-Strand Conformation Polymorphism (SSCP)

In general polyacrylamide gel electrophoresis doesn't allow detection of polymorphism due to difference of one base pair length/type. Single stranded DNA confirmation polymorphism (SSCP) methodology, however, allows detection of polymorphism due to differences of one or more base pairs in the PCR products that is suitable for SNP genotyping. The methodology relies on the secondary structure being different for single strands derived from PCR products that differ by one or more nucleotides at an internal site. For assays using SSCP methodology, PCR products of different genotypes carrying the SNP site are denatured and electrophoretically separated in neutral acrylamide gel. Because of occurrence of SNPs in different genotypes, the gel will show the difference in the length of the resolving fragments. This methodology has been used in several species like *Picea* (Germano and Klein 1999), pearl millet (Bertin et al. 2005), cassava (Castelblanco and Fregene 2006).

5.4.2 High-Throughput Genotyping Assays

Due to the availability of highly parallel genomic assays at present, large scale SNP-based marker genotyping is possible in cost-effective and relatively less time (Gupta et al. 2008). Majority of time, such genotyping assays are available as services offered by companies or genotyping centres.

5.4.2.1 GoldenGate Assay

GoldenGate assay of Illumina Inc. is probably the most popular large scale genotyping assay at present (Fan et al. 2003). The methodology deals with hybridization of allele (SNP)-specific primers directly to genomic DNA immobilised on a solid support. In case of a perfect match the primer is extended and the extension product is ligated to a probe hybridised downstream the SNP position. Subsequently, the ligated product is amplified by PCR using universal primers that are complementary to a universal sequence in the 3□end of the ligation probes and 5□ends of the allele-specific primers, respectively. It is important to note that the ligation probe contains a SNP-specific Tag-sequence while the universal allele-specific primers carry an allele-specific fluorescent label in their 5□end. After performing PCR, the amplified products are captured on beads carrying complementary target sequences for the SNP-specific Tag of the ligation probe. The beads are kept in fiber-optic array bundle that has a compatible format with 96-well microtiter plates. As a result, the GoldenGate assay supports genotyping of 96-, 192-, 384-, 768- and 1536 custom selected SNPs in a single reaction over a 3-day period.

Among plant systems, the GoldenGate assay-based SNP genotyping was undertaken for the first time in barley where barley community, in consultation with Illumina Inc., developed the GoldenGate assay for 1,536 SNPs selected based on EST mining (Rostoks et al. 2006). Inspired by high-throughput and low cost genotyping, the barley community has developed a total of three pilot Illumina oligonucleotide pool assays (OPAs) each containing 1,536 SNPs, under the barley coordinated agriculture project (CAP) in USA. These three OPAs have been used to map three mapping populations and genotype germplasm sets from the United States and Europe. From these three pilot OPAs, over 3,000 high quality SNPs have been used to design two OPAs (3,072 SNPs) for genotyping. These two OPAs will be used for genotyping 960 breeding lines of barley (http://www.barleycap.org/; TJ Close, personal communication). In soybean also, a custom 384-SNP GoldenGate assay was designed using SNPs discovered through the re-sequencing of five diverse accessions (Hyten et al. 2008). Allelic data were successfully generated for 89% of SNP loci (342 of the 384) and finally a gene-based integrated map with 334 SNP loci was prepared.

5.4.2.2 Whole-Genome Genotyping Infinium Assay

Based on Array-CGH (comparative genomic hybridization), Illumina Inc. introduced a very high-density SNP genotyping technology to genomic profiling, termed SNP-CGH, that allows simultaneous measurement of both signal intensity variations and changes in allelic composition. The utility of SNP-CGH was demonstrated with two Infinium whole-genome genotyping BeadChips, assaying 109,000 and 317,000 SNP loci, to detect chromosomal aberrations in samples bearing constitutional aberrations as well tumor samples at sub-100 kb effective resolution in human system (Peiffer et al. 2006).

Under the Infinium assay, *firstly* whole-genome amplification step is used to increase the amount of DNA up to 1,000-fold. Subsequently, the DNA is fragmented and captured on a BeadArray by hybridization to immobilised SNP-specific primers that is followed by extension with hapten-labelled nucleotides. As a result, the primers hybridize adjacent to the SNPs and are extended with a single nucleotide corresponding to the SNP allele. Finally, the incorporated hapten-modified nucleotides are detected by adding fluorescently labelled antibodies in several steps to amplify the signals. Data analysis under Infinium assays is performed using scatter plots as for the GoldenGate assay.

Very recently, Illumina Inc. announced development of the Infinium HD Human1M-Duo (two samples/chip) and the Human610-Quad (four samples/chip) system, featuring up to 2.3 million single nucleotide polymorphisms (SNPs) per BeadChip (www.illumina.com). Both arrays on the Human1M-Duo BeadChip contain markers for more than one million diverse genetic variants, all of which can be used for both whole-genome genotyping and copy number variation (CNV) analysis. The four-sample format of the Human610-Quad BeadChip offers a significant increase in sample throughput and reduced handling in the lab, as it has

550,000 SNPs plus an additional 60,000 genetic markers per sample. Although Infinium assay have not been developed in plant systems so far, availability of sequence data, next generation sequencing technologies for high density SNP discovery in some plant species like rice, maize, soybean may encourage the plant science community to undertake developing and using the Infinium assay soon.

5.5 Applications of Gene-Based Markers in Crop Improvement

Gene-based or functional markers (FMs) can be used for all the applications where traditional markers have been or can be used. In addition, the use of FMs provide added value for a particular application aimed at crop improvement. For instance, on one hand the FMs have been proven the "perfect markers" for foreground selection in marker-assisted selection (MAS), the availability of (or possibility to develop) low cost and high throughput genotyping platforms for gene-based markers (e.g. GoldenGate assays/Illumina arrays) for many crops make these markers the most suitable markers for background selection in marker-assisted breeding (MAB). When these markers are used in the genetic diversity studies, they assay the functional genetic variation in the germplasm collection and therefore can be used for allele mining and association genetics studies. Due to their origin from conserved proportion of the genome, gene-based markers of a species can be used in related species for a variety of applications including enhancing the density of genetic maps (Varshney et al. 2007a) and understanding the genome relationships and evolution (Stein et al. 2007). The utility of gene-based markers has been illustrated in selected two areas in following sections.

5.5.1 Superiority of FMs over Traditional Markers in MAS

Since 1990s, molecular markers have shown their applications for MAS in several crops (see Jain et al.2002 ; Gupta and Varshney 2004; Varshney and Tuberosa 2007). Large scale deployment of molecular markers in public breeding programme was initiated in 1997 at the Australian Molecular Plant Breeding Cooperative Research Centre (MPB CRC) in case of wheat (Eagles et al. 2001; Langridge 2005; Varshney et al. 2007b), shortly afterwards, such programmes were started in USA (MASwheat, that has been transformed into Wheat CAP recently – http://maswheat.ucdavis.edu/), Europe and China. Majority of these programmes have been deploying the SSR or STS/SCAR (developed from RFLPs) markers closely linked to disease resistance genes and agronomic traits.

A large number of markers associated with QTLs/genes for resistance/tolerance to biotic/abiotic stress as well as agronomic traits are reported every year in the form of research publications. However transfer of markers to practical plant breeding,

in the same proportion has not taken place at all or took longer than expected (Tuvesson et al. 2007). As a result, lots of claims have been made about what great things genomics does, but very little has been put into application (Varshney and Tuberosa 2007). One reason for this is the reduced reliability of diagnostic value of linked markers due to genetic recombination between marker and target locus (Bagge and Lübberstedt 2008). Such recombination, majority of times, impairs transfer of marker information from experimental mapping population to unrelated breeding materials. However this is not the case with gene-based and especially FMs that are derived from polymorphic sites within gene coding sequences causally affecting phenotypic trait variation. As a result, the FMs, as compared to anonymous markers including SSRs that were considered as markers of choice till recently, are more reliable for identification and selection of favourable alleles, as absence of recombination between marker and target locus increases the diagnostic power of the marker in the marker-based selection of genotypes.

In view of above, the FMs have been considered as 'good translators' from genomic technologies into improved crop varieties (Thro et al. 2004; Bagge et al. 2007). For example, in case of wheat, a STS marker for polyphenol oxidase (PPO) activity developed from the EST of the *PPO* gene, was found to discriminate accurately between 233 Chinese varieties with low and high PPO activity (Sun et al. 2005). Gene specific markers for waxy starch were used to select the wheat materials in Australia that had the *wx-B* allele, which is associated with good Asian noodle quality (Murai et al. 1999).

As mentioned earlier, development of FMs requires functionally characterized genes, the identification of polymorphic/functional site that affect plant phenotype within the corresponding genes and the validation of association between DNA polymorphisms and trait phenotype (Bagge et al. 2007), FMs have been developed so far only for selected traits and in few crops. For example, in case of rice, the cloning of the gene *xa-5* underlying the bacterial blight resistance (Iyer and McCouch 2004) has made it possible to develop functional markers for *xa5*-mediated resistance (Iyer-Pascuzzi and McCouch 2007). Bagge et al. (2007) and Bagge and Lübberstedt (2008) have recently summarized the current status on cloning of genes in wheat and their potential for functional marker development. A list of cloned genes in some major cereals like rice, wheat, barley and sorghum is available in Varshney et al. (2006) that could be used to develop FMs. Recent advances in the area of genomics like next generation sequencing technologies and high-throughput genotyping platforms mentioned earlier should facilitate the development and application of FMs in several crop species in coming years (Varshney et al. 2009).

5.5.2 Utility of Gene-Based Markers for Allele Mining

Gene-based markers and especially FMs are a better resource for allele mining for the corresponding gene from which the markers developed. The main approaches for allele mining include TILLING (targeting induced local lesions in genomes, see

Till et al. 2007), EcoTILLING (see Till et al. 2007) and candidate gene/FM sequencing. While TILLING approach deals with identification of new alleles after screening the mutant population for the candidate gene/FM, the EcoTILLING and candidate gene/FM sequencing approaches identity the natural allelic variation in a germplasm collection. By using TILLING approach, Slade et al. identified 246 alleles each homoeologue in 1,920 allohexaploid and allotetraploid wheat individuals. These alleles encoded *waxy* enzymes ranging in activity from near wild type to null, and they represented more genetic diversity than had been described in the preceding 25 years. An example of use of EcoTILLING for allele mining can be seen in case of rice for a gene "putative ethylene-responsive element binding protein 3 (ERF3) that colocalized with a drought QTL as 136.6 cM on chromosome 1 (Wang 2005). EcoTILLING in 2 kb ERF3 regions showed 25 SNPs and 5 indels defining 9 haplotypes (hap1 to hap9). Furthermore a significant association was reported between hap9 and yield stability in *indica* rice group.

Sequencing of candidate gene for the corresponding FM in an appropriate germplasm collection is a comprehensive approach for allele mining (Varshney et al. 2005b). For majority of the cereal and legume crops, a vast collection of genomic resources is available in genebanks of international agricultural research centres (IARCs) and national genebanks in different countries. In order to utilize the germplasm of genebank in breeding programme in the efficient way, the manageable collection of germplasm called "core collection", "mini-core collection", "reference collection" etc. have been developed in several crop species (e.g., Varshney et al. 2007c). Sequencing of the corresponding gene(s) for the FM in these germplasm collections provides a range of alleles for the given FM/gene. Phenotyping of the germplasm collection for the respective trait and their analysis with allele data should provide the better alleles for the respective trait (Slade and Knauf 2005). Infact, such allele mining and association genetics approaches are being used in several crop species to link genetic diversity with trait phenotype (Ersoz et al. 2007). This will help breeders to move towards allele-based selection in their breeding programmes.

5.6 Conclusions and Prospects

While development of markers and genetic map was an expensive and time consuming task till few years ago, availability of gene/genome sequence data together with high-throughput marker discovery and genotyping platforms have made the development of genic markers easier and faster. For instance, availability of three OPA assays in barley has provided 2,801 mappable gene-based SNP markers. Indeed, because of the possibility of large scale genotyping (with 1,536 SNP markers) of European barley cultivars, it was demonstrated that whole genome scanning-based association mapping is feasible in self-pollinated species like barley (Rostoks et al. 2006). Similar kinds of mapping efforts were undertaken in

soybean, maize, wheat, etc. and underway in many other crop species. High density marker genotyping is proving useful on one hand to anchor genetic map and physical map, linkage-disequilibrium-based association mapping approach is becoming possible for trait mapping. Among gene-based markers, the FMs are the "perfect markers" for foreground selection in MAS and therefore development of genic markers and especially FMs in different crops in coming future will enhance the application of MAS in breeding programmes and also for allele mining for the corresponding gene(s) using the germplasm collection held in genebanks or mutant populations. Availability of high-throughput genome-wide and low cost genotyping platform provides opportunities to accelerate breeding practices through the use of markers in background selection during marker-assisted breeding.

Recent advances in nanotechnology, nucleic acid chemistry, computational biology and automation indicates that development and application of gene-based markers using high-throughput marker discovery and genotyping assay is still a relatively young field, and more exciting advances are expected in the future. One of the great promises of genic markers, using high-throughput approaches, is that the ability to carry out comprehensive genomic analyses easily, inexpensively, accurately and rapidly with high sensitivity should create a new generation of routine genomic tools to assist the crop breeding.

Acknowledgements Thanks are due to Generation Challenge Programme (GCP), National Fund of Indian Council of Agricultural Research, Pigeonpea Genomics Initiative under Indo-US Agricultural Knowledge Initiative (AKI) and Department of Biotechnology (Government of India) for funding the research of author on development of genic molecular markers in legume species at ICRISAT.

References

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

Bagge M, Lübberstedt T (2008) Functional makers in wheat: technical and economic aspects. Mol Breed 22:319–328

Bagge M, Xia X, Lübberstedt T (2007) Functional makers in wheat. Curr Opin Plant Biol 10:211–216

Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS (2007) SNP discovery via 454 transcriptome sequencing. Plant J 51:910–918

Bertin, Zhu JH, Gale MD (2005) SSCP-SNP in pearl millet-a new marker system for comparative genetics. Theor Appl Genet 110:1467–1472

Bing N, Hoeschele I (2005) Genetical genomics analysis of a yeast segregant population for transcription network inference. Genetics 170:533–542

Borevitz JO, Liang D, Plouffe D, Chang HS, Zhu T, Weigel D, Berry CC, Winzeler E, Chory J (2003) Large-scale identification of single-feature polymorphisms in complex genomes. Genome Res 13:513–523

Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. Science 296:752–755

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* 37:225–232

Castelblanco W, Fregene M (2006) SSCP-SNP-based conserved ortholog set (COS) markers for comparative genomics in cassava (*Manihot esculenta* Crantz). Genet Resour 24:229–236

Ching ADA, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski A (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19

Choi HK, Mun JH, Kim DJ, Zhu H, Baek JM, Mudge J, Roe B, Ellis N, Doyle J, Kiss GB, Young ND, Cook DR (2004) Estimating genome conservation between crop and model legume species. Proc Natl Acad Sci 101:15289–15294

Cui X, Xu J, Asghar R, Condamine P, Svensson JT, Wanamarker S, Stein N, Roose M, Close TJ (2005) Detecting single-feature polymorphisms using oligonucleotide array and robustified projection pursuit. Bioinformatics 21:3852–3858

Das S, Bhat PR, Sudhakar C, Ehlers JD, Wanamaker S, Roberts PA, Cui X, Close TJ (2008) Detection and validation of single feature polymorphisms in cowpea (*Vigna unguiculata* L. Walp) using a soybean genome array. BMC Genomics 9:107

Eagles HA, Bariana HS, Ogbonnaya FC, Rebetzke GJ, Hollamby GJ, Henry RJ, Henschke PH, Carter M (2001) Implementation of markers in Australian wheat breeding. Aust J Agric Res 52(12):1349–1356

Ersoz ES, Yu J, Buckler ES (2007) Applications of linkage disequilibrium and association mapping in crop plants. In: Varshney RK, Tuberosa R (eds) Genomics-assissted crop improvement, vol I. Springer, The Netherlands, pp 97–119

Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, Mcbride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS (2003) Highly parallel SNP genotyping. Cold Spring Harbor Symopsium Quantitaive Biology 68:69–78

Fan JB, Chee MS, Gunderson KL (2006) Highly parallel genomic assays. Nature 7:632–644

Feltus FA, Singh HP, Lohithaswa HC, Schulze SR, Silva TD, Paterson AH (2006) Acomparative genomics strategy for targeted discovery of single-nucleotide polymorphism and conserved-noncoading sequences in orphan crops. Plant Phys 140:1183–1191

Germano J, Klein AS (1999) Species-specific nuclear and chloroplast single nucleotide polymorphisms to distinguish *Picea glauca*, *P. mariana* and *P. rubens*. Theor Appl Genet 99:1–2

Grover D, Woodfield AS, Verma R, Zandi PP, Levinson DF, Potash JB (2007) QuickSNP: an automated web server for selection of tagSNPs. Nucleic Acids Res 35:W115–W120

Gupta PK (2008) Single-molecule DNA sequencing technologies for future genomics research. Trends Biotech 26:602–611

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

Gupta PK, Varshney RK (2004) Cereal Genomics. Kluwer Academic, The Netherlands, pp 639

Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphism: a new paradigm for molecular markers technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–534

Gupta PK, Varshney RK, Prasad M (2002) Molecular markers: principles and methodology. In: Jain SM, Ahloowalia BS, Brar DS (eds) Molecular techniques in crop improvement. Kluwer Academic, The Netherlands, pp 9–54

Gupta PK, Rustgi S, Mir RR (2008) Array-based high throughput DNA markers for crop improvement. Heredity 101:5–18

Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, Fox P, Glasscock JI, Hickenbotham M, Huang W, Magrini VJ, Richt RJ, Sander SN, Stewart DA, Stromberg M, Tsung EF, Wylie T, Schedl T, Wilson RK, Mardis ER (2008) Whole-genome sequencing and variant discovery in *C.elegans*. Nat Methods 5:183–188

Hudson M (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8:3–17

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

Hyten DL, Song Q, Choi IY, Yoon MS, Specht JE, Matukumalli LK, Nelson RL, Shoemarker RC, Young ND, Cregan PB (2008) High-throughput genotypying with the GoldenGate assay in the complex genome of soybean. Theor Appl Genet 116:945–952

Iyer AS, McCouch SR (2004) The rice bacterial blight resistance gene *xa5* encodes a novel form of disease resistance. Mol Plant Microbe Interact 17(12):1348–1354

Iyer-Pascuzzi AS, McCouch SR (2007) Recessive resistance genes and the *Oryza sativa-Xanthomonas oryzae* pv. *oryzae pathosystem*. Mol Plant Microbe Interact 7:731–739

Jain SM, Brar DS, Ahloowalia BS (2002) Molecular techniques in crop improvement. Kluwer Academic, The Netherlands, pp 601

Jayashree B, Hanspal MS, Srinivasan R, Vigneshwaran R, Varshney RK, Spurthi N, Eshwar K, Ramesh N, Chandra S, Hoisington DA (2007) An integrated pipeline of open source software adapted for multi-CPU architectures: use in the large-scale identification of single nucleotide polymorphisms. Comp Funct Genomics 35604:7

Jayashree B, Jagdeesh VT, Hoisington D (2008) cisprimertool: software to implement a comparative genomics strategy for the development of conserved intron scanning (CIS) markers. Mol Ecol Resour 8:575–577

Krist M, Yu Q (2007) Genetical genomics: successes and prospects in plants. In: Varshney RK, Tuberosa R (eds) Genmics assisted crop improvement. Springer, The Netherlands, pp 245–266

Komuri T, Nitta N (2005) Utilization of the CAPS/dCAPS method to convert rice SNPs into PCR-based markers. Breed Sci 55:93–98

Kota R, Varshney RK, Prasad M, Zhang H, Stein N, Graner A (2007) EST-derived single nucleotide polymorphism markers for assembling genetic and physical maps of the barley genome. Funct Integr Genomics. doi:10.1007/s10142-077-0060-9

Kumar R, Qiu J, Joshi T, Valliyodan B, Xu D, Nguyen HT (2007) Single feature polymorphism discovery in Rice. PLoS ONE 2:e284

Langridge P (2005) Molecular breeding of wheat and barley. In: Tuberosa R, Phillips RL, Gale M (eds) In the wake of the double helix: from the green revolution to the gene revolution. Avenue Meida, Bologna, Italy, pp 279–286

Lohithaswa HC, Feltus FA, Singh HP, Bacon CD, Bailey CD, Paterson AH (2007) Leveraging the rice genome sequence for monocot comparative and translational genomics. Theor Appl Genet 115:237–243

Luo ZW, Potokina E, Druka A, Wise R, Waugh R, Kearsey MJ (2007) SFP Genotyping from affymetrix arrays is robust but largely detects *Cis*-acting expression regulators. Genetics 176:789–800

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133–141

Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Van Tassell CP (2006) SNP-PHAGE-High throughput SNP discovery pipeline. BMC Bioinformatics 7:468

May GD, Lekha PT, Kashiwagi J, Huntley JJ, Farmer AD, Cook DR, Varshney RK (2008) Whole transcriptome shotgun sequencing for variant detection and transcript profiling in chickpea (*Cicer arietinum* L.). In: Plant Animal Genomes XVI Conf San Diego USA P385 (http://www.intl-pag.org/16/abstracts/PAG16_P05f_385.html)

Murai J, Taira T, Ohta D (1999) Isolation and characterization of the three *Waxy* genes encoding the granule-bound starch synthase in hexoploid wheat. Gene 234:71–79

Neff MM, Neff JD, Chory J, Pepper AE (1998) dCAPS, a simple technique for the genetic analysis of single nucleotide polymorphisms: experimental applications in *Arabidopsis thaliana* genetics. Plant J 14:387–392

Neff MM, Turk E, Kalishman M (2002) Web-based primer design for single nucleotide polymorphism analysis. Trends Genet 18:613–615

Panitz F, Stengaard H, Hornshøj H, Gorodkin J, Hedegaard J, Cirera S, Thomsen B, Madsen LB, Høj A, Vingborg RK, Zahn B, Wang X, Wang X, Wernersson R, Jørgensen CB, Scheibye-Knudsen K, Arvin T, Lumholdt S, Sawera M, Green T, Nielsen BJ, Havgaard JH, Brunak S, Fredholm M, Bendixen C (2007) SNP mining porcine ESTs with MAVIANT, a novel tool for SNP evaluation and annotation. Bioinformatics 23:i387–i391

Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16:1136–1148

Phan HT, Ellwood SR, Adhikari K, Nelson MN, Oliver RP (2007) The first genetic and comparative map of white lupin (*Lupinus albus* L.): identification of QTLs for anthracnose resistance and flowering time, and a locus for alkaloid content. DNA Res 14:59–70

Rajesh PN, Muehlbauer FJ (2008) Discovery and detection of single nucleotide polymorphism (SNP) in coding and genomic sequences in chickpea (*Cicer arietinum* L.) Euphytica. doi: 10.1007/s10681-008-9675-8

Ronald J, Akey JM, Whittle J, Smith EN, Yvert G, Kruglyak L (2005) Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. Genome Res 15:284–291

Rostoks N, Borevitz JO, Hedley PE, Russell J, Mudie S, Morris J, Cardle L, Marshall DF, Waugh R (2005) Single-feature polymorphism discovery in the barley transcriptome. Genome Biol 6 (6), art. no. R54

Rostoks N, Ramsay L, MacKenzie K, Cardle L, Bhat PR, Roose ML, Svensson JT, Stein N, Varshney RK, Marshall DF, Graner A, Close TJ, Waugh R (2006) Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. Proc Natl Acad Sci USA 103:18656–18661

Savage D, Batley J, Erwin T, Logan E, Love CG, Lim GAC, Mongin E, Barker G, Spangenberg GC, Edwards D (2005) SNPServer: a real-time SNP discovery tool. Nucleic Acids Res 33:W493–W495

Schmitt BA, Costa JH, de Melo DF (2006) AOX -A functional marker for efficient cell reprogramming under stress? Trends Plant Sci 11:281–287

Shah N, Teplitsky MV, Minovitsky S, Pennacchio LA, Hugenholtz P, Hamann B, Dubchak I (2005) SNP-VISTA: An interactive SNP visualization tool. BMC Bioinformatics 6:292

Slade AJ, Knauf VC (2005) TILLING moves beyond functional genomics into crop improvement. Transgenic Res 14:109–115

Smith AD, Xuan Z, Zhang MQ (2008) Using quality scores and longer reads improves accuracy of Solexa read mapping. BMC Bioinformatics 9:128

Stein N, Prasad M, Scholz U, Thiel T, Zhang H, Wolf M, Kota R, Varshney RK, Perovic D, Grosse I, Graner A (2007) A 1000 loci transcript map of the barley genome- new anchoring points for integrative grass genomics. Theor Appl Genet 114:823–839

Sreenivasulu N, Varshney RK, Kavikishore PV, Weschke W (2004) Tolerance to abioitic stress- a functional genomics approach. In: Gupta PK, Varshney RK (eds) Cereal genomics. Kluwer Academic, The Netherlands, pp 483–514

Sun DJ, He ZH, Xia XC, Zhang LP, Morris C, Appeis R, Ma W, Wang H (2005) A novel STS marker for polyphenol oxidase activities in bread wheat. Mol Breed 16:209–218

Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics 7:438

Tang J, Leunissen JA, Voorrips RE, van der Linden CG, Vosman B (2008) HaploSNPer: a web-based allele and SNP detection tool. BMC Genet 9:23

Thiel T, Kota R, Grosse I, Stein N, Graner A (2004) SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. Nucleic Acids Res 32:e5

Thro AM, Parroot W, Udall JA, Beavis WD (2004) The experience of initiative for future agricultural and food systems. Crop Sci 44:1893

Till BJ, Comai L, Henikoff S (2007) Tilling and Ecotilling for crop improvement In: Varshney RK, Tuberosa R (eds) Genomics-assissted crop improvement, vol I, Springer, The Netherlands, pp 333–349

Tuvesson S, Dayteg C, Hagberg P, Manninen O, Tanhuanpää P, Tenhola-Roininen T, Kiviharju E, Weyen J, Förster J, Schondelmaier J, Lafferty J, Marn M, Fleck A (2007) Molecular markers and doubled haploids in European plant breeding programmes. Euphytica 158:305–312

Unneberg P, Strömberg M, Sterky F (2005) SNP discovery using advanced algorithms and neural networks. Bioinformatics 21:2528–2530

Varshney RK, Graner A, Sorrells ME (2005a) Genic microsatellite markers in plants: features and applications. Trends Biotech 23:48–55

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

Varshney RK, Hoisington DA, Tyagi (2006) Advances in cereal genomics and applications in crop breeding. Trends in Biotech 24:490–499

Varshney RK, Beier U, Khlestkina EK, Kota R, Korzun V, Graner A, Börner A (2007a) Single nucleotide polymorphisms in rye (*Secale cereale* L.): discovery, frequency, and applications for genome mapping and diversity studies. Theor Appl Genet 114:1105–1116

Varshney RK, Langridge P, Graner A (2007b) Application of genomics for molecular breeding of wheat and barley. Adv Genet 58:122–155

Varshney RK, Mahender T, Aggrawal RK, Börner A (2007c) Genic molecular markers in plants: development and applications. In: Varshney RK, Tuberosa R (eds) Genomics-Assisted Crop Improvement, Vol I: Genomics Approaches and Platforms. Springer, The Netherlands, pp 13–30

Varshney RK, Nayak S, Jayashree B, Eshwar K, Upadhyaya HD, Hoisington DA (2007d) Development of cost-effecive SNP assays for chickpea genome analysis and breeding. J SAT Agric 3:1. http://www.icrisat.org/Journal/chickpea_pigeonpea3.htm

Varshney RK, Tuberosa R (2007) Genomics-assisted crop improvement: an overview. In: Varshney RK, Tuberosa R (eds) Genomics-Assissted Crop Improvement, Vol I: Genomics Approaches and Platforms, Springer, The Netherlands, pp 1–12

Varshney RK, Nayak SN, May GD, Jackson SA (2009) Next generation sequencing technologies and their implications for crop genetics and breeding. Trends in Biotech, in press

Wang HH (2005) Application of ecoTILLING to relate molecular variation in a rice ethylene response factor (ERF3) gene to drought stress responses. Los Baños, Laguna, Philippines.86 I

Wang L, Liu S, Niu T, Xu X (2005) SNPHunter: a bioinformatic software for single nucleotide polymorphism data acquisition and management. BMC Bioinformatics 6:60

Warthmann N, Fitz J, Weigel D (2007) MSQT for choosing SNP assays from multiple DNA alignments. Bioinformatics 23:2784–2787

Wangkumhang P, Chaichoompu K, Ngamphiw C, Ruangrit U, Chanprasert J, Assawamakin A, Tongsima S (2007) WASP: a Web-based Allele-Specific PCR assay designing tool for detecting SNPs and mutations. BMC Genomics 8:275

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872–876

West MAL, Kim K, Kliebenstein, van Leeuwen H, Michelmore RW, Doerge RW, St Clair DA (2006a) Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in *Arabidopsis*. Genetics 175:1441–1450

West MAL, van Leeuwen H, Kozil A, Kliebenstein DJ, Doerge RW, St Clair DA, Michelmore RW (2006b) High-density haplotyping with microarray-based expression and single feature polymorphism markers in *Arabidopsis*. Genome Res 16:787–795

Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ, Davis RW (1998) Direct allelic variation scanning of the yeast genome. Science 281:1194–1197

Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Züchner S, Hauser MA (2005) SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics 21:4181–4186

Zhang R, Zhu Z, Zhu H, Nguyen T, Yao F, Xia K, Liang D, Liu C (2005) SNP Cutter: a comprehensive tool for SNP PCR-RFLP assay design. Nucleic Acids Res 33:W489–492