# Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants

**P. K. Gupta*, J. K. Roy and M. Prasad****

Molecular Biology Laboratory, Department of Agricultural Botany, Ch. Charan Singh University, Meerut 250 004, India
**Present address: IPK, Corrensstrasse 3, D-06466 Gatersleben, Germany

**Molecular markers are useful for a variety of purposes relevant to crop improvement. The most important of these uses is the indirect marker-assisted selection (MAS) exercised during plant breeding. For this purpose, molecular markers need to be amenable to automation and high throughput approaches. However, the gel-based assays that are needed for most molecular markers are time consuming and expensive, limiting their utility. The new generation molecular markers, called single nucleotide polymorphisms (SNPs) do not always need these gel-based assays. They are also the most abundant of all marker systems known so far, both in animal and plant genomes. A large number of SNPs have already been developed in the human genome, some of them proving useful for diagnosis of diseases. A beginning has also been made in the development and use of SNPs in higher plants, including some crop and tree species. Hopefully in future, they will be used in plants in a big way. Several approaches can be used for discovery of new SNPs and about a dozen different methods are now available for SNP genotyping. Some of these methods are also suitable for automation and high throughput approaches. These methods, in principle, make a distinction between a perfect match and a mismatch (at the SNP site) between a probe of known sequence and the target DNA containing the SNP site. The target DNA in most of these methods is a PCR product, except in some cases like 'invasive cleavage assay', and 'reduced representation shotgun (RRS)' devised and used recently. The different methods of SNP discovery and detection, along with examples of some known uses of SNPs in plant systems are described in this article.**

IN recent years, plant breeding has witnessed a revolution due to emergence of molecular breeding, a subject which deals with all aspects of plant molecular biology, having applications in crop improvement programmes. Molecular breeding comprises two major areas, the transgenic crops and the molecular marker technology. While remarkable success has already been achieved in the production and commercialization of transgenic crops, molecular marker technology has yet to find its rightful place in plant breeding programmes. Molecular marker technology is user-friendly and no biosafety or bioethics questions of the kind raised against transgenic crops, have been raised against this technology. In fact, the lack of cost-effectiveness and non-availability of high throughput approaches for handling large segregating populations have limited the use of molecular marker technology for plant breeding. In recent years, there has been emphasis on the development of newer and more efficient molecular marker systems involving inexpensive non gel-based assays with high throughput detection systems. The impact of these developments will certainly be felt in the next few years. Availability of single nucleotide polymorphisms (SNPs) is one such development that will be addressed in this article.

A series of different molecular marker systems, which became available during the last two decades, can be broadly classified into three classes (for acronyms used, see Table 1): (i) the first generation molecular markers, including RFLPs, RAPDs and their modifications; (ii) the second generation molecular markers, including SSRs, AFLPs and their modified forms, and (iii) the third generation molecular markers including ESTs and SNPs. Since ESTs are mainly used for studies on functional genomics, SNPs are the only new generation molecular markers for individual genotyping needed for molecular marker-assisted selection (MAS). In view of a causal similarity of SNPs with some of these marker systems and the fundamental difference with several other marker systems, the molecular markers have also been classified into SNPs (due to sequence variation, e.g. RFLP) and non-SNPs (due to length variation, e.g. SSRs). There is some evidence that the stability of SNPs and, therefore, the relative fidelity of their inheritance is higher than that of the other marker systems

*For correspondence. (e-mail: pkgupta@ndf.vsnl.net.in)

**Table 1.** List of acronyms used in this article

| |
|---|
| AFLP – amplified fragment length polymorphism |
| GBA – genetic bit analysis |
| DASH – dynamic allele-specific hybridization |
| EST – expressed sequence tag |
| OLA – oligonucleotide ligation assay |
| RAPD – random amplified polymorphic DNA |
| RFLP – restriction fragment length polymorphism |
| SNP – single nucleotide polymorphism |
| SSCP – single strand conformation polymorphism |
| SSR – simple sequence repeat |
| TMHA – temperature modulated heteroduplex analysis |
| TMHC – temperature modulated heteroduplex chromatography |
| MALDI-TOF MS – matrix-assisted laser desorption ionization-time of flight mass spectroscopy |
| RRS – reduced representation shotgun |

like SSRs and AFLPs. SNPs, at a particular site in a DNA molecule should in principle involve four possible nucleotides, but in actual practice only two of these four possibilities have been observed at a specific site in a population[1]. Consequently, SNPs are biallelic as against the polyallelic nature of the once much sought after SSRs. However, the extraordinary abundance of SNPs largely offsets the disadvantage of their being biallelic and makes them the most attractive molecular marker system developed so far. For instance, at least in some parts of the human genome, as well as in the genomes of several crop species, the frequency of SNPs has been shown to be an order of magnitude higher than that of SSRs[2]. According to most recent estimates, one SNP occurs every 100–300 bp in any genome (according to several other estimates, they occur once every 1000 bp), thus making SNPs the most abundant molecular markers known so far. The public database for SNP (dbSNP), that was created a few years ago, already had 8,03,557 SNPs in August 2000, which grew to 25,58,364 in November 2000 (release on 8 December 2000), suggesting rapid growth in SNP database. For human genome, ~300,000 SNPs are already available in public domain and at least another 500,000 (including SNPs from the expressed region of the genome, i.e. copy SNPs or cSNPs) are available with the private companies like CuraGen[3]. Despite this, the enthusiasm for human SNPs seems undiminished, so that the SNP Consortium (TSC) had plans to produce half a million SNPs by the end of the year 2000 (~150,000 SNPs were already produced by March/April 2000), and Japanese Government planned to produce another 200,000 SNPs latest by the year 2002 (ref. 4). It is suggested that these activities may lead to the construction of human SNP map having up to 500,000 to 1,000,000 SNPs (estimates differ according to different workers and organizations), with a density of one SNP every 3–5 kb of the genome. However, according to some recent estimates, 30,000 SNPs should be adequate for scanning the human genome[5,6].

During the last few years, SNPs have also attracted the attention of plant molecular biologists, and one may hope that in future SNP maps will be prepared and used extensively in many plant systems. In plant systems, the SNPs seem to be more abundant than even those in the human genome, so that in preliminary studies conducted in wheat, one SNP per 20 bp (ref. 7), and in the maize genome, one SNP per 70 bp (ref. 8) have been recorded in certain regions of these genomes (see later, for more details). One can only hope that SNPs will be developed expeditiously in all major crops in a large scale and will be extensively utilized in future for a variety of crop improvement programmes, although non-availability of adequate sequence data may limit this activity. In this article, we propose to discuss the methods used for development and detection of SNPs and their major uses relevant to crop improvement programmes. We hope that the information will be of interest not only to the general readers, but also to specialists working in the area of molecular markers. Despite the fact that SNPs in plants in most cases were studied through techniques that were first developed and used for human SNPs, the work on human SNPs will not be covered in this article in any detail, except for the purpose of reference.

## Significance of SNPs

One of the several key issues in the study of heredity and variations at the molecular level is the detection of associations between DNA sequence variation and the heritable phenotypes. A variety of molecular markers have been successfully used for detection of their associations with major traits of economic value in several crop plants[9–11], although very few uses of these associations for MAS have been made by the practising plant breeders. Perhaps soybean cyst nematode (SCN) resistance is one such solitary example, where molecular markers are being used in actual selection during breeding for superior soybean cultivars[12]. This is partly because, the success of MAS in the hands of plant breeders depends on the development of molecular marker systems, that are not only more efficient and cost-effective, but are also amenable to automation and high throughput approaches to handle large segregating populations. SNPs have proved ideal for this purpose, although relative to biomedical research, their adoption for plant breeding has been rather slow. In view of the significance that is being attached to this marker system, 12,400 web pages for SNPs were already available on internet when a search was conducted on 12 January 2001. 'The SNP Consortium (TSC) Ltd' for human SNPs, led by 10 pharmaceutical companies and The Welcome Trust, started in April 1999 and is anticipated

to continue until the end of 2001. Two IT (information technology) companies, namely Motorola and IBM also recently joined TSC. In this connection, the three annual 'International Meetings on SNPs and Complex Genome Analysis' held in 1998, 1999 and 2000 are also relevant, although at these meetings the major thrust was on biomedical research, there being hardly any discussion on SNPs for plant breeding. In USA, the National Center for Biotechnology Information (NCBI), in collaboration with National Human Genome Research Institute (NHGRI) has also established the database dbSNP to serve as a central repository both for SNPs and *indels* (insertions and deletions). As mentioned earlier, this database already has 25,58,364 SNPs in its collection, and more are being added every day. This extensive data on SNPs will be integrated with other NCBI genomic data and will be freely available to the scientific community in a variety of forms.

SNPs may be found both in the non-repetitive coding or regulatory sequences and in the repetitive non-coding sequences. When present in the coding sequences, they may or may not determine the mutant phenotype, but will show 100% association with the trait and will therefore, be very useful, both for MAS and for gene isolation. When found in the proximity of coding sequences, although the association of SNPs with traits will be less than 100%, the association with the economic traits will still allow their use in MAS, and also in positional cloning. In still other cases, these may be away from any gene and therefore may not prove useful. Consequently, one may need to discover many more SNPs, than the number really needed. At this time, it is difficult to guess the proportion of SNPs that will be of immediate use.

## Methods used for discovery and identification of new SNPs

RFLPs, RAPDs and SSRs which were the markers of choice during the last two decades, need gel-based assays and are, therefore, time consuming and expensive. Therefore, emphasis is now shifting towards the development of molecular markers, which can be detected through non gel-based assays. One of the most popular of these non gel-based marker systems is SNP, which represents sites, where DNA sequence differs by a single base. This polymorphism has been shown to be the most abundant, so that at least one million SNPs should be available, only in the non-repetitive transcribed region of the human genome[6,13]. Some of the methods available for the discovery of these SNPs are described in this section, although rarely one or more methods described later for SNP genotyping can also be used for SNP discovery (e.g., Masscode™ genotyping/discovery system of QIAGEN Genomics).

### *Locus specific-PCR amplification*

In this approach, locus specific PCR primers are synthesized from the available genomic sequences and PCR amplification is undertaken using DNA samples from several individuals. PCR amplified products are sequenced and sequence differences are used for discovery of new SNPs. Since prior sequence information is necessary, the method can be used only for genomic regions with known sequences.

### *Alignment among available genomic sequences*

The development of SNPs involving non gel-based assays has recently been facilitated by the availability of genome-wide sequences and EST databases. Alignment of genomic and EST sequences is the most convenient method for discovery of SNPs, if genomic/EST sequences from a heterozygote or from more than one individuals of the same species are available in the databases. The alignment of sequences is automated through computer software, and will allow detection of SNPs in a cost-effective manner. However, a comparison of genomic sequences will detect SNPs, both in coding and non-coding regions, while that of ESTs will detect SNPs only in the coding region. For instance, the software SNP pipeline was recently utilized for the identification of ~3000 candidate SNPs from human EST data[14]. Sequences experimentally obtained from a shotgun library may also be aligned to available genomic sequences to discover new SNPs.

The above method can also be used for detection of SNPs at the regional level within the genome, when trait related SNPs need to be discovered and the gene(s) for the trait are already sequenced and mapped in specific regions of the chromosome. In still other cases, genes for specific traits or diseases are identified using genome-wide candidate gene approach and can be used for SNP discovery. This approach has been successfully used to discover SNPs for a number of genes responsible for human diseases.

### *Whole genome shotgun sequences*

In this approach, random clones from the genomic library prepared from a mixture of DNA from several individuals are sequenced. This should require several fold coverage of the whole genome before SNPs can be detected by alignment of sequences belonging to the same locus.

### *Overlapping regions in BACs and PACs*

This is one of the common methods for the detection of SNPs in organisms that have been used for genome se-

quencing (e.g. humans, *Arabidopsis*, rice). Since the overlapping sequences from BACs/PACs (bacterial/P1 artificial chromosomes) may be derived from genomes of different individuals, an SNP in the overlapping region can be detected by a mismatch[15].

### *Reduced representation shotgum*

Sometimes genomic sequences may not be available, or it may not be desirable to use the available genomic sequences for the discovery of SNPs. In such cases, sequencing and SNP detection can be done in parallel, as recently done using reduced representation shotgun (RRS) approach. This approach uses subsets of genome, each containing manageable number of loci to permit resampling. For this purpose the RRS approach involves the following steps: (i) equimolar quantities of genomic DNA from several unrelated individuals is mixed, digested to completion with an appropriate enzyme and electrophoresed; (ii) a narrow band (500–660 bp) is excised and utilized to make a partial genomic library, whose clones are sequenced; (iii) sequences representing the same loci are aligned to discover candidate SNPs, which are verified by resequencing. Using RRS approach in two recent studies, 47,172 and 65,000 human SNPs were discovered without using genomic sequences available in the databases[16,17]. A map for human chromosome 22 with 2730 SNPs could also be prepared by this approach using the DNA extracted from many copies of the flow sorted chromosome 22.

## Methods used for genotyping individuals at SNP loci

A number of methods for SNP genotyping are now available, although all of them may not be equally useful. The approach used for this purpose, relies on the ability to distinguish a perfect match from a single base mismatch. The instrumentation and the techniques like high density oligonucleotide arrays on DNA chips, MALDI-TOF MS detection system and pyrosequencing, that recently became available, will allow accurate genotyping of individual plants at a large number of these biallelic loci in parallel, since it requires only plus/minus assay, permitting easier automation. The assays used for SNP genotyping can be broadly classified into two groups, the gel-based assays and the non gel-based assays, the latter being preferred in most laboratories to economize on time and money.

### *Gel-based assays for SNP detection*

*RFLP and AFLP-based assay:* The presence of SNPs can be detected by RFLP or AFLP conducted on PCR products, whenever such SNP generates or destroys a specific restriction site for an enzyme. The PCR product in this method is subjected to restriction digestion with individual enzymes, and then used either for RFLP or AFLP to detect differences in patterns, which will be due to changes in the restriction sites. This method of the detection of SNPs still needs gel-based assays, and therefore, has recently been taken over by several other methods involving non gel-based assays, discussed in the next section.

*Single-strand conformation polymorphism:* Single-strand conformation polymorphism (SSCP) technology allows detection of polymorphism due to differences of one or more base pairs in the PCR products and is therefore suitable for SNP detection. The technique 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. In order to detect SNPs by this method, PCR product carrying the SNP site is denatured and electrophoretically separated in neutral acrylamide gel[18]. Any difference between the wild strain and a mutant genotype will suggest the presence of SNP. The technique has been utilized for detection of SNPs in the genus *Picea* (spruce)[19].

*Allele-specific amplification for SNP genotyping:* SNPs can also be detected by designing allele-specific primers for individual SNP sites. Different fluorochromes are attached at the 5′ ends and the polymorphic nucleotides are attached at the 3′ ends of the two primers. These allele-specific primers, when used for PCR with preamplified DNA as the template, the amplified product will be allele-specific and could be identified either on polyacryamide gel or on an automatic sequencer[20–22]. This approach has been utilized for SNP genotyping in barley[21,22] (see later).

### *Non gel-based assays for SNP detection*

The non gel-based assays will differ depending upon whether the SNP site is at the 3′ end of an amplicon or at an internal position. It will be seen that when the SNP is at the 3′ position of an amplicon, it can be detected simply by the failure of PCR. The detection of SNPs at the internal position of an amplicon, however, requires an initial PCR amplification, followed by a non gel-based assay, that discriminates between wild and mutant alleles. The common non gel-based assays for detection of SNPs at the internal sites are based on the detection of a perfect match and a mismatch between the PCR product and an oligonucleotide used as a probe. Following is a summary of the non gel-based assays that are generally used, but there may be others, which have been used only occasionally.

*Mismatch at 3′ end leading to failure in PCR reaction:* If SNP is present at the 3′ end of an amplicon template, it can be detected simply by the failure of amplification due to mismatch between the primer sequence and the binding site in the template, although it may be difficult to distinguish this failure of PCR due to SNP from PCR failure due to other reasons.

*TaqMan assay:* In an assay described as TaqMan, an oligonucleotide TaqMan probe is labelled with a fluorescent reporter molecule (e.g. FAM or TET) at the 5′ end and a quencher (e.g. TAMRA) at the 3′ end. The probe on hybridization to the template DNA is degraded at its 5′ end due to exonuclease activity of *Taq* polymerase (TaqMan), so that the reporter is released leading to a rise in fluorescence signal. However, when due to the presence of an SNP, the probe mismatches with the template leading to failure in duplex formation, no such degradation at the 5′ end of the probe is possible and there is no rise in fluorescence signal. Different combinations of reporters and quenchers will also permit multiplexing so that as many as six SNPs can be scored in a single PCR reaction[23].

*Molecular beacons:* Molecular beacons are allele-specific hairpin-shaped oligonucleotide hybridization probes that become fluorescent upon target binding[24]. The probe (molecular beacon) will be specific for the target SNP sequence, and the sequences at its two ends will be complementary to each other. The two ends of the oligonucleotide are labelled just like the oligonucleotide probe used in TaqMan assay. The probe in isolation (when not forming a duplex with the target DNA) generates a hairpin structure due to self-annealing of its two ends, thus quenching the reporter. But when the probe anneals with the template, it gets linearized, thus separating the reporter from the quencher and permitting fluorescence signal (Figure 1). Several molecular beacons, each designed to use a different target and labelled with a different fluorophore, can be used to distinguish multiple targets in the same reaction mix[25]. However, the number of probes that can be used in such a mix is limited since the monochromatic light that is used in most detection systems does not excite all fluorophores equally well. To overcome this problem, wavelength-shifting molecular beacons were recently used, which had, instead of one, two fluorophores (harvester fluorophore and emitter fluoprophore) arranged serially at one end of the beacon, and a quencher on the other end[26]. This enhances the multiplexing capacity of molecular beacon detection assays. The fluorescence signals, both in TaqMan and molecular beacon can be detected by appropriate sensing devices (e.g. MALDI-TOF MS).

*Oligonucleotide ligation assay:* In this approach, two independent probes (one is 5′ biotinylated and the other 3′ fluorescent-labelled) are used for hybridization with PCR product, so that when the probe matches the product, the two probes anneal with the PCR product and undergo ligation resulting in an oligonucleotide which is biotinylated at the 5′ end and fluorescent-labelled at the 3′ end[27]. The ligation product, which is fluorescent-labelled at the 3′ end, is captured on a solid streptavidin-coated matrix due to biotinylation at its 5′ end, and the signal is detected by autoradiography. However, when there is a mismatch due to the presence of SNP, fluorescent-labelled and biotinylated oligonucleotides are unable to ligate, so that the oligonucleotides captured by streptavidin, carry no signal (Figure 2). The OLA technique can also be suitably modified to allow single well genotyping of two alleles or to develop a multiplex typing system[28].

*DNA chips and microarrays:* DNA chips and microarrays of immobilized oligonucleotides of known sequences, which differ at specific sites of individual nucleotides (at the site of SNP) can also be used for

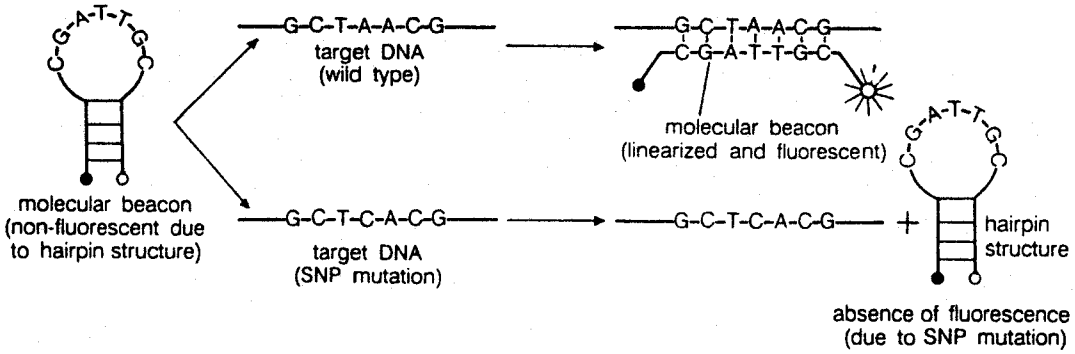


**Figure 1.** Diagrammatic representation of the operation of molecular beacons (see text for details; modified from ref. 25).

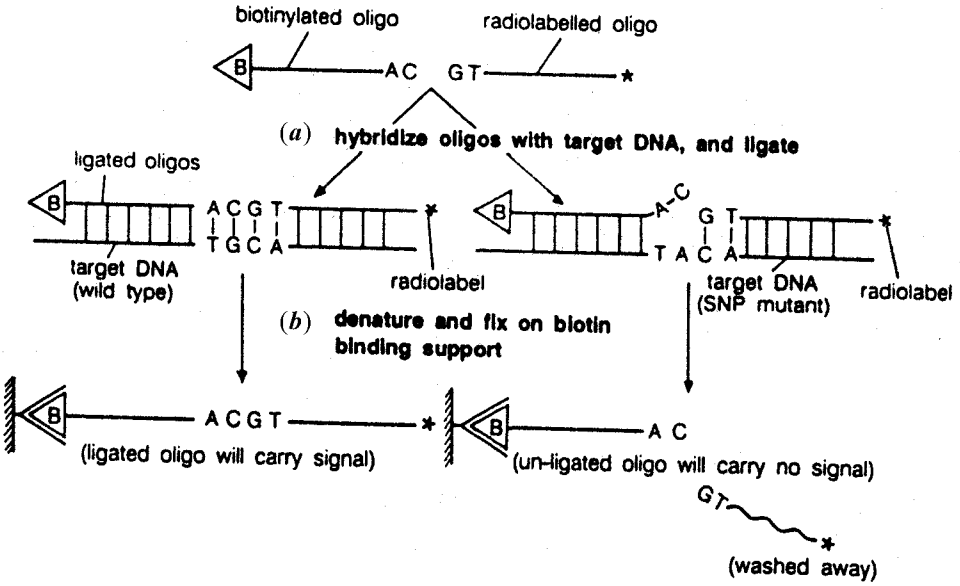


**Figure 2.** Diagrammatic representation of the oligonucleotide ligation assay depicting gene detection through ligation of hybridized oligonucleotide probes (see text for details; modified from ref. 27).

detection of SNPs. The technique is actually suitable to score several SNPs in parallel from each sample in a multiplexed fashion. It makes use of the technique of sequencing by hybridization (SBH) and involves tiling strategy (see Figure 3 in ref. 30). Four oligonucleotides in a column of an array will differ only at the SNP site and only one will be fully homologous. When such an array is hybridized with biotinylated PCR product, the perfect match will allow binding and mismatched products will be washed away. The perfect match in each case can be detected through a detection system (for details see refs 29 and 30).

*Dynamic allele-specific hybridization:* This technique is based on the differences in melting temperatures between duplexes resulting due to perfect match and mismatch between the PCR product and an allele-specific oligonucleotide, 15–21 bases long[31]. The differences in the melting temperatures are detected by a novel approach involving the use of a dye, which intercalates in a duplex DNA molecule and emits fluorescence. For PCR amplification, two primers are used, one of which is biotinylated to allow immobilization of the PCR product on a solid support. The immobilized PCR product is denatured, so that only biotinylated single-stranded DNA is retained on the solid support and the other strand is washed away. The biotinylated single-stranded DNA on the solid support is then hybridized to allele-specific oligonucleotide probe containing the SNP site. The duplex formed with the probe is detected by a low-cost fluorescent intercalating dye specific for double-stranded DNA (Figure 3). The dye emits fluorescence proportionate to the amount of double stranded

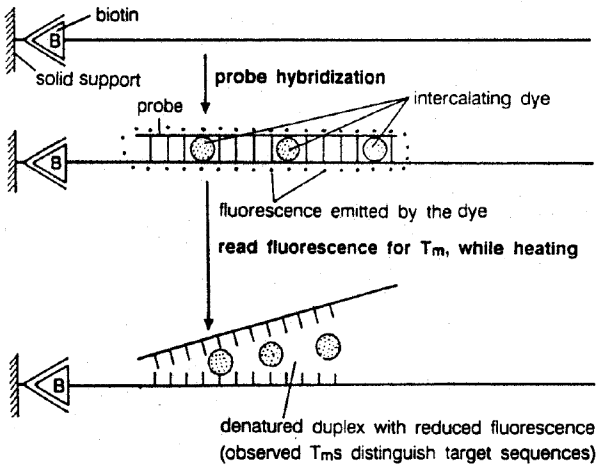


**Figure 3.** Diagrammatic representation of the dynamic allele-specific hybridization, used for SNP genotyping (modified from ref. 31).

DNA produced due to probe–target hybridization. When this hybrid duplex is denatured steadily by increasing temperature, the melting can be followed by the reduction in the above fluorescence. A rapid and sudden fall in fluorescence indicates the melting temperature (Tm) of the duplex. A single base mismatch results in a dramatic lowering of this melting temperature, thus permitting detection of SNP.

*Minisequencing and genetic bit analysis:* SNPs can also be detected by minisequencing (sequencing of a few bases around the SNP site) through Sanger's dideoxynucleotide method, where the oligonucelotide primer has a sequence one or more bases upstream of

the SNP site[32]. A mixture of all the four dNTPs and one of the four possible ddNTPs that corresponds to the SNP locus, is used for primer extension, so that the incorporation of a single ddNTP at the SNP site will terminate the reaction and will allow detection of SNP. Methods have been developed, where extended primers are solid phase purified and detected by any one of the available methods, including radioactivity, colourimetry or MALDI-TOF MS. The detection of extended primer that is available due to SNP mutation will identify the presence of SNP.

In some cases, one-step primer extension is achieved through the use of a primer, which is just upstream of the SNP site, so that incorporation of a single ddNTP will terminate the reaction and will allow the detection of SNP[32]. One such technique, described as genetic bit analysis (GBA), is based on hybridization capture of a single-stranded PCR product to a sequence-specific microplate-bound primer. This is followed by enzyme-mediated single base extension of the captured primer across the polymorphic site, enabling direct determination of SNP through colourimetery[33]. The technique has been used as a diagnostic tool in human paternity tests as well as in pedigree analysis. Among plant systems, the successful use of the technique has been demonstrated in onions for distinguishing plastomes of cytoplasmic male sterile (CMS) and fertile lines, which differ by a single SNP (Figure 4). It has been shown that semi-automated GBA can be conducted using 96-well microtiter plates[34].

*Temperature modulated heteroduplex analysis using dHPLC WAVE™ system:* In several of the above methods of SNP detection, prior knowledge of the DNA sequence in the region surrounding the SNP is needed, which may not be available. Therefore, SNP detection sometimes requires screening for a sequence variant without any *a priori* knowledge of the exact location of a mutation in a given gene. This will be possible if the wild type DNA sample is available for comparison with SNP mutant. For this purpose, denaturing high pressure liquid chromatography (dHPLC) has been used, where each SNP yields a unique chromatography pattern with temperature modulated heteroduplex analysis (TMHA) or temperature modulated heteroduplex chromatography (TMHC) (http://arrayit.com/SNP_Services). The characteristic chromatography pattern can be used not only to identify a novel sequence variant, but also as a diagnostic tool, if an SNP is already characterized. It has been shown that TMHA has extraordinary sensitivity to distinguish heteroduplexes from homoduplexes with perfect accuracy, and it is this property which has been utilized for SNP detection (Figure 5). Transgenomics Inc (San Jose, USA) has developed the dHPLC WAVE™ system which is a fully automated platform for DNA fragment analysis in molecular biology research and diagnostics. This has an autosampler for withdrawing samples, an analytical cartridge oven for maintaining controlled temperature regime of the analytical cartridge in dHPLC system and a UV detector. The use of UV instead of fluorescent labels or radioisotopes for detection makes WAVE™ system extremely cost-effective. A pair of primers is designed to generate a PCR product of up to 800 bp spanning the SNP variant. These PCR products are stored in tubes or wells of 96-well microplates and are used for HPLC loading with the help of autosampler available in the WAVE™ system.

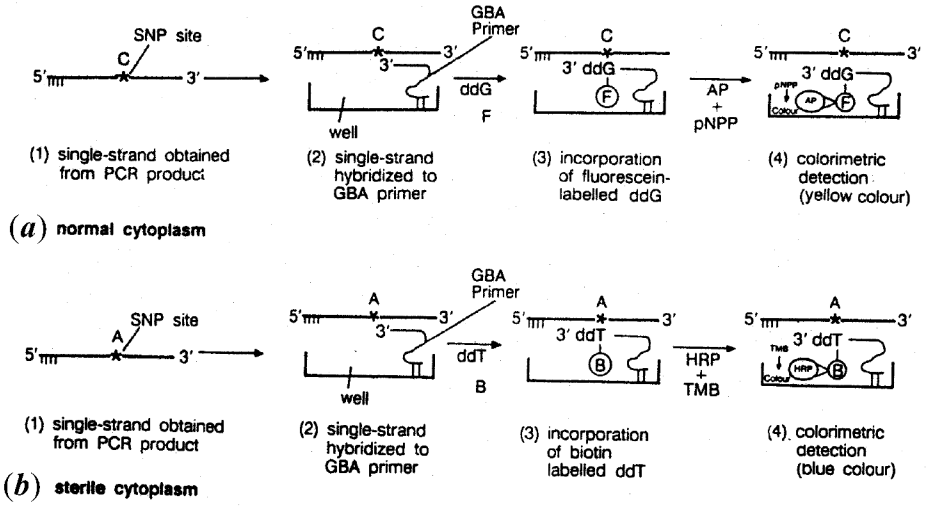


**Figure 4 *a, b*.** Schematic representation of single-nucleotide typing using genetic bit analysis in onion (see text for details; modified from ref. 34).

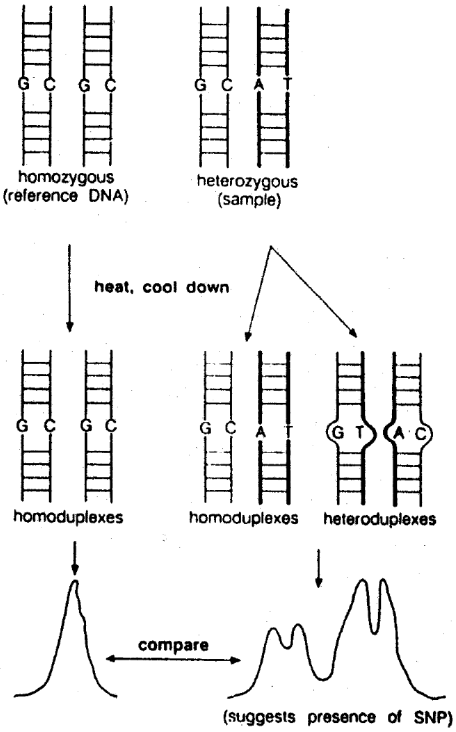


**Figure 5.** Diagrammatic representation of the principle involved in temperature modulated heteroduplex analysis using dHPLC (see text for details; modified from Transgenomics Inc <http://arrayit.com/SNP_Services>).

The WAVE™ system is based on the principle of TMHA for distinguishing heteroduplexes from homoduplexes, representing a sequence that carries the SNP. Individuals which are heterozygotes for SNP will have 1 : 1 ratio of wild type and variant DNA, so that the PCR products will also have 1 : 1 ratio of wild type and variant sequence. In case of homozygous mutant, the DNA will have to be mixed with wild type PCR amplified DNA. In either case, the mixture of wild type and variant DNA is heated and cooled again, so that the sample will then have a mixture of homo- and heteroduplexes. The heteroduplexes partially denature due to single base pair mismatch and can then be distinguished from corresponding homoduplexes by ion-pair reversed-phase liquid chromatography (IP-RP-HPLC) used in the WAVE™ system developed by Transgenomic Inc. (Figure 5).

*Masscode™ system:* This system claims to offer the highest throughput ever for SNP genotyping (40,000 SNPs per day per single Masscode system) and is being marketed by QIAGEN Genomics (http://www.giagengenomics.com). SNP genotyping services are also being provided by QIAGEN Genomics. The technology of dHPLC, discussed above for the WAVE™ system, is also being used in the proprietary Masscode™ SNP discovery system of QIAGEN Genomics (http://www.Qiagen genomics.com).

*Pyrosequencing for SNP genotyping:* Pyrosequencing™ is a proprietary technology developed by the company named Pyrosequencing AB (SSE: PYRO) with its headquarters at Uppsala, Sweden, which also markets the complete pyrosequencing system PSQ™96 (also the kits of consumables), developed and manufactured by them. It is the first dedicated system for SNP detection analysing one SNP every 6 seconds (~600 SNPs per hour), and is therefore being used by a number of laboratories in Europe and North America.

Pyrosequencing is actually a new sequencing method for obtaining sequences of short DNA segments (up to ~20 nucleotides). With the help of PSQ™96, within 15 min, sequences on 96 templates can be obtained simultaneously (http://www.intl-pag.org/pag/8/abstracts/). The method relies on step-wise addition of individual dNTP (with simultaneous release of pyrophosphates, i.e. $PP_i$) and monitoring their template guided incorporation into the growing DNA chain via chemiluminescent detection of the formation of pyrophosphate[35,36]. Incorporation of a nucleotide into DNA will be possible, only if it is complementary to the next base in the template strand and the quantity incorporated will depend on the number of one or more consecutive complementary bases. Unincorporated dNTP is degraded using the enzyme apyrase. Pyrophosphate released is utilized to convert 5′ amino phosphosulfurate (APS) into ATP with the help of the enzyme ATP sulfurylase, and the ATP produced drives luciferase mediated conversion of luciferin into oxyleciferin, generating light. The light produced is proportionate to ATP produced, which in its turn will be directly proportionate to the dNTP consumed. The emitted light is detected by a CCD camera and seen in a pyrogram as a peak, whose height will tell us about the number of molecules of dNTP incorporated. The following reactions are involved in the emission of light.

$$(DNA)_n + dNTP \xrightarrow{\text{polymerase}} (DNA)_{n+1} + PP_i$$

$$APS + PP_i \xrightarrow{\text{sulfurylase}} ATP$$

$$\text{luciferin} + ATP \xrightarrow{\text{luciferase}} \text{oxyluciferin} + \text{light}$$

Pyrosequencing is particularly suitable for SNP genotyping, since genotyping of previously identified SNPs by this method requires sequencing of only a few nucleotides (1–5 bp). Pyrosequencing is being used for SNP genotyping and also for rapid mapping of ESTs in wheat and corn by DuPont and Pioneer Hi-Bred in USA. For this purpose, the following steps are used: (i) procure sequence information generated in the SNP identification programme; (ii) design sequencing primers close to identified SNP sites; (iii) amplify SNP loci using one biotinylated and one standard primer; (iv) separate biotinylated single strands by magnetic strand separation on streptavidin-coated beads; (v) use liquid

phase pyrosequencing machine (PSQ$^{TM}$96), with above primers and the bead-bound template for prosequencing a few bases. The machine will utilize the above reactions and will produce a pyrogram giving sequences of the target sites. The limitations with the above pyrosequencing method, however, include the following: (i) need for PCR (not needed in 'invasive cleavage' discussed below); (ii) need for separation of single-stranded template.

*Invasive cleavage assay – A non-PCR assay:* In all the methods of SNP analysis described above, an initial target amplification using PCR is needed. However, PCR does have significant limitations, when used in a high throughput approach, so that approaches involving simpler and more direct analysis of DNA without prior PCR amplification have been developed. One such approach employs 'invasive cleavage assay' for nucleic acids and a MALDI-TOF-MS detection system[37]. The invasive cleavage assay involves hybridization of genomic DNA with two sequence-specific oligonucleotides, one termed the *invader oligonucleotide*, and the other termed the *probe oligonucleotide*. The *invader oligonucleotide* has a sequence homology with a segment of genomic DNA upstream of the SNP site. The *probe oligonucleotide*, on the other hand, has a segment at its 3′ end that is homologous to the target DNA, and another segment at its 5′ end, that has no homology with the target DNA. On hybridization, a duplex is formed between the homologous segment of the probe oligonucleotide and the target DNA. The invader oligonucleotide now invades into the duplex for at least one nucleotide, thus forming an overlap at this point of invasion. The flap endonucleases (FENs) cleave the unpaired region (including the overlap) available on the 5′ end of the probe, resulting in a 3′-hydroxyl DNA cleavage product.

A modification of the above invader assay, called *invader squared assay* has been used to amplify the signal for the detection of SNPs. This is a two-step reaction, in which primary cleavage product serves as an *invader oligonucleotide* for a secondary invasive cleavage reaction, for which a fresh target and a fresh probe oligonucleotide are supplied to the reaction mix. This produces secondary cleavage products, that are detected by MALDI-TOF-MS (Figure 6). Several cycles of invasive cleavage can be used to amplify the signal further.

## Haplotyping using more than one linked SNPs

The methods for genotyping described above determine the presence of individual SNPs in a diploid individual, but cannot determine which chromosome of a diploid pair is associated with each SNP. Haplotype of an individual, on the other hand, describes specific alleles at a number of linked SNP loci associated with each chromosome of a homologous pair. This can be done at the

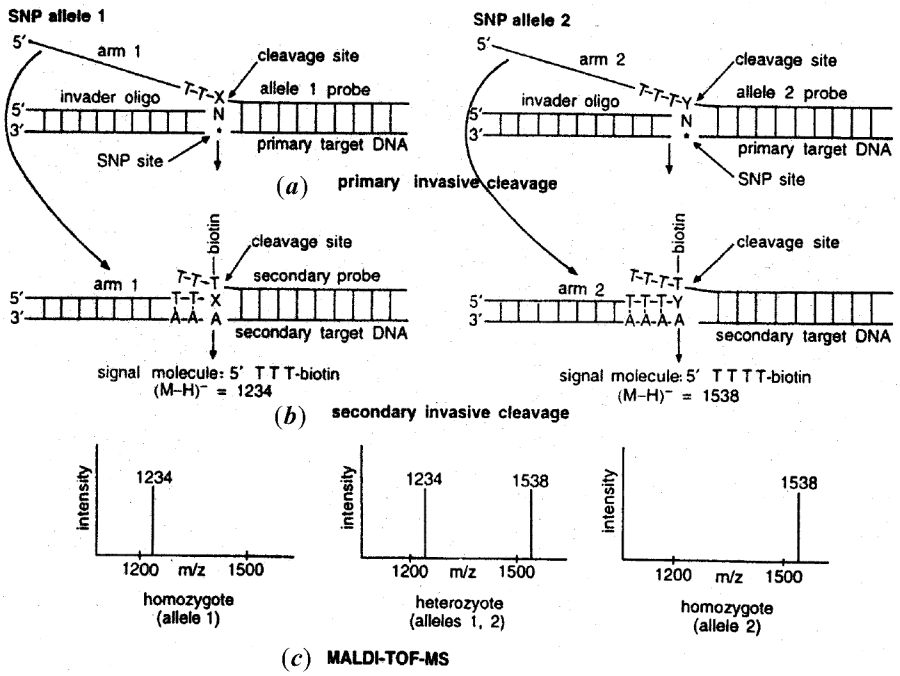


**Figure 6.** Diagrammatic representation of 'invader squared assay' and MALDI-TOF-MS detection system for detection of SNP (see text for details; modified from ref. 38).

level of a gene, a region of chromosome or a long DNA fragment. Haplotyping has been considered to be important in diagnosis of several diseases, because the mutant alleles for two or more SNPs, when present on the same homologue sometimes give a phenotype, which differs from the one that is due to location of these mutants on two different homologues of the same pair. This can be illustrated using a hypothetical example of two biallelic SNP loci (Aa and Bb), which can give four different combinations of haploid genotypes (AB, Ab, aB, ab) on an individual chromosome, and 10 different haplotypes among diploid individuals (AB/AB, AB/Ab, AB/aB, AB/ab, Ab/Ab, Ab/aB, Ab/ab, aB/aB, aB/ab, ab/ab). Of the 10 haplotypes, only two (AB/ab and Ab/aB) represent heterozygosity at both the loci, and the phenotypes may differ, as also shown in the classical *cis-trans* effect, while studying the fine structure of a gene. The situation will be much more complicated, when more than two or even up to a dozen SNPs located on a chromosome are considered together for haplotyping. In view of its importance, several methods have been devised for haplotyping, the most important being the recent use of single-walled carbon nanotube (SWNT) atomic force microscopy (AFM) probes, which allow multiplexed detection of polymorphic sites and determination of haplotypes in DNA fragments of up to 10-kilobase size[38]. The utility of haplotyping over genotyping at individual SNP loci has been demonstrated recently in the field of pharmacogenomics (e.g., asthma; gene *OPRM1* controlling heroin and cocaine dependence)[39].

## SNPs in *Arabidopsis* genome

The *Arabidopsis* Genome Initiative (TAGI) has recently completed its phenomenal work on whole genome sequencing of Columbia accession of *Arabidopsis*, and its publication in 14 December 2000 issue of *Nature* has been widely celebrated. Using these efforts, The Stanford DNA Sequence and Technology Centre (Stanford, USA) earlier deposited a collection of 412 SNPs detected between Landsberg *erecta* and Columbia strains, which were later confirmed, and many of them even mapped[40]. Cereon Genomics (Cambridge, USA) has, however, made an effort to sequence Landsberg *erecta* accession of *Arabidopsis*, using whole-genome shotgun approach[41]. These data of Cereon Genomics have been utilized to identify as many as 40,000 SNPs (one SNP every 3.3 kb) and 20,000 indels (one indel every 6.1 kb) between Landsberg *erecta* and Columbia strains. These extensive data on SNPs in *Arabidopsis* genome have been made accessible for the academic and non-profit organizations through TAIR (The *Arabidopsis* Information Resource) and will be updated with the recent availability of the sequence of the Columbia accession[42]. For the analysis of these SNPs, a recent PCR-based method, described as single-nucleotide amplified polymorphisms (SNAP) has also been devised, which should facilitate map-based cloning using SNPs[43].

## Use of SNPs in crop improvement programmes

SNP can be found within a gene or may be found in its close proximity. When found within a gene, it may or may not be responsible for the mutant phenotype, but in either case, it can be used for positional cloning of the gene in question. Once a large collection of SNPs is available, their use will depend on whether or not genetic determinants for all traits are included in those SNPs. For instance in humans, according to *causal hypothesis* the final collection of SNPs will include all genetic determinants, so that the allelic associations among SNPs will be irrelevant[44]. In contrast to this, according to *proximity hypothesis*, most genetic determinants of diseases will not be included even in a sample of several hundred thousand SNPs, so that allelic association among SNPs can be very important and will be used for positional cloning[45]. We believe that in plants, most SNPs will not be genetic determinants and therefore associations among SNPs and the traits of economic value will be of major interest to the plant breeders.

### *Association of SNPs with genes of economic value*

*SNP marker for waxy gene ($W_x$) controlling amylose content in rice:* Amylose is the principal component controlling the cooking and nutritional properties of cereals. Amylose-free or low amylose starch is considered desirable for certain food and non-food industries[46]. Therefore, it is an important breeding trait for the development of new cultivars in cereals including rice. In most cereals, amylose synthesis is controlled by starch synthase enzyme, which is encoded in waxy ($W_x$) gene. In rice, it has been shown that high and low amylose types can be differentiated on the basis of a single nucleotide difference, i.e. an SNP located near the waxy gene[47]. This marker should prove very useful in MAS exercised for the selection of low amylose at the seedling stage.

*SNP marker for dwarfing gene in rice:* In rice, an important semi-dwarfing gene is *sd-1*. A gDNA clone RG109 was mapped close to this gene, and was converted into PCR-based SSR markers. The sequences flanking this SSR were found to have an SNP which can be used in selection for *sd-1* in a range of rice crosses[48].

*SNP for male sterility in onion:* In onion, Genetic Bit Analyses (GBA$^{TM}$) has been used for discriminating

between SNP alleles in the plastomes, which are responsible for cytoplasmic male sterility and fertility[34]. Onion breeding lines can thus be screened at the seedling stage for accurate prediction of sterility, so that this SNP can be utilized for MAS in onion breeding programme, to save time, labour and economic resources.

*SNP marker for soybean cyst nematode resistance:* In soybean, RFLP and SSR markers are already being used by the practising plant breeders for the selection of SCN resistance[12]. More recently, SNPs have been detected using molecular beacon for the selection of soybean cyst nematode resistance alleles at *rhg1* and *rhg4* loci[49]. These SNP markers should prove superior to the markers that were available earlier and were being used for MAS in actual plant breeding.

### *SNPs for cultivar identification and evaluation of diversity*

*Identification of cultivars in barley:* In a study of a number of barley cultivars, SNPs were detected and allele-specific primers were developed, which enabled identification of these cultivars using SNPs. These SNPs were detected using primers for individual SNP sites that were designed with the help of sequence library generated from several barley cultivars. These primers when used for PCR, the amplified product was allele-specific and could be identified due to different fluorochromes used with two allele-specific primers. For SNP development in barley, an allele-specific database (barley *alleledb*) is also being prepared at the University of Montana, Bozeman[21,22].

*SNPs in tree species (scot pine and spruce):* Scot pines are highly variable and heterozygous. Sequence amplification in this tree species however, can be performed in haploid megagametophytes to avoid problems due to heterozygosity. A number of Scot pine populations were examined for SNPs in genes encoding several important structural proteins or enzymes like phenylalanine-lyase (involved in lignin synthesis for development of secondary xylem), so that selection can be exercised for desirable genotypes using SNPs[50]. Nuclear and chloroplast SNPs have also been developed in the genus *Picea* and can be used to distinguish the different species of this genus (*P. glauca, P. mariana* and *P. rubens*)[19].

*SNP markers for diversity at homoeoloci in wheat:* At DuPont Biotech (Newark, USA), DNA fragments, each 1 kb in length, were amplified from three starch biosynthesis genes in bread wheat. In each case, the amplified product was a mixture of sequences from A, B and D genomes (homoeoloci). The PCR products, when cloned and sequenced (24 clones for each gene), resolved 1 SNP every 20 bp, which will allow development of genome-specific markers. These markers were utilized for evaluation of the level of inter-varietal SNP variation for each genome in a hierarchical manner. This can be related with phenotypic differences, and new sources of desirable allelic variation in land races and wild relatives can be detected for gene introgression.

*SNP allelic diversity in maize:* At DuPont and Pioneer Hi-Bred, the available sequences of maize genome were used for designing the primers to amplify sequences (300 bp long) preceding the poly-A sites at the 3′ untranslated segments of several maize genes. Genomic DNA from 30 maize lines was utilized as template for PCR amplification. The alignment and analyses of products from 20 loci in 30 genotypes as above revealed the occurrence of one SNP per 70 bp, and one *indel* per 160 bp. This study allowed selection of 8 genotypes representing maximum allelic diversity; these genotypes will be used to catalogue SNP alleles at 1000 loci, that were selected from both the ESTs and the genes of interest.

## Conclusion

During the last 3–5 years, SNPs have emerged as the new generation molecular markers, which have already been developed in large number for the human genome. In the next year, while many more human SNPs will be developed and mapped, SNPs will also be produced in several crops. In parallel with their discovery and development, SNPs are already being used in humans for detection of association with a variety of diseases. In crop plants, however, only a beginning has been made in the area of SNP discovery and detection. In future, they will certainly be used in a number of crops, not only for studies involving associations with a number of traits of economic value, but also for the study of genetic diversity and variety identification. Enormous genomic and cDNA sequence data that are accumulating in the databases will be extremely useful in future for discovery of new SNPs. A number of gel-based and non gel-based methods will also be used for detection of already characterized SNPs and for genotyping of populations at these SNP sites. Newer methods will also be developed for this purpose. This will be facilitated due to automation and high throughput approaches, that are already available for work on SNP. In this connection, MALDI-TOF MS will certainly become popular not only for SNP genotyping, but also for a variety of other purposes. Pyrosequencing is another approach that will be increasingly used in future. Therefore, in the coming decade, those working in the area of molecular markers in crops will remain busy with the discovery and use of SNPs in a number of crops for a variety of purposes

relevant to crop improvement. This is certainly true about the laboratories in the developed world. One can only hope that in India also, facilities will be created at least in some select laboratories, to make its share of contribution in this exciting area of research involving SNP discovery and detection, particularly in plant systems.


1. Bookes, A. J., *Gene*, 1999, **234**, 177–186.
2. Kwok, P. Y., Deng, Q., Zakeri, H., Taylor, S. L. and Nickerson, D. A., *Genomics*, 1996, **31**, 123–126.
3. Hodgson, J., *Nat. Biotechnol.*, 2000, **18**, 475.
4. Roberts, L., *Science*, 2000, **287**, 1898–1899.
5. Ott, J., *Proc. Natl. Acad. Sci. USA*, 1999, **97**, 2–3.
6. Collins, A., Lonjou, C. and Morton, N. E., *Proc. Natl. Acad. Sci. USA*, 1999, **9**, 15173–15177.
7. Wolters, P., Powell, W., Lagudah, E., Snape, J. and Henderson, K., in Plant and Animal Genome VIII Conference, 9–12 January 2000, San Diego, USA (http://www.intl-pag.org/pag/8/abstracts/).
8. Bhattramakki, D. *et al.*, in Plant and Animal Genome VIII Conference, 9–12 January 2000, San Diego, USA (http://www.intl-pag.org/pag/8/abstracts/).
9. Mohan, M., Nair, S., Bhagwat, A., Krishna, T. G., Yano, M., Bhatia, C. R. and Sasaki, T., *Mol. Breed.*, 1997, **3**, 87–103.
10. Gupta, P. K., Varshney, R. K., Sharma, P. C. and Ramesh, B., *Plant Breed.*, 1999, **118**, 369–390.
11. Gupta, P. K. and Varshney, R. K., *Euphytica*, 2000, **113**, 163–185.
12. Young, N. D., *Mol. Breed.*, 1999, **5**, 505–510.
13. Wang, D. G. *et al.*, *Science*, 1998, **280**, 1077–1082.
14. Buetow, K. H., Edmonson, M. N. and Cassidy, A. B. *Nat. Genet.*, 1999, **21**, 323–325.
15. Taillon-Miller, P., Gu, Z., Li, Q., Hillier, L. and Kwok, P. Y., *Genome Res.*, 1998, **8**, 748–754.
16. Altshuler, D., Pollara, V. J., Cowles, C. R., Van Etten, W. J., Baldwin, J., Linton, L. and Lander, E. S., *Nature*, 2000, **407**, 513–516.
17. Mullikin, J. C., Hunt, S. E., Cole, C. G. and 40 others, *Nature*, 2000, **407**, 516–520.
18. Orita, M., Suzuki, Y., Sekiya, T. and Hayashi, K., *Genomics*, 1989, **5**, 874–879.
19. Germano, J. and Kleim, S., *Theor. Appl. Genet.*, 1999, **99**, 37–49.
20. Buttema, C. D. and Sommer, S. S., *Mutat. Res.*, 1993, **288**, 93–102.
21. See, D., Kanazin, V., Talbert, H. and Blake, T., *Barley Newsl.*, 1998 (http://wheat.pw.usda.gov/ggpages/BarleyNewsletter/42/post26.html).
22. See, D., Kanazin, V., Talbert, H. and Blake, T., *BioTechniques,* 2000, **28**, 710–716.
23. Lee, L. G., Livak, K. J., Mullah, B., Graham, R. J., Vinayak, R. S. and Woudenberg, T. M., *Biotechniques*, 1999, **2**, 342–349.
24. Tyagi, S. and Kramer, F. R., *Nat. Biotechnol.*, 1996, **14**, 303–308.
25. Tyagi, S., Bratu, D. P. and Kramer, F. R., *Nat. Biotechnol.*, 1998, **16**, 49–54.
26. Tyagi, S., Marras, S. A. E. and Kramer, F. R., *Nat. Biotechnol.*, 2000, **18**, 1191–1196.
27. Landergren, U., Kaiser, R., Sanders, J. and Hood, L., *Science*, 1988, **244**, 1077–1080.
28. Tobe, V. O., Taylor, S. L. and Nickerson, D. A., *Nucleic Acids Res.*, 1996, **24**, 3728–3732.
29. Lemieux, B., Aharoni, A. and Schena, M. *Mol. Breed.,* 1998, **4**, 277–289.
30. Gupta, P. K., Roy, J. K. and Prasad, M., *Curr. Sci.*, 1999, **77**, 875–884.
31. Howell, W. M., Jobs, M., Gyllensten, U. and Brookes, A. J., *Nat. Biotechnol.*, 1999, **17**, 87–88.
32. Syvanen, A.-C., Aalto-Setala, K., Harju, L., Kontula, K. and Soderlund, H., *Genomics*, 1990, **8**, 684–692.
33. Nikiforov, T. T., Rendle, R. B., Goelet, P., Rogers, Y. H., Kotewicz, M. L., Anderson, S., Trainor, G. L. and Knapp, M. R., *Nucleic Acids Res.*, **22**, 4167–4175.
34. Alcala, J., Giovannoni, L. M., Pike, L. M. and Reddy, A. S., *Mol. Breed.*, 1997, **3**, 495–502.
35. Ronaghi, M., Uhlen, M. and Nyren, P., *Science*, 1998, **281**, 363–365.
36. Ahmadian, A., Gharizadeh, B., Gustafsson, A. C., Sterky, F., Nyren, P., Uhlen, M. and Lundeberg, J., *Anal. Biochem.*, 2000, **280**, 103–110.
37. Griffin, T. J., Hall, J. G., Prudent, J. R. and Smith, L. M., *Proc. Nat. Acad. Sci. USA*, 1999, **96**, 6301–6306.
38. Woolley, A. T., Guillemette, C., Cheung, C. L., Housman, D. E. and Lieber, C. M., *Nat. Biotechnol.*, 2000, **18**, 760–763.
39. Davidson, S., *Nat. Biotechnol.*, 2000, **18**, 1134–1135.
40. Cho, R. J. *et al.*, *Nat. Genet.*, 1999, **23**, 203–207.
41. Rounsley, S., Subramaniam, S., Cao, Y., Bush, D., DeLoughery, C., Field, C., Phillips, C., Iartchouk, O., Last, R., Wiegand, R., Fischhoff, D. and Timberlake, B., Proc. 10th Intern Conf. Arabidopsis Res (Abstract 3–1), 4–8 July 1999, University of Melbourne, Australia.
42. Lukowitz, W., Gillmor, C. S. and Scheible, W-R., *Plant Physiol.*, 2000, **123**, 795–805.
43. Drenkard, E., Richter, B. G., Rozen, S., Stutius, L. M., Angell, N. A., Mindrinos, M., Cho, R. J., Oefner, P. J., Davis, R. W. and Ausubel, F. M., *Plant Physiol.*, 2000, **124**, 1483–1492.
44. Risch, N. and Merikangas, K., *Science*, 1996, **273**, 1516–1517.
45. Kruglyak, L., *Nat. Genet.*, 1999, **22**, 139–144.
46. Morell, M. K., Peakall, R., Appels, R., Preston, L. R. and Lloyd, H. L., *Aust. J. Exp. Agric.*, 1995, **35**, 807–819.
47. Ayres, N. M., McClung, A. M., Larkin, P. D., Bligh, H. F. J., Jones, C. A. and Park, W. D., (http://www.nal.usda.gov/ttic/tektran/data/000007/55/0000075521.html).
48. Graland, S. H., Lewin, L., Blakeney, A. and Henry, R., in Plant and Animal Genome VIII Conference, 9–12 January 2000, San Diego, USA (http://www.intl-pag.org/pag/8/abstracts/).
49. Cregan, P. B., *Annu. Meet. Crop Sci. Soc. Am.*, 1999, p. 158.
50. Dvornik, V., Mikkonen, M., Sirvio, A. and Savolainen, O., in Plant and Animal Genome VIII Conference, 9–12 January 2000, San Diego, USA (http://www.intl-pag.org/pag/8/abstracts/).


ACKNOWLEDGEMENTS. The article was written during the tenure of P.K.G. as CSIR-ES awarded by CSIR. J.K.R. is working as SRF in an NATP project sanctioned by ICAR.

Received 25 September 2000; revised accepted 27 November 2000