General and Applied Plant Physiology – 2010, Volume 36 (3–4), pp. 159–166

Available online at http://www.bio21.bas.bg/ipp/

SINGLE NUCLEOTIDE POLYMORPHISM (SNP): A TREND IN GENETICS AND GENOME ANALYSES OF PLANTS

Mandaliya V. B., R. V. Pandya, V. S. Thaker*

Centre for advanced studies in plant biotechnology and genetic engineering, Department of Biosciences, Saurashtra University, Rajkot 360 005, India

Telephone No.: +91 281 2586419

Received: *17 December 2008* Accepted: *14 September 2010*

Summary. SNPs are sequence variations that occur most frequently in the plant genome and have been used as molecular markers in a wide range of studies. Based on its position in plant genome, SNP may have structural and functional effects. For finding the SNP position in plant genome, over the last 20 years, researchers have developed a number of genome scans techniques. The outcomes of this genome scans are documented in the form of databases worldwide. These databases are the resources for various genomic and evolutionary studies. Finally, SNP analysis provides a significant view towards structural and functional genomics and genetic improvement in the plant based future studies.

Key words: Molecular markers; SNP; SNP application; SNP database.

Abbreviations: SNP-single nucleotide polymorphism; PCR-polymerase chain reaction; RFLP-restriction fragment length polymorphism; AP-PCR-arbitrarily primed polymerase chain reaction; CAPS-cleaved amplified polymorphic sequence; STS-Simple Tag Sequence; RAPD-random amplified polymorphic DNA; SCAR-sequence characterized amplified regions; AFLP-amplified fragment length polymorphism; SSAP-sequence-specific amplification polymorphism; SSR-simple sequence repeat; EST-expressed sequence tags.

INTRODUCTION

Numerous studies have demonstrated that crops exhibit great phenotypic and genomic variability. In order to exploit this diversity an efficient genetic marker system is required (Mace et al., 2008). Genetic markers fall into one of the three broad classes: those based on visually assessable traits (morphological and agronomic traits), those based on gene products (biochemical markers), and those relying on a DNA assay (molecular markers) (Semagn et al., 2006). Molecular markers entered a new exciting and progressive era with the promise to

*Corresponding author: casprogramme@gmail.com

significantly enhance efficiency of plant genetics and breeding research. Molecular markers, which are phenotypically neutral and literally unlimited in number, have allowed scanning of the whole genome and assigning landmarks in high density on every chromosome in many plant species (Khlestkina and Salina, 2006). During the past two decades, different types of molecular markers have been developed, evolved, applied to studying patterns of genetic diversity in the various genomic studies. The development of molecular marker techniques over the last two decades (Agarwal et al., 2008) were schematically presented in Fig. 1. According to a broad classification of molecular markers, they are divided into three classes (Khlestkina and Salina, 2006): (i) Hybridization-based molecular markers (e.g., RFLP); (ii) PCR-based molecular markers (e.g., AP-PCR, CAPS, STS, RAPD, SCAR, AFLP, SSAP, SSR, ISSR, EST); and (iii) DNA chip and sequencing-based molecular markers (SNP). RFLP is the most widely used hybridization-based molecular marker. It was initially used for human genome mapping (Botstein et al., 1980), and later adopted for plant genomes (Weber and Helentjaris, 1989). The major strength of RFLP markers are high reproducibility, codominant inheritance, good transferability between laboratories, no sequence information required. There are, however, several limitations for RFLP analysis: it requires the presence of high quantity and quality of DNA (Deborah et al 1991), it is time consuming, laborious, and expensive, usually requires radioactively labeled probes, the level of polymorphism is low, and few loci are detected per assay. PCR is a molecular biology technique for enzymatically replicating (amplifying) small quantities of DNA without using a living organism (Bhatnagar and Khuran, 2003). The major advantages of PCR techniques compared to hybridization-based methods are as follows: a small amount of DNA is required, elimination of radioisotopes in most techniques, no prior sequence knowledge is required for many applications. However, PCR-based markers have limitations such as reproducibility, dominant inheritance, and homology. Public accessibility to the genome sequences has enabled the study

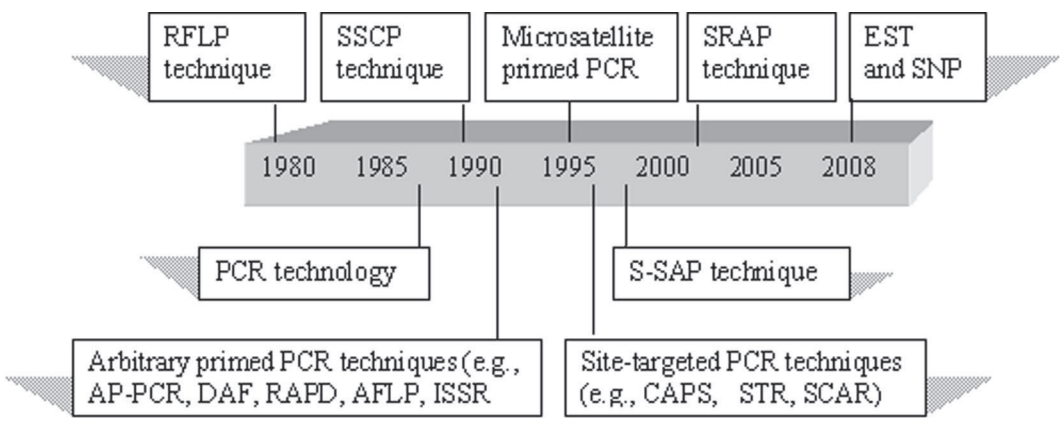

Fig. 1. Schematic presentation showing the development of major molecular marker techniques over last two decades.

of sequence variations between cultivars and subspecies. These studies revealed that single nucleotide polymorphisms (SNPs) are highly abundant and distributed throughout the genome in various species including plants (Batley et al., 2003a). The abundance of these polymorphisms in plant genomes makes the SNP marker system an attractive tool for mapping, marker-assisted breeding and map-based cloning (Batley et al., 2003b). SNPs (sometimes pronounced 'snips') is any of the variations in single nucleotides in a DNA sequence. The genetic code is specified by the four nucleotide "letters" A (adenine), C (cytosine), T (thymine), and G (guanine). SNP variation occurs when a single nucleotide, such as A, replaces one of the other three nucleotide letters - C, G, or T. For example, a SNP might change the DNA sequence AA**G**GCTAA to A**T**GGCTAA (Khlestkina and Salina, 2006). According to the definition given by Brookes (1999) "SNPs are single base pair positions in genomic DNA, at which different sequence alternatives (alleles) exist in normal individuals in some population(s), herein the least frequent allele has an abundance of 1% or greater". For example, by comparing sequences from a japonica rice cultivar to those from an indica cultivar, Yu et al. (2002) identified, on average, one SNP every 170 bp. SNPs are not mutations. Mutations are differences in DNA sequence in an individual that are rare, and may be unique to the individual (or their family line). Polymorphisms are differences in DNA sequence that are found in many individuals, at a specified frequency (usually 1% or greater of a population). Polymorphisms start as mutations, but if they become "fixed" in the population, and achieve sufficient frequency, they become polymorphisms (Kimchi-Sarfaty et al., 2006). On the basis of their position in a genome and their function effects they are classified by various modes of classification. Depending on the SNP positions in a gene, SNPs are classified into noncoding SNP and coding SNP (Yuan et al., 2006). Coding SNPs can be further subdivided into two groups: (i) Synonymous: when single base substitutions do not cause a change in the resultant amino acid; (ii) Non-synonymous: when single base substitutions cause a change in the resultant amino acid. Furthermore, depending on the availability of data on the functional effect of the single-nucleotide substitutions, they are classified into anonymous SNPs (functional effect is unknown), candidate SNPs (presumably having a functional effect), and protein SNPs (single-nucleotide substitutions, resulting in a change in the protein function or expression). Recent studies have shown that SNPs may have functional effects on the following (Yuan et al., 2006): (i) Protein structures, by changing single amino acids; (ii) Transcriptional regulation, by affecting transcription factor binding sites in promoter or intronic enhancer regions; and (iii) Alternative splicing regulation, by disrupting exonic splicing enhancers or silencers.

Main stretegies of SNP genime scans.

Without *a priori* knowledge of DNA polymorphisms finding single nucleotide changes in the any genome seems like a daunting prospect, but over the last 20 years, researchers have developed a number of techniques that make it possible to do that. The following major techniques

enable to find SNP in the any genome (Kwok and Chen, 2003):

1. Denaturing Gradient Gel Electrophoresis (DGGE): DGGE takes advantage of the fact that denaturation of double-stranded DNA is highly dependent on its sequence (Fischer and Lerman, 1983). Since the electrophoretic mobility of a partially-open DNA molecule is greatly retarded, a DNA fragment traversing down a gel stops at the point where one of its ends begins to melt. Therefore, DNA molecules with differences in the low melting domain will have different final positions in the gel.
2. Chemical Cleavage of Mismatch (CCM): A rational strategy for mismatch detection was developed by exposing the mixture of reannealed DNA fragments to the oxidants to modify the cytosines and thymines at the mismatched sites. The chemically modified base is then cleaved by piperidine and the point of mismatch can be ascertained by sizing the cleavage product by gel electrophoresis (Hansen et al., 2006).
3. MutS Protein-binding Assays: The *E. coli* MutS protein recognizes and binds to heteroduplex DNA with up to 3 mismatched bases in a row. A binding assay utilizes MutS protein immobilized on magnetic beads to capture heteroduplex DNA labeled with biotin that is in turn detected by an enzyme-linked immunosorbent assay (ELISA), thereby increasing its sensitivity (Lishanski et al., 1994).
4. Denaturing High Performance Liquid Chromatography (DHPLC): In this method, instead of using a gel and separating the DNA fragments by electrophoresis, a modified resin and HPLC are employed for fragment analysis. When the DNA fragments are separated at elevated temperatures, partial melting occurs and the heteroduplex DNA containing mismatches will have a different retention time than the homoduplex DNA. Because HPLC is a robust technology and autosampling is used routinely, DHPLC is a very simple method to implement (Huber et al., 2001).
5. Direct DNA Sequencing: Until very recently, direct DNA sequencing was laborious and expensive. The presence of polymorphisms is represented by missing or additional bands in the sequencing ladder. The greatest advantage in SNP detection by direct DNA sequencing is the complete information it yields (Hanke and Wink, 1994).

Targeted SNP discovery is still at the stage of scanning relatively small segments of DNA one at a time. Local target, SNP discovery relies mostly on direct DNA sequencing or on denaturing high performance liquid chromatography (dHPLC). Either DNA sequencing becomes a lot cheaper and easier to do, or some new approach must be developed to allow for local SNP discovery on the hundred-kilobase to megabase scale (Kwok and Chen, 2003).

Major databases for SNP analysis.

With the completion of the Human Genome Project, a large number of subtle variations (polymorphisms) among the population have been found. The most abundant type of these variations is the single nucleotide polymorphisms (SNPs),

but currently contains limited annotation information (Karchin et al., 2005). Table 1 (http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi., lastly updated on April 14, 2008, accessed on 15th Nov. 2008) shows SNP reported in NCBI databases and Table 2 shows 8 of 36 popular SNP databases associated to plant varieties.

Table 1. dbSNP Public database.

Organism	No. of Submissions	No. of RefSNP Clusters (# validated)
Salmo salar	755	755 (0)
Cooperia oncophora	426	426 (96)
Ficedula albicollis	37	37 (15)
Ficedula hypoleuca	28	20 (10)
Bison bison	6	6 (2)
Saccharum hybrid cultivar	42853	42853 (0)
Oryza sativa	5872081	5418373 (22057)
Pinus pinaster	1439	32 (0)
Glycine max	281	278 (234)
Arabidopsis thaliana	301	184 (184)
Zea mays	148	146 (80)
Allium cepa	45	50 (0)

Major applications of SNP markers.

Single nucleotide polymorphism (SNP) analysis provides:

1. SNP-based markers introduce polymorphisms that are easy to database, and significantly increase the density of the genetic linkage map, e.g. Grapevine (*Vitis vinifera* L.) (Troggio et al., 2007).
2. A useful tool to quantify linkage disequilibrium (LD), e. g. sunflower elite inbred lines. This includes the estimation of nucleotide diversity, the assessment of linkage disequilibrium structure (LD) and the evaluation of selection processes (Fusari et al., 2008).
3. Fine-mapping of candidate regions and determination of haplotypes associated with traits of interest, e. g. sugar beet alleles of expressed genes are very frequently organized as robust intragene haplotypes (Schneider et al., 2001).
4. In order to understand the genetic basis of phenotypic diversity within

and between populations, e. g. analysis of phenotypic diversity in tomato (Deynze et al., 2007).

5. SNP are useful for evolutionary studies, e. g. elite maize inbred lines. The genetic distance between haplotypes is large and indicative of an ancient gene pool and possible interspecific hybridization events in maize ancestry (Ching et al., 2002).

SNPs are highly stable, diallelic in populations, and their allele frequencies can be estimated easily in any population. Many technologies have been developed to type SNPs in an automated fashion, and many of these yield simple positive or negative outcomes that can be interpreted easily by a computer. SNPs are studied worldwide and documented in the form of databases. These databases are the

Table 2. SNP Database.

No	Database	Web Link	Usage	Details
1	dbSNP	http://www.ncbi.nlm.nih.gov/	Provides the location of a SNP in a gene and its alleles, allele frequency, and context sequence	More than 6 million validated SNPs
2	HapMap	http://hapmap.org/cgi-perl/gbrowse	Provides information about the haplotype and linkage disequilibrium around a SNP	More than 1 million SNPs
3	HGVbase	http://hgvbase.cgb.ki.se/	Provide the publicly available database	2.8 million SNPs
4	GVS	http://gvs.gs.washington.edu/GVS	SeattleSNPs Program for Genomic Applications (PGA)	4.5 million SNPs
5	JSNP	http://snp.ims.u-tokyo.ac.jp/	To identification of disease-related genes	197,000 SNPs
6	SAAP	http://www.bioinf.org.uk/saap/Brainarray/Database/SearchSNP/snpfunc.aspx	It maps individual updated protein residues in the PDB automatically	2384 protein structure data
7	Plant Markers	http://markers.btk.fi	To identify putative SNP, SSR and conserved orthologue set markers	Screening from over 50 plant species
8	rSNP_Guide	http://wwwmgs.bionet.nsc.ru/mgs/systems/rsnp/	For analysis of transcription factor binding to target sequences	Contains 46 entries

resources for various genomic and evolutionary studies. Furthermore, SNP analysis provides the most comprehensive view of the plant genome reported to date and will be relevant for future studies on structural and functional genomics and genetic improvement.

Acknowledgments: The authors would like to thank the State Government of Gujarat for financial support for Centre for Advanced Studies in Plant Biotechnology and Genetic Engineering (CPBGE) programme.

REFERENCES

Agarwal M, N Shrivastava, H Padh, 2008, Advances in molecular marker techniques and their applications in plant sciences, Plant Cell Rep, 27: 617–631.

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

Batley J, R Mogg, D Edwards, HO'Sullivan, KJ Edwards, 2003b, A high-throughput SNuPE assay for genotyping SNPs in the flanking regions of *Zea mays* sequence tagged simple sequence repeats. Mol Breed, 11: 111–120.

Bhatnagar S, P Khurana, 2003, *Agrobacterium tumefaciens*-mediated transformation of Indian mulberry, *Morus indica* cv. K2: a time-phased screening strategy. Plant Cell Rep, 21: 669–675.

Botstein D, RL White, M Skolnick, RW Davis, 1980, Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet, 32: 314 331.

Brookes A, 1999, The Essence of SNPs. Gene, 234: 177–186.

Ching A, KS Caldwell, M Jung, M Dolan, OS Smith, S Tingey, M Morgante, AJ Rafalski, 2002, SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet. 3: 19, PubMed ID: 12366868

Deborah L, A Sieglinde, R Shoemaker, PM Gresshoff, 1991, The genetic locus controlling supernodulation in soybean (*Glycine max* L.) co-segregates tightly with a cloned molecular marker, Mol Gen Genet, 228: 221–226.

Deynze AV, K Stoffel, CR Buell, A Kozik, J Liu, E van der Knaap, D Francis, 2007, Diversity in conserved genes in tomato. BMC Genomics, 8: 465, Doi: 10.1186/1471-2164-8-465.

Fischer SG, LS Lerman, 1983, DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. Proc Natl Acad Sci. USA, 80: 1579–1583.

Fusari C M, VV Lia, HE Hopp, RA Heinz, NB Paniego, 2008, Identification of Single Nucleotide Polymorphisms and analysis of Linkage Disequilibrium in sunflower elite inbred lines using the candidate gene approach. BMC Plant Biol, 8: 7.

Hanke M, M Wink, 1994, Direct DNA sequencing of PCR-amplified vector inserts following enzymatic degradation of primer and dNTPs. BioTechniques, 17: 858–860.

Hansen LL, J Just, A Lambrinakos, RGH Cotton. 2006, Mutation detection by solid-phase chemical cleavage of mismatches using radioactive probes. Cold Spring Harb Protoc, doi:10.1101/pdb.prot4120.

Huber CG, A Premstaller, W Xiao, H Oberacher, GK Bonn, PJ Oefner, 2001, Mutation detection by capillary denaturing high-performance liquid chromatography using monolithic columns. J Biochem Biophys Methods, 47: 5–19.

Karchin R, M Diekhans, L Kelly, DJ Thomas, U Pieper, N Eswar, D Haussler, A Sali, 2005, LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics, 21: 2814–20.

Khlestkina EK, EA Salina, 2006, SNP Markers: Methods of analysis, ways of development, and comparison on an example of common wheat. Russ J Genetics, 42: 585–594.

Kimchi-Sarfaty C, OJ Mi, I Kim, ZE Sauna, AM Calcagno, SV Ambudkar, MM Gottesman, 2006, A "silent" polymorphism in the *MDR1* gene changes substrate specificity. *Science Express, Published on Dec.* 21, 2006.

Kwok, P., X. Chen, 2003, Detection of single nucleotide polymorphisms. Curr Iss Mol Biol, 5: 43–60.

Lishanski A, EA Ostrander, J Rine, 1994, Mutation detection by mismatch binding protein, MutS, in amplified DNA: application to the cystic fibrosis gene, Proc. Nat. Acad. Sci. U.S.A., 91: 2674–2678.

Mace ES, L Xia, DR Jordan, K Halloran, DK Parh, E Huttner, P Wenzl, A Kilian, 2008. DArT markers: diversity analyses and mapping in *Sorghum bicolor*. BMC Genomics, 9, 26, Doi: 10.1186/1471-2164-9-26.

Schneider K, B Weisshaar, DC Borchardt, F Salamini, 2001, SNP frequency and allelic haplotype structure of *Beta vulgaris* expressed genes, Mol Breed, 8: 63-74.

Semagn K, Å Bjørnstad, MN Ndjiondjop, 2006, An overview of molecular marker methods for plants. African J Biotech, 5: 2540-2568Troggio M, G Malacarne, G Coppola, C Segala, DA Cartwright, M Pindo, M Stefanini, R Mank, M Moroldo, M Morgante, MS Grando, R Velasco, 2007, A dense single-nucleotide polymorphism-based genetic linkage map of grapevine (*Vitis vinifera* L.) anchoring pinot noir bacterial artificial chromosome contigs, Genetics, 176: 2637–2650.

Weber D, T Helentjaris, 1989, Mapping RFLP loci in maize using B-A translocations. Genetics, 121: 583–590.

Yu J, S Hu, J Wang, G K Wong, et al. 2002, A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science, 296, 79–92.

Yuan H, J Chiou, W Tseng, C Liu, C Liu, Y Lin, H Wang, A Yao, Y Chen, C Hsu, 2006, FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res, 34: 635–641.