Chapter 2

New Technologies for Ultra-High Throughput Genotyping in Plants

Nikki Appleby, David Edwards, and Jacqueline Batley

Summary

Molecular genetic markers represent one of the most powerful tools for the analysis of plant genomes and the association of heritable traits with underlying genetic variation. Molecular marker technology has developed rapidly over the last decade, with the development of high-throughput genotyping methods. Two forms of sequence-based marker, simple sequence repeats (SSRs), also known as microsatellites and single nucleotide polymorphisms (SNPs) now predominate applications in modern plant genetic analysis, along the anonymous marker systems such as amplified fragment length polymorphisms (AFLPs) and diversity array technology (DArT). The reducing cost of DNA sequencing and increasing availability of large sequence data sets permits the mining of this data for large numbers of SSRs and SNPs. These may then be used in applications such as genetic linkage analysis and trait mapping, diversity analysis, association studies and marker-assisted selection. Here, we describe automated methods for the discovery of molecular markers and new technologies for high-throughput, low-cost molecular marker genotyping. Genotyping examples include multiplexing of SSRs using Multiplex-Ready™ marker technology (MRT); DArT genotyping; SNP genotyping using the Invader® assay, the single base extension (SBE), oligonucleotide ligation assay (OLA) SNPlex™ system, and Illumina GoldenGate™ and Infinium™ methods.

Key words: Diversity array technology, DArT, GoldenGate™, Infinium™, Invader®, Multiplex-Ready™ marker technology, MRT, Oligonucleotide ligation assay, OLA, Simple sequence repeat, SSR, Single Base Extension, SBE, Single Nucleotide Polymorphism, SNP, SNPlex™.

1. Introduction

The application of molecular markers to advance plant breeding is now well established *(1)*. Modern agricultural breeding is dependent on molecular markers for the rapid and precise analysis of germplasm, trait mapping and marker-assisted selection (MAS). Molecular markers can be used to select parental

Daryl J. Somers et al. (eds.), *Methods in Molecular Biology, Plant Genomics, vol. 513*

DOI: 10.1007/978-1-59745-427-8_2

genotypes in breeding programs, eliminate linkage drag in backcrossing and select for traits that are difficult to measure using phenotypic assays. Molecular markers have many other uses in genetics, such as the detection of alleles associated with genetic diseases, paternity assessment, forensics and inferences of population history *(2, 3)*. Furthermore, molecular markers are invaluable as a tool for genome mapping in all systems, offering the potential for generating very high-density genetic maps that can be used to develop haplotypes for genes or regions of interest *(4)*. Insight into the organisation of the plant genome can be obtained by calculating a genetic linkage map using molecular markers. Genetic mapping places molecular genetic markers on linkage groups based on their co-segregation in a population. Markers that are transferable between species also enable studies of synteny and genome rearrangement across taxa. Molecular markers are complementary tools to traditional selection. They can increase our understanding of phenotypic characteristics and their genetic association, which may modify the breeding strategy. DNA-based markers have many advantages over phenotypic markers in that they are highly heritable, relatively easy to assay and are not affected by the environment.

The bulk of variation at the nucleotide level is often not visible at the phenotypic level. This variation can be exploited in molecular genetic marker systems. Two sequence-based marker systems, single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs) (*see* **Note 1**) are the principal markers utilised in plant genetic analysis. These are supplemented by anonymous systems such as amplified fragment length polymorphisms (AFLPs) and diversity array technologies (DArT).

1.1. What are SNPs?

DNA sequence differences are the basic requirement for the study of molecular genetics. SNPs are the ultimate form of molecular genetic marker, as a nucleotide base is the smallest unit of inheritance, and a SNP represents a single nucleotide difference between two individuals at a defined location. There are three different forms of SNPs: transitions (C/T or G/A), transversions (C/G, A/T, C/A, or T/G) or small insertions/deletions (indels) *(5)*. SNPs are direct markers as the sequence information provides the exact nature of the allelic variants. Furthermore, this sequence variation can have a major impact on how the organism develops and responds to the environment. SNPs represent the most frequent type of genetic polymorphism and may therefore provide a high density of markers near a locus of interest *(6)*.

SNPs can differentiate between related sequences, both within an individual and between individuals within a population. The frequency and nature of SNPs in plants is beginning to receive considerable attention. Studies of sequence diversity have recently been performed for a range of plant species and these

have indicated that SNPs appear to be abundant in plant systems, with one SNP every 100–300 bp *(7)*. SNPs at any particular site could in principle involve four different nucleotide variants, but in practice they are generally biallelic. This disadvantage, when compared with multiallelic markers such as SSRs, is compensated by the relative abundance of SNPs. SNPs are also evolutionarily stable, not changing significantly from generation to generation. The low mutation rate of SNPs makes them excellent markers for studying complex genetic traits and as a tool for understanding genome evolution *(8)*.

The high density of SNPs makes them valuable for genome mapping, and in particular, they allow the generation of ultra-high-density genetic maps and haplotyping systems for genes or regions of interest and map-based positional cloning. SNPs are used routinely in crop breeding programs *(1)*, for genetic diversity analysis, cultivar identification, phylogenetic analysis, characterisation of genetic resources and association with agronomic traits *(4)*. The applications of SNPs in crop genetics have been extensively reviewed by Rafalski *(4)* and Gupta et al. *(1)*. These reviews highlight that for several years SNPs will coexist with other marker systems. However, with the development of new technologies to increase throughput and reduce the cost of SNP assays, along with further plant genome sequencing, the use of SNPs will become more widespread.

1.2. Simple Sequence Repeats

SSRs are one of the most powerful genetic markers in biology. They have been found in all prokaryotic and eukaryotic genomes analysed to date and are widely and ubiquitously distributed throughout eukaryotic genomes *(9, 10)*. SSRs are short stretches of DNA sequence occurring as tandem repeats of mono-, di-, tri-, tetra-, penta- and hexanucleotides. They are highly polymorphic and informative markers. The high level of polymorphism is due to mutation affecting the number of repeat units. The value of SSRs is due to their genetic co-dominance, abundance, dispersal throughout the genome, multiallelic variation and high reproducibility. These properties provide a number of advantages over other molecular markers, namely, that multiple SSR alleles may be detected at a single locus using a simple polymerase chain reaction (PCR)-based screen, very small quantities of DNA are required for screening, and analysis is amenable to automated allele detection and sizing *(11)*. The hypervariability of SSRs among related organisms makes them excellent markers for a wide range of applications, including genetic mapping, molecular tagging of genes, genotype identification, analysis of genetic diversity, phenotype mapping and MAS *(12, 13)*. SSRs demonstrate a high degree of transferability between species, as PCR primers designed to an SSR within one species frequently amplify a corresponding locus in related species, enabling comparative genetic and genomic analysis.

Studies of the potential biological function and evolutionary relevance of SSRs is leading to a greater understanding of genomes and genomics *(14)*. SSRs were initially considered to be evolutionally neutral *(15)*; however, recent evidence suggests an important role in genome evolution *(16)*. Early suggestions that the majority of DNA was 'junk' or had no biological function are being challenged by the discovery of new functions for these sequences and various functional roles have now been attributed to SSRs. For example, SSRs are believed to be involved in gene expression, regulation and function *(17, 18)* and there are numerous lines of evidence suggesting that SSRs in non-coding regions may also be of functional significance *(19)*. In addition, SSRs provide hot spots of recombination, a variety of SSRs have been found to bind nuclear proteins and there is direct evidence that SSRs can function as transcriptional activating elements *(20)*.

1.3. Diversity Array Technology

DArT is a generic and cost-effective genotyping method based on hybridising DNA to microarrays *(21)*. It was invented to overcome some of the limitations of other molecular marker technologies, and in particular, it does not require prior sequence information. Other advantages of DArT include high multiplexing level for high-throughput analysis and provision of data at low cost. The main technology applications of DArT include genome profiling, genetic map construction and Quantitative Trait Loci (QTL) identification, genetic diversity analysis and cultivar identification *(22–26)*. DArT works in a similar way to AFLP in reducing the complexity of a DNA sample to obtain a representation of the genome. The preferred method of complexity reduction relies on a combination of restriction enzyme digestion and adapter ligation, followed by PCR amplification *(27)* with subsequent hybridisation-based detection.

1.4. Why Novel Marker Technologies are Required?

During the past two decades, several molecular marker technologies have been developed and applied for plant genome analysis, predominantly assessing the differences between individual plants within a species. These marker technologies have been applied to plant breeding to allow breeders to use the genetic composition or genotype of plants as a criterion for selection in the breeding process. However, because of the relatively high cost associated with the development of this technology, these methods have only been applied to a limited number of crop species, predominantly in developed countries. Even in these situations, the application of molecular markers has tended to focus on a small number of high value traits or genomic regions. The recent application of association mapping via linkage disequilibrium (LD) in plants demonstrates the requirement to be able to identify and screen large numbers of markers, rapidly and at low cost.

The development of technologies that increase marker throughput with reducing cost will broaden the uptake of MAS to include more diverse crops and a greater variety of traits.

2. New Marker Discovery Methods

Large quantities of sequence data are generated through cDNA or genome sequencing projects internationally and these provide a valuable resource for the mining of molecular markers. This will be further accelerated with the application of new sequencing technology from Roche (454), Illumina (Solexa) and Applied Biosystems (SOLiD) (*see* Imelfort et al. this volume).

2.1. In Silico SNP Discovery

The challenge of *in silico* SNP discovery is not the identification of polymorphic bases, but the differentiation of true SNP polymorphisms from the often more abundant sequence errors. High-throughput sequencing remains prone to inaccuracies as frequent as one error every one hundred base pairs. This incorrect base calling impedes the electronic filtering of sequence data to identify potentially biologically relevant polymorphisms. There are several different sources of error which need to be taken into account when differentiating between sequence errors and true polymorphisms. The primary source of sequence error comes from the automated reading of raw data, due to the fine balance between the desire to obtain the greatest sequence length and the confidence that bases are called correctly. Phred is the most widely adopted software used to call bases from Sanger chromatogram data *(28, 29)*. The primary benefit of this software is that it provides a statistical estimate of the accuracy of calling each base, and therefore provides a primary level of confidence that a sequence difference represents true genetic variation. There are several software packages that take advantage of this feature to estimate the confidence of sequence polymorphisms within alignments. Where sequence trace files are available, and nucleotide quality may be measured, software such as PolyBayes and Polyphred are the most efficient means to differentiate between true SNPs and sequence error (*see* **Note 2**). Unfortunately, complete sequence trace file archives are rarely available for data sets collated from a variety of sources. Furthermore, sequence quality scores do not identify errors in the sequence incorporated before the base calling process. The principal cause of these prior errors is the inherently high error rate of the reverse transcription process required for the generation of cDNA libraries for Expressed Sequence Tag (EST) sequencing. Similar errors are also inherent, though to a lesser extent, in any PCR amplification process that may be part

of a sequencing protocol. In cases where trace files are unavailable, the identification of sequence errors can be based on two further methods to determine SNP confidence; redundancy of the polymorphism in an alignment and co-segregation of SNPs to define a haplotype.

The frequency of occurrence of a polymorphism at a particular locus provides a measure of confidence in the SNP representing a true polymorphism, and is referred to as the SNP redundancy score. By examining SNPs that have a redundancy score equal to or greater than two (two or more of the aligned sequences represent the polymorphism), the vast majority of sequencing errors are removed. Although some true genetic variation is also ignored due to its presence only once within an alignment, the high degree of redundancy within the data permits the rapid identification of large numbers of SNPs without the requirement for sequence trace files. However, while redundancy-based methods for SNP discovery are highly efficient, the non-random nature of sequence error may lead to certain sequence errors being repeated between runs around locations of complex DNA structure. Therefore, errors at these loci would have a relatively high SNP redundancy score and appear as confident SNPs. In order to eliminate this source of error, an additional independent SNP confidence measure is required. This can be determined by the co-segregation of SNPs to define a haplotype. True SNPs that represent divergence between homologous genes co-segregate to define a conserved haplotype, whereas sequence errors do not co-segregate with a haplotype. Thus, a co-segregation score, based on whether a SNP position contributes to defining a haplotype is a further independent measure of SNP confidence. By using the SNP score and co-segregation score together, true SNPs may be identified with reasonable confidence.

Three tools currently apply the methods of redundancy and haplotype co-segregation: autoSNP *(30, 31)*, SNPServer *(32)* and autoSNPdb. SNPServer is based on autoSNP and provides a real time Internet-based SNP discovery tool, combining redundancy-based SNP discovery and haplotype co-segregation scoring. Sequences may be submitted for assembly with CAP3 (*33)* or submitted preassembled in ACE format. Alternatively, a single sequence may be submitted for Basic local Alignment Search Tool (BLAST) comparison with a sequence database *(34)*. Identified sequences are then processed for assembly with CAP3, and subsequent redundancy-based SNP discovery. SNPServer has an advantage in being the only real time Web-based tool that allows users to rapidly identify novel SNPs in sequences of interest. The recently developed autoSNPdb combines the SNP discovery pipeline of autoSNP with a relational database, hosting information on the polymorphisms, cultivars and gene annotations, to enable efficient mining and interrogation of the data. Users may search for SNPs within genes

with specific annotation or for SNPs between defined cultivars. AutoSNPdb can integrate both Sanger and pyrosequencing data enabling efficient SNP discovery from next generation sequencing technologies.

2.2. SSR Discovery

Previously, the discovery of SSR loci was limited to the construction of genomic DNA libraries enriched for SSR sequences, followed by DNA sequencing *(35)*. This process is both time-consuming and expensive due to the specific sequencing required. The availability of large quantities of sequence data now makes it more economical and efficient to use computational tools to identify SSR loci. Flanking DNA sequence may then be used to design suitable forward and reverse PCR primers to assay the SSR. Several computational tools are currently available for the identification of SSRs within sequence data as well as for the design of PCR amplification primers. These include SSRPrimer *(36)*, which integrates two such tools, enabling the simultaneous discovery of SSRs within single or bulk sequence data, and the design of specific PCR primers for the amplification of these loci. The Web-based version of SSRPrimer permits the remote use of this package with any sequence of interest. SSR Taxonomy Tree demonstrates the application of SSRPrimer to the complete GenBank database, with the results organised as a taxonomic hierarchy for browsing or searching for SSR amplification primers in any species of interest *(37)*.

Because of the redundancy in EST sequence data, with data sets often being derived from several distinct cultivars, it is now possible to predict the polymorphism of SSRs *in silico*. Using an extended version of autoSNPdb, polymorphic SSRs are distinguished from monomorphic SSRs by the representation of varying motif lengths within an alignment of sequence reads. The identification of SSRs that are predicted to be polymorphic between defined varieties greatly reduces the cost associated with the application of these markers.

3. New Genotyping Technologies

3.1. New Genotyping Technologies for SNPs

Many new marker technologies involve improving the genotyping of SNPs, reflecting the increasing popularity of these markers. SNPs can be identified within a gene of interest, or within close proximity to a candidate gene. Although the SNP may not be directly responsible for the observed phenotype, it can be used for the positional cloning of the gene responsible *(1)* and as a diagnostic marker. Furthermore, SNPs are useful to define haplotypes in regions of interest. The success of the human HapMap project *(38)*, where a very large

number of SNPs were assayed over a range of individuals from different groups, demonstrates the value that can be gained from SNP studies. Reducing costs could enable similar studies to be undertaken to gain a greater understanding of plants.

3.1.1. Invader® Assay

The Invader assay® is a relatively new technology designed specifically for genotyping SNPs *(39, 40)*. In this technology, an oligonucleotide Invader probe is designed to anneal immediately next to the variable site, in the opposite direction to a secondary, allele-specific probe. The secondary probe contains a 5′-flap that is non-complementary to the target DNA and so is unable to hybridise to the target sequence. The 3′-end of the bound Invader probe overlaps the primary probe by a single base at the site of the allelic variant or SNP. A three-dimensional complex is formed by hybridisation of the secondary allele-specific overlapping probe to the target DNA containing a SNP site. This complex is only produced if the secondary probe is complementary to the allele and the Invader probe is present. The annealing of the probe complementary to the SNP allele induces cleavage by a thermostable, structure-specific flap endonuclease (FEN). The cleaved 5′-flap fragment then triggers a secondary cleavage reaction between a quencher molecule, a fluorophore and the cleaved fragment, which results in a fluorescent emission. If the secondary probe is not complementary to the SNP allele and no invasive complex is created, the FEN does not perform cleavage and no fluorescence is observed (**Fig. 1**). There are several different approaches to detect the cleavage. Most commonly this method is detected on a fluorescence resonance energy transfer (FRET™) cassette; however, it can also be detected by fluorescence polarisation probes or by mass spectrometry.

The Invader® assay is a highly accurate method, has a low failure rate, and can detect very small (zeptomol) quantities of target DNA. However, it does require the PCR amplification of the target DNA and the design of a specific secondary probe for each of the SNP alleles. This increases the cost of the method, which makes it unsuitable for high-throughput genotyping. While the assay has traditionally been used to interrogate one SNP in one sample per reaction, novel chip- or bead-based approaches are being tested to make this efficient and accurate assay adaptable to multiplexing and high-throughput SNP genotyping. The Biplex Invader® assay *(41)* was recently developed, which allows the detection of both alleles in the same reaction tube. There are two signal fluorophores attached to two different FRET™ cassettes (FRET 1 and 2) that are spectrally distinct and specific to either allele of the biallelic system. The ratios of the two fluorescent signals allow a genotype to be assigned. This method is very attractive for researchers who want to genotype a small number of SNPs over large populations. The utility of

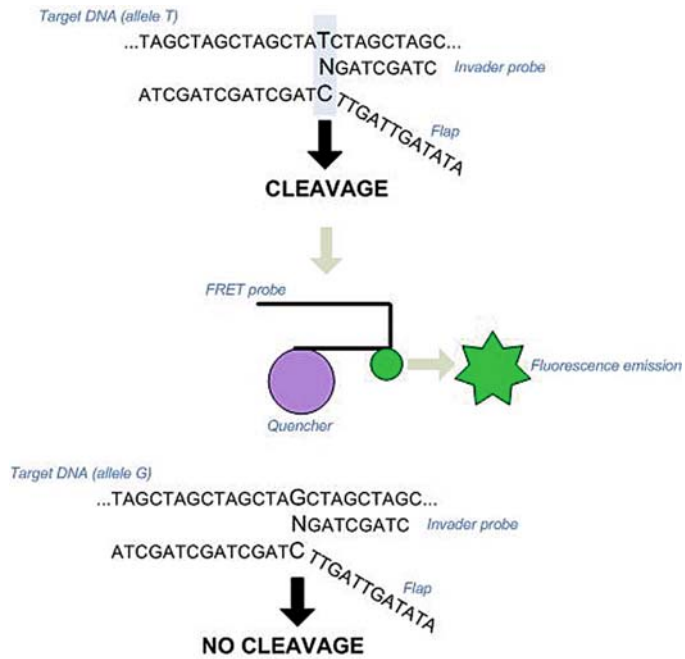

Fig. 1. Overview of the Invader® assay.

this new technology in plants has been demonstrated by Gupta et al. *(42)* for the accurate determination of gene copy number in a molecular breeding program involving both transgenic and non-transgenic plants.

3.1.2. Illumina GoldenGate™ and Infinium™ Assays

The Illumina GoldenGate™ technology is a novel array technology based on microbeads assembled into 96 sample arrays, with redundant bead types for increased confidence calls *(43)*. This technology is particularly suited for high-throughput genotyping *(44)*. The arrays have up to 50,000 beads, each around 3 microns in diameter. The beads are distributed among 1,520 bead types, with each bead type representing a different oligonucleotide probe sequence. This provides 30 copies of each bead type, with the result that a genotype call is based on the average of many replicates. This inherent redundancy increases robustness and genotyping accuracy. The assay performs allelic discrimination directly on genomic DNA, then generates a synthetic allele-specific PCR template before performing PCR on this artificial template. This is a reversal of conventional SNP genotyping assays which usually use PCR to amplify a SNP of interest and carry out allelic discrimination on the PCR product. The Illumina Bead Station GoldenGate™ assay is most suitable for researchers performing large-scale

association studies, such as whole-genome linkage mapping and large-scale fine mapping. It can be carried out in 384, 768 and 1536 sample formats using custom SNP panels. The GoldenGate™ assay was developed specifically for multiplexing to high levels while retaining the flexibility to choose any SNPs of interest to assay.

GoldenGate™ assay technology involves two allele-specific oligonucleotides (ASOs) and one locus-specific oligonucleotide (LSO) for each SNP (**Fig. 2**). The ASOs are designed to have a T_m of 60°C, within the range 57–62°C, and the LSO has a T_m of 57°C, within the range 54–60°C. Each ASO consists of a 3′ portion that hybridises to the DNA at the SNP locus, with the 3′ base complementary to one of the two SNP alleles, and a 5′ portion that incorporates a universal PCR primer sequence (P1 or P2, each associated with a different allele). The LSOs consist of three parts: at the 5′ end is a SNP locus-specific sequence; the middle contains an address sequence complementary to one of the capture sequences on the array; and there is a universal PCR priming site (P3′) at the 3′ end. The genomic DNA is attached to a solid support before the start of the assay, and the oligonucleotides targeted to specific SNPs of interest are then annealed to the DNA. The attachment step is performed to improve assay specificity by removing unbound and non-specifically hybridised oligonucleotides using stringency washes, while the correctly hybridised oligonucleotides remain on the solid phase. Following the annealing and washing steps, an allele-specific primer extension step is carried out, in which DNA polymerase extends the ASOs if their 3′ base is complementary to the SNP *(45)*. This is followed by ligation of the extended ASOs to their corresponding LSOs, which creates the PCR templates. This ligated product is amplified by PCR using universal primers that are complementary to a universal sequence in the 3′-end of the ligation probes and 5′-ends of the allele-specific primers, respectively. The ligation probe contains a SNP-specific tag-sequence, and the universal allele-specific primers carry an allele-specific fluorescent label in their 5′ end. The three universal PCR primers P1, P2 (each fluorescently labelled with a different dye) and P3 associate a fluorescent dye with each SNP allele. After PCR, the amplified products are captured on beads carrying complementary target sequences for the SNP-specific tag of the ligation probe. Each SNP is assigned a different address sequence, which is contained within the LSO. Each of these addresses is complementary to a unique capture sequence represented by one of the bead types in the array. Therefore, the products of the assays hybridise to different bead types in the array, allowing all genotypes to be read simultaneously. The ratio of the two primer-specific fluorescent signals identifies the genotype as either of the two homozygotes or heterozygote. This universal address system, consisting of artificial

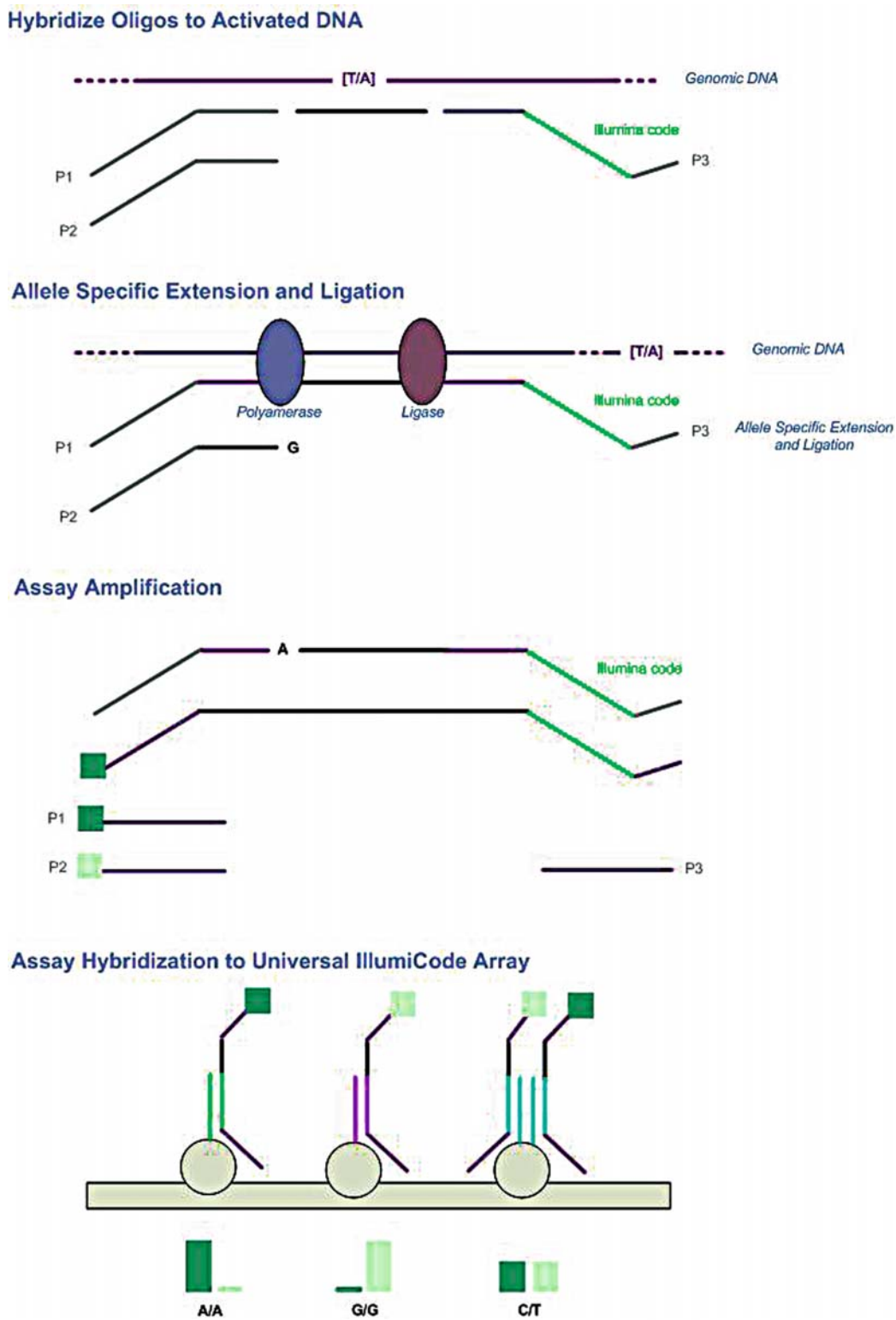

Fig. 2. Overview of the Illumina GoldenGate™ assay.

sequences that are not SNP specific, allows any set of SNPs to be read on a common, standard array, providing flexibility and reducing array manufacturing costs. Custom assays are made on demand by building the address sequences into the SNP-specific assay oligonucleotides.

In order to identify suitable SNPs for the GoldenGate™ assay, only 40 bp of sequence surrounding the SNP is required, and either strand can be chosen for the assay. One major advantage of the GoldenGate™ method is that it requires only three universal primers for PCR, regardless of the number of assays, which saves on costs, and primer sequence-related differences in amplification rates between SNPs are eliminated. This new technology has recently been applied to barley, with the development of a barley Illumina GoldenGate™ assay. This high-throughput SNP platform provides barley researchers with a unique integrated mapping and diversity analysis platform based on more than 3,000 gene-based markers.

Genome-wide genotyping of fixed sets of hundreds of thousands of SNPs is performed using the novel Infinium™ II assays. In this assay, a whole-genome amplification step is used to increase the amount of DNA up to 1,000-fold. The DNA is fragmented and captured on a bead array by hybridisation to immobilised SNP-specific primers, followed by extension with hapten-labelled nucleotides. The primers hybridise adjacent to the SNPs and are extended with a single nucleotide corresponding to the SNP allele. The incorporated hapten-modified nucleotides are detected by adding fluorescently labelled antibodies in several steps to amplify the signals.

3.1.3. Single Base Extension and MALDI–TOF Assays

A popular technology for genotyping SNPs is the minisequencing technique *(8)*, also known as primer extension or single base extension (SBE). In this method, a detection primer is designed to target a sequence immediately upstream of the SNP. The 3′-terminus of the oligonucleotide is then extended, by only one base, by a DNA polymerase using labelled dideoxynucleotide triphosphates (ddNTPs). The terminating fluorescent dye corresponds to a specific ddNTP nucleotide base, making it possible to detect up to four allelic variants at a variable site and discriminate heterozygous from homozygous genotypes. Different detection platforms such as microarrays *(45)*, capillary electrophoresis *(46)*, pyrosequencing *(47)*, flow cytometry *(48)*, mass spectrometry *(49)* or fluorescence plate readers *(50)* can be employed with this minisequencing method, demonstrating its flexibility and adaptation to different analytical technologies.

A novel marker technology, the Sequenom iPLEX™ Assay, uses the SBE coupled with a matrix-assisted laser desorption/ionisation time of flight (MALDI–TOF) mass spectrometer (**Fig. 3**). The iPLEX™ assay begins with PCR amplification of the target region containing the SNP, as with the SBE. However,

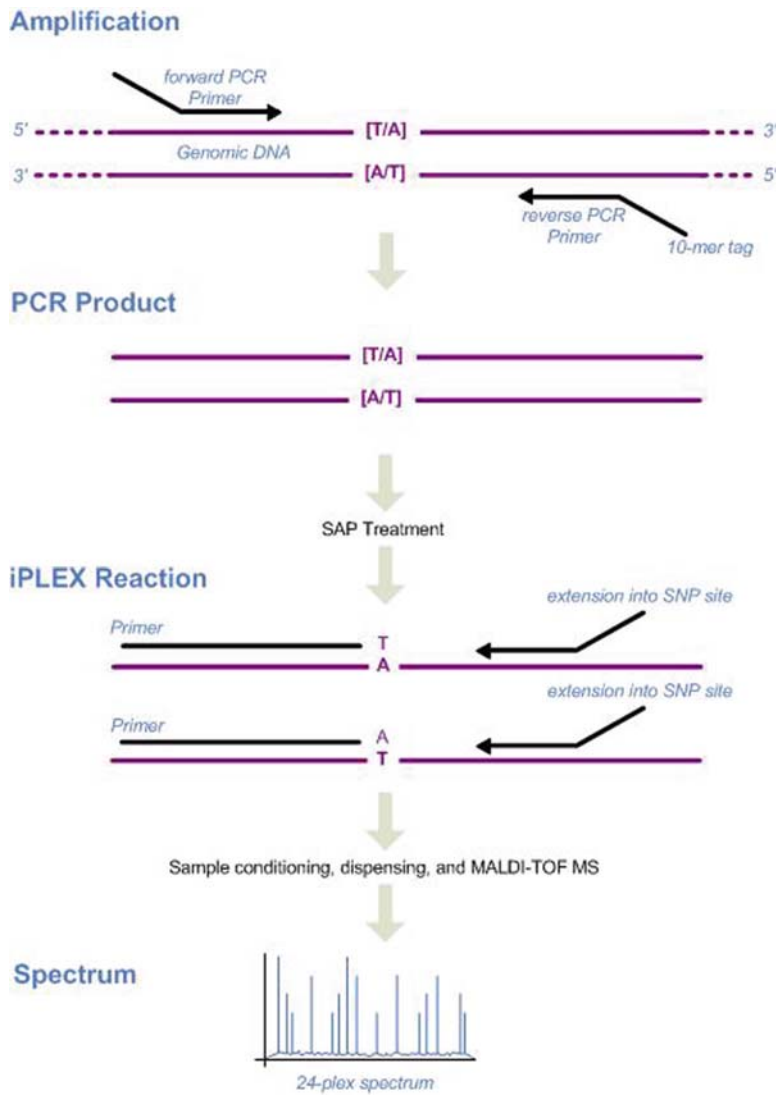

Fig. 3. Overview of the Sequenom iPLEX™ assay.

the PCR primers each have a specific 10-mer tag attached at the 3′ end. The PCR product is treated with Shrimp Alkaline Phosphatase to remove the unincorporated dNTPs, and the multiplex reaction is extended by one base using specific primers. The reaction is desalted to optimise mass spectrometric analysis, and the genotypes are analysed using the MassARRAY workstation. Up to 24 SNPs can be assayed together in one iPLEX™ reaction and this method has been used by Törjek et al. *(51)* to develop a set of 112 SNP markers in *Arabidopsis thaliana*, which suggests that the method can be used as a medium to high-throughput genotyping system.

3.1.4. Oligonucleotide Ligation Assay

A further novel marker technology for genotyping SNPs is the oligonucleotide ligation assay (OLA) *(52)*. This method is based on the properties of an enzymatic reaction in which two adjacent oligonucleotides may be covalently joined by a DNA ligase when annealed to a complementary DNA target. Both of the primers must have perfect base pair complementarity at the ligation site, allowing the discrimination of two alleles at a SNP site. The OLA method has recently been commercialised using the Applied Biosystems SNPlex system, which uses OLA for allelic discrimination and ligation product amplification *(53)*. Genotype information is encoded into a universal set of dye-labelled, mobility-modified fragments, called Zipchute™ Mobility Modifiers, for rapid detection by capillary electrophoresis. The same set of Zipchute™ Mobility Modifiers are used for every SNPlex pool, regardless of which SNPs are chosen

In the first step of the SNPlex, an OLA reaction is performed, where ASO and LSO probes hybridise to the target sequence (**Fig. 4**). These allele-specific and locus-specific probes ligate when they are hybridised to a perfectly matching sequence at the SNP site. At the same time, universal linkers are ligated to the distal termini of the ASO and LSO ligation probes. These linkers contain universal PCR primer-binding sequences and sequences complementary to ASO and LSO probes. A unique ZipCode sequence is attached at the 5′-end of the genomic equivalent sequence within each ASO, allowing the OLA step to encode the genotype information of every SNP into unique ligation products. No optimisation of the OLA is required as all probes are designed to function under the same hybridisation conditions. The unligated probes and linkers, along with any excess genomic DNA, are removed by enzymatic digestion using exonuclease I and lambda exonuclease, to ensure efficiency of the subsequent PCR reaction. This is a simultaneous PCR amplification of the purified ligation products with a single pair of PCR primers, one of which is biotinylated. The use of the universal pair of PCR primers ensures that optimisation of PCR reaction conditions is not required. The biotinylated amplicons are then bound within wells of streptavidin-coated microtitre plates. This allows the non-biotinylated strands to be removed, leaving the single-stranded amplicons bound to the plate. The fluorescently labelled universal ZipChute™ probes then hybridise to the bound single-stranded amplicons. Each ZipChute™ probe contains a sequence complementary to the unique ZipCode sequence within each ASO and contains a mobility modifier, which assigns to each ZipChute™ probe a specific rate of mobility during capillary electrophoresis. The specifically bound ZipChute™ probes are analysed using an Applied Biosystems *3730/3730xl* DNA Analyser. One SNP is typically characterised by two possible alleles, therefore the two fluorescent peaks in an electropherogram represent the two alleles of a specific SNP.

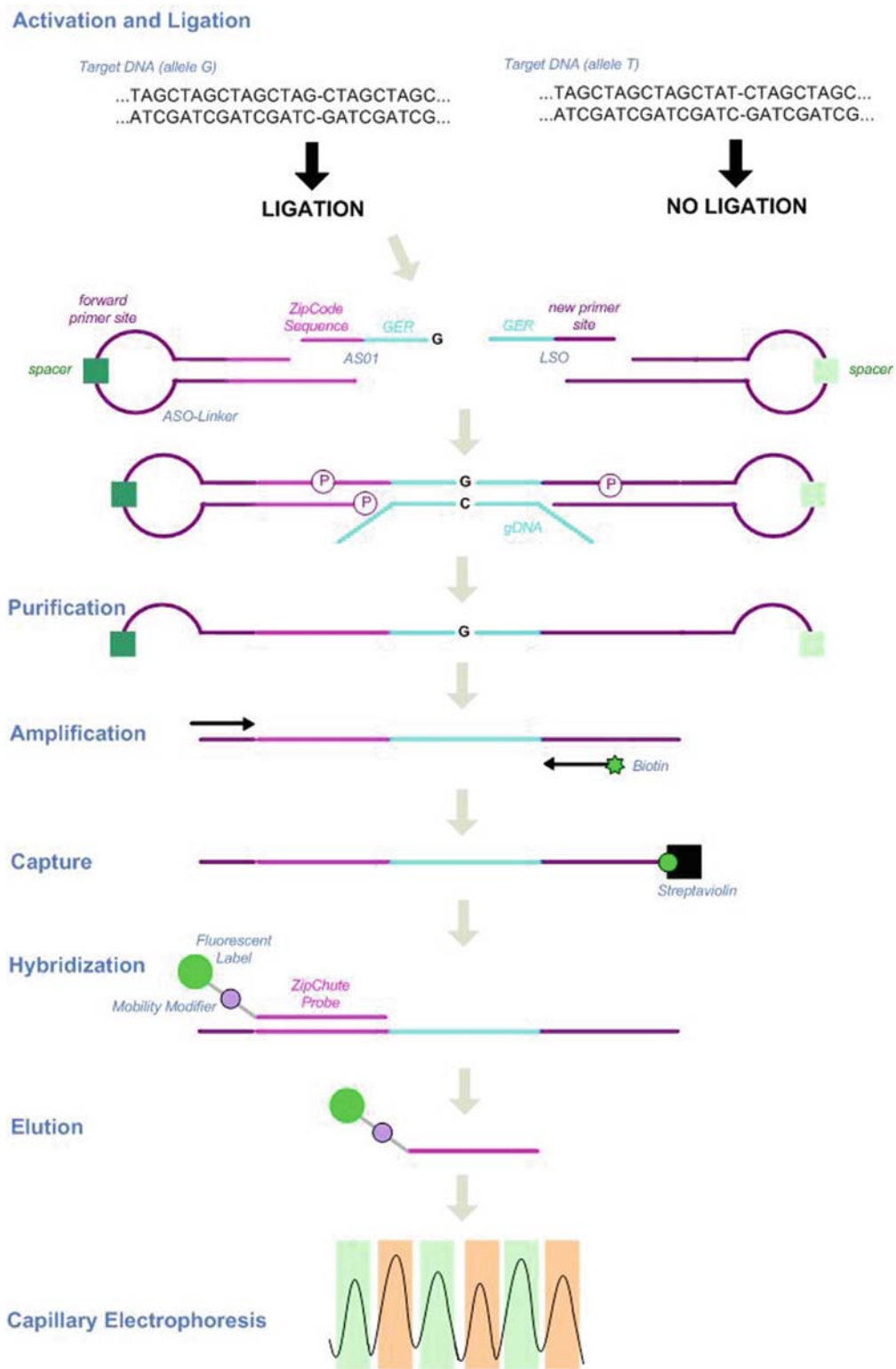

Fig. 4. Overview of the Applied Biosystems SNPlex™ assay.

3.2. New SSRs Technologies

Novel technologies for SSRs have been limited to new approaches to increase the multiplex ratio of the SSRs, to increase throughput and decrease costs. One such technology is the Multiplex-Ready™ Marker technology (MRT), developed at the University of Adelaide. This reduces marker deployment costs for fluorescent-based SSR analysis, and increases genotyping throughput by more efficient electrophoretic separation of the SSRs. MRT is a single step, closed tube assay which involves two PCR steps (**Fig. 5**). In the first step, target loci are amplified using locus-specific primers, tagged at the 5′ end with a defined sequence. This PCR product is used as a template in the second PCR step, in which short dye-labelled primers complementary to the defined sequence amplify the products for automated analysis. The use of the defined primer tag sequence improves automation by providing a consistent PCR yield for markers within a multiplex assay, as well as between reactions. The system is open to flexible dye labelling and has robust tolerance to variations in the concentration and quality of the target DNA. Furthermore, it is compatible with standard capillary electrophoresis instrumentation. The method has been applied for high-throughput analysis of markers used in cereal breeding and is currently being deployed in several Australian cereal research and breeding programs.

3.3. Diversity Array Technology

DArT is diversity array technology *(21)* that assays for the presence of a specific DNA fragment within a representative sample from total genomic DNA. The method does not require prior sequence knowledge, so can be used for plants for which little or no sequence information is available (*see* **Note 3**). The method consists of several steps. The first steps involve complexity reduction in the DNA of interest; creation of a library, which is then arrayed onto a glass slide; followed by hybridisation of fluorescently labelled DNA onto the slides; and lastly detection of the hybridisation signal.

DArT reduces the complexity of a DNA sample to obtain a representation. This involves restriction enzyme digestion and adapter ligation, followed by amplification *(27)*. The genomic representation contains two types of fragments, constant fragments, found in any representation which is prepared from a DNA sample from an individual belonging to a given species, and polymorphic fragments, only found in some but not all of the samples. These polymorphic fragments are the informative DArT markers. Their presence or absence in a sample is assayed by hybridising the representation to a DArT array consisting of a library of that species. The library creation involves generating genomic representations from a pool of individuals covering the genetic diversity of the species that is being studied. These fragments are cloned into a vector and transformed into *Escherichia coli*.

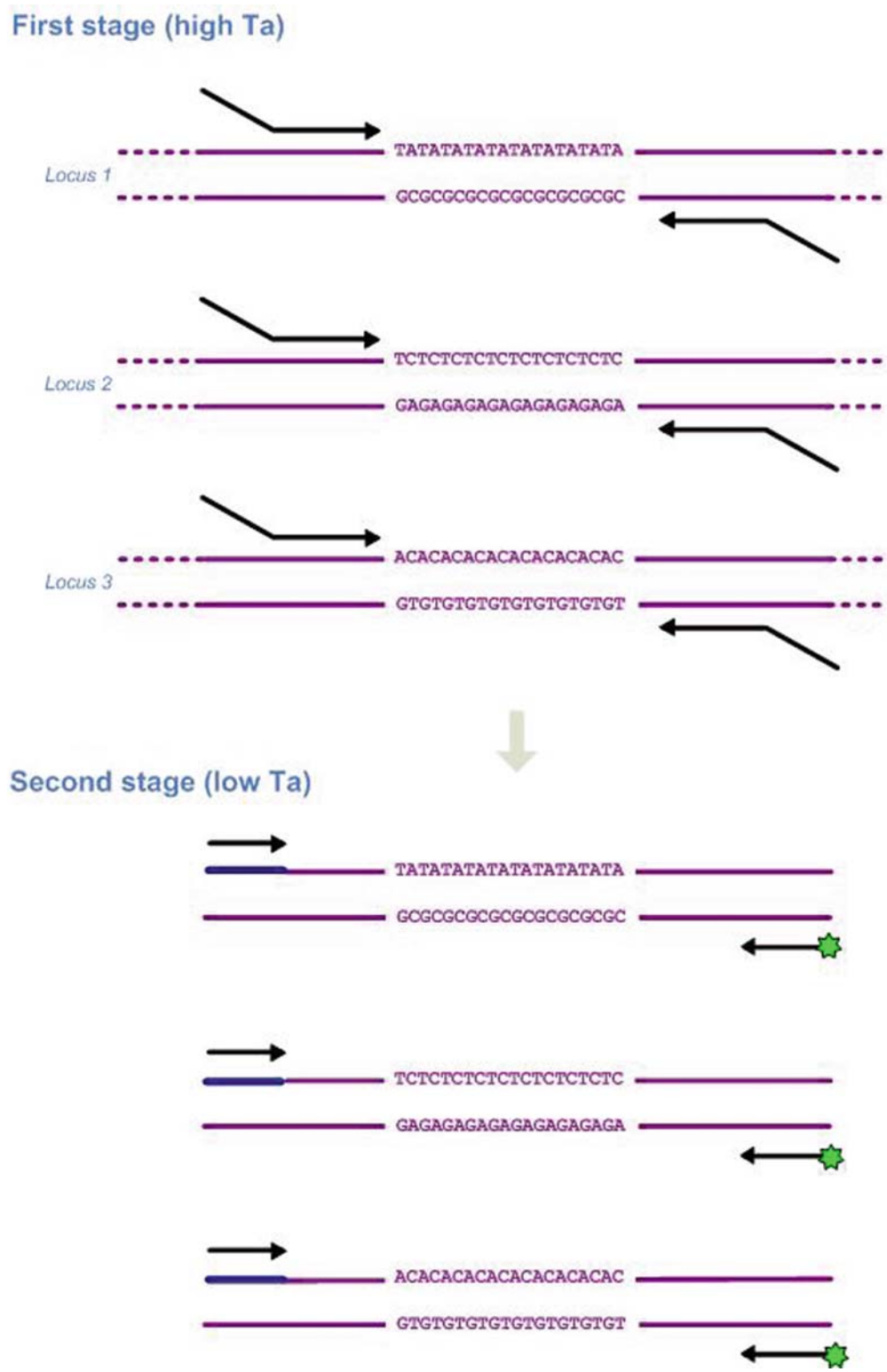

Fig. 5. Overview of the Multiplex-Ready™ Marker assay.

Within the library, each colony contains one of the fragments from the genomic representation. A selection of clones from the library are arrayed into 384-well plates. The fragments within the library are then amplified and spotted onto glass slides using a microarrayer to form the genotyping DArT array. The genotyping arrays are hybridised with genomic representations of individual DNA samples prepared using the same complexity reduction method. These representations are labelled with one fluorescent label, while the vector fragment is labelled with a different fluorescent label to act as a reference. Each individual representation will only hybridise to matching fragments on the genotyping array, thereby displaying a unique hybridisation pattern. The hybridised slides are washed to remove unbound labelled DNA and then scanned to detect the fluorescent signal emitted from the hybridised fragments.

There have been many applications of DArT in plant genomics. A comprehensive collection of DArT markers that are polymorphic for wheat and barley germplasm has been assembled, with over 1,000 markers for barley and 2,000 for wheat. Services are also offered for other crops such as apple, cassava, tomato, sorghum, ryegrass, chickpea, sugarcane, lupin, banana and coconut.

4. Conclusions

Molecular markers have many applications in plant breeding, and the ability to detect the presence of a gene (or genes) controlling a particular desired trait has given rise to MAS. These new technologies make it possible to speed up the breeding process. For example, a desired trait may only be observed in the mature plant, but MAS allows researchers to screen for the trait at a much earlier growth stage. Further advantages of molecular markers are that they make it possible to select simultaneously for many different plant characteristics. They can also be used to identify individual plants with a defined resistance gene without exposing the plant to the pest or pathogen in question. In order to increase throughput and decrease costs, it is necessary to eliminate bottlenecks throughout the genotyping process, as well as minimise sources of variability and human error to ensure data quality and reproducibility. These new technologies may be the way forward for the discovery and application of molecular markers and will enable the application of markers for a broader range of traits in a greater diversity of species than currently possible.

Notes

1. SSRs are also referred to as microsatellites following the method of their initial identification. They are now more commonly called SSRs.
2. PolyPhred integrates phred base calling and quality information within phrap-generated sequence alignments *(54)*. The alignments are viewed and marked for inspection using Consed *(55)*. This method has now been extended to include Bayesian statistical analysis. PolyBayes *(56)* is a fully probabilistic SNP detection algorithm that calculates the probability that discrepancies at a given location of a multiple alignment represent true sequence variations as opposed to sequencing errors. This calculation takes into account the alignment depth, the base calls in each sequence, the base quality values, the base composition in the region and the expected a priori polymorphism rate.
3. Where there is a large amount of sequence data is available for a species, markers such as SNPs and SSRs will provide more information and should be used. In species for which there is limited sequence available, anonymous markers such as DArT may be more cost-effective.

References

1. Gupta, P.K., Roy, J.K., and Prasad, M. (2001) Single nucleotide polymorphisms: A new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr. Sci.* 80, 524–535.
2. Brumfield, R.T., Beerli, P., Nickerson, D.A., and Edwards, S.V. (2003) The utility of single nucleotide polymorphisms in inferences of population history. *Trends Ecol. Evol.* 18, 249–256.
3. Collins, A., Lau, W., and De la Vega, F.M. (2004) Mapping genes for common diseases: The case for genetic (LD) maps. *Hum. Hered.* 58, 2–9.
4. Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5, 94–100.
5. Edwards, D., Forster, J.W., Chagné, D., and Batley, J. (2007) What are SNPs?, in *Association Mapping in Plants* (Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., and De Silva, H.N., eds.), Springer, NY, 41–52.
6. Batley, J. and Edwards, D. (2007) SNP applications in plants, in Association Mapping in Plants (Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E. and, De Silva, H.N. eds.) Springer, NY, 95–102.
7. Edwards, D., Batley, J., Cogan, N.O.I., Forster, J.W., and Chagné, D. (2007) Single nucleotide polymorphism discovery, in Association Mapping in Plants (Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E., and De Silva, H.N. eds.) Springer, NY, 53–76.
8. Syvanen, A.C. (2001) Genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* 2, 930–942.
9. Tóth, G., Gáspári, Z., and Jurka, J. (2000) Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* 10, 967–981.
10. Katti, M.V., Ranjekar, P.K., and Gupta, V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.* 18, 1161–1167.
11. Schlötterer, C. (2000) Evolutionary dynamics of microsatellite DNA. *Nucleic Acids Res.* 20, 211–215.
12. Tautz, D. (1989) Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res.* 17, 6463–6471.

13. Powell, W., Machray, G.C., and Provan, J. (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.* 1, 215–222.
14. Subramanian, S., Mishra, R.K., and Singh, L. (2003) Genome-wide analysis of microsatellite repeats in humans: Their abundance and density in specific genomic regions. *Genome Biol.* 4, R13.
15. Awadalla, P. and Ritland, K. (1997) Microsatellite variation and evolution in the *Mimulus guttatus* species complex with contracting mating systems. *Mol. Biol. Evol.* 14, 1023–1034.
16. Moxon, E.R. and Wills, C. (1999) DNA microsatellites: Agents of evolution. *Sci. Am.* 280, 94–99.
17. Kashi, Y., King, D., and Soller, M. (1997) Simple sequence repeats as a source of quantitative genetic variation. *Trends Genet.* 13, 74–78.
18. Gupta, M., Chyi, Y.-S., Romero-Severson, J., and Owen, J.L. (1994) Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. *Theor. Appl. Genet.* 89, 998–1006.
19. Mortimer, J., Batley, J., Love, C., Logan, E., and Edwards, D. (2005) Simple sequence repeat (SSR) and GC distribution in the *Arabidopsis thaliana* genome. *J. Plant Biotechnol.* 7, 17–25.
20. Li, Y.-C., Korol, A.B., Fahima, T., Beiles, A., and Nevo, E. (2002) Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Mol. Ecol.*11, 2453–2465.
21. Jaccoud, D., Peng, K., Feinstein, D., and Kilian, A. (2001) Diversity arrays: A solid state technology for sequence information independent genotyping. *Nucleic Acids Res.* 29, e25.
22. Xia, L., Peng, K., Yang, S., Wenzl, P., de Vicente, C., Fregene, M., and Kilian, A. (2005) DArT for high-throughput genotyping of cassava (*Manihot esculenta*) and its wild relatives. *Theor. Appl. Genet.* 110, 1092–1098.
23. Yang, S., Pang, W., Ash, G., Harper, J., Carling, J., Wenzl, P., Huttner, E., and Kilian, A. (2006) Low level of genetic diversity in cultivated pigeonpea compared to its wild relatives is revealed by diversity arrays technology (DArT). *Theor. Appl. Genet.* 113, 585–595.
24. Xie, Y., McNally, K., Li, C.Y., Leung, H., and Zhu, Y.Y. (2006) A high-throughput genomic tool: Diversity array technology complementary for rice genotyping. *J. Integr. Plant Biol.* 48, 1069–1076.
25. Akbari, M., Wenzl, P., Vanessa, C., Carling, J., Xia, L., Yang, S., Uszynski, G., Mohler, V., Lehmensiek, A., Kuchel, H., Hayden, M.J., Howes, N., Sharp, P., Rathmell, B., Vaughan, P., Huttner, E., and Kilian, A. (2006) Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. *Theor. Appl. Genet.* 113, 1409–1420.
26. Wenzl, P., Li, H., Carling, J., Zhou, M., Raman, H., Paul, E., Hearnden, P., Maier, C., Xia, L., Caig, V., Ovesna, J., Cakir, M., Poulsen, D., Wang, J., Raman, R., Smith, K.P., Muehlbauer, G.J., Chalmers, K.J., Kleinhofs, A., Huttner, E., and Kilian, A. (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and phenotypic traits. *BMC Genom.* 7, 206.
27. Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., and Kilian, A. (2004) Diversity arrays technology (DArT) for whole-genome profiling of barley. *PNAS.* 101, 9915–9920.
28. Ewing, B. and Green, P. (1998a) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8, 186–194.
29. Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998b) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* 8, 175–185.
30. Barker, G., Batley, J., O'Sullivan, H., Edwards, K.J., and Edwards, D. (2003) Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP. *Bioinformatics* 19, 421–422.
31. Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., and Edwards, D. (2003) Mining for single nucleotide polymorphisms and insertions/deletions in maize expressed sequence tag data. *Plant Physiol.* 132, 84–91.
32. Savage, D., Batley, J., Erwin, T., Logan, E., Love, C.G., Lim, G.A.C., Mongin, E., Barker, G., Spangenberg, G.C., and Edwards, D. (2005) SNPServer: A real-time SNP discovery tool. *Nucleic Acids Res.* 33, W493–W495.
33. Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.* 9, 868–877.
34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
35. Edwards, K.J., Barker, J.H.A., Daly, A., Jones, C., and Karp, A. (1996) Microsatellite libraries enriched for several microsatellite sequences in plants. *Biotechniques* 20, 758–760.
36. Robinson, A.J., Love, C.G., Batley, J., Barker, G., and Edwards, D. (2004) Simple sequence repeat marker loci discovery using SSRPrimer. *Bioinfomatics* 20, 1475–1476.
37. Jewell, E., Robinson, A., Savage, D., Erwin, T., Love, C.G., Lim, G.A.C., Li, X., Batley, J.,

Spangenberg, G.C., and Edwards, D. (2006) SSR Primer and SSR Taxonomy Tree: Biome SSR discovery. *Nucleic Acids Res.* 34, W656–W659.

38. Hapmap, C.A. (2003) The International HapMap Project: The International HapMap Consortium. *Nature* 426, 789–796.
39. Mein, C.A., Barratt, B.J., Dunn, M.G., Siegmund, T., Smith, A.N., Esposito, L., Nutland, S., Stevens, H.E., Wilson, A.J., Phillips, M.S., Jarvis, N., Law, S., De Arruda, M., and Todd, J.A. (2000) Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Res.* 10, 330–343.
40. Olivier, M. (2005) The Invader® assay for SNP genotyping. *Mutat. Res.* 573, 103–110.
41. Olivier, M., Chuang, L.M., Chang, M.S., Chen, Y.T., Pei, D., Ranade, K., de Witte, A., Allen, J., Tran, N., Curb, D., Pratt, R., Neefs, H., de Arruda, M., Law, S., Neri, B., Wang, L., and Cox, D.R. (2002) High-throughput genotyping of single nucleotide polymorphisms using new biplex invader technology. *Nucleic Acids Res.* 30, e53.
42. Gupta, M., Niaunsuksiri, W., Schulenberg, G., Hartl, T., Novak, S., Bayan. J., Vanopduop, N., Bing, J., and Thompson, S. (2008) A non-PCR-based Invader® assay quantitatively detects single-copy genes in complex plant genomes. *Mol. Breeding* 21, 173–181.
43. Fan, J.-B., Oliphant, A., Shen, R., Kermani, B.G., Garcia, F., Gunderson, K.L., Hansen, M., Steemers, F., Butler, S.L., Deloukas, P., Galver, L., Hunt, S., McBride, C., Bibikova, M., Rubano, T., Chen, J., Wickham, E., Doucet, D., Chang, W., Campbell, D., Zhang, B., Kruglyak, S., Bentley, D., Haas, J., Rigault, P., Zhou, L., Stuelpnagel, J., and Chee, M.S. (2003) Highly parallel SNP genotyping. *Cold Spring Harb. Symp. Quant. Biol.* 68, 69–78.
44. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G., and Chee, M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.* 37, 549–554.
45. Pastinen, T., Kurg, A., Metspalu, A., Peltonen, L., and Syvänen, A.-C. (1997) Minisequencing: A specific tool for DNA analysis and diagnostics on oligonucleotide arrays. *Genome Res.* 7, 606–614.
46. Batley, J., Mogg, R., Edwards, D., O'Sullivan, H., and Edwards, K.J. (2003). A high-throughput SNuPE assay for genotyping SNPs in the flanking regions of Zea mays sequence tagged simple sequence repeats. *Mol. Breeding* 11, 111–120.
47. Ekstroem, B., Alderborn, A., and Hammerling, U. (2000) Pyrosequencing for SNPs. *Proceedings of SPIE—The International Society for Optical Engineering* 3926, 134–139.
48. Chen, J., Iannone, M.A., Li, M.-S., Taylor, J.D., Rivers, P., Nelsen, A.J., Slentz-Kesler, K.A., Roses, A., and Weiner, M.P. (2000) A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res.* 10, 549–557.
49. Haff, L.A. and Smirnov, I.P. (1997) Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass spectrometry. *Genome Res.* 7, 378–388.
50. Hsu, T.M., Chen, X., Duan, S., Miller, R.D., and Kwok, P.-Y. (2001) Universal SNP genotyping assay with fluorescence polarization detection. *BioTechniques* 31, 560–570.
51. Törjek, O., Berger, D., Meyer, B.C., Müssig, C., Schmid, K.J., Sörensen, T.R., Weisshaar, B., Mitchell-Olds, T., and Altmann, T. (2003) Establishment of a high-efficiency SNP-based framework marker set for *Arabidopsis. Plant J.* 36, 122–140.
52. Landegren, U., Kaiser, R., Sanders, J., and Hood, L. (1988) A ligase-mediated gene detection technique. *Science* 241, 1077–1080.
53. Tobler, A.R., Short, S., Andersen, M.R., Paner, T.M., Briggs, J.C., Lambert, S.M., Wu, P.P., Wang, Y., Spoonde, A.Y., Koehler, R.T., Peyret, N., Chen, C., Broomer, A.J., Ridzon, D.A., Zhou, H., Hoo, B.S., Hayashibara, K.C., Leong, L.N., Ma, C.N., Rosenblum, B.B., Day, J.P., Ziegle, J.S., de la Vega, F.M., Rhodes, M.D., Hennessy, K.M., and Wenz, H.M. (2005) The SNPlex genotyping system: A flexible and scalable platform for SNP genotyping. *J. Biomol. Tech.* 16, 398–406.
54. Greeen, P. (1994) Phrap. unpublished. www. Phrap.org.
55. Gordon, D., Abajian, C. and, Green, P. (1998) Consed: A graphical tool for sequence finishing. *Genome Res.* 8, 195–202.
56. Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z.J., Zakeri,, H., Stitziel,, N.O., Hillier, L., Kwok,, P.Y. and, Gish,, W.R. (1999) A general approach to single nucleotide polymorphism discovery. *Nat. Genet.* 23, 452–456.
57. Chagné, D., Batley, J., Edwards, D., and Forster, J.W. (2007) Single nucleotide polymorphisms genotyping in plants, in Association Mapping in Plants (Oraguzie, N.C., Rikkerink, E.H.A., Gardiner, S.E. and, De Silva, H.N., eds.), Springer, NY, 77–94.