Sequence-Based Markers

4

4.1 Introduction

The DNA markers like RFLPs, AFLPs, and SSRs were extensively used in various biological investigations and for marker-assisted selection (MAS) in both animals and plants. However, the development of many of these markers, e.g., RFLPs and SSRs, is demanding and expensive as it involves time-consuming cloning, construction of probe libraries, and/or sequencing for primer design. In addition, scoring of a number of these markers across many individuals is also expensive, labor intensive, and time-consuming. Therefore, continuous efforts were made to develop such DNA markers that are reliable, abundant, almost evenly distributed throughout the genome, and relatively cheaper, developed with minimum effort and time and are amenable to automation and high-throughput genotyping. The genome sequence data generated by the human genome-sequencing project revealed that bulk of sequence variation among different individuals was due to changes at single-base positions distributed throughout the genome. The variation in single base pairs of DNA is known as *single nucleotide polymorphism* (*SNP*). Subsequently, SNPs were found to be universal and the most abundant markers; they constitute ~90 % of the genetic variation in any organism. This marker system yields reliable and reproducible results and is amenable to automation and high-throughput genotyping (Mammadov et al. 2012).

The discovery of SNPs involves sequencing of genomic DNA or cDNA (complementary or copy DNA) from two or more individuals/lines of a given species and comparing these sequences using a suitable computer program. SNPs may also be discovered by in silico alignment and analysis of genomic/EST sequence data available in the databases of the concerned species. In either case, once SNPs are discovered, they can be genotyped using any one of more than 30 different detection methods based on one or more of the following reactions: (1) DNA hybridization, (2) primer extension, (3) oligonucleotide ligation, and (4) DNA replication. Several of these methods have been automated and scaled up for high-throughput SNP genotyping; some of these technologies are considered in some detail in Chap. 13. In this chapter, we shall discuss DNA sequencing, methods for SNP discovery, and small- to moderate-scale SNP genotyping strategies.

4.2 DNA Sequencing

The determination of base sequence of a DNA fragment is called *DNA sequencing*. DNA sequencing became feasible due to the following important developments: (1) availability of restriction enzymes, (2) development of electrophoresis techniques capable of separating DNA fragments differing by a single nucleotide, and (3) gene cloning and PCR techniques that make available very large number of copies of

B.D. Singh and A.K. Singh, *Marker-Assisted Plant Breeding: Principles and Practices*,
DOI 10.1007/978-81-322-2316-0_4, © Author(s) 2015

individual DNA fragments required for sequencing. Initially, two methods, a chemical and an enzymatic method, of DNA sequencing were developed; these methods are popularly termed as *first-generation DNA sequencing procedures*. Soon the *second-* or *next-generation DNA sequencing* (*NGS*) *methods* were developed, which use PCR for in vitro cloning in the place of in vivo cloning and are much faster and cheaper (Pandey et al. 2008; Schendure and Ji 2008; Edwards 2013). At present, the *third-generation DNA sequencing* (*TGS*) *methods* are becoming commercially available; these methods sequence single DNA molecules without any cloning (Schadt et al. 2010).

4.2.1 First-Generation DNA Sequencing Methods

The *chemical method of DNA sequencing* uses specific chemical modifications of DNA bases, ultimately, leading to breaks in DNA strands at the sites occupied by the modified bases. Four separate reactions are set up for the modification of different bases, and gel electrophoresis, followed by autoradiography, allows deduction of the base sequence of the DNA strand. The *enzymatic method of DNA sequencing*, also called *Sanger–Coulson method* (Sanger et al. 1977), on the other hand, uses single-stranded DNA fragments for DNA replication catalyzed by the Klenow fragment of *E. coli* DNA polymerase I. Often the two complementary strands of a DNA fragment are sequenced in separate reactions for an enhanced reliability of the sequence data. For each strand, four separate reactions are set up. In each reaction mixture, the DNA strand, a suitable primer, the Klenow fragment, the four deoxyribonucleotides (dNTPs, viz., dATP, dGTP, dCTP, and dTTP), and the other reagents required for DNA replication are provided; at least one of the four dNTPs is radioactive to allow radioautographic imaging of the bands after gel electrophoresis. In addition, in each reaction mixture, a different 2′,3′-dideoxynucleotide (ddNTP) is also added at a concentration of about 1/100 of that of the normal deoxyribonucleotides used in the reaction.

The ddNTPs do not have a free 3′-OH group. Therefore, when a ddNTP is incorporated at a site into a growing polynucleotide chain, there is no further addition of nucleotides to the chain beyond this site. Therefore, ddNTPs are called *chain terminators* or simply *terminators*. At the concentration used here, a given ddNTP would cause chain termination at any one of all the possible sites, at which its complementary base occurs in the template DNA strand. In the end, therefore, the mixture will contain partially synthesized polynucleotide chains of different lengths produced by chain termination at every point where the base complementary to the given ddNTP is present in the template DNA strand. The DNA duplexes formed in the four reaction mixtures are denatured; the mixtures are loaded in gel lanes and subjected to electrophoresis. The bands formed in the gel lanes are visualized by radioautography, and the base sequence is read by comparing the band positions in the four lanes. This procedure enables sequencing of fragments of up to 700–800 bases.

The Sanger–Coulson method was automated to support the various genome-sequencing projects. The automated procedure uses fluorescent labels (a different label for each of the four ddNTPs) in the place of radioactivity, capillaries in the place of routine gels for electrophoresis, and computer-based sequence detection, data storage, and processing. These automated sequencers have been in use for over 30 years, and until recently most genome-sequencing projects were exclusively based on this technology. The current *read lengths*, i.e., the lengths of sequences of single fragments, are up to 1,000 bp with an error rate of 0.001 %. However, whole-genome sequencing required several sequencers located at a large center, having highly automated template preparation and other supporting facilities. In addition, the sequencing process is highly demanding in terms of both infrastructure and processing efforts, and the sequencing costs are rather high (Deschamps and Campbell 2010; Edwards 2013).

4.2.2 Next-Generation DNA Sequencing Methods

The next-generation DNA sequencing methods, also called massively parallel sequencing (MPS) technologies, are faster and cheaper and require much less template preparation than the Sanger–Coulson method. The NGS methods use PCR amplification for template preparation (in vitro *library preparation*), which takes merely 2 h, and they are amenable to very high throughput. Further, they allow simultaneous sequencing of hundreds of thousands to hundreds of millions of different DNA fragments (Schendure and Ji 2008). At present, there are five NGS methods, namely, (1) 454 sequencing, (2) Solexa method, (3) ion semiconductor sequencing, (4) Polony method, and (5) massively parallel signature sequencing (MPSS). The first three methods (454, Solexa, and ion semiconductor sequencing methods) use DNA synthesis for sequencing (*sequencing by synthesis*, *SBS*), while the Polony and MPSS methods employ oligonucleotide hybridization to the template followed by ligation to the growing chain. The MPSS is suited for quantification of gene expression; it uses multiple cycles of enzymatic cleavage and ligation to determine 17–20-bp-long "signature" sequences from the ends of cDNA molecules to distinguish and quantify the different RNA species present in the sample. The 454, Solexa, ion semiconductor, and Polony methods have already been commercialized for high-throughput sequencing and are briefly described in the following sections.

According to a survey, the use of NGS technologies in public and private sequencing laboratories of the USA and Europe had gone up to 56 % by 2010. The most frequent application of these technologies was mRNA expression profiling, followed by biomarker discovery, resequencing, diagnostics, and targeted resequencing. In 2011, Illumina HiSeq 2000 and Illumina GAIIx platforms were the market leaders in terms of sales. However, SOLiD 3 Plus was rated to have the highest accuracy as compared to the Illumina GAIIx and GS FLX systems. It is projected that NGS and TGS technologies will eventually replace the established techniques like targeting-induced local lesions in genomes (TILLING), TILLING in wild populations (Eco-TILLING), and endonucleolytic mutation analysis by internal labeling (EMAIL).

4.2.2.1 Template Preparation

The template for sequencing is single-stranded DNA (ssDNA), which can be prepared from genomic DNA, BAC clones, PCR products, and cDNA. Genomic DNA and BAC clones are randomly sheared by sonication, nebulization (mechanical shearing), or enzymatic digestion by DNase I to produce fragments of suitable size, while PCR products and cDNA may not need fragmentation. Often one may need to sequence some specific regions identified by linkage studies. In such cases, methods like "enrichment," "genome partitioning," or "genome capture" can be used for template sample preparation. These methods involve mRNA extraction, hybridization to preselected probes, or attachment of barcodes/index sequences to the fragments. After fragmentation, DNA fragments of 300–800 bp (used for shotgun sequencing) or 3–20 kb (used for paired-end sequencing) are separated. For *shotgun sequencing*, short adapters (adapters A and B) specific for both 3′ and 5′ ends are attached to each fragment; these adapters facilitate purification, amplification, and sequencing (Fig. 4.1). The fragments are now made single-stranded, and one single strand is attached to a single capture bead. These beads, along with the amplification reagents and the enzymes, are then enclosed in droplets of water-in-oil mixture. The emulsion around each bead forms a micro-reactor isolated from all other such beads. PCR amplification produces millions of copies of the single fragment attached to each bead, and all these copies become attached to the same capture bead. These beads form the in vitro library used for sequencing (Fig. 4.1).

In the case of *paired-end sequencing*, adapters are added to both the ends of the much larger fragments to facilitate their circularization. The

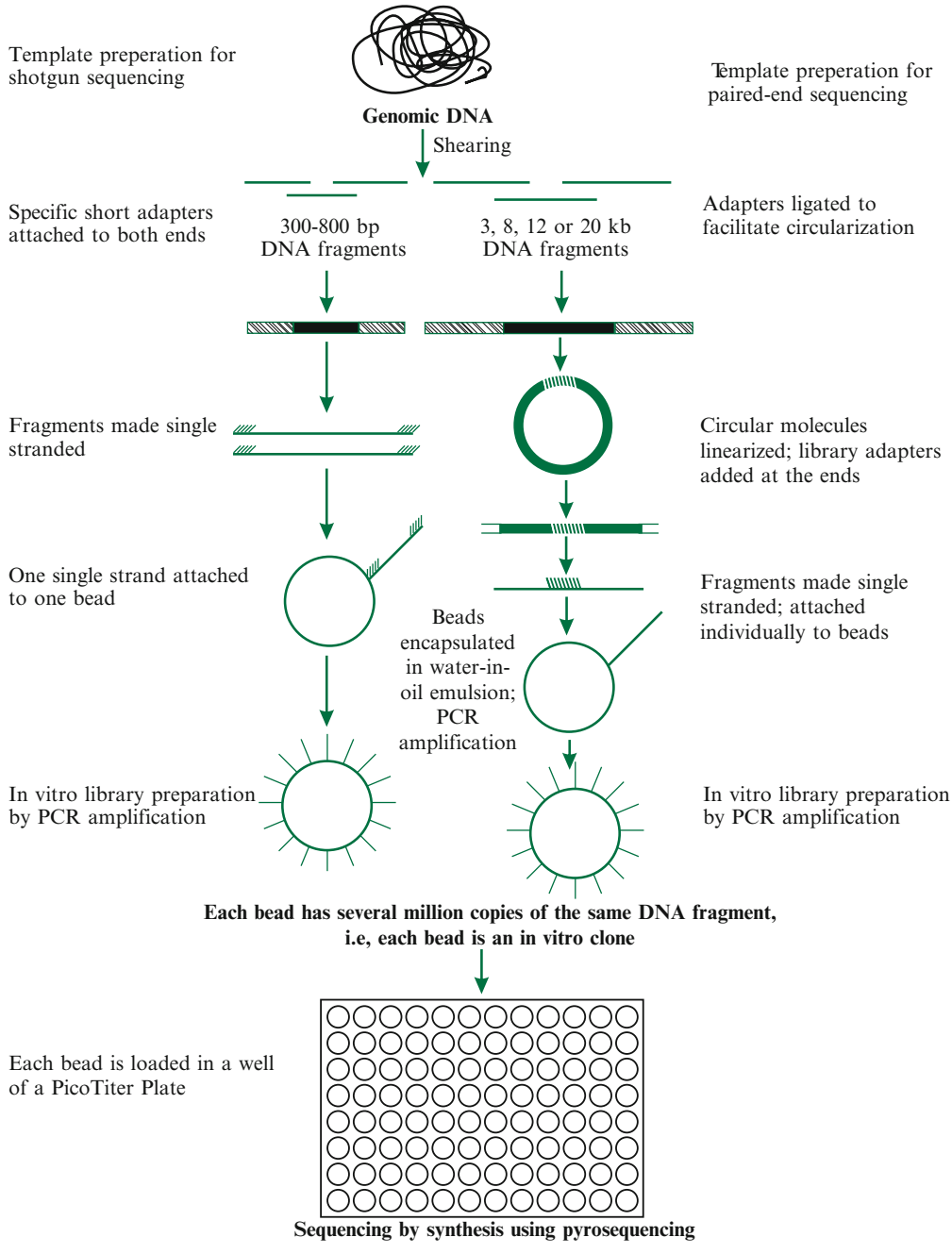


Fig. 4.1 A generalized schematic representation of the 454 sequencing method. The template preparation in other NGS methods is generally similar

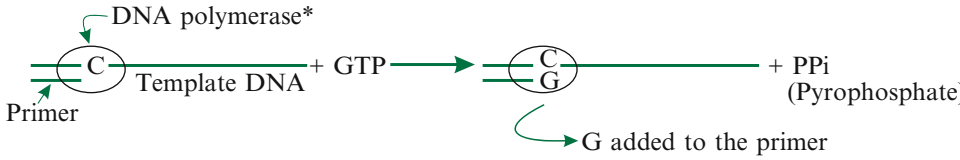
circularized DNA is fragmented, linear fragments containing the adapters are separated, their ends are polished, and library adapters (adapters A and B) are linked to both their ends. Thus each fragment has library adapters at the two ends, followed by a short segment corresponding to the ends of the genomic fragment, and finally the adapter sequence located in

the middle region of the fragment. These fragments are made single-stranded, and one single strand is attached to each capture bead. The beads are then processed in the same way as in the case of shotgun sequencing (Fig. 4.1). This description is based on template preparation for the 454 sequencing method, but the other NGS technologies also use similar strategies.

4.2.2.2 The 454 DNA Sequencing Method

This method was the first NGS technology to be commercialized in 2005 by 454 Life Sciences (now Roche Diagnostics), USA. The currently available 454 platforms are Genome Sequencer (GS) FLX System and GS FLX Titanium series. After template preparation (Sect. 4.2.2.1), the capture beads along with the attached DNA fragments are removed from the emulsion and loaded into the wells of a PicoTiter Plate. The size of wells is such that only a single bead can be loaded in each well. DNA sequencing is achieved by the pyrosequencing method. The reagents are flowed in a specific order across the plate, and the chemiluminescence signal is sensed by a sensitive CCD (charge-coupled device) camera. The computer software uses the chemiluminescence data to deduce the base sequence of the template DNA segment attached to every bead.

In *pyrosequencing*, the reaction mixture contains the template DNA, sequencing primer, APS (adenosine-5′-phosphosulfate), luciferin, the Klenow fragment, ATP sulfurylase, luciferase, and apyrase. The nucleotides dCTP, dGTP, and dTTP, and dATPαS (deoxyadenosine α-thiotriphosphate) are added to this reaction mixture sequentially one after the other. dATPαS is used in the place of dATP because it can be used by luciferase for light generation only after it has been used for DNA synthesis. In contrast, dATP will be used for producing light even when it is not used for DNA synthesis. When a dNTP, say, dGTP, is added to the 3′ end of the primer or the growing chain, one pyrophosphate (PPi) moiety is released (Fig. 4.2a). This PPi is used by ATP sulfurylase to convert APS into ATP (Fig. 4.2b), which is then used by luciferase to



(A) Release of pyrophosphate by DNA polymerase action

Adenosine 5′-phosphosulfate (APS) + PPi $\xrightarrow{\textbf{ATP sulfurylase}}$ **ATP + SO_4^-**

(B) ATP synthesis by ATP sulfurylase

Luciferin + ATP $\xrightarrow{\textbf{Luciferase}}$ **Oxyluciferin + ADP + Pi (yields light)**

(C) Luciferase uses ATP to generate visible light

dNTP (including ATP) $\xrightarrow{\textbf{Apyrase}}$ **dNDP + dNMP + Phosphate (Pi)**

(D) Pyrase action degrades dNTPs

Fig. 4.2 The various reactions catalyzed by the four enzymes used in pyrosequencing. *PPi* pyrophosphate, *Pi* inorganic phosphate, *ATP* adenosine triphosphate, * Klenow fragment (Based on de Vienne et al. 2003)

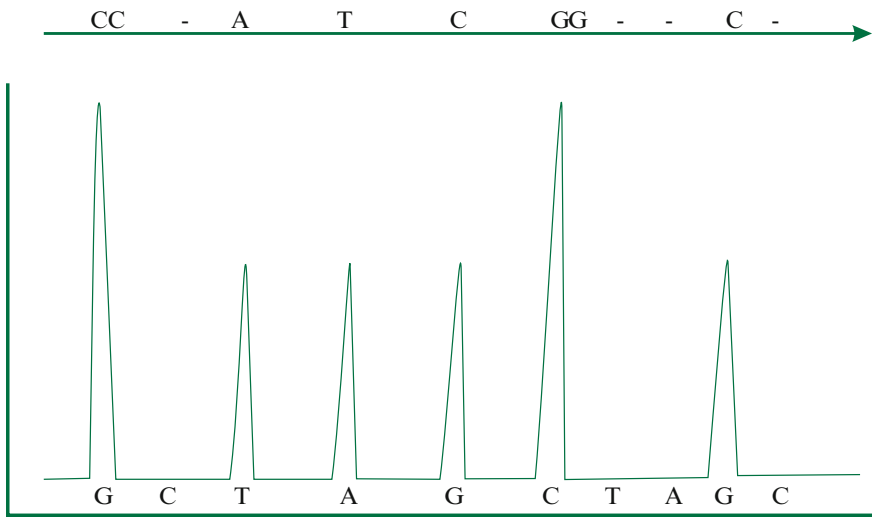


Fig. 4.3 A pyrogram pattern, and the nucleotide sequence deduced from this pattern. Production of light indicates nucleotide incorporation into the primer/growing chain. A stronger light signal, e.g., in response to the addition of the first G and the second C, reveals incorporation of the concerned nucleotide at two consecutive sites. But a lack of light signal, e.g., for the first C, shows lack of incorporation of the concerned nucleotide (Based on de Vienne et al. 2003)

generate visible light (Fig. 4.2c) (Ronaghi et al. 1996). A sensitive CCD camera detects the light, and the template nucleotide at this position is deduced (it will be C in this case). The intensity of light generated is proportional to the amount of PPi generated, i.e., the amount of nucleotide added to the primer/growing chain. Therefore, the light signal will be twice as intense if the same nucleotide occurs at two consecutive sites (Fig. 4.3). However, if the G from the dGTP was not added to the growing chain, no light will be generated. The enzyme apyrase continues to hydrolyze the unincorporated dNTPs (Fig. 4.2d) as well as the ATP produced by the ATP sulfurylase action. As a result, soon the ATP is exhausted and light production ceases; the next dNTP can now be added to the reaction mixture. Since the rate of dNTP degradation by apyrase is slower than its incorporation into the growing chain, sufficient dNTP remains available for DNA synthesis. Similarly, ATP degradation by apyrase is slower than ATP production by ATP sulfurylase so that enough ATP becomes available for light production when a dNTP is used for DNA replication.

The GS FLX system can process over one million beads at a time, and one run takes about 10 h, including template preparation. The data from paired-end sequencing can be combined with that from shotgun sequencing to readily generate a high-quality draft genome of large complex organisms. The average read length (length of individual sequences) in shotgun sequencing is ~400 bases, but bulk of the reads are of 500 bases; the GS FLX+ can now give reads of up to 1,000 bases (Edwards 2013). Read accuracy of GS FLX is over 99.6 %, while consensus accuracy is more than 99.99 %. *Read accuracy* is the accuracy of the sequence of individual reads, while *consensus accuracy* is the accuracy of the sequence of a fragment obtained as consensus of the sequences of all the reads of the fragment. In this and the Sanger–Coulson method, the error rate increases with the position of the base in the fragment due to a reduction in enzyme efficiency/concentration, leading to a reduced light signal-to-noise ratio. GS FLX can generate 400-Mb sequence in a 10-h run at a cost of US $ 5,000–7,000, while GS FLX Titanium XL+ can produce one million reads of up to 1,000 bp each (total sequence 1 Gb).

This technology can be used for de novo sequencing and assembly, genome sequencing and mapping, transcriptome analysis, analysis of epigenetic changes, etc. As the 454 method does not use chain terminators, a base will become incorporated as many times in a single cycle as its complementary base occurs consecutively in the template strand. When the same base occurs several (usually, >6) times consecutively (e.g., AAAAAAA) in the template, occasionally it is read one base less than the actual number, i.e., $n - 1$ times in the place of n times. This may lead to errors in base sequences of those stretches of template DNA, in which a base occurs more than once in tandem. Further, artifacts of single base pair deletions or insertions can be generated by signal-to-noise threshold problems.

4.2.2.3 The Illumina Sequencing Method

Illumina, USA, commercialized the Solexa NGS technology in 2007 (Bentley and Smith 2008), which is the most widely used NGS technology. The recent platforms of the series are Illumina Genome Analyzer 1 Gb and HiSeq 600 Gb. The sample DNA is fragmented, and two different adapters are ligated to their 5′ and 3′ ends. The fragments are attached to an especially prepared substrate on a flow cell, which contains a dense lawn of primers to be used in the next step of solid phase PCR. Fold-back PCR or bridge PCR produces up to 1,000 identical copies of each DNA fragment. All the copies of one fragment form an isolated cluster of molecules on the flow cell, and together they represent the in vitro clone of the fragment. All the clusters formed on a flow cell together represent the in vitro library (Fig. 4.4). The sequencing primer is now attached to the free ends of the fragments. The four dNTPs used for DNA synthesis have fluorophores linked to them; these fluorophores also serve as chain terminators. The dNTPs are added one at a time, and a CCD camera records their incorporation at the 3′ end of the sequencing primer/growing chain as fluorescence from the fluorophores attached to them. The fluorophore terminator is removed from the dNTP that has just been added to the primer/growing chain, making this nucleotide available for further DNA synthesis. A new dNTP is now added to the reaction mixture, it is incorporated at the ends of the growing chains, the fluorescence is recorded, and then the fluorophore is removed. In this way, the sequence of each DNA fragment is determined. The use of fluorophore chain terminators linked to the dNTPs eliminates the error in base sequence determination when the same base is present at two or more consecutive positions in the template strand.

Usually, read length ranges from 35 to 150 bases, and the accuracy is greater than 98.5 %. The total error-free read given by Illumina HiSeq 2000 is over 400 Gb in one run, which takes 7–8 days (Edwards 2013). MiSeq and HiSeq 2500 systems generate read lengths of up to 250 bp and have improved data capture and greater flexibility. The Illumina system can be used for de novo genome sequencing; genome resequencing for the analysis of SNPs, InDels, copy number variation (CNV), and structural variation; transcript profiling; etc. However, the PCR amplification step introduces a high error rate. The fluorescence properties of the four dyes used in this method tend to produce substitutions of A for C, G for T, and vice versa in the sequence data. In addition, the terminators of some nucleotides may not function properly so that a second nucleotide may be added to the growing chain in the same reaction cycle, generating a deletion of one base pair. However, base substitution errors are more common than insertion/deletion errors.

4.2.2.4 The ABI SOLiD Technology

The Applied Biosystems, USA, commercialized the Polony method in 2005 as SOLiD 3.0 platform (Schendure et al. 2005). SOLiD stands for "sequencing by oligonucleotide ligation detection" since this method achieves DNA sequencing by detecting oligonucleotide ligation. The DNA sample is fragmented (fragment size 600 bp to 6 kb) and processed in a manner similar to that for paired-end sequencing (Fig. 4.1). The beads along with the attached DNA molecules are immobilized in a single layer in an acrylamide matrix on a glass slide. An anchor primer is then hybridized to the adaptor sequence attached

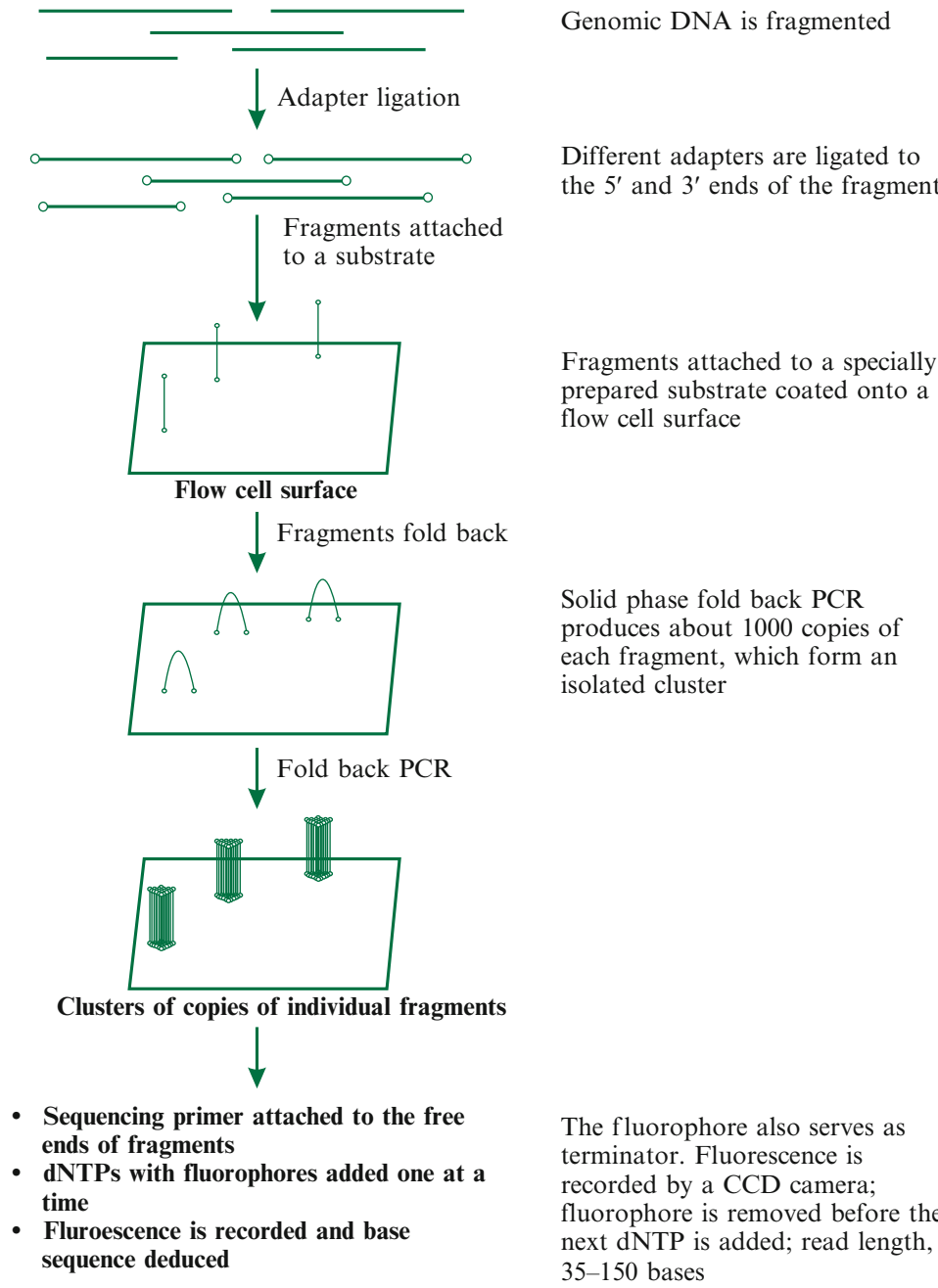


Fig. 4.4 A simplified schematic representation of the Illumina sequencing method (Based mainly on Bentley and Smith 2008)

to the template DNA. Sequencing is done by using a set of 16 oligonucleotides for hybridization with the template DNA and ligation to the 5′ end of the anchor primer/elongating chain. Each oligonucleotide is 8 bases long and is labeled with fluorophore at the 5′ end, and each member of a set of 16 oligos has a unique combination of two nucleotides at its 3′ end.

At a given time, four specific oligonucleotides of the set, each labeled with a different fluorophore, are added and allowed to pair at their 3′ ends with the template DNA. The

3′ ends of the oligonucleotides paired with the template DNA are ligated to the 5′ ends of the anchor primer molecules, the color of fluorescence is recorded, and the unpaired 5′ ends of the oligonucleotides are removed. A new set of four oligonucleotides is now added and the steps of the first cycle are repeated. After five cycles of oligonucleotide hybridization and ligation, the DNA is melted and the newly synthesized DNA strands are removed. A new anchor primer is now added that is one base shorter than the adaptor. Therefore, hybridization will begin one base upstream of the site it began in the first cycle and into the adaptor sequence. Again five cycles of hybridization and ligation are carried out, and fluorescence from each cycle is recorded. The data from the two repeats of ligation reactions are compared and analyzed to obtain the base sequence of the template strand. The repeat hybridization run using one base shorter primer allows each base to be examined twice and to fill any gaps that may remain after the first run.

The SOLiD 3.0 platform gives sequence reads of ~50 bases and generates over 20 Gb of total sequence per run, and each run takes about 6–7 days. In 2011, SOLiD 5500 and SOLiD 5500 XL systems were introduced; these systems give sequence data of up to 300 Gb per run at 99.9 % accuracy (Edwards 2013). The average error rates are lower when a good quality reference genome sequence is available and is used for error correction. In the absence of a reference genome, the error rate is higher than that for Illumina GA. Errors in base sequence arise from PCR amplification, beads carrying a mixture of fragments, incomplete dye removal, etc. The essential features of the three common NGS technologies, viz., the 454, Illumina, and ABI SOLiD technologies, are summarized in Table 4.1.

Table 4.1 A comparison among the three common NGS technologies: the 454, Illumina, and ABI SOLiD technologies

Feature	454	Illumina	ABI SOLiD
Sequencing reaction	Pyrosequencing (sequencing by synthesis)	Sequencing by synthesis	Oligonucleotide hybridization
Terminator	Not used	Used	Used
Detection based on	Luminescence generated by luciferase	Fluorescence from fluorophore	Fluorescence from fluorophore
Major error in base calling	InDels	Base substitutions	Base substitutions
Chief cause of error	Incorrect deduction of homo-polymorphic length from intensity of luminescence	Asynchronous DNA synthesis in the later cycles	Bias in fluorescence intensities in later machine cycles
Template DNA fragments attached to	Beads in microtiter plate wells	A specific substrate on a flow cell	Beads in an acrylamide matrix
Run duration[a]	10 h	7–8 days	6–7 days
Average read length (shotgun sequencing)	400 bases (GS FLX +, 1,000 bases)	35–150 bases (up to 250 bases by Hi Seq 2500)	~50 bases (SOLiD 3.0)
Total sequence data/run	400 Mb (GS FLX), ~1 Gb (GS FLX Titanium +)	400 Gb (Hi Seq 2000)	300 Gb (SOLiD 5500, SOLiD 5500 XL)
Read accuracy (%)	99.6 (99.9)[b]	98.5	–
Template preparation	Shotgun, paired end	Shotgun, paired end	Paired end
Each base examined	Once	Once	Twice
Improved base-calling algorithm[c]	Pyrobayes	Ibis and BayesCall	Rsolid
Draft genome preparation	Yes	Yes	–
Current platforms	GS FLX, GS FLX Titanium	Genome Analyzer 1 Gb, Hi Seq 600 Gb	SOLiD 5500, SOLiD 5500 XL

[a]Including template preparation
[b]The figure within parenthesis is the consensus accuracy
[c]These algorithms reduce error rates by ~5–30 % over the base-calling methods developed by the manufacturers

4.2.2.5 Ion Semiconductor Sequencing

In the case of *ion semiconductor sequencing*, a semiconductor sensing device or ion chip senses the H^+ ions produced during DNA synthesis by DNA polymerase (Schadt et al. 2010; Edwards 2013). The signals from H^+ ions are used for direct nonoptical identification of the bases present in the DNA template. The ion chip has 3.5-μm-diameter wells, each of which is located directly over each sensor. As a result, the wells confine the DNA fragments and the reagents for DNA synthesis directly over the sensors. The sequencing equipment comprises primarily an electronic detection system that is interfaced with the chip, a microprocessor to process the signals, and a fluidics system for regulating the reagent flow over the chip. The genomic DNA is fragmented, ligated to adapters, and attached to 2-μm-diameter acrylamide beads, and the DNA fragment attached to each bead is amplified by PCR (Fig. 4.5). The DNA polymerase and the sequencing primers are now attached to each template DNA molecule already attached to the beads, and the beads are then pipetted into the loading port of the chip. The well depth and the bead diameter ensure that only a single bead is loaded in each well. The four dNTPs are now added one at a time. When the DNA polymerase adds a dNTP to a primer/growing chain, there is net release of one proton (H^+), which produces a

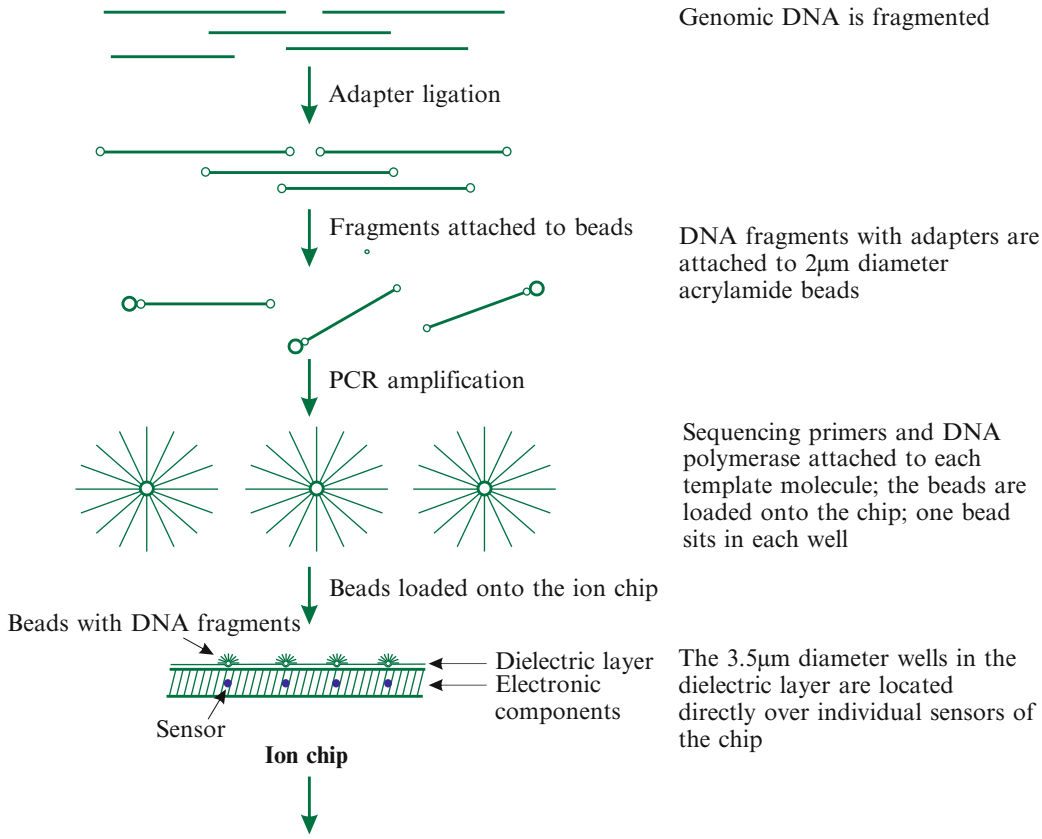


Fig. 4.5 A schematic representation of the ion semiconductor sequencing method. The Ion Torrent Proton platform generates up to 10 Gb sequence data per run; read length, 100–200 bases (Based on Schadt et al. 2010)

change in the pH of the surrounding solution. The sensor located at the bottom of each well detects this change in the pH and the signals are ultimately digitized. In case a dNTP is added to the primer/growing chain more than once due to the occurrence of its complementary base in the template DNA at more than one consecutive position, the change in pH is proportional to the number of nucleotides incorporated (0.02 pH units for each dNTP molecule added). The signal generation and detection takes ~4 s. The unused nucleotides are removed by washing before the new dNTP is added; this takes about one-tenth of a second.

Ion Torrent (acquired by Life Technologies), USA, has commercialized this technology as Ion Torrent PGM (Personal Genome Machine) and Ion Torrent Proton sequencing platforms. The Ion Torrent PGM produces 10 Mb–1 Gb sequence data per run with either 100 or 200 bases read-length protocols and sample multiplexing. But the Ion Torrent Proton platform generates up to 10 Gb of sequence data per run with 100 bp or 200 bases read-length protocols and sample multiplexing. A typical run lasts just 2 h.

4.2.2.6 Limitations of the NGS Methods

The NGS methods generate short length reads that are not easy to assemble as genome sequences because plant genomes contain extensive repeat sequences. In view of this, various sample preparation strategies like mate pair libraries/large insert libraries, paired-end reads, preparations from sorted chromosome, RNA-Seq data, optical mapping, reduced representation libraries, and information from genetic mapping are used to facilitate genome assembly. In fact, few plant genome sequences of high quality have been completed using NGS technologies. These methods use PCR for generating copies of the DNA fragments. This step inevitably introduces bias so that the quality of coverage of different genomic regions is not uniform. In addition, sequencing is based on synthesis or hybridization reaction that uses as template millions of copies of a given fragment. It is expected that reactions at all the copies of a single template fragment will occur in synchrony. This, however, may not happen and some copies may fall out of synchrony; this would introduce error in the base sequence of the given fragment. Each NGS platform provides its own software package for signal acquisition and "base calling" (deduction of bases on the basis of light color and intensity signals) with minimum error rates. In addition, several other base-calling algorithms have been developed (Table 4.2) that reduce base-calling error rates by ~5–30 % over the methods provided with the NGS platforms. But the software packages provided with the NGS platforms are the most widely used.

Sample/template preparation for NGS technologies takes several days, which often involves additional equipment costs, chemicals and other consumables, and physical space. Although NGS technologies generate sequence data at a lower cost per base sequenced, they have greatly increased the size of projects due to, among other things, the huge amounts of sequence data generated, which has created challenges for their storage, analysis and management. The third-generation sequencing methods are based on single DNA molecules, and they do not suffer from the above limitations.

4.2.3 The Third-Generation DNA Sequencing Methods

The *third-generation sequencing methods* do not use PCR amplification for template preparation because they sequence single DNA molecules (Schadt et al. 2010). For this reason, they are often called *single-molecule sequencing* (*SMS*) *methods*. The technologies being developed for TGS are quite diverse and include captured DNA polymerase, nanopores, electronic detection, fluorescence energy transfer, and transmission electron microscopy. Two of these methods emerged as feasible DNA sequencing options during 2011. Some of the TGS technologies are briefly described in the following sections.

Table 4.2 Some of the freely available NGS data analysis and SNP and genotype-calling software packages

Software	Available from	Prerequisites for application	Remarks/functions available
Single-sample calling			
SOAP2	http://soap.genomics.org.cn/index.html	High-quality variant database[a]	NGS data analysis; includes genotype caller SOAPsnp
realSFS	http://128.32.118.212/thorfinn/realSFS/	Aligned reads	SNP and genotype calling; uses allele frequencies
Multi-sample calling			
Samtools	http://samtools.sourceforge.net/	Aligned reads	NGS alignments; computation of genotype likelihoods (samtools); SNP and genotype calling (bcftools)
GATK	http://www.broadinstitute.org/gsa/wiki/index.php/The_Genome_Analysis_Toolkit	Aligned reads	NGS data analysis; SNP and genotype calling (UnifiedGenotyper), SNP filtering (variant filtration); SNP quality recalibration (variant recalibrator)
Multi-sample and LD-based calling			
Beagle	http://faculty.washington.edu/browning/beagle/beagle.html	Candidate SNPs, genotype likelihoods	Imputation, phasing, and association, including genotype calling
IMPUTE2	http://mathgen.stats.ox.ac.uk/impute/impute_v2.html	Candidate SNPs, genotype likelihoods, fine-scale linkage map	Imputation, phasing, and association, including genotype calling
QCall	ftp://ftp.sanger.ac.uk/pub/rd/QCALL	"Feasible" genealogies at a dense set of loci, genotype likelihoods	SNP and genotype calling, generating candidate SNPs without (NLDA) and with (LDA) LD information
MaCH	http://genome.sph.umich.edu/wiki/Thunder	Genotype likelihoods	SNP and genotype calling; generating candidate SNPs without (GPT_Freq) and with (thunder_glf_freq) LD information

[a]For example, dbSNP

4.2.3.1 Helicos Genetic Analysis System

In this method, 100–200-bp-long template fragments are subjected to tailing to generate over 50-nucleotide-long poly(dA) tails at their 3′ ends, followed by blocking of the 3′ ends with a suitable treatment. These fragments are now hybridized with primers [50-nt-long poly (dT)] immobilized on a proprietary substrate within a glass microfluidics cell having 25 channels (Fig. 4.6). The dNTPs used for DNA synthesis are labeled with a bright fluorophore, e.g., Cy3 and Cy5, so that the dNTPs incorporated into single growing chains are readily detected. The four labeled dNTPs (blocked with virtual terminators) are added sequentially, one at a time. When molecules of a given dNTP are added, they will be incorporated at the 3′ ends of those primers/growing chains that are associated with the template molecules having the base complementary to the given dNTP at the proper site. The fluorescence from the incorporated nucleotide is recorded separately for each template molecule. The fluorophores of the incorporated nucleotides and the terminators are removed, and the next dNTP along with DNA polymerase is added. In this way, base sequence of each template molecule is determined.

The length of each read is ~35 bases, and up to one billion reads (and 35 Gb sequence data) can be obtained in one run. Since a virtual terminator is used, a dNTP can be incorporated only at a single site in a template during each reaction cycle even when its complementary base occurs at two or more consecutive sites in the template. The raw read error rate is generally 0.5 %, but the finished/consensus error rate tends to be much lower. Helicos BioSciences Corporation, USA, has commercialized this process as Helicos Sequencer, HeliScope™. This system generates 1 Gb usable sequence data per day (~100 times greater than the first-generation sequencers).

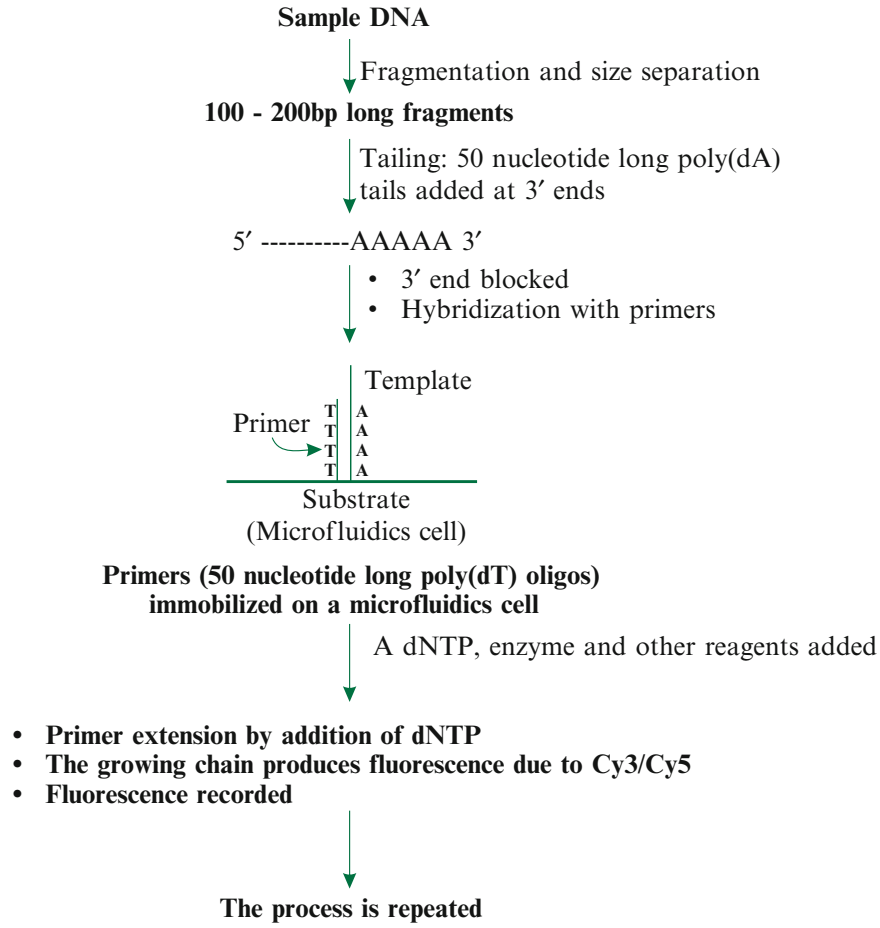


Fig. 4.6 A schematic representation of the Helicos third-generation sequencing method. Use of bright fluorophores (Cy3, Cy5) allows signal detection from replication of a single DNA molecule. This technology has been adapted for direct sequencing of RNA molecules without production of cDNA. In the case of RNA species without 3′ poly(A) tails, poly(A) tails are added to their 3′ ends. Primers (50-nucleotide-long poly(dT) oligos) immobilized on a microfluidics cell; dNTPs labeled with Cy3 or Cy5 and virtual terminators; read length ~35 bases

4.2.3.2 Single-Molecule Real-Time Technology

The *single-molecule real-time* (*SMRT*) *technology* was developed by Pacific Biosciences, USA, and was commercialized as PACBIO RS. This is the most revolutionary approach as it is based on single molecules of DNA polymerase immobilized (by biotin–streptavidin interaction) in zepoliter (10^{-21} L) wells of nanometers in diameter and depth. Each well provides a detection volume of only 20 zepoliters. High concentrations of the four dNTPs labeled with different fluorophores are used for rapid DNA replication. Each DNA polymerase molecule will use a single DNA fragment as template to add the fluorophore-labeled dNTPs to the primer/growing chain (Fig. 4.7). A highly focused detection system continuously records the fluorescence from the nucleotides added to the growing chain in each well. Since the fluorophore is attached to the phosphate moiety, it is automatically removed as the next nucleotide is added, and it diffuses out of the vicinity of DNA polymerase molecule. Since the detection system is focused onto the DNA polymerase molecule, the liberated fluorophore molecules do not interfere with the detection process. The DNA polymerase can sequence the DNA

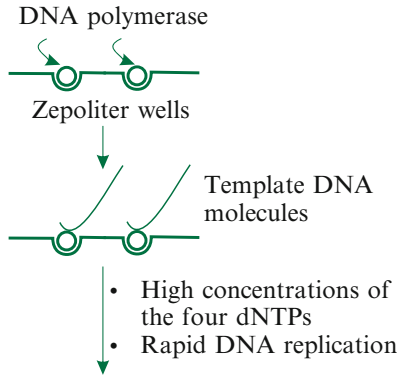


Fig. 4.7 A simplified representation of the single-molecule real-time (SMRT) method of DNA sequencing. Zepoliter $= 10^{-21}$L (Based on Schadt et al. 2010)

fragment more than once, producing multiple coverage of the same molecule (Schadt et al. 2010; Deschamps and Campbell 2010).

The sequencing platform generates 20 Gb sequence data per 30 min. The average read length is ~1,000 bp, while the maximum read length is over 10,000 bp. But an improved technology allows sequencing of up to 20 kb fragments, and efforts are being made to increase it to 40 kb. The raw read error rates may exceed 5 % mainly in the form of insertions and deletions. The use of SMRT bell sample preparation system allows sequencing of both the strands of a DNA molecule in a single cycle, which increases the consensus accuracy of sequence data. It can be used for detection of DNA methylation pattern by using suitable software and for direct RNA sequencing without the need for cDNA preparation. This method uses minimum amounts of reagents and does not require template preparation, and there are no PCR, scan, and wash steps.

4.2.3.3 The Nanopore Sequencing Technologies

In the case of most nanopore sequencing technologies, the DNA molecule and its component bases are passed through an extremely narrow hole (a nanopore), and the component bases are detected by the changes in an electrical current or optical signal caused by them (Schadt et al. 2010). Genetically engineered proteins or a suitable chemical compound may be used to construct the nanopores. The Oxford Nanopore Technologies, UK, uses BASE technology that creates the nanopore by an engineered protein (α-hemolysin). Around 2,000–8,000 nanopores are placed in a lipid bilayer built on a special application-specific integrated circuit chip. At the extracellular face of the nanopore, an exonuclease is attached, while a synthetic cyclodextrin-based sensor is linked at its inside surface; the cyclodextrin acts as the binding site for DNA bases (Fig. 4.8). The DNA sample to be analyzed is restriction digested, the digest is placed onto the chip, and one DNA fragment associates with each nanopore. An enzyme separates the two strands of the DNA duplex, and the exonuclease digests one strand, one base at a time, and passes these bases through the nanopore. Each base sequentially binds to the cyclodextrin located on the inside of the nanopore. This binding creates a disturbance in the electric current passing through the nanopore, which

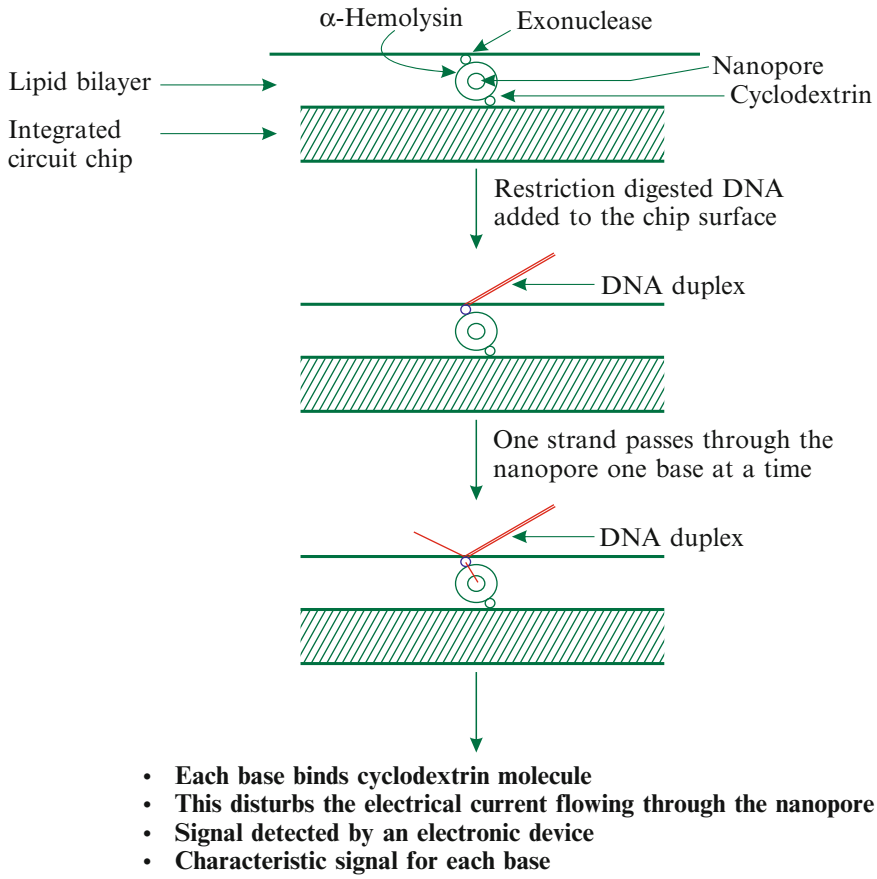


Fig. 4.8 A schematic representation of a nanopore sequencing technology (developed by Oxford Nanopore Technologies, UK). The nanopore is created by an engineered α-hemolysin; exonuclease cleaves the terminal bases one by one and passes them through the nanopore; cyclodextrin binds to the base; this creates disturbance in the electrical current flowing through the nanopore (Based on Schadt et al. 2010)

generates characteristic signal for each DNA base. This signal is sensed by an electronic device and is converted into base sequence data. This technology can detect cytosine methylation without any special chemical processing of the template.

Oxford Nanopore Technologies is preparing to launch two models, namely, MinION and GridION, for sales. MinION USB stick DNA sequencer is the size of a USB drive, is projected to cost less than US $ 1,000, works with a PC, has a lifetime of 6 h from activation, and would generate up to 150 Mb sequence data. The GridION system is designed for bigger runs, uses a standalone machine, and would be able to analyze RNA and protein as well. These systems have an error rate of 4 %.

4.2.3.4 Other Third-Generation Sequencing Technologies

Several other highly innovative third-generation sequencing technologies are in different stages of development, some of which are briefly mentioned here. IBM is developing a DNA transistor that would electronically identify individual bases in a single DNA molecule. NABsys is trying to develop the existing solid-state technologies for whole-genome sequencing based on electronic detection of bases. Genia, on the other hand, is developing a nanopore technology that relies on electrical real-time sequencing of single DNA molecules. The Starlight technology uses fluorescence resonance energy transfer (FRET) for real-time sequencing

of single DNA molecules. Another technology uses a specialized technique with a high-resolution (sub-angstrom) transmission electron microscope for identification of the DNA bases by direct imaging of the base sequence (Edwards 2013).

4.2.4 Comparison Between NGS and TGS Sequencers

The NGS sequencers are simpler to use, very fast, extremely high throughput and comparatively much cheaper, the Illumina Genome Analyzer being the cheapest. In addition, they do not require in vivo cloning and carry out the necessary template preparation in a matter of hours. Finally, they are versatile and can be used for a variety of analyses. The TGS technologies sequence single DNA molecules, are faster and cheaper, and enable a much higher throughput than the NGS sequencers. The error rate of the TGS methods is higher because the opportunity for error removal on the basis of sequencing of multiple copies of each fragment is not available. The NGS sequencers yield shorter read lengths due to the degrading effects of lasers on DNA and enzymes. Further, the washing, which must be done after each cycle, slowly reduces the amount of DNA available for sequencing. Finally, in the case of NG sequencers, asynchronous reactions may increase the error rate, which builds up through the cycles.

4.3 RNA Sequencing

Usually, RNA sequencing involves the production of cDNA by reverse transcription PCR (RT-PCR) and then sequencing of the cDNA product. If the primers for RT-PCR were correctly designed, only the desired mRNA species will be copied as DNA. Initially, Sanger–Coulson method of DNA sequencing was used for sequencing of cDNA/EST (expressed sequence tag) libraries. But this approach does not have high throughput, is expensive, and does not permit quantitative analysis of gene expression. In addition, approaches based on this strategy were generally unable to distinguish among different splicing isoforms.

4.3.1 RNA-Seq

The NGS technologies enable sequencing of complete transcriptomes in almost any population or tissue; this approach is referred to as *RNA-Seq*. RNA-Seq is used for both qualitative and quantitative analyses of genome-wide gene expression. It has also been used to discover up to hundreds of thousands of SNPs (Chepelev et al. 2009; Wang et al. 2009c) at costs similar to those from reduced representation and low-coverage methods (Sects. 13.4 and 13.6). However, RNA-Seq is more likely to discover functional SNPs than other SNP discovery methods. In general terms, the procedure for RNA-Seq consists of isolation of RNA, production of cDNA by reverse transcription, and then sequencing the cDNA population using a suitable NGS technology. The RNA preparation may comprise the total RNA or it may be a specific fraction of the total RNA, e.g., mRNA with poly(A) tails (Fig. 4.9). In the case of long RNA molecules, the RNA molecules themselves or their cDNAs are fragmented to produce, ultimately, cDNA fragments of sizes suitable for NGS sequencing. The fragments are ligated with adapters at one or both the ends, and each fragment is sequenced at one end (single-end sequencing) or both the ends (paired-end sequencing). Typically, reads of 30–400 bases long are obtained, depending on the NGS technology and the sequencing strategy (single-end/paired-end sequencing) used.

The sequence reads are aligned to a reference genome sequence to generate a *genome-wide transcription map* depicting the transcriptional status of all the genes present in the genome as well as their expression levels. But when a reference genome sequence is not available, the reads can still be assembled to produce the transcription map de novo. The available software for mapping of the reads include ELAND, SOAP,

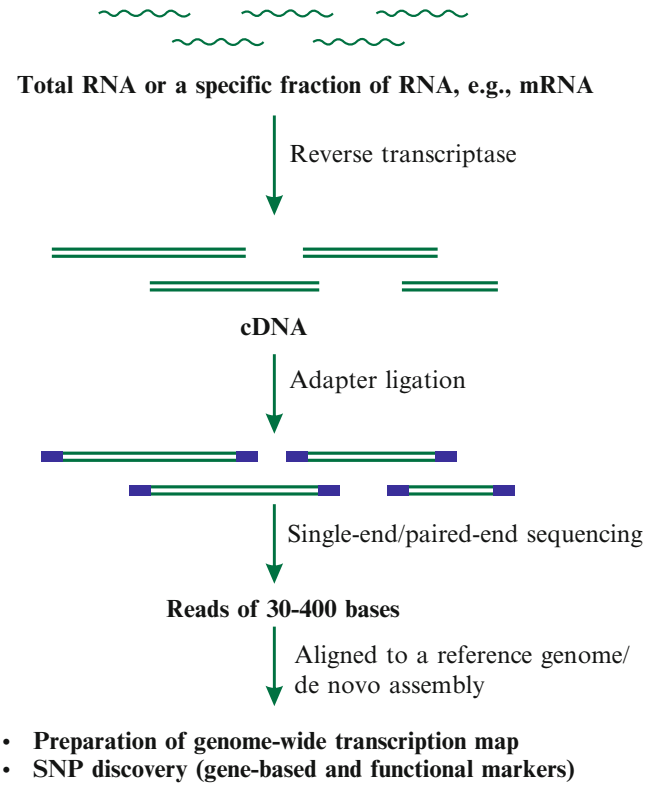


Fig. 4.9 A schematic representation of RNA-Seq technology. The RNA sample may be the total RNA or a specific fraction of RNA, e.g., mRNA. Longer RNA molecules are fragmented either as RNA or as cDNA to generate fragments of suitable size for the NGS technology to be used (Based on Chepelev et al. 2009; Wang et al. 2009c)

MAQ, and RMAP. RNA-Seq allows high-throughput qualitative as well as quantitative analyses of the entire transcriptome. It has revealed many novel features of eukaryotic genomes like overlapping in the 3′-regions of many yeast genes, novel transcribed regions in every genome studied, new splicing isoforms of known genes, 5′ and 3′ boundaries of the transcribed regions of many genes, etc.

RNA-Seq has high resolution, sensitivity, and reproducibility, generates very low background noise, and yields highly accurate quantitative data on gene expression. It requires relatively small quantities of RNA and is particularly suited for transcriptome analysis in non-model organisms. The chief limitation of this approach is the difficulty in inferring genotypes from expression data; this is complicated due to alternative splicing that produces multiple RNA molecules from a single primary RNA transcript. Further, bias may be introduced by cDNA preparation, RNA/cDNA fragmentation, and PCR amplification of the cDNA. For example, the nascent cDNA being synthesized by reverse transcriptase may dissociate from the template RNA molecule and anneal to a new RNA molecule that has a sequence similar to that of the first RNA template. This event, called *template switching*, generates a cDNA molecule made up of the 3′ region of the first RNA template and the 5′ region of the second RNA template. Reverse transcriptases can cause self-priming and, thereby, generate up to 10 % random cDNAs, which are a major source of error. Reverse transcriptases are error prone as they lack proofreading ability. Often the range of dynamic expression may need to be normalized; this becomes problematic when a reference genome is not available. Finally, efficient methods are required for storage, retrieval, and processing of large datasets and for reducing the base sequence errors.

4.3.2 Single-Molecule Direct RNA Sequencing

Helicos BioSciences, USA, has developed and commercialized the technology for direct sequencing of single RNA molecules in a massively parallel sequencing operation by the Helicos® Genetic Analysis System (Ozsolak and Milos 2011). This technology does not involve conversion of RNA to cDNA, PCR amplification, or ligation, uses only minute quantities (several femtomoles) of RNA, and provides deep sequence coverage of the transcriptome. For applications like expression profiling of poly(A)$^+$ RNA encoding genes or mapping of polyadenylation sites, the RNAs are directly used for sequencing. But in studies involving RNA species without poly(A) tails, 3′ polyadenylation of the RNA molecules is carried out. The RNA molecules are now hybridized with the poly(dT) primers immobilized onto the flow cell; the RNA molecules are "filled and locked" and sequenced by synthesis. The read lengths are up to 55 nucleotides (average, 33–34 nt). Each run may yield 800,000–8,000,000 reads per channel, and there are up to 50 channels in the 2 flow cells that can be run simultaneously. Total raw base error rate is 4–5 % (primarily deletions and insertions). A sequence aligner freely available from the Helicos BioSciences HeliSphere significantly reduces the InDel error rates.

4.4 Single-Nucleotide Polymorphisms

Single-nucleotide polymorphisms (*SNPs*, pronounced as "snips") describe variation among different individuals of a species for single base pairs at the corresponding sites of their genomes. Thus a *SNP locus* is a specific position in the genome, at which different nucleotides occur in the same DNA strand of different individuals of the species. *Therefore*, *each SNP locus has to be defined by the sequence flanking the polymorphic nucleotide*. Often insertions and deletions (InDels) are also analyzed as SNPs. The nucleotide polymorphism at a genomic position is considered as SNP only when the least frequent allele has a frequency of 1 % or more. A SNP locus can have four alleles, each allele being represented by one of the four DNA nucleotides (Fig. 4.10). However, many SNP loci have three or even two alleles; in fact, two-allele SNP loci predominate in humans. In any case, SNPs are usually scored as biallelic markers. SNPs are produced by either transition (C to T, A to G, and vice versa) or transversion (A or G to C or T and vice versa). In general, transitions seem to be more frequent than transversions. At least a proportion of C to T (and, consequently, G to A) transitions is produced due to deamination of 5-methylcytosine; this is more likely to occur in genomic regions rich in CpG dinucleotide sequence.

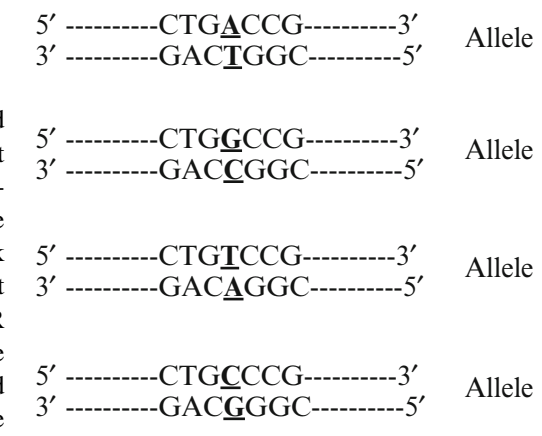


Fig. 4.10 The four possible alleles at a SNP locus. In humans, most SNP loci have only two alleles

SNPs are extremely abundant (about one SNP every 100–300 bp of plant genomes), have relatively low mutation rates, and are relatively easy to detect. SNP density varies among genomes of different species and among different genomic regions of the same species. In general, SNPs are more frequent in noncoding regions than in the coding regions due to a lack of selection pressure in the former. SNPs may generate phenotypic effects by altering either the amino acid sequence of the encoded protein or the splicing pattern of the RNA transcripts. They may also

affect promoter activity and, thereby, generate phenotypic effects. SNPs have proved ideal for automation, and high-throughput marker discovery and analysis; in addition, strategies for combining these two activities have also been developed (Chap. 13). Modern SNP genotyping platforms are supported by improved bioinformatics tools that afford robust automated allele calling and generate high-quality data. This data can be easily shared across groups and stored in common databases irrespective of the genotyping platform used. Many SNP genotyping platforms are capable of efficient, fast, and high-throughput sample processing at increasingly lower cost per data point. However, in the context of plant breeding activities, genotyping cost per sample is more relevant than per data point. Therefore, *a breeder may select an optimal number of SNP loci for each application*; *this might markedly reduce the total genotyping cost*.

The chief limitations of this marker system are the high equipment cost particularly for high-throughput genotyping. The marker development involves resequencing of even whole genomes, which is rather costly. The genotyping procedure is technically demanding and may not be a feasible proposition for many breeding programs. In such cases, it may be desirable to use commercial SNP genotyping services provided adequate funds are available.

4.4.1 Types of SNPs

SNPs are classified in a variety of ways based on different criteria, including genomic location, the effect on phenotype, etc. SNPs located in the noncoding regions of genome are called noncoding SNPs (ncSNPs), while the ncSNPs located in introns are known as intronic SNPs. Exonic SNPs or coding SNPs are found in exons and are comparable to copy SNPs (cSNPs or cDNA SNPs: SNPs discovered in cDNAs). An exonic SNP that does not lead to a change in the amino acid sequence of the concerned protein is called a synonymous SNP (synSNP), while a nonsynonymous SNP (nsSNP) alters the amino acid sequence. A human nsSNP is known as diagnostic SNP, when it is involved in a genetic disease. However, genic SNPs occur in genes and would include intronic and exonic SNPs, as well as the SNPs located in the promoter region of the concerned gene (promoter SNPs or pSNPs). Some of the genic SNPs will affect the function of the concerned gene and would give rise to phenotypic effects; these are termed as functional SNPs or candidate SNPs. But anonymous SNPs do not affect the function of a gene and do not produce a phenotypic effect; most SNPs belong to this category. A reference SNP (refSNP, rsSNP, rsID) is a SNP that serves as a reference point for defining neighboring SNPs. Each refSNP is assigned an rsID number when it is submitted to a databank, e.g., dbSNP. SNPs discovered by mining ESTs or genomic databases are generally called in silico SNPs (isSNPs) or electronic SNPs (eSNPs); these are "virtual" polymorphisms and must be validated by resequencing. Many SNPs located close to each other tend to be inherited together. The alleles of such SNPs located in the same chromosome together constitute SNP haplotype; such SNPs are referred to as haplotype-tagged SNPs. Generally, genotyping only a small number of carefully selected SNP loci from a haplotype block allows the deduction of genotypes at the remaining SNP loci of the block; these SNPs are termed as "tag" SNPs.

The SNPs in polyploid species have been classified as simple SNPs and hemi-SNPs or homoeo-SNPs. A simple or true SNP detects allelic variation between homologous loci of the same genome present in the same or different polyploid species, and it does not detect differences in their other genome(s). This group of SNPs would show typical diploid segregation in most mapping populations, is quite frequent (10–30 %), and would be the most useful for mapping. But hemi-SNPs or homoeo-SNPs, on the other hand, detect homoeologous/paralogous loci in the two or more genomes of the polyploid species and of their diploid progenitors. Therefore, these SNPs are of limited value for mapping (Deschamps and Campbell 2010).

4.5 Methods for Discovery of SNPs

It may be pointed out that all SNPs are initially discovered by sequencing, which remains the method of choice. Sequencing may involve the whole genome, a specific region of the genome, or the transcriptome. One of the major problems in SNP discovery has been the predominance of highly repetitive sequences in plant genomes. Therefore, early efforts at SNP discovery attempted to avoid repetitive sequences by resequencing unigene-derived amplicons and in silico SNP discovery by mining the EST databases, followed by their PCR-based validation. However, these approaches detected gene-based SNPs and did not discover SNPs in the noncoding regions of genes and the intergenic spaces. In addition, amplicon resequencing is expensive and labor intensive. Similarly, many of the SNPs discovered from EST databases were non-allelic in several crops because these SNPs represented paralogous sequences produced by duplication of the concerned genomic regions.

Prior to the development of NGS technologies, whole-genome sequencing was a daunting task. Therefore, it was highly desirable to minimize the sequencing effort by focusing on the genomic regions of interest; amplicon sequencing (Sect. 4.5.1) and sequence capture (Sect. 4.5.6) strategies serve this purpose. But the emergence of NGS technologies has made SNP discovery by whole-genome sequencing a feasible option. In addition, several reduced representation strategies aim at combining SNP discovery with SNP genotyping, using a suitable NGS technology, at reasonable costs (Chap. 13). Further, huge amounts of genome and EST sequence data have accumulated in various databases, which can be mined for SNP discovery. Transcriptome sequencing by RNA-Seq technology is also being used for SNP discovery.

4.5.1 Amplicon Sequencing

In this approach, a pair of specific primers is used for PCR amplification of the desired genomic region, and the PCR product (*amplicon*) is sequenced for identification of SNPs and InDels. This strategy limits the sequencing and analysis efforts to the genomic region of interest and, thereby, reduces the workload. When Sanger–Coulson sequencing is used, separate amplification of each amplicon is necessary. Further, in the case of heterozygotes or when PCR amplification is based on pooled DNA, the amplicons have to be cloned (to separate the amplicons representing the two alleles present in the heterozygotes or in different individuals) before they are sequenced. But the NGS technology has rendered these steps unnecessary since each read is generated from a single amplicon. Therefore, NGS technology permits pooling of tissues, genomic DNAs/cDNAs, or amplicons from different individuals. This approach reduces the quantum of work for template preparation and permits the discovery of all the SNPs, including the rare alleles. In case of pooling, a greater depth of sequencing should be used; in general, the depth would be greater for shorter read lengths. According to an estimate, at least 34×, 101×, and 110× sequencing depth would be needed with 454, ABI SOLiD, and Illumina GA, respectively, for separating sequencing errors from genuine SNPs. However, pooling does not permit determination of marker genotypes and haplotypes of the individuals/lines. In addition, PCR amplification of pooled DNA may lead to preferential priming of certain alleles. These difficulties can be removed by separate PCR amplification of each individual/line, using separate barcodes for each of the amplicons and then pooling the amplicons before sequencing. This approach increases the amount of work for template preparation as well as the total cost. However, it increases the usefulness of data as it combines marker discovery with genotyping of the individuals/lines.

Read length and sequencing depth are critical for detection of rare alleles, identification of InDels, and for eventual marker development. Deep sequencing minimizes false negatives, and ensures detection of genuine SNPs and discrimination of rare alleles from sequencing errors. Short reads enable discovery of InDels of 1–8 bp, while longer reads from 454 sequencing

platform permit identification of InDels of 1 to over 97 bp. Amplicon sequencing can be extended to even such species, for which sequence information is not available. For achieving this, trans-specific or universal primers are designed on the basis of conserved sequences of the target genes extracted from a related species for amplification of orthologous genes in the uncharacterized species. This approach has been successfully used in some plant species. *Orthologous genes* are those genes of different species that perform the same function. In contrast, *paralogous genes* are the genes present in the genome of the same species and have the same function; these genes are produced by duplication, polyploidization, or both.

The limitations of amplicon sequencing include size limit of amplicons (10–20 kb for long-range PCR), base substitutions due to PCR, requirement of sequence information for primer designing, amplification of paralogues by the specific primers based on sequences conserved among paralogues, overrepresentation of amplicon ends in reads, and uneven coverage of internal regions of amplicons. Many of these problems can be mitigated by suitable strategies, including rigorous quality control during sample preparation and bioinformatics tools.

4.5.2 SNP Mining

The simplest, most convenient, and highly efficient method for SNP discovery is bioinformatics analysis of the ever-increasing genomic and/or EST sequences of different individuals available in the databases of the concerned species. In addition, an investigator may sequence genome/ESTs of a genotype/line/individual of interest and analyze the sequence so obtained along with the sequences available in the database. Bioinformatics tools like PolyPhred are used for deducing the base sequence of fragments, assembly of the deduced sequences into contigs, and editing of the contigs. Suitable computer software like SNP Pipeline are then used to align the sequences and detect SNPs. Sequencing errors present in the database, particularly in the genomic regions that are not well characterized, may lead to discovery of false SNPs. Special software like POLYBAYES help minimize false discovery of SNPs due to sequencing errors. The analysis of genomic sequences will identify SNPs located in both coding and noncoding regions of the genome, while EST analysis will discover SNPs only in the coding regions. At present, most of the SNP mining activity is directed at EST databases, possibly for the above reason. Further, the search may be focused at specific regions of the genome that have been either known to be associated with the traits of interest or to contain genes with specific functions. The SNPs discovered by SNP mining are often termed as in silico SNPs (isSNPs) or electronic SNPs (eSNPs). However, *these SNPs must be validated by resequencing*.

4.5.3 Transcriptome Sequencing

Transcriptome sequencing allows rapid and inexpensive discovery of genic SNPs and avoids highly repetitive genomic regions. The NGS, RNA-Seq, and direct RNA sequencing technologies can be used for transcriptome sequencing. The sequence reads are aligned to a reference genome or to EST sequences to discover SNPs and InDels. In case a reference genome is not available, genome sequence of a related species or of the parental species may be used for sequence alignment and marker discovery. Alternatively, the sequence reads can be assembled de novo using appropriate bioinformatics tools. Analysis of EST/transcriptome sequence data also permits discovery of SSR markers. For example, an analysis of watermelon EST sequences obtained from an experiment and the EST datasets obtained from the GenBank enabled the discovery of 5,000 SSRs. Useful markers can also be found in the 3′ UTRs (untranslated regions) of mRNAs. In general, longer sequence reads are preferred for marker discovery as they facilitate sequence alignment and discrimination among paralogues in the case of polyploid species. Paired-end reads overcome

to some extent the above limitations of short reads.

Markers discovered by transcriptome sequencing will be, of necessity, gene-based markers, and a proportion of them will be functional markers. But many QTLs, regulatory sequences like enhancers, locus control regions, etc., are located in noncoding regions of the genome. As a result, it will be unable to discover markers useful for mapping of such QTLs and regulatory elements. Transcriptome sequencing coupled with appropriate experimental design would permit the determination of allele-specific differences in gene expression, estimation of the parental contributions to heterosis, and the role of genetic imprinting in development and performance. However, transcriptome analysis-based marker discovery is limited to only those genes that are transcribed in the concerned tissue/organ during the given developmental stage and under the environmental conditions prevailing at the time of sample collection. Therefore, a fair number of organs/tissues, developmental stages, and environments should be sampled to ensure the representation of most, if not all, of the genes present in the genome of the concerned species. In contrast, sequencing of hypomethylated partial restriction genomic libraries (Sect. 4.5.5) provides a more complete representation of SNPs located in genes than transcriptome sequencing and allows the detection of SNPs situated in introns, regulatory regions, and non-transcribed genes.

Assembly and analysis of NGS data requires appropriate software programs, for which a variety of options are available. A de novo assembly of RNA-Seq sequence data yields contigs, which are called *tentative ESTs* or *tentative unique sequences* (*TUSs*). Bioinformatics tools are used to filter the SNPs discovered from RNA-Seq data, and the filtered SNPs are usually validated by Sager–Coulson sequencing. For example, in one study in maize, transcriptome analysis of shoot apical meristems from two inbreds permitted the detection of 36,000 putative SNPs; these were reduced to 7,000 after stringent processing of the sequence data. Sager–Coulson sequencing was used for confirmation of a sample of 110 from these SNPs, and 85 % of them were successfully validated. Transcriptome analysis of polyploid species requires a more complex experimental design and a comparison with the diploid ancestral species for assigning the tentative ESTs to the homoeologous chromosomes. Unlike the concerned genomes, the RNA transcripts rarely contain repetitive sequences, which is a definite advantage in proper sequence alignment. The error rate of NGS sequence data is rather high, but availability of a good quality reference genome considerably reduces the error rate. It is important to use an optimum sequencing depth because a low sequencing depth would lead to higher error rate and "false negatives." Transcriptome sequencing has been successfully used in several crop species, including maize, canola, sugarcane, wheat, etc.

4.5.4 Whole-Genome Sequencing

Often SNP discovery has been based on whole-genome sequencing of a small number of selected individuals/lines; this approach remains the method of choice wherever resources and other considerations do not preclude this option. *SNP discovery is greatly facilitated by the availability of a good quality reference genome sequence*. One may reduce the sequencing effort by pooling DNAs from the selected individuals/lines and constructing a genomic library from the pooled DNA. Random clones may be picked from this library and used for sequencing. The shotgun sequences so obtained are processed, using appropriate bioinformatics tools for discovery of SNPs. It may be pointed out that the sequencing depth should be large enough (at $>20\times$ coverage) not only to yield sequence data with minimum error but also to ensure that sequence of a given genomic region is available from enough number of individuals/lines to allow SNP discovery. The term sequencing depth may refer to specific nucleotides or to the entire genome. *Sequencing depth for a specific nucleotide* represents the total number of all reads, in which a given genomic position (or base pair) from a given individual is

represented; these reads may be obtained from a single sequencing experiment or from a series of experiments. But *sequencing depth for the whole genome* is the average number of times each base of the genome (the entire genome) of an individual has been sequenced. The sequencing depth for the whole genome is generally referred to as *coverage of sequencing* and is denoted as 10×, 20×, 30×, etc., coverage or depth. The general formula for coverage (C) is $C = LN/G$, where L is the read length, N is the total number of reads, and G is the length (in bp) of the haploid genome of the concerned species. It should be kept in mind that coverage denotes average sequencing depth of the genome as a whole; therefore, some genomic regions would be sequenced at much higher depth, while some others would be sequenced at much lower depth than the coverage level. The minimum coverage level required for a study depends on many factors, including the type of study, gene expression level, the trends in published literature, etc.

The analysis of sequence data for SNP discovery proceeds in several steps. In the case of NGS data, the first step involves image analysis and base calling with the minimum error rate. This can significantly reduce false-positive SNP calls and facilitate sequence assembly, particularly when the coverage is low to moderate. The short sequence reads are then aligned onto the reference genome whenever it is available; this is known as *read alignment* or *read mapping*. The alignment algorithms should be able to handle both sequencing errors, as well as potentially real sequence differences, in the form of base substitutions and InDels, between the reference and the newly sequenced genomes. In addition, the aligners should generate well-calibrated alignment quality values, which are important for variant calling, i.e., determining the genomic positions at which at least one base differs from the reference genome. It has been recommended that Novoalign or Stampy should be used as aligners, and GATK or SOAPsnp should be used for recalibration of per base quality scores (Nielsen et al. 2011). This is followed by realignment of reads, removal of duplicate reads, and a recalibration of the quality scores for each base. Both SNP and genotype calling at a given genomic position depend on the accuracy of base calls as well as the per-base quality scores of the reads overlapping the genomic position.

SNP calling is the determination of the genomic positions at which nucleotide polymorphisms occur. It can be based on data from a single individual/line (*single-sample calling*) or it may simultaneously use data from all individuals in the sample (*multi-sample calling*). As far as possible, multi-sample calling should be used, and the calling methods should involve likelihood ratio tests or Bayesian procedures. Similarly, *genotype calling*, i.e., assigning of SNP alleles to different individuals in the sample, should be done by combining the data from all the individuals in a Bayesian framework, and information on known SNPs (e.g., those listed in dbSNP), linkage disequilibrium (LD), etc., should be included to improve the accuracy of genotype and SNP calls. A number of filtering steps based on a variety of criteria like generally low-quality scores, systematic differences in quality scores of major and minor alleles, aberrant LD patterns, strand bias, etc., may be implemented to improve the accuracy of SNP and genotype calls. Most of the software used for NGS processing carry out both SNP and genotype calls (Table 4.2). Several additional steps like local realignments, combining results from multiple SNP- and genotype-calling algorithms, etc., can be implemented to improve genotype calls. Finally, uncertainty should be incorporated in the subsequent statistical procedures for analyzing the data. It may be pointed out that analysis of NGS data is evolving rapidly, and new tools for data analysis are being continuously developed. Therefore, the choice of most suitable software package for any task will keep on changing with time.

4.5.5 Reduced Representation Approaches

Genome sequencing of a sample of individuals/lines has to be resorted to when either genome sequences are not available or it is desirable to

use genome sequences of a set of new lines/individuals. Sequencing of whole genomes is the ideal strategy, but it involves considerable time, effort, and financial and other resources. Further, sequencing of whole genomes may not be necessary for many types of studies. In view of the above, many strategies for simultaneous SNP discovery and genotyping have been developed (Sects. 13.3–13.6). In general, these methods sample a fraction of the whole genome for sequencing so that the cost and effort for marker discovery and genotyping are greatly reduced. One approach aims to enrich the sampled fragments with gene-rich regions by construction of a hypomethylated partial restriction (HMPR) library as follows. The genomic DNA of the target individual/line is digested completely with a 5-methylcytosine-sensitive restriction enzyme like *Hpa*II (5′ C/CGG 3′) with a 4 bp recognition sequence. The digest is subjected to electrophoresis; fragments of 100–600 bp are separated and used for sequencing by an NGS technology. The genomic regions having repetitive DNA are usually hypermethylated; consequently, they will be present as much larger fragments and will be excluded. This approach may eliminate ~95 % of the maize genome and enrich the selected fragments four- to five-fold for genic sequences. Sequencing of the gene-enriched fragments from two maize inbreds allowed the identification of a large number of putative SNPs. However, it restricts SNP discovery to the regions near the recognition sites of the enzyme used for digestion. Therefore, two or more 5-methylcytosine-sensitive restriction enzymes with distinct recognition sequences should be used to get a more complete representation of the genic regions.

4.5.6 Sequence Capture

Sequence capture is a targeted SNP discovery strategy applied to specific genomic regions. This strategy can be applied when the genomic region of interest is known and a closely related reference genome sequence is available. It involves designing of oligonucleotide probes or primers specific for the genomic regions of interest to permit their separation or amplification before sequencing. There are three main strategies for sequence capture, viz., SureSelect from Agilent, SeqCap from NimbleGen, and the Targeted Sequencing System from RainDance (Davey et al. 2011). All the three technologies are proprietary, require the customer to provide the target region sequences, and use in-house bioinformatics tools to design probes/primers for the target regions. The *NimbleGen SeqCap technology* uses oligonucleotide probes synthesized on microarray slides, and the lengths of the probes are adjusted to obtain a uniform melting temperature. The genomic DNA fragments are hybridized with the microarray and the captured fragments are used for sequencing. The *Agilent SureSelect* method, on the other hand, implements in-solution target sequence capture using biotinylated RNA probes of 120 nt. The genomic DNA fragments already ligated to sequencing adapters are hybridized with the probes, and the hybridized fragments are separated by exploiting the high affinity of biotin for streptavidin. In both these technologies, repeat sequences are filtered out from the probe set by using specific software programs.

The *RainDance Targeted Sequencing System*, in contrast, uses two rounds of PCR to specifically amplify fragments from the targeted genomic region. This is achieved by designing a set of PCR primer pairs using proprietary software so as to cover most of the genomic region of interest. Each primer pair of the set has at its 3′ end the sequences specific for a segment of the target genomic region, while its 5′ end comprises partial sequence of the adapter for the selected NGS technology. The target-specific primers are used for the first round of PCR amplification. In the second round of PCR, universal primers with the partial NGS adaptor sequences at their 3′ ends are used. The PCR products generated from the second round of PCR are directly used for sequencing. The SureSelect and SeqCap methods capture about 90 % of the targeted genomic region, while the Targeted Sequencing System may capture over 95 % of the region. These

technologies do not appear to introduce a substantial bias in allele representation in the sequence data. The reference genome used for designing the probes/primers should be of high quality and closely related to the population under study.

4.5.7 Validation of Discovered SNPs

Once a group of SNPs has been discovered, each locus should be evaluated to ascertain the following: (1) that it is a true SNP and not a product of sequencing error, faulty read alignment, etc., (2) that its alleles represent homologous genomic regions and not paralogous/homoeologous regions, and (3) that it segregates in a typical Mendelian fashion. The above evaluation is generally referred to as *SNP validation*. One approach for SNP validation is to resequence the concerned genomic regions of carefully selected individuals/lines to confirm that the discovered SNPs represent true polymorphisms. A more informative validation process involves designing a suitable assay for the discovered SNPs and to apply this assay to evaluate a set of diverse germplasm lines or, preferably, a segregating population. This procedure will reveal the discovered SNPs to be real or false, their ability to discriminate among the germplasm lines, and their segregation pattern in the segregating population. The choice of assay will depend mainly on the number of SNPs to be validated. The assays in common use for a large number of SNPs are Illumina's GoldenGate (Sect. 13.2.8) and Infinium assays, TaqMan OpenArray Genotyping system (Sect. 13.2.4), and Kompetitive Allele-Specific PCR (KASP) assay (Sect. 13.2.3). The length of SNP context sequence, the total number of SNPs to be genotyped, and the available funds would have to be considered while selecting one of these assays. *It may be pointed out that SNP validation still remains a cost-intensive procedure*. SNP validation in allopolyploids would be facilitated by the use of haplotype and allele frequency information, and application of bioinformatics tools like HaploSNPer. This strategy would be useful even for diploid species like barley that have highly repetitive genomes.

4.6 Methods for SNP Genotyping

The various SNP genotyping methods/platforms range from scoring of a single SNP marker to a very large number of markers assayed using high-density SNP chips, and they are suited for a wide range of applications. These methods rely on strategies that are able to distinguish between a perfect match from a single base mismatch between an oligonucleotide and the template DNA strand. These strategies are based on nucleic acid hybridization, primer extension, oligonucleotide ligation, DNA replication, or single-strand invasion coupled with cleavage of the displaced strand (Sobrino et al. 2005). The different genotyping methods include allele-specific PCR, 5′-nuclease assay, high-density oligonucleotide arrays or DNA chips, bead-based techniques, primer extension, invasive cleavage or invader technology, MALDI-TOF MS-based homogeneous MassEXTEND (hME) assay, and pyrosequencing. In addition, PCR products can be subjected to restriction enzyme digestion (cleaved amplified polymorphic sequences, CAPSs; Sect. 3.14), electrophoresis of single strands (single-strand conformation polymorphism, SSCP; Sect. 3.15), or denaturing gradient gel electrophoresis (D/TGGE; Sect. 3.16) for genotyping of known SNPs. These techniques can be broadly classified into gel-based and gel-free assays; the latter group of assays is preferred because the methods in this group are amenable to high-throughput analysis leading to economy of time and other resources.

4.6.1 Allele-Specific PCR

Allele-specific PCR is designed to amplify only one of the alleles at a SNP locus (Okayama et al. 1989). It uses a pair of primers, one of which is based on a conserved sequence present in all the alleles. The other primer of the pair is specific to the genomic region having the SNP

locus, and the base at its 3′-end corresponds to the SNP locus. When the 3′ terminal base of the second primer is complementary to the SNP allele, it pairs with the allele and supports amplification of the genomic region and yields a PCR product detectable by gel electrophoresis (Fig. 4.11a). A mismatch at the 3′ end of the primer greatly reduces the chances of amplification so that no amplification product would be detectable (Fig. 4.11b). Therefore, allele-specific PCR generates a dominant STS marker scored as "present"/"absent." But sometimes a single-base mismatch at the 3′ end of a primer is unable to prevent amplification. On the other hand, sometimes amplification may fail due to an error in the setting up of the experiment. The first difficulty is resolved by introducing a mismatch at the second base from the 3′ end of the primer (Fig. 4.11c). This mismatch will still allow amplification of the allele for which the primer has been designed but will effectively prevent amplification of the other alleles. The second problem can be overcome by using four different primers, i.e., one primer for each SNP allele, and screen every

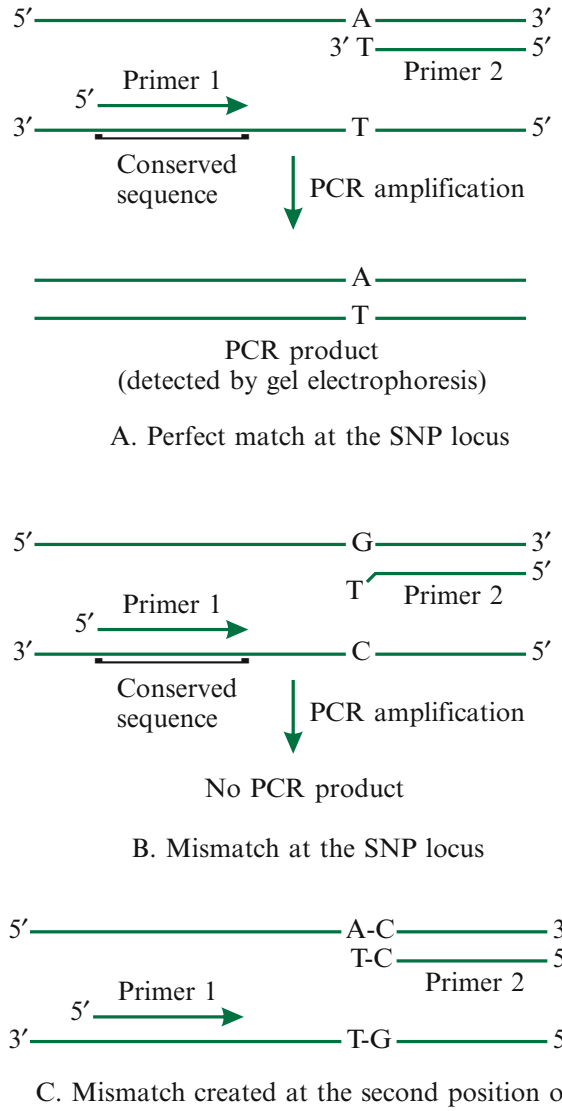


Fig. 4.11 A simplified representation of allele-specific PCR for genotyping of a SNP locus. The mismatch at the second base from the 3′ end of the primer 2 increases effectiveness of allele discrimination. This does not prevent amplification in case the 3′ terminal base is matched with the SNP allele, but it definitely prevents amplification in case of a mismatch

individual with all the four primers. If there is an experimental error, amplification will fail with all the four primers. But when there is no error, one primer is expected to generate amplification product in every individual. The four primers can be designed in such a way that each of them amplifies product of a different length, or a different fluorophore may be attached to each primer. This would allow all the four primers to be used in a single PCR tube for each individual. The allele-specific PCR is a user-friendly approach for SNP analysis by any laboratory with PCR facility. However, the overall throughput is low, and only a small number of SNPs can be analyzed by this approach. This strategy has been modified as KASP™ genotyping assay for a high-throughput SNP and InDel genotyping (Sect. 13.2.3).

4.6.2 5′-Nuclease Assay (TaqMan® Assay)

This technique gets its name from the fact that it uses the 5′-nuclease activity of Taq polymerase in real-time PCR to quantify the hybridization of allele-specific oligonucleotides with the genomic DNAs of the test individuals and deduces the SNP allele from this information. It uses two PCR primers for amplification of the target sequence, i.e., the genomic region containing the SNP locus, and a specifically designed probe, TaqMan™ probe, complementary to that region of the target sequence that has the SNP locus (Livak 1999). This probe has a fluorescent dye attached to its 5′ end and a quenching dye linked to its 3′ end. As long as the fluorescent dye molecule is located near the quenching dye molecule, there will be no fluorescence due to the quenching action of the latter. The base at the 5′ end of the probe is complementary to the SNP allele it detects. In case the 5′ end of the probe is paired properly with the SNP allele present in the target sequence, the 5′-nuclease activity of Taq polymerase will cleave the whole probe beginning at its 5′ end. This will free the fluorescent dye molecule, which will diffuse away from the quenching dye; as a result, it will now generate fluorescence (Fig. 4.12). But if the base at the 5′ end of the probe is not complementary to the SNP allele, there will be mismatch, and Taq polymerase will not be able to cleave the probe at its 5′ end. As a result, there will be no fluorescence. One may design one TaqMan probe for each allele at an SNP locus, label them with different fluorophores, and use them in a single PCR tube. In such a case, the ratios between the fluorescence of different colors will permit a highly reliable scoring of the SNP alleles.

The TaqMan® assay is homogeneous, quick (on an average, 2 h per run), and simple, PCR and data calling occur simultaneously in real-time mode, and the throughput is high. The assay generates 2,000 data points per day per person in a monoplex mode; it can also be run in a duplex mode to generate up to 3,000 data points per day per person. However, the procedure is based on a costly real-time PCR machine, and the costs of labeled probes and other consumables are high. The TaqMan® assay has been commercialized as the high-throughput TaqMan OpenArray Genotyping system (Sect. 13.2.4) by Applied Biosystems, USA. It has also been adapted for a cost-effective medium multiplexing, high-throughput SNP genotyping platform using nanofluidic dynamic arrays (Sect. 13. 2.6).

4.6.3 Molecular Beacons

Molecular beacons are specially designed oligonucleotide hybridization probes used for identification of SNP alleles. The central region of a molecular beacon is complementary to the sequences flanking the target SNP locus, including the SNP allele to be detected (Sobrino et al. 2005). The sequences on either side of the central region are universal sequences, and they are complementary to each other. A fluorophore is attached to the 5′ end of the probe, while a quenching dye is attached to its 3′ end. The probe molecules will form a hairpin structure due to pairing between their 3′ and 5′ end regions. This pairing will bring the quenching dye in close

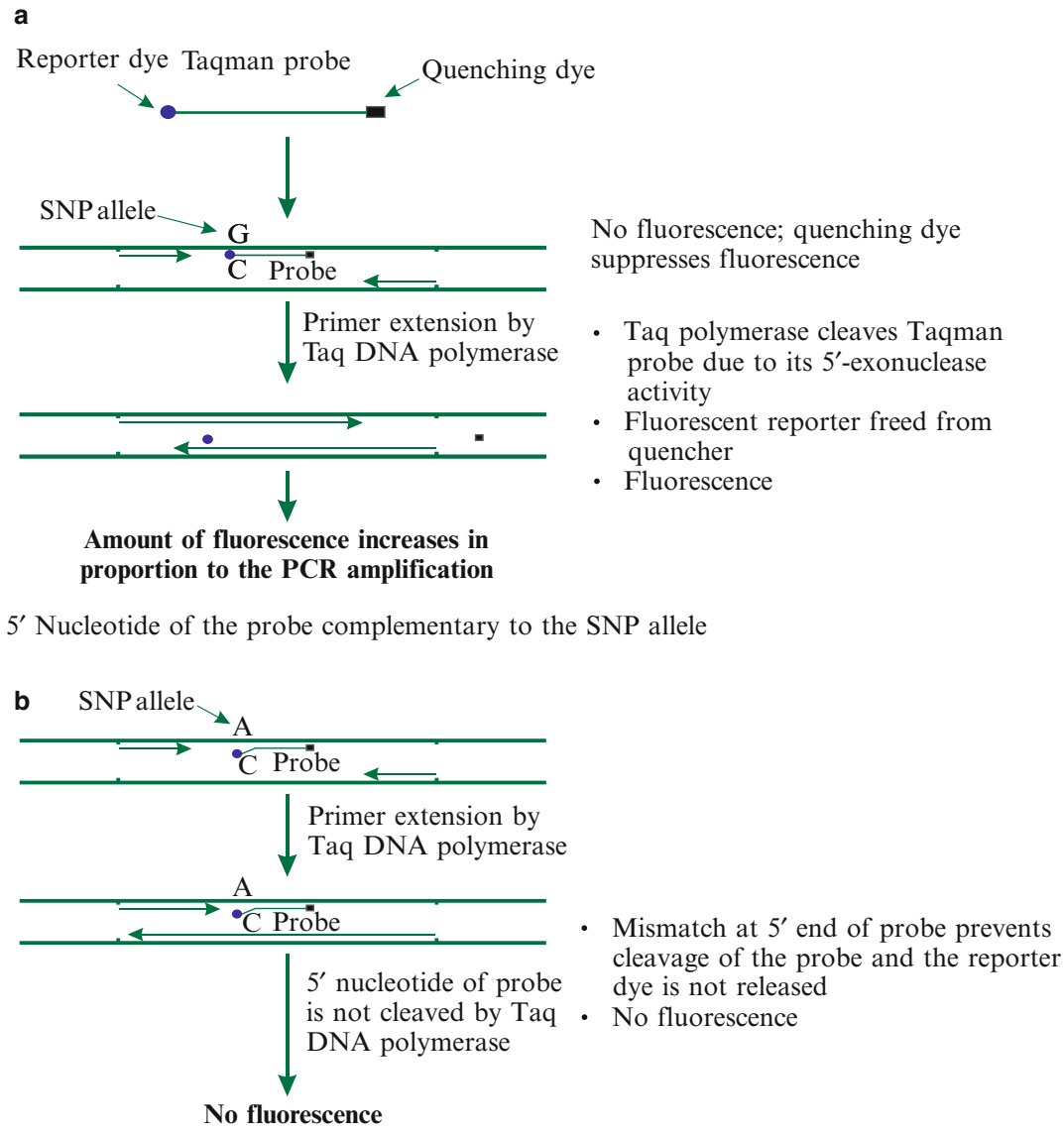


Fig. 4.12 Use of TaqMan™ probe to quantify PCR products. Primer 1 and Primer 2 are the two PCR primers; Probe is TaqMan™ probe that has a fluorescent reporter at its 5′ end and a quencher at the 3′ end. (**a**) 5′ Nucleotide of the probe is complementary to the SNP allele: fluorescence is produced as the reporter dye is released by 5′ exonuclease action of Taq DNA polymerase. (**b**) 5′ Nucleotide of the probe is not complementary to the SNP allele: mismatch at 5′ end of the probe prevents its cleavage and the release of the reporter dye. Therefore, fluorescence is not produced

proximity to the fluorophore, due to which there will be no fluorescence (Fig. 4.13a). But when the probe base pairs with the specific SNP allele, it becomes linear and the quenching dye becomes separated from the fluorophore, and fluorescence is generated (Fig. 4.13b). A molecular beacon is mixed with denatured PCR product representing the concerned genomic region of the test individual/line and allowed to anneal. If the allele at the target SNP locus is complementary to the beacon, the two will base pair and there will be fluorescence (Fig. 4.13b). But if the SNP allele were not

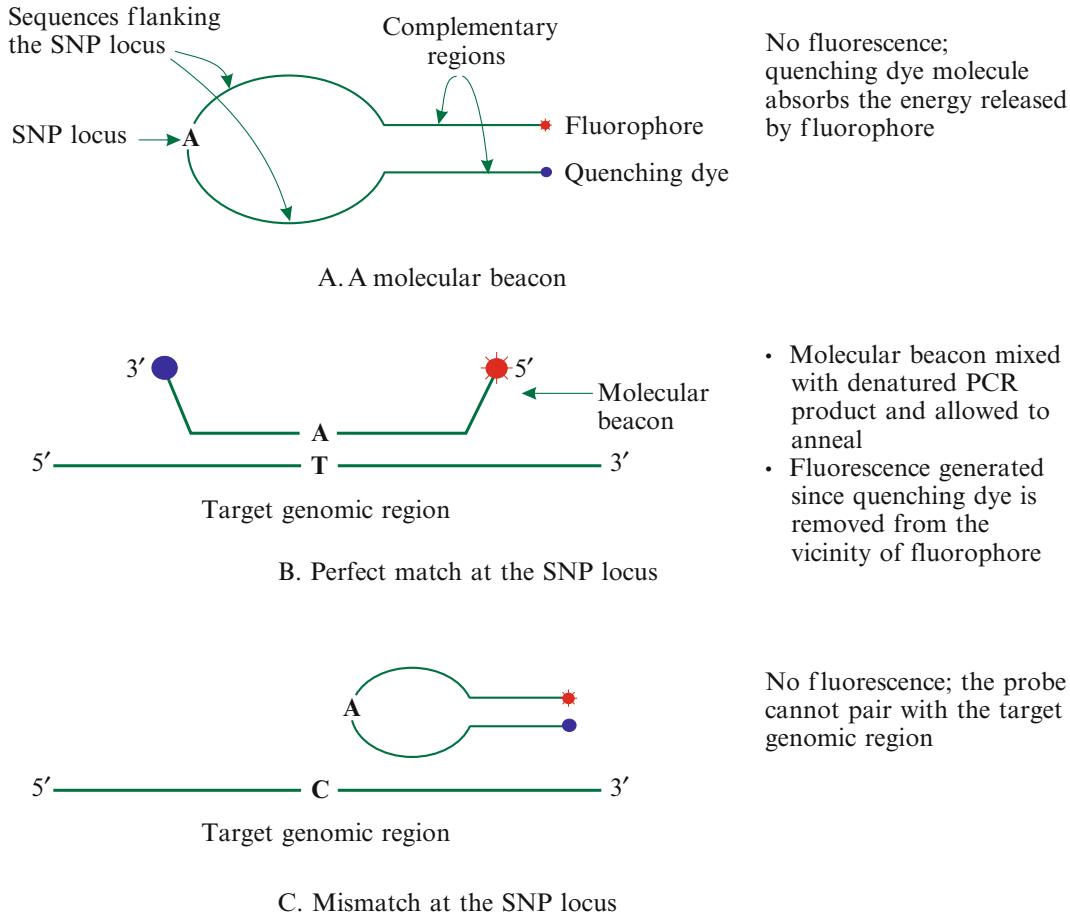


Fig. 4.13 A schematic representation of molecular beacon and its use for detection of SNP alleles. Molecular beacon is mixed with denatured PCR product representing the concerned genomic region, allowed to anneal, and then fluorescence is monitored (Based on Sobrino et al. 2005)

complementary to the beacon, there will be no pairing and fluorescence (Fig. 4.13c). A suitable sensing device detects the fluorescence signal, which is used to deduce the SNP allele.

Some degree of multiplexing can be achieved by labeling two or more molecular beacons, specific for different SNP loci, with different fluorophores and using them in a single reaction vessel. However, most detection systems use monochromatic light for excitation of fluorophores, which limits the number of different fluorophores that can be assayed together efficiently. One strategy to overcome this difficulty employs two fluorophores, one harvester fluorophore and one emitter fluorophore arranged serially, at the 5′ end of the probe in the place of single fluorophores used normally.

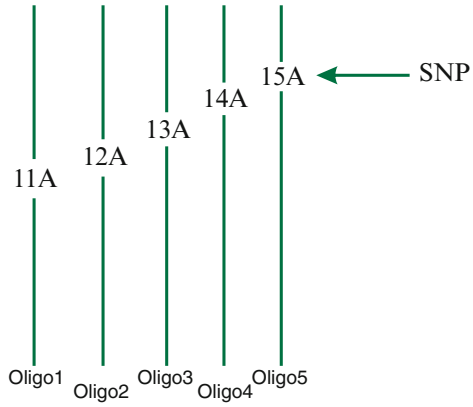
4.6.4 Microarray-Based SNP Genotyping

Microarray-based SNP genotyping requires the development of SNP microarrays/DNA chips for simultaneous genotyping at several SNP loci. A subset of the polymorphic SNP loci is selected mainly on the basis of their position in the genome, the level of polymorphism and suitability for the assay, and used to construct a microarray. A *microarray* or *DNA chip* is a

small plaque/wafer of silicon, glass, or metal, onto which one end of a large number of different single-stranded DNA molecules is covalently linked and arranged in spots (Appendix 2.3). Each spot has several copies of a single DNA molecule of 25 nt representing the SNP locus and includes the nucleotide involved in the SNP around its middle position. In order to ensure high reliability, each SNP allele is represented by five different oligonucleotides; in each of these oligonucleotides, the variable SNP is located at a different position, ranging from two bases on one side of the central base to two bases on the other side (Fig. 4.14a). At the same time, each of the oligonucleotides is spotted at two to three different locations (Fig. 4.14b), which serve as replications and help eliminate false-positive signals (possibly due to nonspecific hybridization). It may be pointed out that the SNP locus and the sequences surrounding this locus influence the hybridization efficiency. Therefore, it is very difficult to optimize the conditions for detection of a panel of SNPs using an array, and ingenious approaches are used to overcome this difficulty (Sobrino et al. 2005).

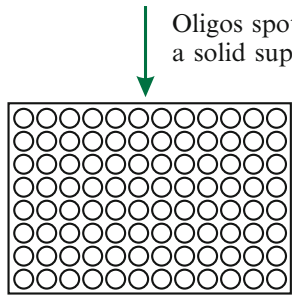
Genomic DNA from each individual to be genotyped for SNPs is used for a series of PCR reactions to amplify all the short genomic regions having the different SNPs. For this reason, each SNP locus is first converted into an STS by designing a pair of primers for its reliable amplification. The PCR products are labeled by fluorescence, and all the PCR products from a single individual are pooled and used for hybridization with the DNA chip (Fig. 4.14). The non-hybridized PCR products are removed by washing under such conditions that permit only perfectly base-paired PCR products to remain associated with the oligonucleotides spotted onto the chip. A fluorescence scanner is used to measure fluorescence at each spot on the chip, and the data are analyzed with the help of image analysis software. Since the position of each oligonucleotide on the chip is known, the alleles present at different SNP loci are readily deduced. This approach simultaneously analyzes all the SNP loci of the test individual/line. Microarrays can also be used for simultaneous genotyping of a large number of individuals/lines at a given SNP locus. This type of assay may be needed in certain situations, e.g., during MAS. In such a case, the PCR products representing the concerned SNP locus from individual plants of the relevant segregating generation are spotted onto a glass slide. This microarray is hybridized with labeled probes representing the alternative alleles of the concerned SNP locus, and plants with the desirable SNP allele are identified. This technique is referred to as *tagged microarray marker approach*; it has been successfully used in the case of humans and pea.

Wang et al. (1998) were the first to use DNA chips for SNP genotyping in humans. The microarrays have to be custom made for every species and whenever the panel of SNP loci is altered. The development of chips involves considerable amount of work like designing of STS primers for each SNP locus and construction of the oligos for every locus. Since hybridizations for all the SNPs are carried out simultaneously under the same conditions, the oligos must be designed with great care so that they all have identical requirements for perfect hybridization. This requires considerable expertise, and specialized software have been developed for this purpose. The synthesis of oligos onto the chips requires the construction of expensive "masters" for each set of oligos. Therefore, the initial development of SNP chip for a species is very costly, but the subsequent production of a large number of identical chips may be much cheaper. As a result, SNP chips are relevant only for large projects. The efficiency of discrimination between completely matched and mismatched oligos in hybridization is much lower than the ability of DNA polymerases or DNA ligases to distinguish between them. This problem is particularly aggravated in the case of microarrays since many different oligos need to be hybridized under a single set of conditions; this adversely affects the accuracy of genotype calls. Therefore, universal microarrays that can be used in any species with any set of SNPs have been developed, e.g., the Illumina's "Sentrix Array Matrix" for the GoldenGate assay (Sect. 13.8).



The numbers 11, 12, 13, etc. denote the base position corresponding to the SNP allele in the 25 nucleotide long oligos

A. Five oligos representing a single SNP locus



The five oligos are spotted in three replicates

B. Microarray

- **PCR amplification of the genomic regions corresponding to all SNP loci of an individual/line**
- **PCR products labeled with fluorescence**
- **PCR products from one individual/line pooled and hybridized with probes on the chip**
- **Washing leaves only perfectly paired PCR products hybridized with the probes**
- **Fluorescence measured at each spot of microarray**
- **Fluorescence data analyzed to deduce SNP alleles**

Fig. 4.14 A simplified schematic representation of microarray-based genotyping of SNP loci on the basis of hybridization of PCR products with probes on the microarray (Based on Sobrino et al. 2005)

4.6.5 Bead-Based Techniques

The *bead-based techniques* are similar to microarray method, but they use oligos attached to fluorescent microbeads of 3–5 μm diameter for hybridization (de Vienne 2003). The microbeads are coated with a combination of two fluorescent dyes (red and orange). Different concentrations of the two dyes are combined to generate beads of several different types. The bead types can be distinguished from each other by flow cytometry on the basis of intensity and wavelength of the fluorescent light emitted by them. One can generate 100 different types of microbeads by combining 10 different fluorescence intensities with 2 different wavelengths. To each bead type, several copies of an oligo representing a specific allele of a particular SNP locus are attached. Each SNP locus is represented by two oligos corresponding to the two alleles of the locus,

which are attached to two different bead types. Thus the set of 100 bead types will enable simultaneous analysis of 50 SNP loci, each having two alleles. All the bead types are pooled and distributed into different tubes prior to hybridization.

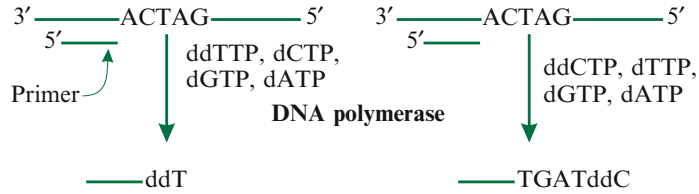
Fluorescence-labeled PCR products corresponding to the SNP loci represented on the microbeads from an individual are hybridized with the pooled microbeads. The non-hybridized PCR products are removed by washing. The beads are passed in a single row through the capillary of a flow cytometer, where they are exposed to two laser beams. The data on the levels of fluorescence in response to the two laser beams are recorded. These data enable the identification of the microbead type and, thereby, the SNP locus and its allele being examined. In addition, the fluorescence level of the PCR product, i.e., the genotyping signal, is also recorded. This signal reveals the "presence" or "absence" of the particular SNP allele. The flow cytometer can examine thousands of microbeads in a few seconds. Data from a large number of beads are collected, and the mean values of fluorescence of the PCR products for each bead type are calculated. This allows deduction of the alleles at the different SNP loci.

The technique has high-throughput potential but has the same limitations as DNA chips. The level of multiplexing is limited by the availability of only green color for the genotyping signal. At present, the use of a 96-well flow fluorometer would permit scoring of thousands of genotypes in a single 96-well format reaction. The bead-based approach has been successfully used for genotyping on the basis of allele-specific hybridization, allele-specific primer extension, single-base extension, and oligonucleotide ligation assay. The microarray- and bead-based techniques are not freely available as they are "closed" or proprietary technologies.

4.6.6 Primer Extension

The *primer extension* method involves annealing of a specially designed primer to the target PCR product, extension of the primer by one to few nucleotides using DNA polymerase (Sokolov 1990; Braun et al. 1997), and analysis of the products of extension to deduce the allele at the SNP locus. This primer is so designed that the base at its 3′ end is complementary to the base just preceding the polymorphic base of the SNP locus present in the PCR product (Fig. 4.15a). As a result, the first nucleotide added to the primer will be complementary to the polymorphic base of the SNP locus. Initially, one ddNTP and the remaining three dNTPs were used in a reaction mixture for primer extension. As a result, for each PCR product, four separate reactions, each using a different ddNTP, had to be set up. In case the ddNTP present in a reaction mixture was the first nucleotide to be added to the primer, there will be no further extension of the primer. But if one of the dNTPs was the first to be added, the primer extension will continue up to the point, at which the base complementary to the concerned ddNTP occurs in the PCR product (Fig. 4.15a). The products of primer extension are analyzed by either electrophoresis in an automated DNA sequencer or by MALDI-TOF MS (matrix-assisted laser desorption ionization time of flight mass spectrometry). The ddNTP permitting addition of only a single base to the primer is identified; the base complementary to this ddNTP will be present at the SNP locus. The primer extension approach has been developed as the homogeneous MassEXTEND (hME) assay for high-throughput SNP genotyping (Sect. 13.2.5; de Vienne 2003).

Alternatively, the four ddNTPs are used in a single reaction mixture for each PCR product so that the primer will be extended by a single nucleotide only (*single-base extension*, *SBE*). Phosphodiesterase II digestion is used to trim the 5′ ends of the products of primer extension, and the molecular weights of the shortened products are determined by MALDI-TOF MS. This permits an accurate identification of the ddNTP added to the extended primer and deduction of the SNP allele. MALDI-TOF MS analysis takes merely 4 s per sample, but the equipment is very expensive, and it requires high expertise. In addition, an extremely



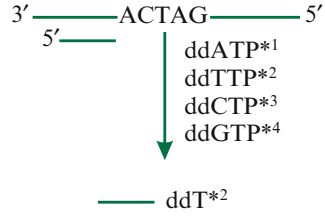
For each PCR product, four reactions are set up; in each reaction a different ddNTP and the remaining three dNTPs are included to support DNA synthesis

- Only one nucleotide (ddTTP) added to the primer
- No further extension of the primer is possible due to the addition of ddTTP
- SNP allele deduced to be A

- dTTP is the first nucleotide added
- DNA synthesis continues till ddCTP is added to the primer
- The product is much longer in this case
- SNP allele is not G

- The products from the other two reactions (with ddGTP/ddATP) will also be longer than those from the first reaction
- Products analyzed by electrophoresis/MALDI-TOF-MS

A. Primer extension (the initial scheme)



A single reaction set up for each PCR product; ddATP, ddTTP, ddCTP and ddGTP, each labelled with a different fluorophore, included in the reaction mixture; primer extended by a single nucleotide (ddTTP*2), which is identified by the fluoroscence

- **Added nucleotide identified by the fluorophore**
- **SNP allele deduced to be A**

B. Single base extension (SBE)

Fig. 4.15 A schematic representation of primer extension and its modification called single-base extension (*SBE*). The first A in the sequence ACTAG of the PCR product represents the SNP locus. *1, *2, *3, and *4, the four distinct fluorophores used to label the ddNTPs. ddNTP, dideoxynucleotide; *ddATP* 2′,3′-dideoxyadenosine triphosphate, *ddTTP* 2′,3′-dideoxythymidine triphosphate, *ddCTP* 2′,3′-dideoxycytidine triphosphate, *ddGTP*, 2′,3′-dideoxyguanosine triphosphate, *MALDI-TOF-MS*, matrix-assisted laser desorption ionization time of flight mass spectrometry (Based on de Vienne et al. 2003)

sophisticated laboratory setup is essential for an optimum use of the mass spectrometer. On a smaller scale, the four ddNTPs can be labeled by different fluorophores, each giving a different color on fluorescence. Since in a given reaction mixture, only one of the four ddNTPs will be added to the primer, the fluorescence color of the product will permit easy identification of the added ddNTP and, thereby, the deduction of the allele present at the SNP locus (Fig. 4.15b).

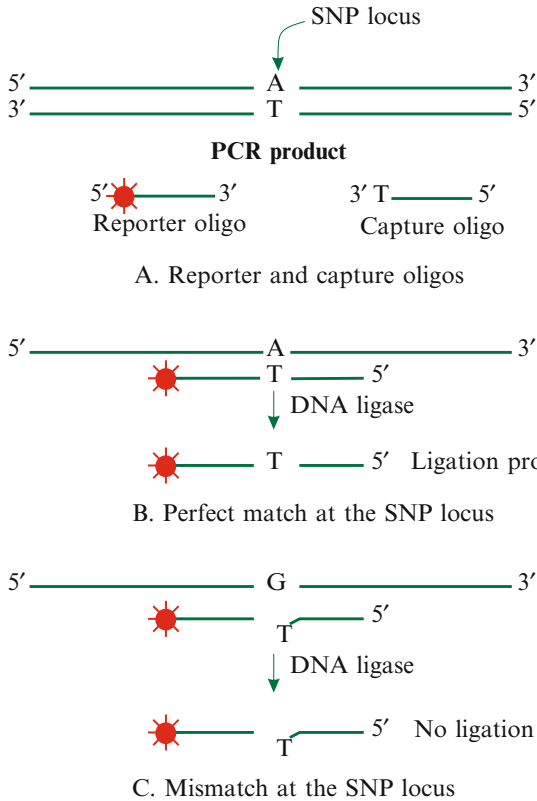
SBE approach has been used to develop diagnostic assays and microarrays for high-throughput genotyping. The SBE assay is also called *genetic bit analysis* (GBA) or *mini-sequencing*. SBE has been used to develop a diagnostic tool, in which the primer is bound to a microtiter plate well. The PCR product is denatured and allowed to anneal to the bound primer. DNA polymerase adds a single nucleotide, corresponding to the SNP site, to the primer, which allows direct determination of the SNP allele. Applied Biosystems, USA, has used this strategy for its 5–10-plex, medium-throughput genotyping system called SNaPshot®. Multiplexing is achieved by using primers of different lengths (from 23 to 60 nt). The primers for different loci differ by four to five nucleotides, and detection is based on capillary electrophoresis. The use of a 96 capillary system allows one person to generate over 10,000 data points per day.

4.6.7 Pyrosequencing

Pyrosequencing has been used for NGS technology (the 454 sequencing technology), which is suitable for both SNP discovery and SNP genotyping. The use of this technology for SNP genotyping is considered in some detail in Chap. 13 (Sect. 13.2.2).

4.6.8 Oligonucleotide Ligation Assay

Oligonucleotide ligation assay (*OLA*) utilizes a pair of oligos for hybridization with the PCR products, followed by ligation of the two oligos by DNA ligase (Landergren et al. 1988). One of the two oligonucleotides is complementary to the SNP locus, i.e., it includes the polymorphic nucleotide, and the sequence on the upstream of the SNP locus; this is known as *capture oligo*. The other oligo is complementary to the sequence on the downstream side of the SNP locus and does not include the SNP locus; this is called *reporter oligo*. The reporter oligo is labeled with a fluorophore. The pair of oligos thus represents contiguous regions including the SNP locus (Fig. 4.16a). The two oligos, the PCR product representing the target genomic region of an individual, and DNA ligase are added to a reaction mixture, heated to denature the DNA, and then cooled to permit their annealing. The two oligos would pair perfectly to the PCR product if the base at the 3′ end of the capture oligo were complementary to the SNP allele in the PCR product. DNA ligase will ligate the two oligos to generate a product having the combined lengths of the two oligos (Fig. 4.16b). However, if the base at the 3′ end of the capture oligo were



The target region is amplified by PCR. Capture oligo represents the SNP locus at its 3′ terminus, while reporter oligo is labelled and corresponds to the 5′region next to the SNP locus

PCR product, the two oligos heated to denature DNA, cooled to allow annealing of oligos to the template, DNA ligase ligates the two oligos. The quantity of ligation product can be increased by LCR

There is mismatch at the SNP locus. Therefore, DNA ligase fails to ligate the two oligos

Fig. 4.16 A simplified schematic representation of oligonucleotide ligation assay. The 3′ terminal base of capture oligo represents the base involved in the SNP. Steps of LCR are similar to those of PCR, viz., denaturation, annealing, and ligation. A thermostable DNA ligase like Taq DNA ligase is used for LCR. LCR, ligase chain reaction (Based on Sobrino et al. 2005)

not complementary to the SNP allele, there will be mismatch at this base, and the ligation reaction will be highly inefficient. Therefore, a negligible amount of the ligation product will be produced (Fig. 4.16c).

The quantity of ligation product can be greatly increased by using a thermostable ligase like Taq DNA ligase in a ligase chain reaction (LCR) procedure that is similar to PCR. The reaction mixture is repeatedly heated to denature the DNA and then cooled to allow hybridization of the two oligos with the PCR product, followed by ligation of the two oligos to generate the product. The OLA procedure can be used in combination with the DNA chip or bead-based techniques to overcome the difficulty in designing of the oligos with the same optimum hybridization conditions. However, the combined procedure is quite complex and demanding. Ligation-based assays are more amenable to multiplexing than primer extension-based assays since ligation is less prone to interference between primers. But OLA tends to be more expensive due to the use of SNP-specific fluorescent primers, while the single-base extension reaction uses a common set of fluorescent ddNTPs for all the SNP loci. The OLA assay system has been modified to develop the 96- and 192-plex assay system SNPlex™ that exploits the specificities of different DNA ligases. OLA is also used for the Illumina's highly multiplexed GoldenGate™ assay (Sect. 13.2.8; Sobrino et al. 2005).

4.6.9 Dynamic Allele-Specific Hybridization

The *dynamic allele-specific hybridization* (*DASH*) uses specific probes for hybridization with the target PCR products (de Vienne et al. 2003). It discriminates between perfect pairing and mismatch at the SNP locus of the PCR product on the basis of relative melting temperatures of the duplexes so produced and thus deduces the SNP allele. One of the two primers used to amplify the PCR product is conjugated with biotin. This PCR product is added to a microtiter plate well coated with streptavidin, to which biotin binds. Thus one strand of the PCR product remains attached with the microtiter plate well, while the other strand is washed away with alkali (Fig. 4.17). This single-stranded preparation is hybridized at a low temperature with an oligonucleotide probe specific for one allele of the SNP locus. An intercalating dye specific for double-stranded DNA (dsDNA) is added into the well. The intensity of fluorescence generated by this dye will be proportionate to the amount of dsDNA. The microtiter well is now gradually heated, and the fluorescence intensity is continuously monitored. There would be a rapid fall in fluorescence intensity as the dsDNA begins to denature. Under appropriate conditions, mismatch at a single base pair, i.e., the SNP locus, leads to an easily detectable lower melting temperature than that with perfect pairing. The sequence of the oligo used for hybridization with the PCR product together with the relative melting temperature of the duplex so formed allows deduction of the SNP allele at this locus. This assay procedure is quick and can be used for reliable scoring of all SNP types, and a suitable device for its implementation is available.

4.6.10 Denaturing High-Performance Liquid Chromatography

In the *denaturing high-performance liquid chromatography* (*dHPLC*) procedure, ion-pair reversed-phase high-performance liquid chromatography is used to separate perfectly matched DNA homoduplexes from heteroduplexes having one or more mismatched base pairs (de Vienne et al. 2003). The PCR product of a test individual is mixed with the PCR product of a reference individual that has a known allele at the SNP locus. The mixture is heated to denature the DNA and then cooled to permit renaturation (Fig. 4.18a). If the test PCR product is exactly the same as the reference PCR product, all DNA duplexes will be perfectly matched, and only one peak of elution will be detected. But if the SNP allele in the test PCR product is different from

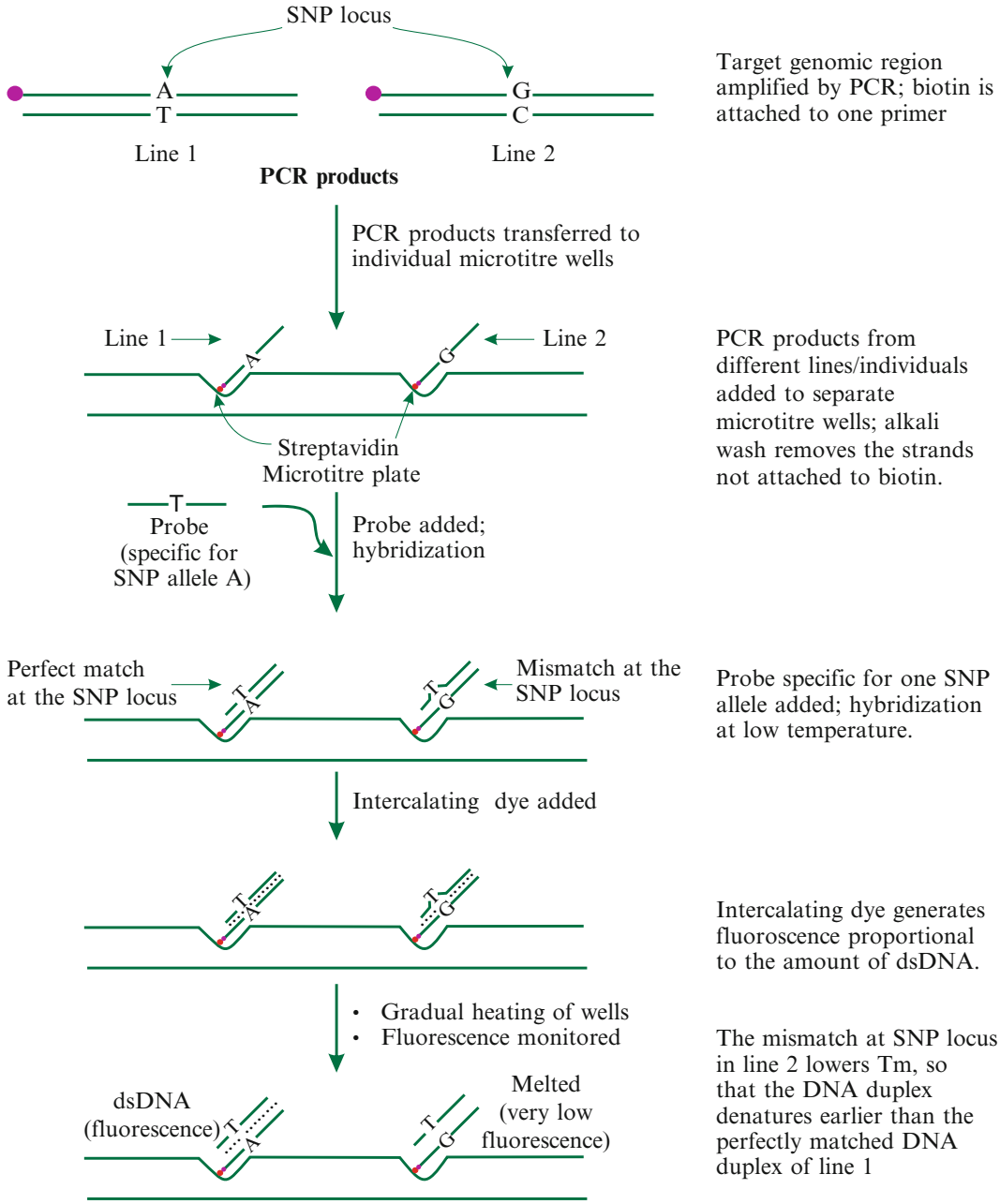


Fig. 4.17 A simplified schematic representation of dynamic allele-specific hybridization (*DASH*). Biotin specifically binds streptavidin; as a result, the biotinylated strand is retained in the microtiter wells. Each well has several copies of the concerned strand of the PCR products (only one strand is shown in each well here) (Based on de Vienne et al. 2003)

that in the reference PCR product, renaturation will produce the two homoduplexes (corresponding to the two PCR products) as well as two heteroduplexes formed by pairing between the strands of the two PCR products (Fig. 4.18b). As a result, there will be two peaks of elution in this case, one for the two homoduplexes and the other for the two heteroduplexes. The procedure requires precise control of temperature and gradient conditions.

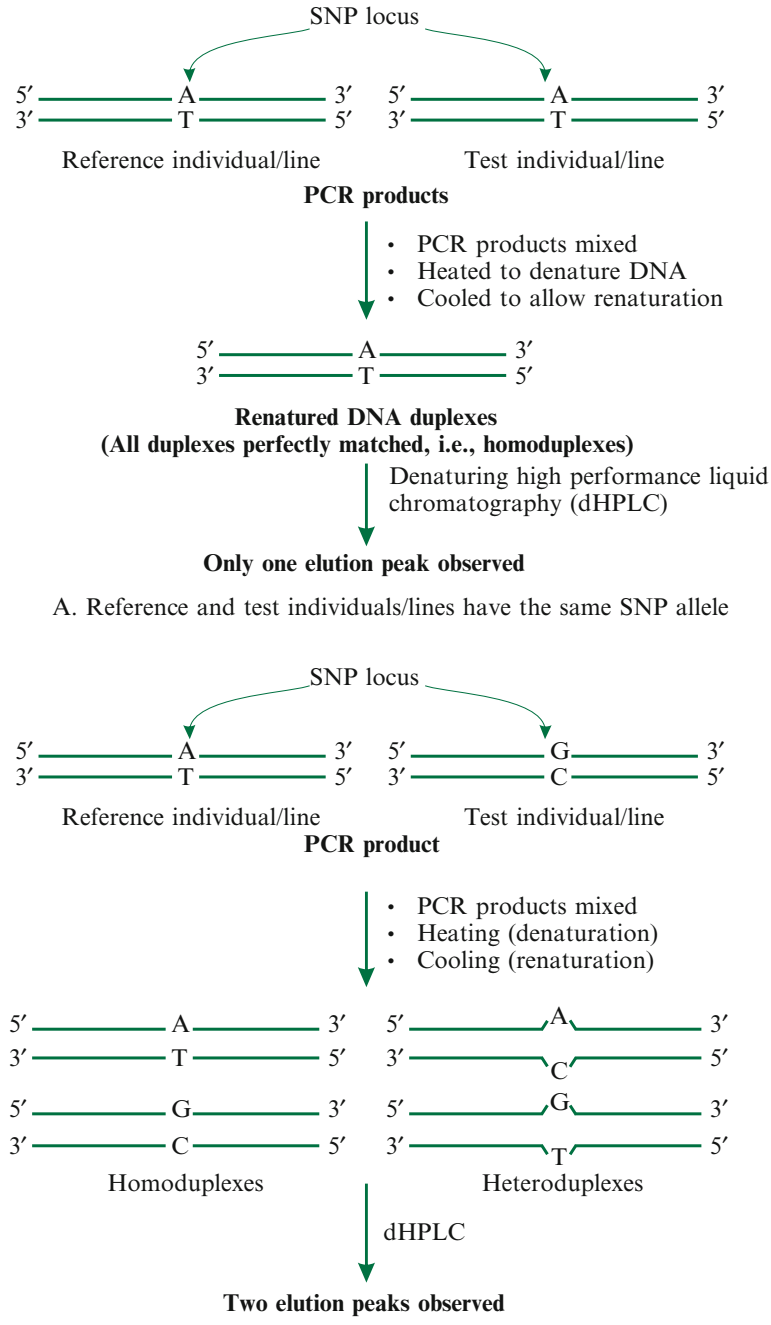


Fig. 4.18 A simplified representation of denaturing high-performance liquid chromatography (*dHPLC*). When the test and reference individuals/lines have different alleles, two elution peaks are observed: one peak corresponds to the two homoduplexes, while the other peak is due to the two heteroduplexes (Based on de Vienne et al. 2003)

Transgenomics, Inc. (San Jose, USA), has developed the fully automated dHPLC WAVE™ system for the analysis of PCR products. The dHPLC WAVE™ system has been used to develop the Masscode™ system by QIAGEN Genomics for high-throughput SNP genotyping as well as SNP discovery (http://www.qiagengenomics.com).

4.6.11 InDels as Molecular Markers

InDels are generally scored as SNPs, but Salathia et al. (2007) developed an InDel array for accurate InDel genotyping. They constructed the array using 70-nt-long oligos representing 240 unique InDel polymorphisms between two *A. thaliana* accessions. InDels of >25 bp were selected to maximize differential hybridization. For each InDel locus, 40 bp of sequence on both sides from the center of the insertion was used to derive the best 70-bp-long oligo; the GC content of the oligo was kept close to 50 %. The test DNA was sonicated, and the genomic fragments were directly labeled with Cy3 and Cy5 fluorophores. Competitive hybridization with the InDel array oligos was performed using 6 μg of the labeled genomic DNA fragments of each of the two accessions. The slides were washed to remove the free probes and the probes involved in nonspecific hybridization. The fluorescence signals were recorded with a sensitive detector, and the data were processed using appropriate bioinformatics tools to deduce InDel genotypes. The InDels were readily recognized with great precision so that there was no need for array replicates and complex statistical analyses. The InDel markers were distributed over the *A. thaliana* genome at an average distance of ~500 kb. Multi-well chips would allow groups of 16 lines to be genotyped in a single experiment. Shotgun sequencing or even partial genomic sequences should permit the application of this approach to non-model organisms for which reference genomes are not available. InDel polymorphisms have also been used for accurate mapping of recessive mutations in *A. thaliana*, identifying alternative expression isoforms of genes in *indica* and *japonica* rice and QTL mapping in salmon. Bulk of the InDels are of 1 bp, and those of 2–4 bp are the second most frequent category, while the frequency of 5 bp or longer InDels is ~10 % or less.

4.7 Epigenetic Markers

Epigenetics is the study of a change in gene function without any change in the gene base sequence. *Epigenetic changes* involve DNA methylation, RNA interference, and histone modification (acetylation, methylation, phosphorylation, and ubiquitination); these changes are also known as *epigenetic marks* (Edwards 2013). A genome-wide study of the epigenetic marks is referred to as *epigenomics*. The sites of cytosine methylation in the genome can be determined by bisulfite sequencing. In this strategy, the template DNA is treated with sodium bisulfite prior to sequencing. This treatment causes deamination of cytosine, thereby converting it to uracil. But when cytosine is methylated at 5 C, it is protected from deamination by the bisulfite treatment. Therefore, bisulfite sequencing will read normal cytosine as thymine, while methylated cytosine will be read as cytosine. The third-generation sequencing technologies are able to directly detect methylation sites. An analysis of the DNA methylation patterns in specific regions of the genome and in the genome as a whole would help understand their role in normal development and in disease. Epigenomic analyses will also elucidate the role of epigenetic changes in environmental adaptation, heritable genetic variation generated by epigenetic changes (*epimutation* and somaclonal variation), and agronomic performance of elite lines developed by breeding programs. *Somaclonal variation* is the heritable variation generated in cells and tissues grown in vitro, in the plants regenerated from them, and in the progeny of these plants.

4.8 Use of Genomics, Transcriptomics, Proteomics, and Metabolomics in Marker Development

The term *genome* denotes the complete set of nuclear and cytoplasmic genes present in an organism. *Genomics* is the field of study concerned with analysis of whole genomes in terms of their organization, including sequence, and function, including metabolic pathways and their interactions. Genomics is generally divided into the following two domains: (1) structural and (2) functional genomics. *Structural genomics* deals with determination of the complete

genome sequence and the complete set of proteins produced by an organism. *Functional genomics*, on the other hand, is the study of the gene expression patterns and the functioning of metabolic pathways. *Transcriptome* is the full complement of RNA molecules, including their quantities, produced by a cell during a specific developmental stage and exposed to a given environment. Thus *transcriptomics* aims to catalogue all the species of RNA transcripts expressed in a tissue/organ; their expression levels, splicing patterns, etc.; and the effects of developmental stages and environmental conditions on their expression.

The term *proteome* refers to the complete set of proteins produced in a cell during a specific developmental stage and under the given environmental conditions. *Proteomics*, thus, is the study of proteome using a diverse array of techniques starting with simple genetic analysis to mass spectrometry. Proteomics is usually classified into structural, functional, and expression proteomics. The discipline of s*tructural proteomics* deals with mapping of the 3-D structure of proteins and analyzing the nature of protein complexes present in a specific cell/organelle. The use of proteomics techniques for analyzing the characteristics of protein networks operating in a living cell constitutes *functional proteomics*. *Expression proteomics*, on the other hand, refers to a comparative quantitative analysis of the expression patterns of proteins between samples differing by some variable. *Metabolome* comprises all the metabolites representing the end products of cellular processes present in a cell, tissue, organ, or organism. Therefore, *metabolomics* is the systematic study of the characteristic small-molecule metabolite profiles generated by the various cellular metabolic processes.

Thus *genomic resources* of a species comprise the sum total of information about the structural and functional aspects of its genome. These resources include detailed high-density genetic maps, contig-based physical maps (including draft/completed genome sequences and their annotations), deep-coverage large-insert libraries, ESTs, gene expression levels and patterns (transcriptome), proteome, and the metabolome. The genome-sequencing projects dramatically accelerated the pace of developments in various areas of genomics, and vast amounts of data have been/are being generated for an increasingly large number of plant species. Some plant species like *A. thaliana* and rice have been investigated far more extensively and intensively than others so that the information accumulated about them is much more complete than that for other plant species. *A. thaliana*, a member of the Cruciferae family, is considered as a model dicot plant for molecular biology studies since it has a small genome size (125 Mb), low content of repetitive sequences, and short generation time and generates a large number of progeny per plant. Similarly, rice has emerged as the model monocot crop species due to its relatively small genome size and conservative genome organization. The genomes of these species have been sequenced and extensively annotated, and functions of a large number of their genes have been experimentally determined. Therefore, the genomes of these and other extensively studied species serve as reference for a variety of investigations, including identification of genes/gene families with specific functions, determination of conserved orthologous set of genes, etc.

Comparative genetic mapping of molecular markers revealed that the gene order is largely conserved (collinear or syntenic) among related plant species, e.g., among the species of grass family, and to some extent even across angiosperms. But comparisons among genome sequences revealed a much lower extent of collinearity of genes, since small-scale sequence rearrangements and InDels disturb the collinearity even between such species that are closely related. For example, comparisons of sequence-based maps reveal extensive breakdown of collinearity between wheat and rice, maize and rice, and sorghum and rice genome sequences. Knowledge of the extent of synteny and the locations of syntenic genomic regions and the patterns of chromosomal rearrangements would enable the transfer of genomic information from one species to the other. This would also

facilitate marker development for the whole genome as well as for specific genomic regions of a species based on the genomic information from a related species. Genome sequences can be analyzed with the help of suitable computer programs to identify molecular markers. For example, SSR markers can be developed by mining the end sequences of bacterial artificial chromosome (BAC) clones and screening of EST databases. All SNP markers are discovered by comparing genome and/or EST sequences of two or more lines/individuals, and SNP genotyping assays are designed on the basis of sequences flanking the SNP loci (Sects. 4.5 and 4.6). For example, comparison of genome sequences of the *indica* and *japonica* subspecies of rice has revealed several SNPs, including InDels. Similarly, the conserved orthologous sequences (COSs) are identified by comparing EST databases of a group of related species against a reference genome like that of *A. thaliana* (Sect. 3.20). Single feature polymorphisms (SFPs) are discovered by using either microarrays developed for gene expression analysis or designing microarrays based on sequences of all the annotated genes, unigenes, and ESTs of the species (Sect. 2.8). By screening the consensus EST sequences or the unigene sequences from many plant species, it is feasible to predict molecular markers like SSRs, SNPs, and COSs that could be developed as functional markers. However, all the predicted functional markers need to be confirmed and validated by appropriate genetic analyses and, ultimately, genetic transformation.

Transcriptome analysis generates a large collection of ESTs, and EST databases exist for most of the important species of plants. But the EST data have several limitations, including unidentified contaminants, chimeric sequences, paralogous and/or homoeologous sequences, and ESTs representing putatively nonfunctional transcripts. Moreover, EST databases lack the non-transcribed *cis*-acting elements and genes expressed at very low levels. However, the EST databases do serve as a rich and invaluable sequence resource for the transcribed regions of the genomes that have been exploited for a variety of purposes. Analysis of transcriptome data pertaining to segregating populations has enabled the identification of *expression QTLs* (*eQTLs*), i.e., QTLs concerned with regulation of expression levels of the genes analyzed in the study. In case a high-quality complete genome sequence is available for a plant species, annotation of the genomic regions harboring eQTLs will facilitate the identification of genes and *cis*-acting sequences involved in the regulation of gene expression relevant for various phenotypes. Efforts are being made to use metabolite levels as markers for the prediction of performance and to assess their usefulness as selection criteria.

4.9 Polymorphic Information Content of Marker Loci

The chief function of molecular markers is a clear-cut and reproducible classification of individuals/lines on the basis of DNA sequence variation. As a result, molecular markers also serve the purpose of reliable identification, based on close linkage, of the genes present in different individuals/lines. A codominant marker would also reveal the allelic states of these genes in the individuals/lines irrespective of whether they are heterozygous or homozygous for these genes. In contrast, a dominant marker will correctly identify the homozygotes but will fail to differentiate the heterozygotes from the dominant homozygotes. For this reason, codominant markers are considered to be more informative than dominant markers. Further, the usefulness of any marker locus for discrimination among different individuals/lines depends on the degree of polymorphism exhibited by the locus in the given population.

Polymorphic information content (*PIC*) of a marker locus is a measure of the degree of its polymorphism and is indicative of its usefulness in linkage and other studies. The PIC has been defined in various ways mainly depending on the biological material in which the marker locus is present and the particular use to which the marker is to be put. A simple and generalized definition of *PIC* is as follows: it is the

probability of a marker locus being polymorphic between two random individuals/lines selected from a given population. It can be readily shown that in a population homozygous for a biallelic marker locus, the PIC for the locus will equal $2pq$, where p and q are the frequencies of the marker alleles a_1 and a_2, respectively. In a homozygous population like a recombinant inbred line (RIL), there will be only two genotypes for the marker, viz., a_1a_1 and a_2a_2, and the frequencies of these genotypes will be p and q, respectively. Therefore, the probability that any two individuals randomly chosen from this population will differ at the marker locus will equal the product of the frequencies of the two genotypes multiplied by two, i.e., $2pq$. It has been shown that the same will be the situation, i.e., PIC $= 2pq$, in the case of a random mating population and in an F_2 population provided the marker locus is in Hardy–Weinberg equilibrium and the marker is codominant.

Since $2pq$ is also the frequency of heterozygotes in a random mating population, PIC is often referred to as *expected heterozygosity* (H_e) for the marker loci. The term expected heterozygosity is in use for the following reason as well. In human linkage studies, analysis of progeny from a parent heterozygous for the marker locus and affected by a dominant disease may allow one to infer the marker allele most likely linked with the disease allele. The value of PIC for a biallelic marker ranges between 0 (only one marker allele present in the population, i.e., $p = 1$ and $q = 0$ or vice versa) and 0.5 ($p = q = 0.5 = 2pq$). But as the value of p (and, consequently, that of q) deviates from 0.5, the PIC value decreases. For example, when values of p and q are 0.4 and 0.6, respectively, the PIC value declines to 0.48 ($=2 \times 0.4 \times 0.6$), while it drops down to merely 0.18 when the values of p and q are 0.1 and 0.9, respectively.

It can be readily shown that in the case of a multiallelic marker locus, the value of PIC would equal $1 - \sum {p_i}^2$, where p_i is the frequency of i^{th} allele at the marker locus. This is because the value of $\sum {p_i}^2$ would equal the sum of the frequencies of homozygotes for all the alleles at the marker locus present in the population, and that of $1 - \sum {p_i}^2$ will be the same as $\sum 2p_iq_i$. The value of PIC for a multiallelic locus ranges between zero (only one allele present in the population) and 1 (infinite number of alleles present in the population). For example, the PIC score for a marker locus with five alleles, each allele having the frequency of 0.2, will be 0.8 [$=1 - (5 \times 0.04)$]. *Thus, the PIC estimate is the property of a specific marker locus in a given population and depends on the number and frequencies of the marker alleles in the population. Therefore, PIC estimates will differ for different loci of a single marker system and for different populations for the same marker locus.*

In any study, several marker loci of a marker system are analyzed. The information from all the loci scored for the marker system may be pooled to estimate the average PIC score for the marker system. It can be shown that the average PIC score (H_{av} = average heterozygosity) for all the polymorphic markers scored for a marker system will equal $\sum H_{ei}/n_p$, where H_{ei} is the expected heterozygosity or the PIC score of the ith marker locus and n_p is the number of polymorphic loci present in the population. However, some of the marker loci may not be polymorphic in the population, but they should be taken into account while estimating the average PIC for the marker system. This can be done by multiplying the average PIC score with $ß$, i.e., the ratio of polymorphic marker loci to the total number of loci scored. Thus the PIC for the marker system would equal $ß\,H_{av}$. *However, the H_{av} estimate for a marker system is applicable to a particular population, from which it is estimated, and it may be only of limited value in other populations.*

A single assay for some marker systems permits the scoring of a single locus, while each assay for some other marker systems evaluates several marker loci. The average number of markers scored per assay of a marker system is described as its *multiplex ratio*. This ratio is different from the extent of multiplexing possible for a marker system in that it indicates the number of different markers analyzed by a single assay without application of any multiplexing strategy (Sects. 3.3.3 and 3.12.2). The multiplex ratio will be one or close to one for markers like

SSRs, SCARS, CAPSs, etc., but will be much larger for marker systems like AFLPs, RAPDs, ISSRs, etc. *Marker index* for a marker system is estimated as the product of multiplex ratio and the average PIC score for the marker system in the given population. Marker index, thus, reflects the degree of polymorphism that would be detected by each assay of the given marker system in the population. Similar to the PIC score, these indicators of the marker usefulness also will be applicable to the concerned population and would merely serve as rough indicators for other populations.

A comparison among different marker systems has been done in several crops, including soybean, barley, and wheat. In a study with soybean, SSR markers were found to have the highest expected heterozygosity, while AFLP markers had the highest multiplex ratio and the highest marker index. In comparison, RAPD markers were intermediate in terms of both expected heterozygosity and multiplex ratio, whereas RFLP markers were moderate with respect to expected heterozygosity (Powell et al. 1996). Studies with other crops have also revealed a similar picture.

4.10 Marker System Selection

RFLPs were the first DNA markers to be developed, and they were extensively used in various biological investigations, including plant breeding. But with the development of more user-friendly PCR-based markers during the 90s, the interest in RFLPs declined, and soon SSRs became the most widely used molecular markers. The dominance of SSRs began to be challenged by SNPs about a decade ago, and since then the latter have rapidly emerged as the marker of choice in view of their abundance and almost uniform distribution throughout the genome. However, the search for new marker systems continues, and so far nearly two-dozen different marker systems have been developed. The salient features of some of the common marker systems are compared in Tables 4.3, 4.4, and 4.5. It would be seen that each marker system has some desirable features that favor its plant breeding application, but some of its other features limit its usefulness. For example, RAPD technique is relatively simple and straightforward and requires much less time than RFLPs and AFLPs, but this marker system has moderate to poor reproducibility. SSR markers are highly polymorphic, PCR based, easily detectable, and codominant, but their development requires considerable time and effort. Similarly, AFLPs are highly reproducible and can be applied to any species since there is no specific marker development step, but they are dominant and anonymous, and their detection requires much more skill and instrumentation than that of RAPDs and SSRs.

The selection of a DNA marker system for a plant breeding application depends on several factors, including the objectives of the project, the financial resources available to the project, availability of the desired marker system for the concerned species, and the reproducibility of the marker system. The objective of the project would determine the scale of operations in terms of the numbers of markers and the samples to be scored during a cropping season (Table 2.4). In view of the above, the research worker has to critically evaluate each marker system for its potential utility to his/her project and select the most suitable marker system. In general, the choice will be influenced by the following features of the marker systems: degree of polymorphism, dominance/codominance of marker alleles, simplicity and speed of detection procedures, amenability for multiplexing and automation, need for prior sequence information and the amount of work required for marker development, and above all the reproducibility of the marker system. For genetic mapping, the genotyping procedure should be simple and cost-effective, and the information content of the marker should be moderate to high. In addition, the marker should be abundant and distributed across the whole genome. Cost of genotyping would depend on the amount of DNA needed for analysis, need for cloning and sequencing, the amount of potentially useful genetic information acquired per data point, the type of genetic information needed, dominance relationship of

Table 4.3 A comparison among different marker systems

Marker system	Abundance	Reproducibility	Degree of polymorphism	Locus specificity	Technical requirement	DNA quantity	Automation	Genotyping cost	Major application
RFLP	High	High	Medium	Yes	High	High	Low	High	PM[a]
RAPD	High	Low	Medium	No	Low	Low	Medium	Low	LM
SSR	Medium	Medium	Medium	No	Medium	Low	Medium/high	Low	DA
SSCP	Low	Medium	Low	Yes	Medium	Low	Low	Medium	LM
CAPS	Low	High	Low	Yes	High	Low	Low	Medium	LM
SCAR	Low	High	Medium	Yes	Medium	Low	Medium	Low	LM, PM
AFLP	High	High	Medium	No	Medium	Medium	Medium/high	Medium	LM
IRAP/REMAP	High	High	Medium	Yes	High	Low	Medium/high	Low	DA
RAMPO	Medium	Medium	Medium	Yes	High	Low	–	–	DA
SRAP/EST	Medium	High	Medium/high	–	Medium	Medium	Medium/high	Low	LM
ISSR	High	High	Medium	No	Low	Low	Medium	Low	DA
SNP	Very high	High	Medium	Yes	High	Medium	High	Low	LM

Based mainly on Meksem and Kahl (2005) and Agarwal et al. (2008)
[a]*PM* physical mapping, *LM* linkage mapping, *DA* genetic diversity analysis

Table 4.4 A rough classification of the different marker systems on the basis of their various features

	Level		
Feature	Low	Moderate	High
Detection			
Equipment cost	RAPD, SCAR, SSR, ISSR, COSs	AFLP, RFLP	SNP, SFP, DArT
Technical expertise	RAPD, SCAR, SSR, ISSR, COSs	AFLP	SNP, SFP, DArT, RFLP
Throughput	RFLP	RAPD, SCAR, SSR, ISSR, COSs, AFLP	SNP, SFP, DArT
Automation, including data acquisition and processing	RFLP, RFLP, RAPD, ISSR,	SSR, AFLP, COSs, SCAR	SNP, SFP, DArT
Assay time		RAPD, SCAR, SSR, ISSR, COSs, AFLP, SNP, SFP, DArT	RFLP
Cost per data point	RAPD, SCAR, SSR, ISSR, COSs, DArT, SNP, SFP	CAPS, AFLP	RFLP
Marker development			
Time and effort	RAPD, ISSR, SRAP	SCAR, AFLP	SNP, SFP, DArT, RFLP, SSR, COSs
Need for sequence information	RAPD, ISSR, RFLP, DArT, SRAP (not required)	SCAR	SNP, SFP, SSR, COSs
Use of bioinformatics tools	RAPD, ISSR, RFLP, DArT, SRAP (not required)	SCAR	SNP, SFP, SSR, COSs
Other features			
Reproducibility/reliability	RAPD		SNP, SFP, DArT, RFLP, SSR, COSs, AFLP
Scale of operation	RFLP	RAPD, SCAR, SSR, ISSR, COSs, AFLP	SNP, SFP, DArT
Plant material required	RAPD, SSR, ISSR, SCAR, COSs	RFLP, SNP, SFP, DArT	

Table 4.5 A summary of differences among different array-based techniques for detecting DNA polymorphisms. All the markers are scored as presence/absence and are regarded as cost-effective

	Marker system			
Parameter	SNP	SFP	DArT	RAD tag
Sequence information	Required	Required	Not required	Not required
Markers represent	Random genomic regions	Genic regions	Random genomic regions	Random genomic regions
PCR amplification	Required in some assays like MIP and GoldenGate	Not required	Required	Required
Number of markers scored per assay	High	High	Moderate	Moderate
Type of array used	Tag array on beads/glass, oligonucleotide array/ GeneChip	High-density oligonucleotide array/ GeneChip	Glass-spotted DNA microarray	Tiling microarray, oligonucleotide array/ GeneChip
Resolution	High	High	Moderate	Moderate

Based on Gupta et al. (2008)

marker alleles, amenability to automation, and the proprietary status of the technique for marker detection.

A discussion on the selection of a suitable marker system can be only in general terms, and it may not be possible to provide specific recommendations. We may begin our discussion with reference to large-scale breeding projects with adequate financial resources. In such cases, one would need a marker system capable of high to very high throughput and automated data acquisition and analysis. Four marker systems, namely, SNPs, DArT, SFPs, and RAD markers, satisfy these criteria. All these marker systems require considerable laboratory infrastructure and sophistication and moderate to large amount of marker development effort. However, SFP and SNP markers are sequence based and either good quality genome sequences should be available or de novo sequencing would be necessary for their development. In contrast, DArT and RAD markers are anonymous and their development does not require sequence information; as a result, they can be developed for any crop species irrespective of the availability of genomic resources. Therefore, the choice among them will depend mainly on the considerations of marker density requirement, cost per data point, and the availability of the marker systems for the concerned species. At present, SNPs are the preferred markers and almost all large-scale breeding programs are routinely using them. DArT markers are steadily gaining in popularity for fingerprinting, diversity studies, selection of parents, and linkage mapping, while SFPs and RAD markers have also been used.

In the case of breeding programs of small to moderate size, most of the DNA markers are available for application. However, RAPDs have limited reliability, and RFLPs are not user-friendly. Therefore, even when RFLP markers are available for achieving the desired goals, other marker systems would be preferable. When the financial resources are adequate and the desired markers are available, the choice will have to be between SSRs and SNPs. In most situations, these markers can be assayed in the laboratory, and where required SNP genotyping services are commercially available. Further, a moderate-sized breeding program can be hardly expected to de novo develop SSR and SNP markers. In case SSR and SNP markers are not available for the desired goal and genomic and/or financial resources do not support their de novo development, one has to select a marker system like AFLP, DArT, SRAP, or SCoT that does not require prior sequence information for marker development. DArT is a proprietary technology, and its development as well as detection would require substantial expenditure on equipment or the activity will have to be outsourced. AFLPs do require some expenditure on equipment, but this will be much less than that for DArT. The SRAP, SCoT and other similar markers are in experimental stages, but appear to be quite promising.

The objectives of the program also influence issues like marker density and the genomic regions to be targeted for marker genotyping. For example, a much higher marker density would be needed for association studies and genomic or genome-wide selection than those for linkage mapping and MAS. Further, even in the case of association studies, a much higher marker density would be needed in a cross-pollinated species than in a self-pollinated species. Therefore, SNPs become the preferred marker system for programs like association studies and genomic selection. It has been argued that a much higher density of SNP markers would compensate for their lower PIC as compared to that of SSRs. Similarly, when a specific region of the genome is to be targeted and/or fine mapping is to be done, an abundant marker system like SNP is preferable to the others.

Questions

1. Explain the features that make NGS technologies faster and cheaper than the first-generation technologies.
2. Briefly describe the procedure of one of the NGS technologies, and discuss the applications of the NGS technologies.

3. How do third-generation sequencing technologies differ from the NGS technologies, and what advantages do they offer in comparison to the latter?
4. Explain the meaning of PIC and discuss its significance for a marker system.
5. What are the various issues relevant to the selection of a suitable marker system for marker-assisted selection?
6. Discuss the usefulness of genomic resources in the development of molecular markers, especially single nucleotide polymorphism.
7. Briefly explain the use of primer extension for determining the SNP alleles at a given locus.
8. Discuss the use of microarrays for SNP genotyping