Plant Genotyping II

SNP Technology

This page intentionally left blank

PLANT GENOTYPING II

SNP Technology

Edited by

Robert J. Henry

Centre for Plant Conservation Genetics
Southern Cross University
Lismore, New South Wales, Australia

www.cabi.org

CABI is a trading name of CAB International

CABI Head Office
Nosworthy Way
Wallingford
Oxfordshire OX10 8DE
UK

Tel: +44 (0)1491 832111
Fax: +44 (0)1491 833508
E-mail: cabi@cabi.org
Web site: www.cabi.org

CABI North American Office
875 Massachusetts Avenue
7th Floor
Cambridge, Massachusetts 02139
USA

Tel: +1 617 395 4056
Fax: +1 617 354 6875
E-mail: cabi-nao@cabi.org

A catalogue record for this book is available from the British Library, London, UK.

Library of Congress Cataloging-in-Publication Data

Plant genotyping II : SNP technology / edited by Robert J. Henry.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-84593-382-1 (alk. paper)
1. Plant genome mapping. 2. Genetic polymorphisms. I. Henry, Robert J.

QK981.45.P565 2008
581.3'5--dc22

2007041585

ISBN: 978 1 84593 382 1

Typeset by SPi, Pondicherry, India.
Printed and bound in the UK by Biddles Ltd, King's Lynn.

Contents

Contributors

G. Ablett, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: gary.ablett@scu.edu.au*

G.L. Barker, *University of Bristol, School of Biological Sciences, Woodland Road, Bristol BS8 1UG, UK. E-mail: Gary Barker@bristol.ac.uk*

L.M.E. Bradbury, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: l.bradbury.10@scu.edu.au*

P.C. Bundock, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: peter.bundock@scu.edu.au*

N.O.I. Cogan, *Department of Primary Industries, Biosciences Research Division, Victorian AgriBiosciences Centre, La Trobe Research and Development Park, Bundoora, Victoria, 3083, Australia. E-mail: noel.cogan@dpi.vic.gov.au*

G. Cordeiro, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: giovanni@cordeiros.net*

M.J. Cross, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: mike.cross@scu.edu.au*

M.P. Dobrowolski, *Department of Primary Industries, Biosciences Research Division, Hamilton Centre, Hamilton, Victoria 3300, Australia and Molecular Plant Breeding Cooperative Research Centre, Australia. E-mail: mark.dobrowolski@dpi.vic.gov.au*

K.J. Edward, *University of Bristol, School of Biological Sciences Woodland Road, Bristol BS8 1UG, UK. E-mail: K.J.Edwards@bristol.ac.uk*

F. Eliott, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: frances.eliott@scu.edu.au*

J.W. Forster, *Department of Primary Industries, Biosciences Research Division, Victorian AgriBiosciences Centre, La Trobe Research and Development Park, Bundoora, Victoria, 3083, Australia. E-mail: john.forster@dpi.vic.gov.au*

M.G. Francki, *Department of Agriculture and Food Western Australia, 3 Baron-Hay Court, South Perth, Western Australia 6151 and State Agricultural Biotechnology Centre, Murdoch University, Murdoch, Western Australia 6983, Australia and Molecular Plant Breeding Cooperative Research Centre, Australia. E-mail: mfrancki@agric.wa.gov.au*

R.J. Henry, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: robert.henry@scu.edu.au*

D. Irwin, *Sequenom Herston, Queensland, Australia. E-mail: dirwin@sequenom.com*

L. Izquierdo, *Griffith University, Lismore, New South Wales, Australia. E-mail: izquierdoliz@hotmail.com*

S. Kasem, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: s.kasem.11@scu.edu.au*

T. Pacey-Miller, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: toni.pacey-miller@scu.edu.au*

J.A. Pattemore, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: j.pattemore.16@scu.edu.au*

R.L. Poole, *University of Bristol, School of Biological Sciences Woodland Road, Bristol BS8 1UG, UK. E-mail: R.L.Poole@Bristol.ac.uk*

A. Rafalski, *DuPont Crop Genetics Research, Experimental Station E353, Route 141 and Henry Clay Road, Wilmington, Delaware. E-mail: antoni.rafalski@cgr.dupont.com*

N. Rice, *Australian Plant DNA Bank, Southern Cross University, Lismore, New South Wales, Australia. E-mail: nicole.rice@scu.edu.au*

K.F. Smith, *Department of Primary Industries, Biosciences Research Division, Hamilton Centre, Hamilton, Victoria 3300, Australia and Molecular Plant Breeding Cooperative Research Centre, Australia. E-mail: kevin.f.smith@dpi.vic.gov.au*

G.C. Spangenberg, *Department of Primary Industries, Biosciences Research Division, Victorian AgriBiosciences Centre, La Trobe Research and Development Park, Bundoora, Victoria, 3083, Australia. E-mail: german.spangenberg@dpi.vic.gov.au*

S. Tingey, *DuPont Crop Genetics Research, Experimental Station E353, Route 141 and Henry Clay Road, Wilmington, Delaware. E-mail: Scott.V.Tingey@USA.dupont.com*

M. Trau, *Centre for Nanotechnology and Biomaterials University of Queensland, St Lucia, Queensland, Australia. E-mail: m.trau@uq.edu.au*

D.L.E. Waters, *Centre for Plant Conservation Genetics, Southern Cross University, Lismore, New South Wales, Australia. E-mail: daniel.waters@scu.edu.au*

Preface

Plant Genotyping: The DNA Fingerprinting of Plants was published in 2001. The book was based upon a workshop held at Southern Cross University in 2000 with additional contributions from some authors who did not attend the workshop. The techniques available for plant DNA analysis have advanced considerably in the time since the original volume was published. *Plant Genotyping II: SNP Technology* aims to describe some of the important recent developments in this field. This book is based upon a second workshop held recently to review progress in this area. Recent developments focus on high-throughput methods and generally target single nucleotide polymorphism (SNP) discovery and analysis.

R.J. Henry
Centre for Plant Conservation Genetics
Southern Cross University

This page intentionally left blank

1 SNP Discovery in Plants

K.J. Edward, R.L. Poole and G.L. Barker

Introduction

Marker-assisted selection (MAS) using DNA-based molecular markers is now used by virtually all commercial plant and animal breeding companies (reviewed in Henry, 2001). Molecular markers can be used to identify both useful genotypes for inclusion in breeding programmes and interesting progeny for further study. However, the cost of MAS is a limitation when compared to phenotypically driven conventional breeding. In plants microsatellite markers are currently the most popular type of marker system; however, their development is resource-intensive and they are difficult to multiplex. Recently there has been considerable interest in the development of single nucleotide polymorphism (SNP pronounced SNiP)-based marker systems. More recently the use of SNPs has been fuelled by the explosion in the number of expressed sequence tags (ESTs) available in the various databases and this trend is set to continue for the foreseeable future (Varshney *et al*., 2004). SNPs are the most common form of sequence variation between individuals within a species (reviewed in Brookes, 1999). New SNPs are continually arising within every cell of an organism but most are removed through the action of the enzymatic process known as mismatch repair (MMR), and hence SNPs which become fixed within a germ line and a population are simply those mutations which have escaped the repair process (reviewed in Kunkel and Erie, 2005). As mutations are the source of all SNPs, the number of SNPs within a specific population reflects the forward and reverse mutation rates, the number of generations that separate the population from its closest relative and the number of independent progenitors of that population. The number and distribution of SNPs are also influenced by the past and present selection pressures applied to the population including any bottlenecks through which it has passed. Hence, while mutations probably occur at broadly similar frequencies in most species, the number of SNPs present

 Plant Genotyping II: SNP Technology (ed. R.J. Henry)

within a population will often show considerable variation, and it should come as no surprise to any plant breeder that different plant species will have different SNP frequencies and that these frequencies will vary within different regions of the genome. Because of this, the task of 'discovering' SNPs within small highly selected populations (such as many crop species) is likely to require considerably more effort than that required to discover SNPs in a large natural population of a plant species undergoing limited selection. Although this argument is a simple one it is surprising how many plant researchers are disappointed by the frequency of SNPs in the various (but not all) crop species; such difficulties merely reflect the material being investigated and not the methods employed. In such cases it is interesting to note that recently technologies have been applied to numerous crop species to create mutant populations (via chemical and radioactive mutagenesis). Such mutant populations are by definition a rich source of new and novel SNPs, albeit SNPs which do not exist within the wider population (Slade *et al*., 2004).

SNP Characteristics

Brookes defined SNPs as 'single base pair positions in genomic DNA at which different sequence alternatives (alleles) exist in normal individuals in some population(s), wherein the least frequent allele has an abundance of 1% or greater' (Brookes, 1999). The definition suggested by Brookes indicates that SNPs can be bi-, tri- or tetra-allelic polymorphisms and excludes single base insertion/deletion variants (indels). A typical SNP is shown in Fig. 1.1.

By the above definition a single sequence examined in isolation cannot be said to have a SNP; for instance, in Fig. 1.1 individuals have either a 'G' SNP or an 'A' SNP at positions 11 and 19. However, other scenarios are also possible; for instance, a 'G or C' SNP scenario, or a 'G or T' SNP scenario.

Not all SNP pairings occur with similar frequencies; for instance, in humans two out of every three SNPs result from a C/T change (a transition), whereas in wheat and maize C/T transitions represent between 45% and 55% of the observed SNPs (Edwards group, University of Bristol, 2006, unpublished data). Why are C/T SNPs so common compared to, say, C/G SNPs (transversions)? The answer lies in the chemical structure of DNA and its four bases; for instance, C/T SNP transitions (and their opposite-strand counterpart G/A SNPs) are thought to occur through the deamination of (normally) methylated dC, resulting in it being read as a T by DNA polymerase during replication. Understanding the nature and frequency of the different types of SNPs can provide valuable help when trying to determine if an identified sequence change is real or simply a result of an error in sequencing. Using this general rule the majority of SNPs observed should be transitions (C/T or G/A SNPs) and, furthermore, if the SNP occurs in the coding region of a gene, it is most likely to be synonymous, i.e. not result in a change in the amino acid sequence of the encoded protein. In contrast, transversion non-synonymous SNPs leading to a change in amino acid content are likely

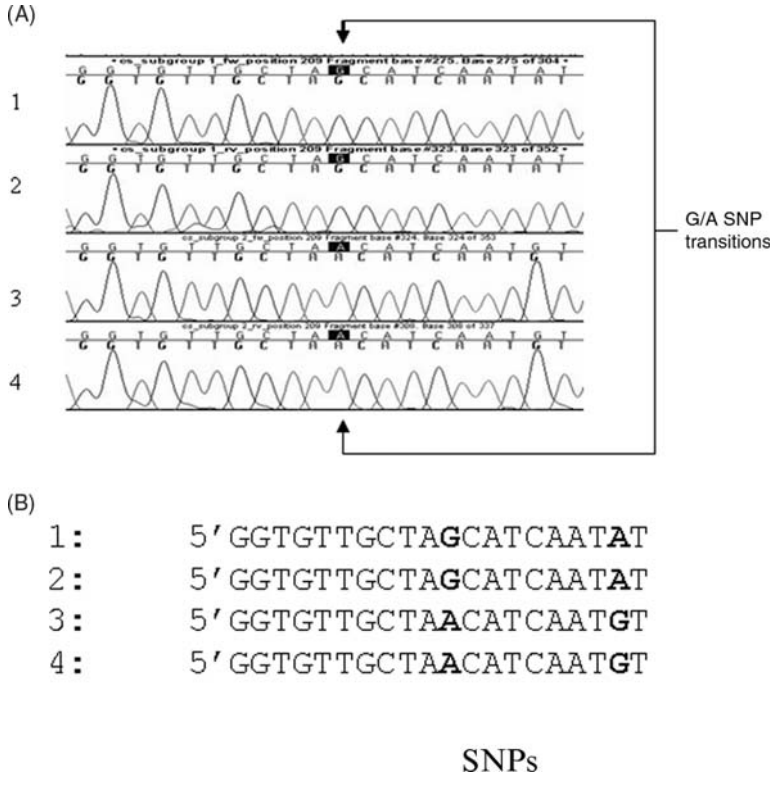


Fig. 1.1. (A) Two G/A SNP transitions in four different wheat lines as discovered by the direct sequencing of PCR products. In the example shown the first two individuals (1 and 2) have the same two SNPs when compared to the third and fourth individuals (3 and 4), which have an 'A' base. (B) Sequence representation of the four chromatograms showing the two G to A transitions.

to be rare. It is important to note that while non-synonymous SNPs are relatively rare, as they result in changes to the amino acid sequence of the encoded protein, they are more likely to lead to changes in protein function and hence phenotype, and therefore more likely to be of considerable interest to plant geneticists.

In addition to considering individual SNPs, researchers also need to consider SNPs in the context of the term 'haplotype'. In this context 'haplotype' is defined as any set of closely linked SNPs inherited as a unit through a lack of recombination between the individual SNPs (more often referred to as linkage disequilibrium (LD); reviewed in plants in Flint-Garcia *et al*., 2003). Hence, within a haplotype the specific combinations of SNPs can be considered to be 'in association' with each other and therefore for genotyping purposes regarded as a single unit (Buntjer *et al*., 2005). Figure 1.2 shows a scenario in which six SNPs and four haplotypes were identified within eight varieties across a 280 base pair (bp) region of the wheat genome. In Fig. 1.2, five of the six SNPs are transitions and only one is a transversion.

	28	48	75	145	233	277	Haplotype
1	C	T	C	T	C	C	1
2	C	C	T	C	T	C	2
3	C	C	T	C	T	C	
4	C	C	T	C	T	C	
5	G	C	C	C	C	T	3
6	G	C	C	C	C	T	
7	G	C	C	C	C	T	
8	C	C	C	T	C	C	4

Fig. 1.2. Representation of the six SNPs and four haplotypes found in eight wheat genotypes. The columns represent the six SNPs identified by direct sequencing in 280 bp of wheat DNA (at positions 28 to 277). The rows correspond to the genotype of the eight individual wheat varieties (1–8). In this region it appears that the SNP co-segregates as one of four haplotypes, each of which is represented by a different colour.

Understanding LD and how it relates to individual SNPs can be useful in both the design of genotyping experiments and identifying how many SNPs are required to cover a particular genome. In the example shown in Fig. 1.2, although six SNPs are present within this particular linkage block, genotyping with just three of the SNPs (those found at positions 28, 48 and 75) would be sufficient to define the four haplotypes. Of course the haplotypes and hence the LD that exists between them will decay over time as recombination events split up the various blocks and generate new haplotype variants. If LD extends over a large physical distance, few SNPs will be required to cover the genome but the resolution of the resulting linkage map will be low. Conversely where LD decays rapidly many SNPs will be required for genome coverage but the resulting map will have high resolution.

As previously suggested, SNPs are the most abundant genetic polymorphism in higher eukaryotic genomes including plants. Numerous studies have suggested that the frequency of SNPs in plants range from one SNP per 21 bp in potato to one SNP per 7000 bp in tomato (Table 1.1). Why the difference? As defined previously, SNP frequency will be determined by the biology and the history of the species, with crops being generally less diverse (due to greater levels of selection and less time since the last common ancestor) than wild species. However, two further reasons for the differences in frequency, seen in the species presented in Table 1.1, are both the regions targeted for study and the diversity of the individuals used. For instance, the study by Gilchrist *et al*. (2006) in cottonwood found SNP frequencies ranging from one SNP per 64 bp for non-coding regions to one SNP per 229 bp in coding regions. Hence, targeting non-coding or intron-containing regions might be a more logical and profitable route to SNP discovery than screening coding regions.

Table 1.1. Published abundance of SNPs in various plant genomes.

Species	SNP abundance	References
Arabidopsis	From 1 SNP per 36 bp to 1 SNP per 3.3 kb	Jander *et al*., 2002
Tomato	1 SNP per 7kb	
	1.5–7.3 SNPs per kb	Nesbitt and Tanksley, 2002 Labate and Baldo, 2005
Potato	1 SNP per 21 bp	Rickert *et al*., 2003
Barley	1 SNP per 78 bp	Russell *et al*., 2004
Rice	1 SNP per 232 bp	Feltus *et al*., 2004
Wheat	1 SNP per 540 bp	Somers *et al*., 2003
	1 SNP per 370 bp	Khlestkina and Salina, 2006
Maize	1 SNP per 47.7 bp for non-coding sequences and 1 SNP per 130.5 bp in coding sequences	Ching *et al*., 2002
	1 SNP per 83 bp	Bhattramakki *et al*., 2002
Loblolly pine	1 SNP per 63 bp	Brown, 2004
Soybean	1 SNP per 272.5 bp	Zhu *et al*., 2003
Western black cottonwood	1 SNP per 64 bp for non-coding regions and 1 SNP per 229 bp for coding regions	Gilchrist *et al*., 2006

What Is So Special About SNP Discovery in Plants?

As plant geneticists, it is not surprising we think that plants are special. From the very beginning humans have exploited plants to provide them with food and shelter and today a handful of crops, mainly cereals, provide most of our calorific requirements. Exploitation of these plants has invariably resulted in strong selection pressures being applied such that as Darwin suggested '[d]omesticated races show adaptation, not indeed to the animals or plants own good, but to man's use or fancy' (Darwin, 1859). In first cultivating and then domesticating plants we have subjected them to intense selection pressure with the result that a reduced number of alleles, compared to unselected material, remain within the populations. From the point of view of this chapter, fewer alleles means fewer SNPs and hence one should assume that SNPs will be hard to find in previously selected material, like that currently being grown in intensive farming systems. In addition, the ability of modern plant breeders to screen many tens of thousands of plants for very specific phenotypes such as yield or flower colour continue to influence, and potentially reduce, the number of alleles and therefore the number of SNPs at any given locus. That said, more recently a better understanding of the possible consequences of genetic bottlenecks in

agriculture is beginning to result in new strategies being employed such that allele/SNP diversity might now begin to show signs of recovery (Cox *et al*., 1986). However, the advent of genetic modification could pose new genetic diversity-based problems as genetic modification usually results in the modification of a single individual representing a single genotype. As the transgene itself then becomes, at least in part, the subject of selection, it is clear that any genes/alleles linked to the insertions site are likely to be favourably transmitted to the progeny and future generations via linkage drag. Thus, the availability of molecular markers such as SNPs from regions immediately surrounding the insertion site could be of considerable use in identifying those rare recombination events that could reduce the linkage drag and hence help to 'move' the transgene to a wider variety of chromosomal environments (Boerma and Walker, 2005).

In addition to the considerations above, there are several reasons why plants pose particular challenges when it comes to SNP discovery.

Resource requirements

SNP discovery (including validation) is expensive and time-consuming; recent work in various laboratories including our own in the UK suggests that in most plant species the cost of discovery and validation for each individual SNP is in the region of US$200. Therefore, a collection of 1000 SNPs might cost US$200,000 to develop. While this figure might appear to be high, the majority of the costs are in the validation rather than the discovery phase; however, the precise balance of the costs varies depending upon the procedures used. While these costs are high they are similar to the cost of developing microsatellite markers or of converting other marker systems to PCR-based, single locus markers. Although the development costs of SNPs are high, once developed and validated a polymorphic SNP is a valuable asset for the laboratory responsible for developing it. In this context, public SNP consortia, where the results immediately enter the public domain, appear to be a very cost-efficient route to developing large numbers of SNP markers for previously under-resourced plant species. An excellent example of such a consortium effort can be found in the National Science Foundation-funded project 'Haplotype polymorphism in polyploid wheat and their diploid ancestor' (http://wheat.pw.usda.gov/SNP/new/project.shtml), which at its end in June 2006 had resulted in 3326 polymorphic SNP loci in either hexaploid wheat or its diploid ancestors.

Unfortunately, the funds available to discover SNPs in most plant species are considerably less than those available for SNP discovery in humans or other higher mammals. In addition to the direct costs, successful SNP discovery and validation also requires considerable infrastructure, especially if one wishes to undertake large-scale genome-wide SNP discovery. This is because there is a requirement to have access to high throughput sequencing and genotyping facilities as well as having access to a dedicated expert in bioinformatics to manage the resulting databases. Without access to such skills it is unlikely that the average researcher would be able to develop and utilize all but a handful of SNPs at any one time.

Vegetative propagation and self-pollination

Plants have the almost unique ability to either undergo vegetative propagation or maintain themselves through self-pollination. Clearly material derived by vegetative propagation will be genetically identical and in the case of self-pollination show reduced genetic diversity and increased levels of homozygosity. In the case of vegetative propagation, screening for polymorphic markers between individuals will probably be a waste of time. Unfortunately, vegetative propagation is widely used in the horticultural industry and to a lesser extent in forestry. For instance in the UK, genotyping studies in several tree species have revealed surprisingly low levels of genetic diversity and evidence for the past use of clonal material in woodland and hedgerow planting (Winfield *et al*., 1998). Hence, before starting out on a programme of SNP discovery and validation it is logical to establish the provenance and diversity of the material to be used by carrying out a preliminary genotyping study, for instance by using other less expensive marker systems.

Polyploidy

Approximately 75% of plant species are polyploids (Soltis and Soltis, 1995); while polyploidy is not unique to the plant kingdom, the extent of polyploidy means that virtually all species of interest to plant geneticists will show evidence of either current or ancient polyploidy. The presence of multiple, usually related, genomes within the plant cell presents a significant problem for the identification of SNPs. In genotyping experiments it means that one has to distinguish between alleles, paralogues (members of a multigene family within a single genome) and homoeologues (genes present on the different progenitor species genomes in an allopolyploid; reviewed in Small *et al*., 2004). Hence, the issue for plant geneticists working with polyploids is: 'How can one differentiate between SNPs in the homologous, paralogous and homoeologous copies of a particular gene?' The ability and efforts of plant geneticists to solve this problem will be discussed throughout this chapter.

Methods for SNP Discovery in Plants

SNP discovery and SNP validation go hand in hand. Without validation one cannot be sure whether an identified SNP is present or simply an artefact of the identification procedure, for instance a sequencing error. However, discovery and validation are not the same as SNP detection. SNP detection is usually taken to mean the technologies and procedures put into place to screen various genotypes for the presence or absence of specific previously identified and validated SNPs. This is an important distinction to make; many chapters and reviews supposedly dedicated to SNP discovery are no more than a catalogue of technologies for SNP detection. In this chapter we will focus our efforts on SNP discovery and where necessary SNP validation with just the occasional mention of SNP detection.

What do we mean by SNP discovery? Again many procedures have been developed which identify that a SNP is probably present within a DNA fragment (compared to an alternative form of that fragment) without identifying the SNP itself. The detection of fragments containing SNPs is central to the process of efficient SNP discovery and as such some time will be spent describing some of the methods employed; however, following on from this kind of preliminary detection it is important to realize that such procedures all require further effort, sometimes considerable further effort, before the nature of the SNP can be determined, validated and then adopted in any one of the myriad SNP detection procedures.

Methods for Identifying DNA Fragments Containing SNPs

Although methods for identifying DNA fragments containing SNPs do not, without further work, identify the actual SNP, they are of considerable use in narrowing down the number of fragments subjected to more detailed analysis. Such procedures are therefore of use when one has to screen many hundreds or thousands of genotypes for rare alleles or when one is required to screen thousands of fragments across a small number of genotypes.

Strand-length polymorphisms

The first generation of DNA-based plant geneticists identified differences in defined genomic fragments through the use of restriction fragment length polymorphisms (RFLPs; Tanksley *et al*., 1989); however, since then more efficient PCR-based technologies including amplified fragment length polymorphisms (AFLPs; Vos *et al*., 1995) or cleavage amplification polymorphisms (CAPs; Konieczny and Ausubel, 1993) are more likely to be used to identify DNA fragments showing linkage to a trait of interest. However, in all cases the detection of a fragment length polymorphism or the absence/presence of a particular band on an agarose or polyacrylamide gel is taken to indicate the presence of a sequence variation, which might be SNP-based. In many cases there is no need to identify the underlying sequence variation; it is simply a case of using the assay as a means of determining which form exists within a particular individual. However, to convert the RFLP, AFLP or CAP to a polymorphism which can be used in a standard SNP detection procedure further work will be required as described later on in this chapter.

Strand conformational polymorphism

Single-strand conformational polymorphism (SSCP) is one of a number of techniques capable of detecting differences in the conformation of small (100–700 bases) single- or double-stranded DNA molecules (Orita *et al*., 1989). Other related techniques include heteroduplex analysis (Hauser *et al*., 1998)

and temperature/denaturing gradient gel electrophoresis (Myers *et al*., 1988). All of these techniques require the production of sequence-specific PCR primers capable of amplifying the appropriate fragments, which are then subjected to electrophoresis through a polyacrylamide type matrix; however, the requirement for specific PCR primers, often labelled with fluorescent dyes, does limit the procedure to at most a few hundred genes. In the case of SSCP, electrophoresis enables the operator to probe the shape of single-stranded DNA molecules as they migrate through the matrix. Two or more DNA molecules of the same size, but with slightly different sequences, will migrate through the matrix at different speeds, even if they differ by as little as a single nucleotide (Michaud *et al*., 1992; Sheffield *et al*., 1993).

A significant advantage of SSCP is that it can simultaneously distinguish differences between several fragments of the same or different sizes and hence can be used to distinguish related fragments (paralogues and homoeologues), which is especially useful when working with polyploids. Also, via the use of various fluorescent dyes, numerous reactions can be multiplexed (Larsen *et al*., 1999). The ability of SSCP to discriminate between related fragments means that it has recently become the method of choice to detect fragments containing SNPs in various crop species including allohexaploid wheat. To increase the possibility of finding both SNPs and sequence length variants in polyploids it has now become common practice to use primers derived from ESTs but which flank introns. However, the precise identification of introns when only EST sequences are known is time-consuming and requires that the sequence from a related and sequenced species, such as rice in the case of wheat, is available (Bertin *et al*., 2005). It should be noted that SSCP does not locate either the exact or even the approximate location of any SNP within the fragment; this can only be achieved by follow-on sequencing of the amplified fragments.

Targeting induced local lesions in genomics

Targeting induced local lesions in genomics (TILLING) is a procedure designed to detect DNA polymorphisms using the mismatch-specific endonuclease CEL1. The procedure has been extensively reviewed by Gilchrist and Haughn (2005) and will be described in detail in Chapter 4 (this volume); hence, there is no need to describe it in detail here. The procedure has gained much support from researchers looking for either natural variation in target genes (via Ecotilling) or induced mutations in reverse genetics-based procedures (Slade *et al*., 2004). The overall technology is not a simple procedure to master and it is therefore recommended that inexperienced researchers either seek the support of experienced laboratories or use the SSCP-based procedures described above. A significant advantage of TILLING over SSCP is that it is capable of scanning large (up to 1.5 kb) fragments for polymorphisms, and following cleavage the resulting fragment sizes do indicate the approximate region of the polymorphism. However, as for the conformational assays TILLING is incapable of defining the precise nature of the SNP and again this must be determined by further fragment manipulation and sequence analysis.

Microarray-based SNP discovery

The work of Winzeler *et al*. (1998) highlighted the potential of using oligonucleotide-based microarrays for SNP detection and, with further work, SNP discovery. Their work showed that sequence polymorphisms in genomic DNA could be detected via hybridization to species-specific oligonucleotide microarrays such as the 25-mer oligonucleotide-based Affymetrix GeneChip. In this procedure the various microarray features are each designed to sequences from a specific allele. It is not necessary, nor is it usually possible, to fit all of the features on a single microarray from a single genotype; what is important is that the allele specificity of each of the features is known. As described in Fig. 1.3, genomic DNA prepared from two or more genotypes is sheared, labelled and hybridized to the array under conditions that promote specific hybridization. If a particular sequence from one of the genotypes

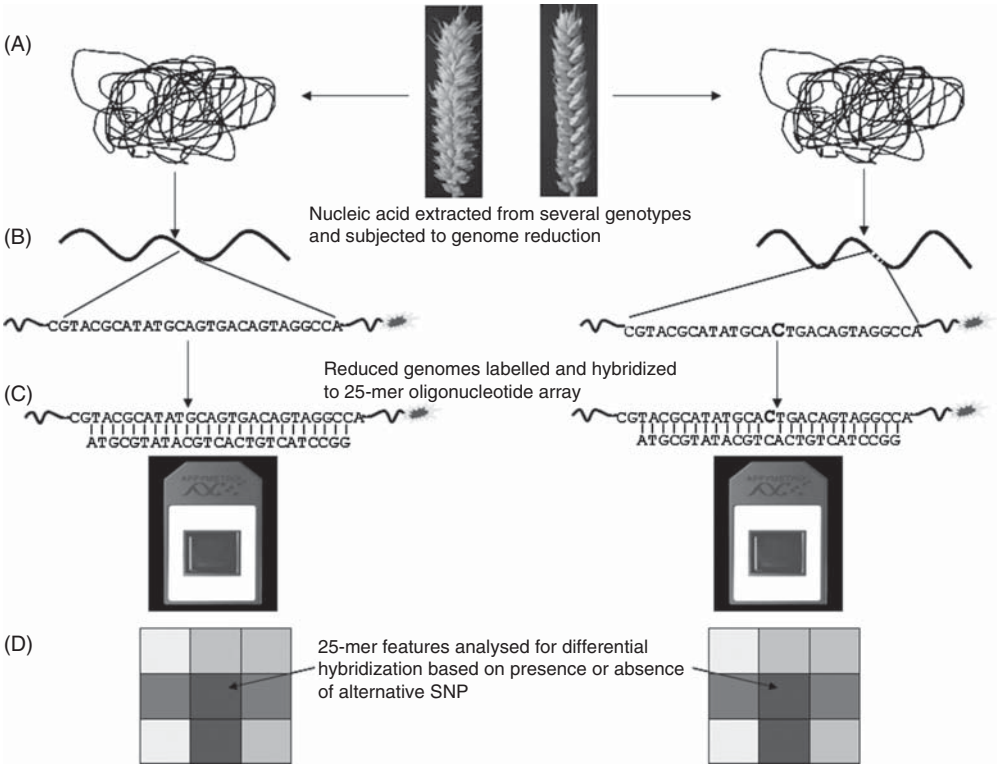


Fig. 1.3. Procedures for genome reduction and the detection of single-feature polymorphisms in plants. In the example given here genomics DNA is first extracted from the plant genotypes to be compared (A) and subjected to genome reduction and fluorescent labelling (B). The DNA fragments are then hybridized to a suitable array format (here we have used the Affymetrix GeneChip) (C). Following washing and scanning any features showing differential hybridization are identified and analysed further (D).

shows homology to the sequence on a feature there will be hybridization, whereas if there are SNPs within the region of similarity no hybridization will take place. To denote the fact that the polymorphic feature is based upon differential hybridization of an as yet unknown SNP (or indel) the term single feature polymorphism (SFP) has been coined (Borevitz *et al*., 2003). Hence, for each SFP the researcher is able to decide if the DNA from a particular genotype is either homologous or non-homologous to the sequence on the microarray. Unfortunately, in this type of experimentation not all mismatches will have an equal effect on reducing the efficiency of hybridization; as described by Gresham *et al*. (2006) mismatches at either end of a 25-mer feature are likely to have less effect in reducing the extent of hybridization than mismatches in the centre of the feature. Therefore, while the procedure is theoretically capable of detecting sequence variants at any one of the 25 positions, it is probable that only variants within the central ten bases can be detected with any certainty. However, even with these limitations the procedure offers considerable advantages; for instance, by using the standard Affymetrix GeneChip configuration it is possible to screen up to 11 × 65,000 25-mer sequences or 17.8 million bases for polymorphisms in a single experiment. Even if one is only able to detect polymorphisms in the core ten bases of the feature this still means that 7.1 million bases can be screened in each experiment.

As originally described by Winzeler *et al*. (1998) the procedure does have a single big disadvantage, which without solution means that it is unsuitable for most plant species including the major crops. The work conducted by these authors utilized *Saccharomyces cerevisiae*, an organism which has a small genome and for which a complete genome sequence exists. A small genome means that each individual fragment used to probe the oligonucleotide array can be labelled with high efficiency and the presence of relatively low numbers of fragments ensures clean and strong hybridization signals. While the procedure has been used with some success in *Arabidopsis*, it cannot be used with large genome species such as maize or wheat (Borevitz *et al*., 2003). In genomes such as maize and wheat the presence of large numbers of competing fragments, such as paralogues and homoeologues, results in poor hybridization signals and little or no discrimination between similar sequences. Recently problems with complex genomes have been partially overcome through the use of 'genome reduction' techniques such that once again microarray-based procedures offer significant potential for the large-scale discovery of SFPs (Fig. 1.3).

As the term implies genome reduction is the process by which the complexity of the genome is reduced by further processing. Here reduction results in fewer fragments remaining in the DNA mix and therefore being available for hybridization. The nature of the remaining fragments is dependent upon the reduction procedures employed.

Genome reduction via restriction digestion

It is well known that restriction sites for many restriction enzymes are not uniformly distributed throughout the genome. This is certainly the case for those restriction enzymes that are sensitive to methylation, for instance *Pst*1,

*Mlu*1 and *Hpa*II. Digestion of plant genomic DNA with any or a combination of these restriction enzymes often results in two categories of fragments, those with low (<1 kb) and those with high (>1 kb) molecular weights. Low molecular weight fragments are often enriched for genic sequences and therefore by harvesting the low molecular weight fragments, via gel exclusion and using these as substrates for hybridization to the microarrays, it is possible to obtain meaningful signals capable of discriminating between SFP types. Further refinements utilizing PCR to amplify subsections of the genome using adaptors and primers similar to those used in AFLP-based studies have also been recently employed (Syvänen, 2005).

Interestingly, extreme genome reduction technologies similar to those described above have recently been used to directly identify SNPs through very high throughput sequencing protocols, thus missing out the microarray step (Altshuler *et al*., 2000). The direct identification of SNPs via sequence analysis of highly reduced genomes relies on the ability to process large numbers of clones and sequences.

Genome reduction via the transcriptome

Several reports have suggested that transcriptome analysis in conjunction with oligonucleotide arrays might be a suitable procedure for the detection of SFPs. As for genomic DNA-based hybridization, this procedure relies on the differential hybridization of fluorescently labelled cDNA fragments to 25-mer oligonucleotide features on a microarray. A significant advantage of using cDNA for such hybridizations is that one is able to specifically target genes. Through the use of material prepared from tissues at different developmental stages or following specific treatments, it is also possible to screen those genes involved in specific biological processes. However, there are also considerable problems in using material derived from the transcriptome for the detection of polymorphic sequences through differential hybridization. A signal difference between two genotypes might be for reasons other than simple sequence variation, such as differential expression of transcripts as a result of the plants being grown under slightly different conditions, for example when they are placed in different regions of the same growth chamber or due to differences in a controlling transcription factor rather that the mRNA sequences themselves. Many of these problems can be overcome with the use of suitable controls and a significant number of replicates (Cui *et al*., 2006). Somers *et al*. (2003) have recently reported the use of the wheat Affymetrix GeneChip to define 1455 probe sets which differ in their hybridization intensity and which define 542 quantitative trait loci (QTL) in a single mapping population (Jordan *et al*., 2007).

Summary of the Methods Used for Identifying DNA Fragments Containing SNPs

The above procedures, and many other variants, are designed to identify the regions of the genome which might contain SNPs and which therefore could

become the focus of further characterization. Some of these procedures are described in more detail in later chapters; however, the fact remains that while the various procedures can be used in low to medium throughput procedures to genotype lines via SNP detection, without further work the nature of the SNP remains unknown. Without knowing the precise characteristics, it remains difficult to adapt the SNP for use in any of the very high throughput procedures, many of which are again described in later chapters. Hence, for the remaining part of this chapter we will spend some time discussing the various processes required to identify the exact nature of the SNPs identified.

True SNP Discovery

Identifying fragments derived from RFLPs, AFLPs, CAPs, SSCPs or features on a microarray that are likely to contain SNPs, while being important, does not in itself 'discover' the underlying nature of the SNP. In some cases identifying the actual SNP responsible for the observed fragment polymorphism is not required as the fragment polymorphism can be used directly as the tool of SNP detection. While at first glance it may seem straightforward to discover a SNP when the fragment containing it is known, it is not a procedure that lends itself to high throughput SNP discovery. So far such SNP conversions have been limited to cases where the underlying SNP is important, for instance due to it resulting in, or being linked to, phenotypic change. Because the procedures required are not straightforward, because they are often encountered in plants and because they often throw up polyploid-linked complications, it is worth examining the process in some detail.

Isolation of SNPs from fragments showing migration polymorphism

In the case of RFLPs, CAPs, SSCP, TILLING and microarrays the specific probe containing the (internal) SNP is of course already available and further material can easily be amplified and directly sequenced via the use of the PCR primers. However, due to the nature of AFLPs and related techniques, the SNP underlying the migration polymorphism (or presence/absence status) will almost certainly lie at the end of the fragment and therefore be inaccessible without further work. In this case the AFLP band of interest will need to be carefully excised from an agarose or polyacrlyamide gel using one of the myriad techniques described in the literature and then subjected to Vectorette-PCR or inverse-PCR to isolate and sequence the flanking regions. Once this is done these flanking regions can then be sequenced and the sequence used to design PCR primers flanking the original, now internal polymorphism. At this stage AFLP-derived probes become equivalent to probes derived from any of the other technologies described above.

In the cases described above, direct sequencing of the fragments containing the polymorphism from several different genotypes, followed by sequence alignment using one of the many alignment programs such as SEQUENCHER or

DNAMAN, should identify any individual SNP(s) and the likely haplotypes. However, it is at this stage that most plant researchers begin to encounter problems; for many plant species this procedure results in poor-quality sequences with many ambiguous bases. Usually poor-quality or difficult-to-read sequences are obtained from PCR-generated templates due to the presence of multiple copies of related but non-homologous sequences (paralogous and homoeologous sequences), often because the species is either a polyploid or, in the case of AFLP-derived fragments, the sequences are members of a repetitive element family. In some cases size heterogeneity of the PCR products can be observed on agarose or polyacrylamide gels, but this happens only when the non-homologous sequences are of different sizes, possibly due to the primers flanking an intron which show size heterogeneity within the gene family. In either case it then becomes necessary to first clone the mixed PCR fragments (from a single genotype) and sequence multiple clones. Figure 1.4 suggests the various steps and decision points often encountered during the conversion of such fragments to SNPs in plants.

Following the cloning, sequencing and sequence alignment steps it should then be possible to design 'family'-specific primers which can be used to amplify homoeologous/paralogous-specific fragments from a single genotype. Once specific amplification using these primers has been confirmed in a

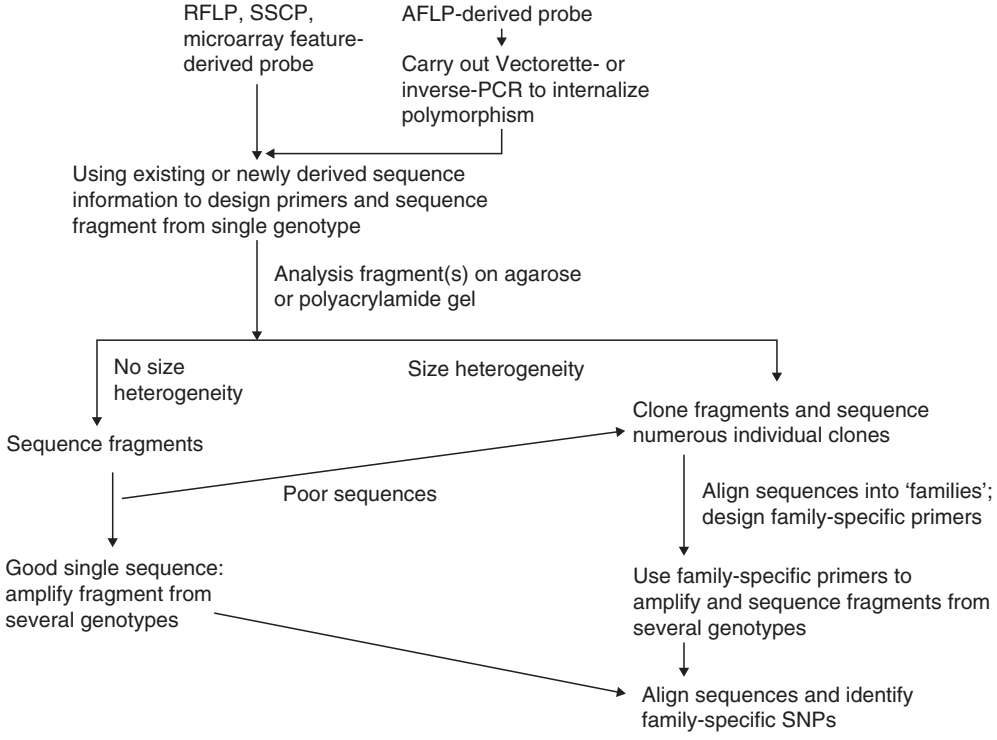


Fig. 1.4. Flow diagram of the decision points encountered when converting fragments to SNPs in plants.

single genotype, they can then be used to generate family-specific fragments from multiple genotypes, which in turn can be sequenced and aligned to identify homoeologous/paralogous-specific SNPs and subsequently inter-genotype SNPs (Fig. 1.4). Figure 1.5 shows the sequence of three different but related sequences derived from a single wheat genotype Chinese Spring.

In the case described in Fig. 1.5, a gene thought to be present as a single copy in wheat was amplified from a single variety (Chinese Spring). The resulting PCR products appeared as a single band following agarose gel electrophoresis but generated relatively poor sequence when direct sequencing of the PCR product was carried out. Following this the PCR products were cloned and sequenced as numerous individual clones, a step which indicated that the single PCR reaction contained three related but distinct sequences. As the PCR was carried out using a single genotype it was assumed that the various sequences represent individual homoeologues or paralogues. As wheat is known to be an allohexaploid it was logical to assume that the three sequences represent the three homoeologues, although this could not be confirmed without homoeologue-specific mapping. In this particular case it is interesting to note that the three SNPs are present within a region of 60 bp, a figure close to that of one in 24 nucleotides reported by Somers *et al*. (2003) for homoeologous SNPs in wheat. This example clearly shows that what generally might be

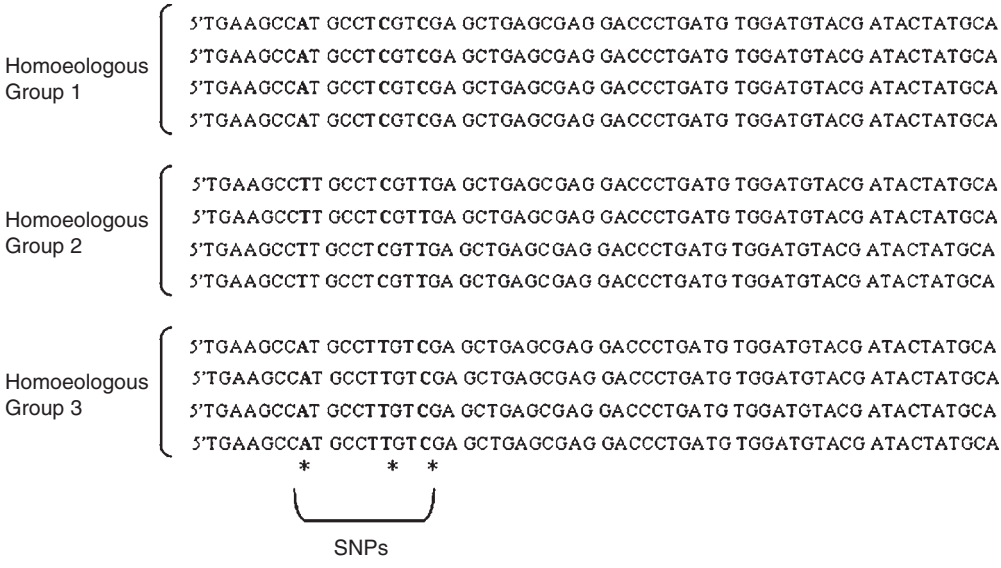


Fig. 1.5. Sequence analysis of the three wheat PCR products derived from a single set of PCR primers. In this example three different sequences were obtained through amplification of the Chinese Spring genotype with a single pair of PCR primers. The three SNPs which differentiate the three 'sequence types' are highlighted by *. In this particular case the sequence variations seen between the homoeologues can be used, in conjunction with other SNPs (not shown), to generate homoeologous-specific PCR primers.

considered to be a relatively simple procedure to discover SNPs within individual fragments is complicated by the presence of multiple copies of related but different genes in both ancient and current polyploid genomes.

This approach described what steps need putting in place if no prior sequence information is known and where the polymorphism is only identified by differential migration during gel electrophoresis. More commonly the researcher already has the primary sequence of the gene and hence it might be assumed that SNP discovery would be straightforward. However, even here this is not necessarily the case; possession of the primary sequence does enable the researcher to bypass the original sequencing step, but if the work involves a polyploid species, the researcher will still have to go through the various steps described in Fig. 1.5 to generate homoeologous/paralogous-specific primers before screening a number of genotypes for varietal SNPs. Such a complicated approach would suggest that even here the process is not amenable to scaling up beyond a few tens of genes. However, through NSF funding, a consortium led by Dvorak recently completed a project to discover SNPs in both allohexaploid wheat and its progenitors using this approach. In this particular case the group used PCR primers to amplify, clone and sequence regions from both the wheat progenitors and several allohexaploid wheat varieties. Following alignment of the sequences obtained from the progenitors and allohexaploid genotypes, it was possible to either identify varietal SNPs or to design homoeologous-specific primers for further rounds of amplification and sequence alignment (http://wheat.pw.usda.gov/SNP/new/project.shtml). While such an approach is admirable, it did require considerable resources and several years to complete; however, it resulted in several thousand SNPs being made available to the wheat community.

In silico SNP discovery

The term '*in silico* SNP discovery' was first used by Picoult-Newberg *et al*. (1999) to describe the process of using sequence information in the public domain to identify SNPs. With some foresight Picoult-Newberg suggested that 'the use of existing resources can rapidly and efficiently lead to high-volume, cost-effective SNP discovery', a statement that has proved remarkably accurate for a number of species, including several plant species. The ability of researchers to utilize publicly available sequences, usually ESTs, to identify SNPs now means that even researchers working on species that have little economic value can generate sufficient numbers of SNPs for simple diversity studies. However, with larger community efforts the possibilities of generating large number of SNPs and SNP-based maps becomes a real possibility. In this part of the chapter we will consider the two possible scenarios of either small or large numbers of sequences being available.

Approaches to in silico *SNP discovery in a sequence-poor environment*

Although the public sequence databases continue to expand at an ever-increasing rate, it is still unfortunately true that for the majority of plant

species the number of sequences is still the limiting factor when it comes to *in silico* SNP discovery. As can be seen from Table 1.2 the number of sequences for each of the designated plant species varies considerably from 6232 for oil palm to 3,480,660 for maize. In addition, within each of the individual sequence type categories there is also considerable variation; for instance, cottonwood has large numbers of core nucleotide sequences (mostly consisting of incomplete genomic sequences) but virtually no ESTs or genome survey sequences (GSSs; consisting of random single pass read genome survey sequences, cosmid/BAC/YAC end sequences, exon-trapped genomic sequences, Alu PCR sequences and transposon-tagged sequences), whereas sugarcane has relatively large numbers of ESTs but few genomic sequences.

In addition to the total number of sequences available within any one species it is also important to consider the number of different genotypes from which the sequences have been collected, because efficient SNP discovery requires large numbers of sequences from several different and diverse genotypes. This is not always the case; for instance, when one examines the situation in cottonwood it can be seen that the majority of sequences are derived from one genotype and hence they are of limited value for *in silico* SNP discovery. However, in such situations it is still possible to utilize the public sequences for SNP discovery in conjunction with amplification and sequencing of further genotypes. For instance, Labate and Baldo recently reported the discovery of 62 SNPs by the alignment of existing tomato ESTs

Table 1.2. Number of core nucleotide, ESTs and GSS records in NCBI on 2 July 2007.

Species	Core nucleotides	ESTs	GSS	Total
Arabidopsis (all)	262,706	1,463,468	1,077,734	2,803,908
Barley	17,813	467,715	2,063	487,591
Brassica (all)	9,587	362,160	807,823	1,179,570
Cottonwood	91,585	9	0	91,594
Fescue	150	44,379	5,320	49,849
Lolium	1,744	12,398	83	14,225
Maize	143,306	1,272,632	2,064,722	3,480,660
Oat	979	7,959	37	8,975
Oil palm	540	5,616	76	6,232
Pea	4,463	87,189	158	91,810
Pepper	7,908	29,411	52	37,371
Pine	13,993	498,706	1,797	514,496
Potato	9,102	292,094	141,484	442,680
Rice	388,576	1,270,898	298,354	1,957,828
Soybean	12,468	438,013	368,350	818,831
Sugarcane	2,322	255,965	3	258,290
Sunflower	5,036	404,170	1	409,207
Tomato	16,629	349,591	320,480	686,700
Wheat	12,765	1,086,302	45,560	1,144,627

in conjunction with further sequencing of an extra three genotypes (Labate and Baldo, 2005). As described by Labate and Baldo, their original sequences of interest were used to screen the public EST databases for orthologues from related species. All the related sequences were then aligned with a consensus sequence, including any ambiguities. Following the guidelines set out in Fig. 1.5, PCR primers were designed within conserved regions that would amplify across predicted variable regions and therefore regions likely to yield SNPs between the more closely related genotypes. Proof of principle tests were carried out on a few genotypes and successful primer combinations used to extend the study to amplify and sequence the polymorphic fragments from a range of diverse genotypes. The advantage of this approach is that by using the available sequence data in combination with simple alignment programmes and further sequencing of diverse genotypes it is possible to generate SNPs for specific genes within a relatively short period of time. However, it is also clear that such an approach, without considerable resources, is unlikely to generate more than a few hundred SNPs, and hence the only really efficient method for large-scale SNP discovery remains the generation of large numbers of sequences from several diverse genotypes.

Approaches to in silico *SNP discovery in a sequence-rich environment*

In those plant species such as *Arabidopsis*, barley, the brassicas, maize, rice and wheat, for which large numbers of sequences have been generated, usually through consortium efforts, there exists the possibility of carrying out large-scale SNP discovery projects using largely automated procedures. Given sufficient computing power and some expertise in computer programming to be able to link the various steps it is certainly possible for most organizations to develop at least a semi-automated procedure which utilizes the publicly available sequences. In addition to these in-house efforts the SNP community has already made available several protocols and program suites, some of which are web-based for *in silico* SNP discovery. Many of these share aspects in common and hence as an example we will describe the features of one such software suite called 'AutoSNP' which was developed in our laboratory and designed to process cereal ESTs and SNPs (Barker *et al*., 2003; http://www.cerealsdb.uk.net/discover.htm). The overall schema employed in AutoSNP is described in Fig. 1.6.

In AutoSNP the available ESTs for the species of interest are first harvested en masse from the public databases in Genbank format and automatically converted to a 'rich' FASTA format, which conserves the cultivar, tissue and development stage annotations in the FASTA header. These FASTA sequence files are then crossmatched against the UNIVEC database to remove vector contamination as well as low-complexity sequences, including polyN tracts. The cleaned-up sequences are then clustered using the TIGR Gene Indices Clustering tool (TGICL) and the resulting precontigs aligned using the program cap3. Putative SNPs are then identified in the various contigs using redundancy (for the program to assign a SNP, each SNP allele must be present in at least two independent sequences). The program then accesses the extent of co-segregation between multiple SNPs in the sequence alignment using permutation tests to find the putative haplotypes and the overall

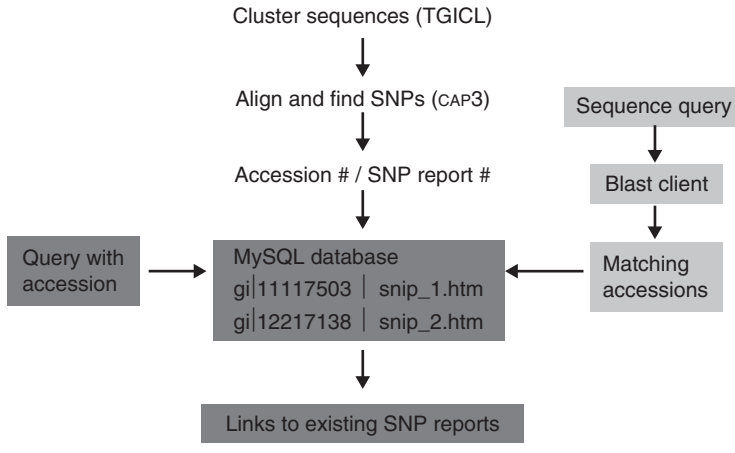


Fig. 1.6. Schema for AutoSNP procedure.

results displayed as an HTML file. Public access to AutoSNP (and the other similar web-based facilities such as SNPSERVER and Quality SNP; Savage *et al*., 2005; Tang *et al*., 2006) is via either Accession number searches or via a sequence query. A typical output from AutoSNP is shown in Fig. 1.7.

There are many aspects of the various *in silico* SNP discovery programs such as AutoSNP which require further description and discussion; for instance, all the various programs use public sequences generated by several, sometimes hundreds of, researchers who employ different (sometimes no) levels of quality control. Hence, the quality of the sequences going into AutoSNP will be variable. A small number of protocols try to overcome this problem by insisting that all sequences meet a minimum quality threshold and that the individual researchers provide the original sequence trace file so that discrepancies can be investigated and resolved (http://wheat.pw.usda.gov/genome/). However, while this is possible with sequences derived locally or from a small number of sources and for which chromatograms are available, it is not usually possible with multi-sourced data such as that contained in the public databases. Such a scenario can lead to several problems; first, poor sequences or long runs of single nucleotides such as polyA tracts derived from mRNA might cause alignment problems during contig formation and, secondly, sequencing errors might suggest the presence of a SNP when none is present (Fig. 1.8).

AutoSNP and similar programs overcome these sequence quality-related problems in several ways; first, they remove simple sequences such as polyA tracts which are either remnants of the mRNA poly A or a result of poor sequencing. Secondly, they only consider polymorphisms which are represented by two or more independent sequences. Unfortunately this does mean that rare alleles could be discarded, especially if the number of sequences in the database from that particular genotype is limited or a particular sequence is represented in the databases only a few times. However, this approach

<previous SNP report 2 next>

cluster_6_contig_2

3 putative SNPs, mean score 2.00

Key to sequences:

A gi | 13115430 | gb | BG313627 . 1 | BG313627
B gi | 19956408 | gb | BJ225577 . 1 | BJ225577
C gi | 20096630 | gb | BJ271067 . 1 | BJ271067
D gi | 9838548 | gb | BE585605 . 1 | BE585605

Summary of SNPs:

Base	ABCD	Min. informative	Co-segregation	Weighted co-segregation (%)
320	AGAG	2	3/3	100
488	GAGA	2	3/3	100
506	AGAG	2	3/3	100

Fig. 1.7. An example SNP report from the wheat data set. This report shows that there are three putative SNPs in the current alignment of four sequences, with a mean informative score of 2. The sequence key assigns a single letter to each sequence accession. The key is followed by a summary table showing the position in the alignment of each SNP, together with the nucleotide present at this position in sequences A–D. For example, in this report, sequence A (gi|13115430) has a guanine (G) at position 488, whereas sequence B has an adenine (A). There are currently three SNP scores to help users decide if a SNP is real. The minimum informative score is simply the minimum number of sequences representing the least common SNP at the current position. In this SNP report, all three SNPs get a co-segregation score of 3/3, because all SNPs agree that sequences A and C represent one haplotype, and sequences B and D another. The weighted co-segregation is 100% because there are no missing data for any of the sequences across the three SNPs.

Fig. 1.8. True SNP or sequencing error? AutoSNP and similar programs require that any sequence variation is present in at least two independent sequences. In this example the G to A transition is contained within two separate sequences and is therefore scored by AutoSNP as a true SNP; however, in the putative C to T transition only one sequence contains the 'T' form and therefore AutoSNP disregards this SNP and will not report it.

does permit the rapid and accurate identification of large numbers of candidate SNPs with a high level of confidence in their validity.

Early on in the use of automated SNP generation programs such as AUTOSNP it became apparent that both the number of sequences and the diversity of genotypes used to generate those sequences were critical to the success of the procedure; for instance, as shown in Fig. 1.9A the number of SNPs clearly increases exponentially as the number of sequences and the contig size increase. However, the relationship between *in silico* SNP discovery and numbers of sequences and genotypes is not a simple direct relationship. For instance, although simulation studies show that low numbers of sequences lead to the elimination of many SNPs due to their single representation in the databases, they also show that the rate of 'false positive SNP' discovery increases, eventually to an unacceptable level, with

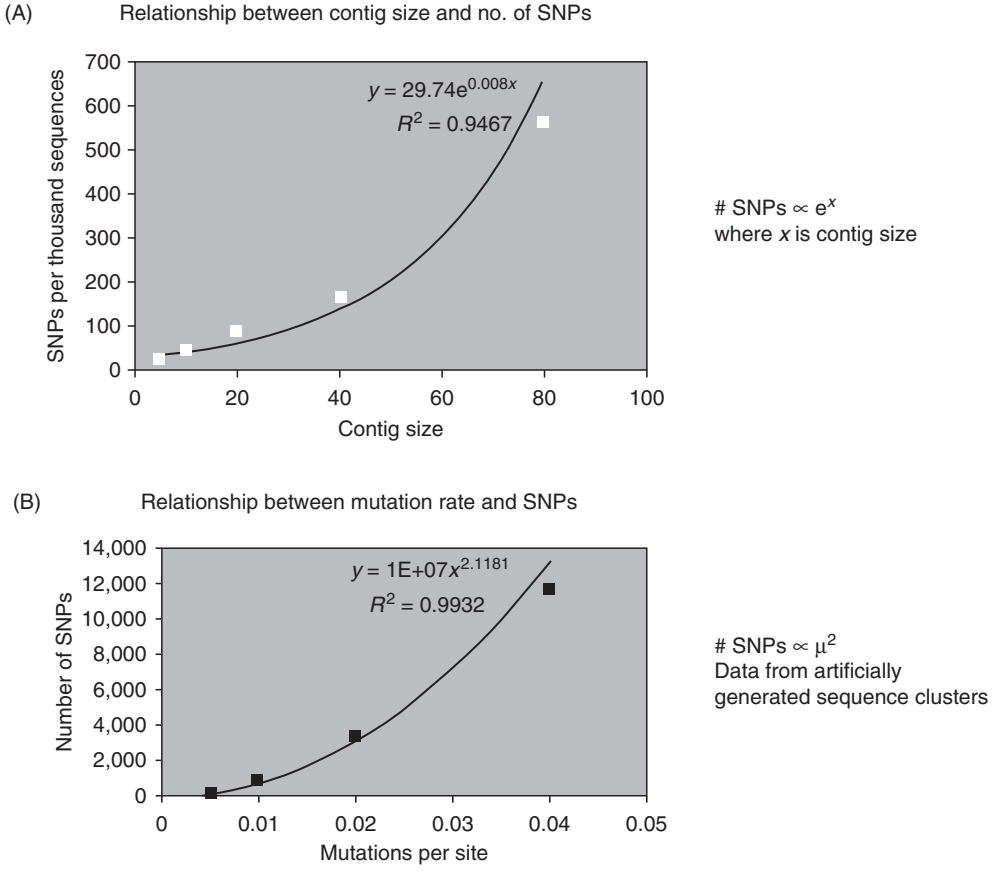


Fig. 1.9. (A) Relationship between contig size and the number of SNPs and (B) exponential growth of false positive SNPs in relation to increasing contig size. In this example the number of SNPs 'discovered' is a result of introducing a specific number of artificial mutations within the members of the contig.

increasing numbers of sequences (Fig. 1.9B). This is because the more times a DNA fragment is sequenced, the likelihood of duplicate sequencing errors also increases, especially in the case of specific sequence motifs that are known to interfere with the sequencing process, such as GC-induced compression.

To address this problem there is a need to use statistical analysis of SNP co-segregation to identify such sequencing errors. The approach used in AutoSNP to solve this problem is based on the assumption that where there are multiple candidate SNPs within a contig it can be assumed that they will be in linkage equilibrium and therefore form specific haplotypes. Hence, where SNPs are 'real' any software should be able to predict the sequence of a polymorphic site using the knowledge of its neighbouring SNPs, whereas a sequencing error-induced SNP would have no such relationship with those flanking it. To carry out this analysis it is necessary to measure the degree of co-segregation between all possible pairwise combinations of SNPs within a contig and to provide statistical support by comparing this measurement with that obtained after randomizing the SNP data from the same contig. Those SNPs with co-segregation scores at the extreme of the randomized data set can then be deemed significant ($p < 0.05$), highly significant ($p < 0.01$) or very highly significant ($p < 0.001$). The results of including such analysis within AutoSNP can be seen in Fig. 1.10.

Using such protocols as that described above it becomes possible to rapidly screen large sequence collections to harvest the available SNPs and to have a degree of confidence that the SNPs identified will be validated. To confirm that this is indeed the case we undertook a rapid analysis of the SNP content within the public sequences for the four major cereals (Table 1.3).

Table 1.3. SNP abundance in the four major cereal taxa.

	Wheat	Rice	Maize	Barley
Sequences[a]	845,616	1,168,062	735,925	383,546
Contigs of 4 or more	38,416	41,182	34,063	14,985
Contigs of 4 or more with ≥1 SNP	13,243	14,957	16,858	5,406
% contigs with ≥ SNP	34	36	49	36
Total SNPs found	122,658	67,625	137,004	24,252
Total co-segregating SNPs	71,245	12,524	70,749	9,728
% co-segregating/total	58	18	51	40
Mean cultivars per locus	6.1	3.9	4.7	5.6
Estimated genome coverage (Mb)	20.1	18.7	18.7	7.2
Total SNPs per kb	6.1	3.6	7.3	3.4
Co-segregating SNPs per kb	3.5	0.7	3.8	1.3

[a]Sequences available after filtering for low complexity and vector contamination. Results as of January 2007.

Base	ABCDEFGHIJKLM	Min. information	Co-segregation	Weighted co-segregation (%)
210	..GGGGGGGC.C.	2	8 out of 8	69
211	..AAAAAAAGGGG	4	8 out of 8	85
213	..CCCCCCCGGGG	4	8 out of 8	85
216	..AAAAAAAGGGG	4	8 out of 8	85
307	GGCCCCCCCGGGG	6	8 out of 8	100
308	AATTTTTTTAAAA	6	8 out of 8	100
454	TTCCCCCCCTTTT	6	8 out of 8	100
505	..GGGGGGGCC.C	3	8 out of 8	77

Haplotype info

Haplotype	....GAT.	has 2 sequences AB
Haplotype	.GGGGATC	has 2 sequences KM
Haplotype	CGGGGAT.	has 1 sequence L
Haplotype	CGGGGATC	has 1 sequence J
Haplotype	GACACTCG	has 7 sequence CDEFGHI

Haplotype distances

	AB	CDEFGHI	J	KM	L
AB	–	1.00	0.00	0.00	0.00
CDEFGHI	1.00	–	1.00	1.00	1.00
J	0.00	1.00	–	0.00	0.00
KM	0.00	1.00	0.00	–	0.143
L	0.00	1.00	0.00	0.143	–

((((AB J)KM)L)CDEFGHI)

Fig. 1.10. Results of using AUTOSNP to screen for SNPs and haplotypes in a contig consisting of 13 wheat sequences (A–M). In this example, eight putative SNPs (at positions 210 to 505) are contained within five haplotypes. Note that due to incomplete overlap between the various sequences in the contig, the weighted co-segregation of the SNPs ranges from 69% to 100%.

Polyploids and *In Silico* SNP Discovery

At the beginning of this chapter we highlighted the fact that the majority of plants were polyploids and that the presence of homoeologue and paralogue sequences complicated SNP discovery. This is just as true for *in silico* SNP discovery as it is for experimental SNP discovery; hence, most *in silico* SNP discovery protocols simply report the SNP as a SNP and make no distinction between homologous, homoeologous and paralogous SNPs. For

instance, when AutoSNP is used to screen for SNPs in both the small and large subunits of the ADPG-P Pyrophosphorylase genes, in both cases it identified three potential haplotypes. However, closer examination of the varietal distribution of these haplotypes strongly suggests that in several genotypes all three haplotypes can be found, and therefore the three haplotypes are almost certainly homoeologous copies and the SNPs are homoeologous-specific SNPs rather than genotype-specific SNPs. In fact, results to date indicate that the majority of *in silico* discovered SNPs from polyploids, such as allohexaploid wheat, are in fact homoeologous/paralogous SNPs rather than genotype-specific SNPs. Disappointingly, recent work by our group suggests that only one SNP in every 30 is a genotype-specific SNP. Such a scenario means that genotype-specific SNPs are extremely valuable but they are also hard to find and validate. The QualitySNP software developed by Tang *et al*. (2006) does suggest that automated mechanisms for discriminating between homologous and homoeologous SNPs might be possible and work by our group also suggests that such a development is possible by incorporating the procedures described in Fig. 1.11.

In the approach described in Fig. 1.11, as in the standard AutoSNP protocol, a sequence contig is generated using all the available similar sequences. From these a consensus sequence is generated, which includes all the ambiguities present in the starting sequences. This consensus sequence is then used in a sequence similarity search against a database of sequences derived from a single variety for which a large number of sequences are present; for instance, in wheat we used the variety Chinese Spring. Using PhyLIP and AutoSNP the resulting contigs are subjected to subgroup clustering to identify the homoeologues and paralogues (within the single variety used) and the homoeologous and paralogous SNPs noted. The original (consisting of all varieties) sequence contig is then used again to search for highly (>97%) similar sequences using the BLASTN algorithm. Using the BLASTN results, sequences can then be assigned to their 'best match' between the various (in Fig. 1.11 there are three) potential homoeologous/paralogous subgroups. The individual subgroups can then be used to re-screen the EST database for varietal haplotypes. In the example shown in Fig. 1.11, the varietal-specific SNPs are highlighted in the three boxes, whereas the homoeologous/paralogous SNPs are open. Although currently undergoing beta testing, we have begun to use this approach to identify genotype-specific SNPs and to date have been successful in identifying and validating several tens of genotype-specific SNPs in wheat; hence, such an approach holds much promise for the future discovery of genotype-specific SNPs in polyploids.

Future Prospects for SNP Discovery in Plants

The active discovery of SNPs in plants is less than 10 years old. In this time protocols have been developed that allow researchers to discover SNPs both experimentally and *in silico*. Further to this, more recent work has begun to allow researchers to discriminate between varietal SNPs, which are useful for diversity

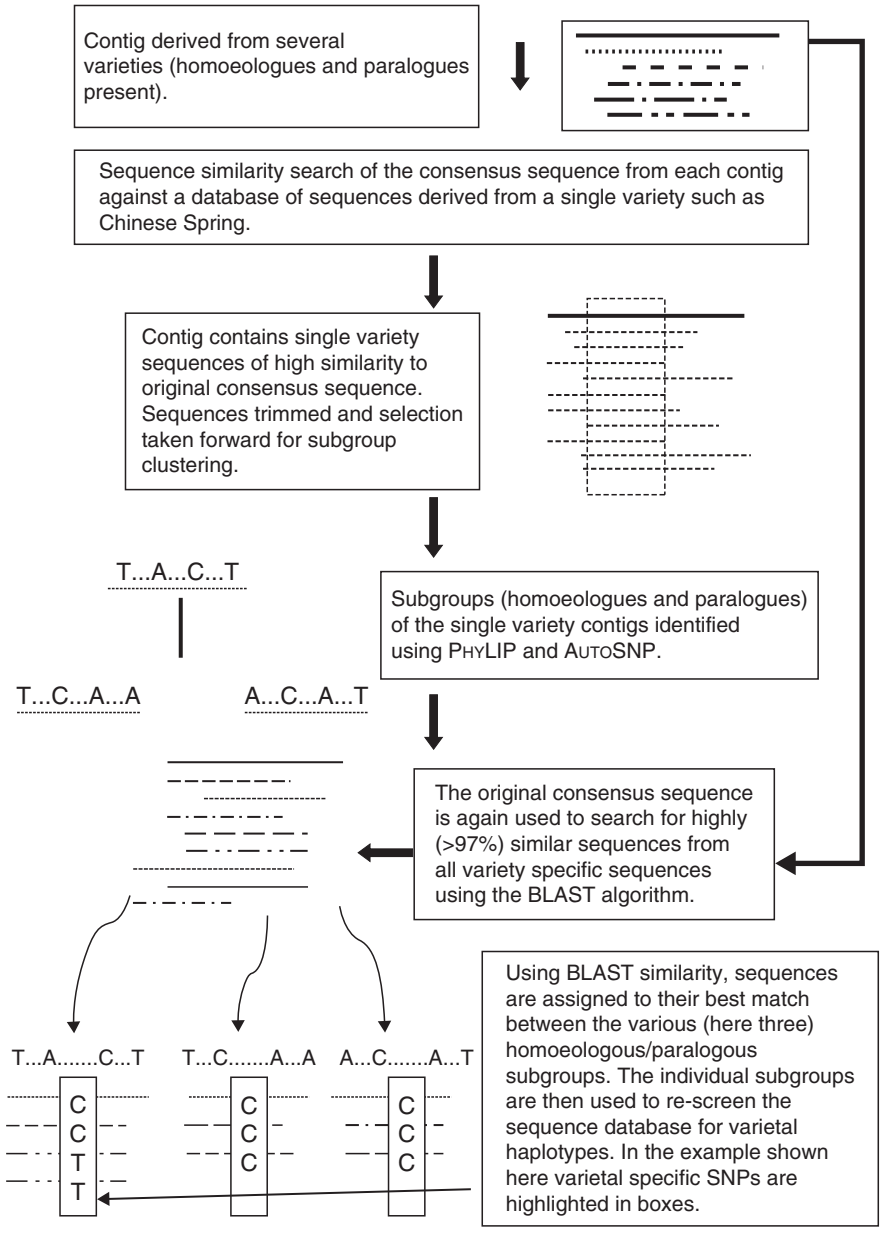


Fig. 1.11. *In silico* protocol for discriminating between homoeologous and genotype-specific SNPs in allohexaploid wheat. (Redrawn and adapted from O'Sullivan, 2006.)

and mapping studies, and homoeologous/paralogous SNPs, which are useful for investigating genome structure and the interaction between the various partners in the polyploid nucleus. In addition, studies are also enabling researchers to focus on non-synonymous SNPs, which are responsible for changes in the amino acid content of proteins and hence are potentially involved in

phenotypic changes. However, all of these past developments and certainly all future developments rely on accurate, well-annotated sequence information being available in the public domain. It is certainly true that individual groups, often based in well-resourced industrial laboratories, are capable of establishing significant SNP discovery programmes without reference to public information. However, these are in the minority and appear to be diminishing as time goes by. Hence, it is clear that as SNP discovery and sequencing technology go hand in hand, the more rapid the sequencing technology, the more rapid the pace of SNP discovery should be. Thus, with more recent technological developments such as Pyrosequencing (reviewed in Ronaghi, 2001), 454 sequencing (http://www.454.com/) and Solexa sequencing (http://www.illumina.com/), it is likely that the amount of sequence information will continue to increase exponentially, at least for the foreseeable future, and therefore so will the rate of SNP discovery. As the volume of sequence data increases rapidly, the computer power needed to handle the data also increases. In the case of *in silico* discovery it seems that the increase in desktop computing power is not keeping pace with these demands. As a result, we increasingly need to use massively parallel computer clusters to process sequence data in a timely fashion. As discussed, efficient SNP discovery is not simply about the amount of sequence information available, but also about sequence annotation and validity. In our studies, on both maize and wheat, it has become clear that incorrect annotation of sequence information, for instance mistakes arising from the purity of the seed stock used, is now becoming a major issue in assigning SNPs to particular genotypes. In part these problems are associated with the correct maintenance of germplasm and in part they are associated with sufficient time being spent in ensuring that the sequences entered into the public databases are clean of vector sequences and are annotated to the maximum extent.

Finally, most of the currently available SNPs have been derived from ESTs, and this is not surprising as for most plant species ESTs make up the bulk of the available sequences; however, this approach has its limitations as it means that only those sequences which are highly or moderately expressed are represented by the available SNPs, and sequences not expressed, such as promoters or sequences expressed at specific times or at low levels, are absent. With the advent of very high throughput sequencing technologies and their application to the sequencing of both the major crop species and, in time, all plant species, it is likely that more genome-wide SNPs will become available, at which time SNP discovery will become truly universal.

References

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L. and Lander, E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407, 513–516.

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.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register III, J.C., Tingey, S.V. and Rafalski, A. (2002) Insertion–deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Molecular Biology* 48, 539–547.

Bertin, I., Zhu, J.H. and Gale, M.D. (2005) SSCP-SNP in pearl millet – a new marker system for comparative genetics. *Theoretical Applied Genetics* 110, 1467–1472.

Boerma, H.R. and Walker, D.R. (2005) Discovery and utilization of QTLs for insect resistance in soybean. *Genetica* 123, 181–189.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research* 13, 5130–5523.

Brookes, A.J. (1999) The essence of SNPs. *Gene* 234, 177–186.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H. and Neale, D.B. (2004) Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proceedings of the National Academy of Sciences of the USA* 101, 15255–15260.

Buntjer, J.B., Sørensen, A.P. and Peleman, J.D. (2005) Haplotype diversity: the link between statistical and biological association. *Trends in Plant Sciences* 10, 466–471.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M. and Rafalski, A.J. (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics* 3, 19.

Cox, T.S., Murphy, J.P. and Rodgers, D.M. (1986) Changes in genetic diversity in the red winter wheat regions of the United States. *Proceedings of the National Academy of Sciences of the USA* 83, 5583–5586.

Cui, X., Affourtit, J., Shockley, K.R., Woo, Y. and Churchill, G.A. (2006) Inheritance patterns of transcript levels in F_1 hybrid mice. *Genetics* 174, 627–637.

Darwin, C.R. (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.

Feltus, F.A., Wan, J., Schulze, S.R., Estill, J.C., Jiang, N. and Paterson, A.H. (2004) An SNP resource for rice genetics and breeding based on subspecies indica and japonica genome alignments. *Genome Research* 14, 1812–1819.

Flint-Garcia, S.A., Thornsberry, J.M. and Buckler, E.S. (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54, 357–374.

Gilchrist, E.J. and Haughn, G.W. (2005) TILLING without a plough: a new method with applications for reverse genetics. *Current Opinion in Plant Biology* 8, 1–5.

Gilchrist, E.J., Haughn, G.W., Ying, C.C., Otto, S.P., Zhuang, J., Cheung, D., Hamburger, B., Aboutorabi, F., Kalynyak, T., Johnson, L., Bohlmann, J., Ellis, B.E., Douglas, C.J. and Cronk, Q.C.B. (2006) Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology* 15, 1367–1378.

Gresham, D., Ruderfer, D.M., Pratt, S.C., Schacherer, J., Dunham, M.J., Botstein, D. and Kruglyak, L. (2006) Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA Microarray. *Science* 311, 1932–1936.

Hauser, M.T., Adhami, F., Dorner, M., Fuchs, E. and Glossl, J. (1998) Generation of co-dominant PCR-based markers by duplex analysis on high resolution gels. *Plant Journal* 16, 117–125.

Henry, R. (ed.) (2001) *Plant Genotyping: The DNA Fingerprinting of Plants*. CAB International, Wallingford, UK.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M. and Last, R.L. (2002) *Arabidopsis* map-based cloning in the post-genome era. *Plant Physiology* 129, 440–450.

Jordan, M.C., Somers, D.J. and Banks, T.W. (2007) Identifying regions of the wheat genome controlling seed development by mapping expression quantitative trait loci. *Plant Biotechnology Journal* 5, 442–453.

Khlestkina, E. and Salina, E. (2006) SNP markers: methods of analysis, ways of development, and comparison on an example of common wheat. *Russian Journal of Genetics* 42, 585–594.

Konieczny, A. and Ausubel, F.M. (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-based markers. *Plant Journal* 4, 403–410.

Kunkel, T.A. and Erie, D.A. (2005) DNA mismatch repair. *Annual Review of Biochemistry* 74, 681–710.

Labate, J. and Baldo, A. (2005) Targeted discovery of highly polymorphic genes in tomato cultivars. *Molecular Breeding* 16, 343–349.

Larsen, L.A., Christiansen, M., Vuust, J. and Andersen, P.S. (1999) High-throughput single-strand conformation polymorphism analysis by automated capillary electrophoresis: robust multiplex analysis and pattern-based identification of Allelic variants. *Human Mutation* 13, 318–327.

Michaud, J., Brody, L.C., Steel, G., Fontaine, G., Martin, L.C., Valle, D. and Mitchell, G. (1992) Single-strand conformational polymorphism analysis: efficacy of detection of point mutations in the human ornithine d-aminotransferase gene. *Genomics* 10, 389–394.

Myers, R.M., Sheffield, V.C. and Cox, D.R. (1988) Detection of single base changes in DNA: ribonuclease cleavage and denaturing gradient gel electrophoresis. In: Davies, K.E. (ed.) *Genome Analysis: A Practical Approach*. IRL, Oxford, pp. 95–139.

Nesbitt, T.C. and Tanksley, S.D. (2002) Comparative sequencing in the genus Lycopersicon: implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics* 162, 365–379.

Orita, M., Suzuki, Y., Sekiya, T. and Hayashi, K. (1989) Rapid and sensitive detection of point mutations and SNA polymorphisms using the polymerase chain reaction. *Genomics* 5, 874–879.

O'Sullivan, H.D. (2006) Association genetics in wheat and development of single nucleotide polymorphic markers from expressed sequence tags. PhD thesis, University of Bristol, UK.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nicerson, D.A. and Boyce-Jacino, M. (1999) Mining SNPs from EST databases. *Genome Research* 9, 167–174.

Rickert, A.M., Jeong, J.H., Meyer, S., Nagel, A., Ballvora, A., Oefner, P.J. and Gebhardt, C. (2003) First generation SNP/InDel markers tagging loci for pathogen resistance in the potato genome. *Plant Biotechnology Journal* 1, 399–410.

Ronaghi, M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Research* 11, 3–11.

Russell, J., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G. and Powell, W. (2004) A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. *Genome* 47, 389–398.

Savage, D., Batley, J., Erwin, T., Logan, E., Love, C.G., Lim, G.A., Mongin, E., Barker, G., Spangenberg, G.C. and Edwards, D. (2005) SNPServer: a real-time SNP discovery tool. *Nucleic Acids Research* 33, W493–W495.

Sheffield, V.C., Beck, J.S., Kwitek, A., Sandstrom, D.V. and Stone, E.M. (1993) The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions. *Genomics* 16, 325–332.

Slade, A.J., Fuerstenberg, S.I., Loeffler, D., Steine, M.N. and Facciotti, D. (2004) A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. *Nature Biotechnology* 23, 75–81.

Small, R.L., Cronn, R.C. and Wendel, J.F. (2004) Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* 17, 145–170.

Soltis, D.E. and Soltis, P.S. (1995) The dynamic nature of polyploid genomes. *Proceedings of the National Academy of Sciences of the USA* 92, 8089–8091.

Somers, D.J., Kirkpatrick, R., Moniwa, M. and Walsh, A. (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome* 46, 431–437.

Syvänen, A.C. (2005) Toward genome-wide SNP genotyping. *Nature Genetics* 37, 5–10.

Tang, J., Vosman, B., Voorrips, R.E., van der Linden, C.G. and Leunissen, J.A.M. (2006) QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. *BMC Bioinformatics* 7, 438.

Tanksley, S.D., Young, N.D., Paterson, A.H. and Bonierbale, M.W. (1989) RFLP mapping in plant breeding: new tools for an old science. *Nature Biotechnology* 7, 257–264.

Varshney, R.K., Graner, A. and Sorrells, M.E. (2004) Genic microsatellite markers in plants: features and applications. *Trends in Biotechnology* 23, 48–55.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M. and Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23, 4407–4414.

Winfield, M.O., Arnold, G.M., Cooper, F., Le Ray, M., White, J., Karp, A. and Edwards, K.J. (1998) A study of genetic diversity in *Poplus nigra betulifolia* in the upper severn area using AFLPs. *Molecular Ecology* 7, 3–10.

Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J. and Davies, R.W. (1998) Direct allelic variation scanning of the yeast genome. *Science* 281, 1194–1197.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D. and Cregan, P.B. (2003) Single-nucleotide polymorphisms in soybean. *Genetics* 163, 1123–1134.

2 SNPs and Their Use in Maize

A. Rafalski and S. Tingey

Introduction

Writing this chapter gives us an opportunity to summarize several years' worth of experience discovering and using SNP markers for a variety of projects including genetic mapping, positional cloning, evaluation of genetic diversity and genetic association analysis. Rather than giving a comprehensive review of the literature, we will focus on conclusions and advice that may be helpful to those embarking on a new project involving SNP markers.

SNPs and Other Genetic Markers

Single nucleotide polymorphisms (SNPs) along with insertions/deletions represent at the most fundamental level the genetic diversity within a species. As such, SNPs are, at least in principle, the ultimate genetic markers. PCR technology and access to DNA sequencing at a reasonable cost have made SNP markers popular, especially in human genetics (Gray *et al*., 2000; The International HapMap Consortium, 2005).

The choice of markers should, of course, also be driven by pragmatic considerations. Other marker types, such as simple sequence repeats (SSRs), justifiably remain popular. Their high informativeness (PIC value or expected heterozygosity), which can reach 0.6–0.8, is a significant advantage. In contrast, randomly chosen SNPs in maize have the PIC value of about 0.2 (Ching *et al*., 2002). It is possible to choose SNP markers which are more informative in the germplasm collection of interest, but this may increase the cost of SNP discovery severalfold. In addition, choosing especially informative SNPs may have undesirable consequences. For example, markers with large allele frequency differences between heterotic groups, but less informative within heterotic groups, may be chosen inadvertently. Choosing markers that are

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

highly informative in a particular collection of germplasm will result in an ascertainment bias, and such markers will skew phylogenic relationships (Clark *et al*., 2005).

SSR markers may misrepresent the underlying SNP haplotype structure of the population because of the very high mutation rate of SSRs. This may result in homoplasy, in that different underlying DNA sequence alleles appear to have the same SSR size variant, or the opposite situation, SSR heteroplasy, where identical sequence alleles (haplotypes) have different SSR length alleles (Peakall *et al*., 1998; Chen *et al*., 2002; Palaisa *et al*., 2003b). It should be stressed that this situation is of concern in population genetic studies, and in population-based genetic association studies, but have no consequence where markers are simply used for mapping in biparental populations.

Because of the issues with marker homoplasy/heteroplasy, when comparing individual SSR marker alleles in the context of their phenotypic effects, it is beneficial to confirm the identity by descent of the relevant alleles. This may not always be possible, if the pedigree information is incomplete and the alleles/individuals cannot be traced to a common ancestor. The use of DNA sequence haplotypes provides at least a partial solution. In the presence of appreciable linkage disequilibrium (LD) (Flint-Garcia *et al*., 2003), a haplotype encompassing several adjacent SNP polymorphism is highly likely to be fully diagnostic for the allelic composition of a much larger surrounding segment of DNA (up to tens of kilobases; Nordborg *et al*., 2002; Jung *et al*., 2004; Hyten *et al*., 2007). If so, barring long-range effects (Wang *et al*., 1999; Clark *et al*., 2004, 2006) the identity in state (SNP haplotype) is likely to be equivalent to identity by descent, simplifying the analysis of population association data (see below).

Sources of SNP Markers

In the past, the need to derive SNP polymorphism information from direct sequencing of alleles of interest was considered to be a major disadvantage of SNP markers. However, sequencing technology has now become easily accessible and relatively inexpensive, facilitating SNP discovery, either by examination of already available sequence data, or by *de novo* sequencing.

Several of the more common sources of SNP polymorphism information are:

- Analysis of available DNA sequences, for example cDNAs or genomic sequences from genetically distinct individuals. ESTs are frequently used for this purpose, if the sequenced libraries have been obtained from two or more genotypes (Useche *et al*., 2001; Barker *et al*., 2003; Batley *et al*., 2003). In maize, we and others took advantage of the cDNAs from public lines B73 and Mo17, the parents of high-resolution genetic mapping populations. These inbreds have also been used to build high-quality physical (BAC contig) maps, on which over 10,000 ESTs have been located by overgo hybridization (Gardiner *et al*., 2004b). The B73 genome is being sequenced by a public consortium (http://www.maizegdb.org/genome/).

These resources greatly facilitate placement of the SNPs on the genetic and physical maps of maize even before the genome sequence is completed. Soon, for at least some organisms, complete genome sequences of several accessions or cultivars may become available. This is the case in rice (*indica* and *japonica* genomic sequences), and in maize, the sequencing of some chromosomes of Mo17 inbred has been proposed. The advent of inexpensive whole genome sequencing ('the $1000 genome') will further increase attractiveness of this approach (Huang *et al*., 2006).

- DNA sequencing of PCR amplicons from several individuals representing the genetic diversity of interest. The PCR amplicons are easy to design from available EST or genomic sequences. Large collections of ESTs are available for many species, or may be obtained at a moderate expense. Care should be taken to select amplicons which are specific to a single gene. In maize, we routinely design primer sets to amplify 350–700 bp and pre-screen these amplicons on a set of maize–oat addition lines (Ananiev *et al*., 1997; Kynast *et al*., 2001) rejecting all amplicons which amplify a product from more than a single maize chromosome. While this is not a perfect test (multiple gene copies located on a single chromosome cannot be discerned), in practice it drastically reduces the occurrence of amplification from multiple loci. Chosen amplicons are then sequenced from a set of eight or 12 highly diverse inbreds, and the sequences examined for the presence of SNPs. At the same time, a rough estimate of allele frequency is also obtained. A number of different software products are available to facilitate discovery of polymorphisms in the aligned sequences. In maize, abundance of indels (Bhattramakki *et al*., 2002) somewhat complicates the task in that multiple alignment or sequence assembly software usually does not handle indels well, necessitating manual editing of the alignment. In species with a low frequency of SNP polymorphism the EST-derived amplicon approach may not be cost-effective. In this case, the solution may be to continue using SSRs, or develop SNPs from longer EST or genomic sequences. Also, this approach does not yield good results in polyploid species, where an amplicon may represent several homoeologous genes. For example, in wheat, it would be necessary to clone each amplicon prior to sequencing to identify the three constituent homoeologues.
- Genomic or EST sequencing. New high-throughput sequencing methods, such as pyrosequencing (Ronaghi *et al*., 1998)-based '454' (http://www.454.com/), may be used to sequence cDNA or genomic libraries. For many genomes good saturation is easy to achieve with this technology and alignment of the resulting data sets will lead to the discovery of a large number of SNPs. This or other deep redundancy technologies could be used to simultaneously sequence encoded reduced representation libraries from two or more genotypes (Fig. 2.1). If a high-quality reference sequence for the organism is available this approach has a lot of power. For *de novo* sequencing approaches, the relatively high error

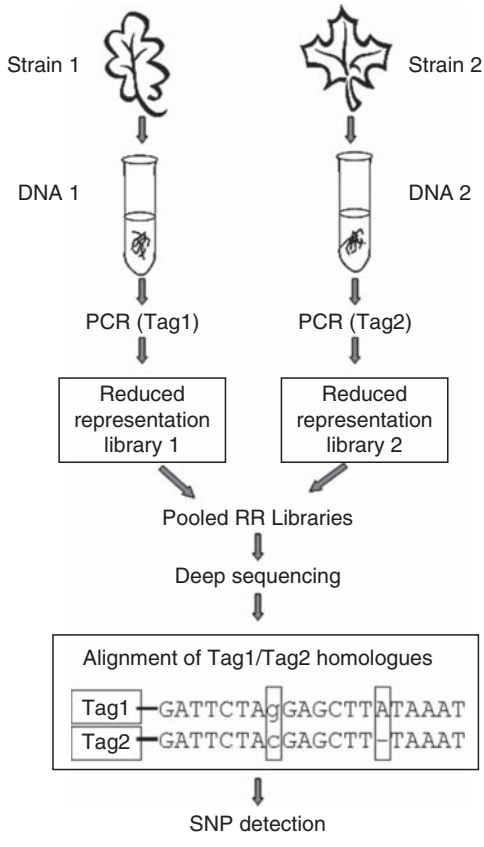


Fig. 2.1. SNP discovery by sequencing of reduced representation genomic libraries. Reduced representation libraries are prepared from two genetically distinct individuals. Each library is distinguished by a different sequence tag introduced during the PCR step. The libraries are pooled and sequenced. The sequences bearing different tags are aligned, and searched for the presence high-quality differences corresponding to SNPs or other polymorphisms.

rate of this method of sequencing is counterbalanced by high sequencing coverage, resulting in sequence redundancy which allows robust SNP calls.

- A number of other methods of identifying genome regions which contain polymorphisms have been proposed (the earliest being RFLP), such as digestion of DNA heteroduplexes with mismatch-recognizing enzymes or chromatographic analysis of heteroduplexes formed with a reference DNA sample (Hsia *et al*., 2005). These methods may be useful in species, where polymorphisms are very rare, but the decreasing costs of DNA sequencing are making them less cost effective.

Genotyping Methods

Many genotyping methods have been developed, and their systematic review is beyond the scope of this discussion (Gut, 2001; Kwok, 2001; Chen and Sullivan, 2003). In general, these methods may be divided into those that are suitable for genotyping a few individuals at a large number of loci, and those that are better suited for genotyping many individuals at few loci. At one extreme, high multiplex ratio technologies (Powell *et al*., 1996) such as Affymetrix chips allow simultaneous genotyping of thousands or hundreds of thousands of loci in a single hybridization. Such methods are suitable for fingerprinting or for whole genome scans (see below). In contrast, in some breeding applications, it may be necessary to follow the segregation of one or a few markers in a very large segregating population (Fig. 2.2). To this end, low multiplex ratio methods, such as TaqMan (Applied Biosystems) or Invader assay (Third Wave Technologies; Olivier, 2005), are more appropriate. These categories are, however, becoming less distinct, as lower-cost, intermediate multiplex ratio methods are becoming available. For example, it is possible to pack a large number of miniaturized arrays into a microplate format, thus providing intermediate level of multiplexing of loci and of individuals in a single device (e.g. Fan *et al*., 2006). The progress in multiplexing by locus has been very impressive. In contrast, multiplexing by individual is to some extent limited by the DNA isolation methodology and by the miniaturization of PCR reactions and fluidics.

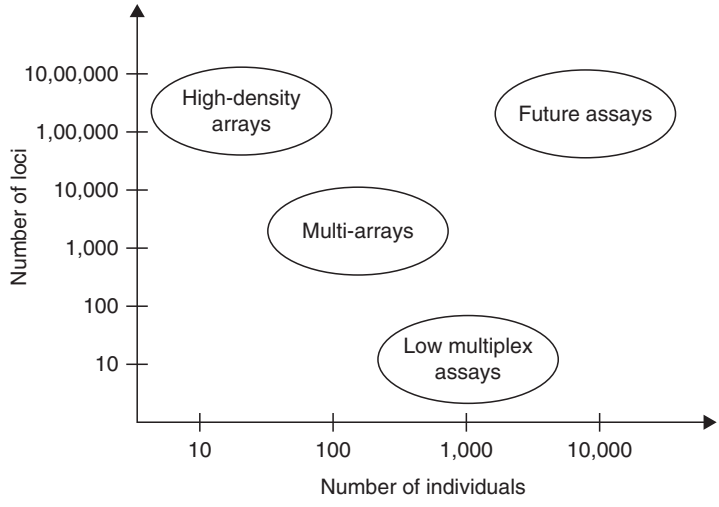


Fig. 2.2. Schematic classification of SNP diagnostic methods into those of very high multiplex ratio but modest throughput (e.g. high-density Affymetrix chips), those with intermediate throughput and multiplex ratio (e.g. arrays of arrays), and those with very high throughput, but limited capacity for multi-locus multiplexing (e.g. Invader assay).

Genetic Mapping in Segregating Populations

Creation of genetic maps or mapping of simple traits in biparental populations is one of the simplest applications of molecular markers. Any type of DNA-based markers will do. Facility and low cost of scoring marker segregation are frequently the most relevant criteria for choosing markers. For example, choosing a subset of SNPs which can be detected by restriction cleavage of PCR products may be appropriate (Konieczny and Ausubel, 1993).

With the rapidly falling costs of DNA sequencing, direct sequencing of PCR products becomes competitive with other methods of scoring SNPs especially if markers are unlikely to be reused. The cost of developing an assay is avoided, but the scoring of the polymorphisms directly on the alignment of sequences from the segregating individuals may add to the cost especially if heterozygous individuals have to be identified (usually requiring examination of sequence traces). Bioinformatic tools to facilitate this have been developed (Marth *et al*., 1999).

When the assay is to be used repeatedly, the high cost of developing a more sophisticated assay will be easily amortized over its lifetime.

We find that direct sequencing is the most straightforward approach to rapidly saturate a small region of the genome with markers. To this end, we use single-copy DNA sequences which have been placed by various means (such as BAC end sequencing) on a highly saturated, integrated physical-genetic map of maize (Coe *et al*., 2002; Cone *et al*., 2002).

It is frequently convenient to start the mapping process by developing a large population sufficient to achieve desired resolution (up to a few thousand individuals, if the eventual objective is cloning the gene), but initiating the mapping in a small subset (100–150) of individuals. Once the approximate map position has been determined, perhaps within 20 cM, markers flanking the locus can be used on the rest of the population to identify recombinants in the region of interest. In principle, this could be done using only two markers, but some level of redundancy (four markers) is advantageous in case of assay failure. The recombinant individuals are then grown to maturity, propagated by selfing or outcrossing as appropriate, and phenotyped. This approach limits the phenotyping to a subset of the population creating a significant saving in time and effort. The recombinant individuals have to be genotyped further with markers distributed within the region of interest only. Here it may not be possible to use pre-existing already developed markers as the resolution increases. At this stage it is then most appropriate to switch to custom-developed markers targeted to the region of interest. This is relatively easy if, as is the case in maize, where the genetic map is well aligned with the physical map. Such markers are likely to be used rarely so it may be easiest to directly sequence PCR amplicons in order to determine the genotype saving the cost of custom marker development.

Fingerprinting

DNA fingerprinting for the purposes of identification of individuals or varieties, plant variety protection or for forensic applications requires stable and

highly informative markers. For that application, SSRs are very appropriate. While mutation rate of SSRs is on the order of 10^{-4} (Vigouroux *et al*., 2002), in practice instability of SSR allele sizes is rarely observed. High PIC value of SSRs, up to threefold higher than SNPs, has been an advantage. A distinct disadvantage of SSRs is, however, the need for size separation of PCR products and scoring of allele sizes usually requiring quality control by highly trained individuals. Therefore, for high-throughput applications SNPs are gaining in popularity. To facilitate adoption of a common set of SNP markers in maize, Pioneer Hi-Bred International is making a collection of highly informative maize SNPs available to the community (S. Smith, Johnston, Iowa, 2007, personal communication).

Diversity Analysis

A common application of genetic markers is the analysis of genetic relationships and diversity within and between populations. If the objective is to understand the diversity at the DNA sequence level, SNP markers are preferred. Faster evolving SSRs may be occasionally more informative in resolving more recently evolved lineages. When using SNPs to analyse diverse populations of different origin (such as comparing elite material with accessions of the undomesticated relative) care should be taken to avoid ascertainment bias. Ascertainment bias would result if the SNP loci chosen for analysis have been sampled non-randomly favouring SNPs originally identified as informative in one of the populations under consideration (Clark *et al*., 2005). When using SNPs discovered by re-sequencing loci from several elite breeding lines for the analysis of undomesticated accession it is quite likely that diversity will be underestimated. This is because many loci monomorphic in the elite population, but polymorphic in the accessions, have been excluded from consideration at the beginning. DNA sequencing of randomly selected EST loci across individuals sampled from both populations is one of the strategies that will avoid ascertainment bias, but it may considerably increase SNP discovery cost, especially in species with a low frequency of SNPs. Even this approach may have a subtle bias, in that alleles that do not amplify from all geneotypes being considered, perhaps due to a polymorphism overlapping one or both primers, will be rejected because of their 'excessive' polymorphism.

Genetic Association Mapping

One of the most promising applications of SNP markers is genetic association mapping. The objective of association mapping is to test the hypothesis that in the population of interest, individuals carrying different alleles (haplotypes) at a locus (or loci) differ in a specific phenotype. This approach, also called LD mapping, has been conceptually developed by human geneticists (Gibbs and Singleton, 2006). The first successful applications in plants have been in maize (Remington *et al*., 2001; Thornsberry *et al*., 2001). Here, we will briefly discuss

some aspects of genetic association mapping without attempting a comprehensive review. Association mapping methods could be divided into two partially overlapping methodologies: candidate gene association and whole genome scan.

Candidate gene associations

In the simplest case, one or a few candidate genes could be tested for evidence of genetic association with a trait. The gene candidates may be selected on the basis of the understanding of the biochemistry of the trait. For example, anthocyanin biosynthesis genes would be reasonable candidates for an association with aleurone layer colouration in maize. All or a fragment of the candidate gene is re-sequenced (as PCR products) across the germplasm collection/population of interest to identify and catalogue polymorphisms present in the population. The same population is scored for the phenotype. Finally, the distribution of the phenotypic scores in individuals carrying allele A is compared using appropriate statistical methodology to the distribution of phenotypic scores of individuals carrying allele B. In addition to testing individual SNPs, it is also common to combine the SNPs within the gene into haplotypes and in turn test each haplotype for association with the trait. A non-parametric statistical test for association is the Kolmogorov–Smirnov test (http://www.physics.csbsju.edu/stats/KS-test.html), which compares a set of numbers representing phenotypic scores of individuals carrying allele A to a set of numbers representing phenotypes of individuals carrying allele B. More sophisticated tests of association that take population structure into account have been developed (e.g. Yu *et al*., 2006).

Theoretical arguments may be made that either individual SNPs or SNP haplotypes may have more power to detect associations so both should be tested.

For example, Palaisa *et al*. (2003a, 2004) analysed the relationship between the haplotype structure of the phytoene deasturase gene in maize (*Y1*) and endosperm colour (white or yellow), and found strong evidence for the association of this phenotype with certain polymorphisms within the *Y1* gene.

A potential pitfall of a candidate gene approach is that the choice of candidates limits the researcher to known genes, missing potentially important aspects of the genetic control of the trait. Therefore, given adequate resources, a whole genome scan approach is preferable. Evidence for the association of a candidate gene with a phenotype has to be interpreted cautiously. It is possible that the candidate gene represents a genetic marker linked tightly with another gene which is responsible for the phenotype. Direct evidence, such as the identification of informative recombination breakpoints, or direct observation of the effect of overexpression or suppression of the gene in transgenic plants, helps to solidify the evidence for the candidate gene itself.

Whole genome scan

To avoid the bias of candidate gene testing towards known genes, it would be preferable to test for association with every gene in the genome. This is rapidly becoming feasible, although it is still costly. In some species there is sufficient LD (non-random association of alleles at two or more loci, in a population) that a subset of all genes may be tested providing close to complete coverage of the genome at a cost of lost resolution. Naturally selfing species such as Arabidopsis or soybean characteristically have strong LD in the population extending sometimes to more than a 100 kb (Hyten *et al*., 2007). In outcrossing species such as maize (or humans) LD decays more rapidly with distance, although the rate of the decay depends strongly on the nature of the population being studied (Gaut and Long, 2003). In the collection of genetically diverse accessions of maize, rapid decay of LD has been reported (Tenaillon *et al*., 2001). However, in elite cultivated maize inbred lines LD is more extensive (Jung *et al*., 2004). Thus, whole genome scans may be possible with several thousand markers in an appropriate set of germplasm.

For example, it is known that colour of cob and pericarp in maize is determined by the allele present at the *P1* locus on chromosome 1 (Grotewold *et al*., 1994). As a proof of principle this locus was mapped by whole genome scan with several thousand markers to a single BAC clone (physical resolution of ~150 kb, Fig. 2.3; O. Smith, S. Luck and A. Rafalski, Wilmington, Delaware, 2007, unpublished data).

SNPs in Positional Cloning

Positional cloning begins with low- to medium-resolution mapping of a trait to a region of perhaps 20 cM, followed by a high-resolution mapping. It is in the second step, in which the available genetic markers may be exhausted

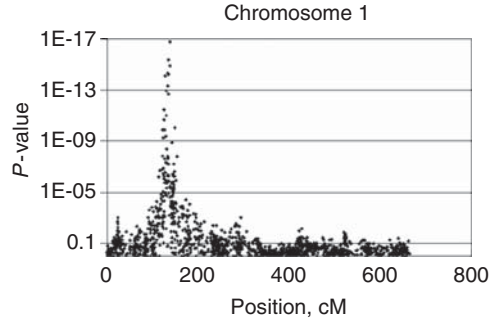


Fig. 2.3. Whole genome scan association mapping of *P1* (red pericarp) locus. SNP and SNP haplotypes of available markers on chromosome 1 of maize were tested for association with cob colour (red or white). The peak of *P*-value for association (*y*-axis) corresponds to the location of *P1* gene.

rapidly as the resolution increases, when custom-designed SNP markers are essential. If a good physical (BAC contig) map of the region exists, as is the case for the B73 maize inbred, it is possible to identify single-copy genic sequences or BAC ends within the interval of interest. In addition to the numerous BAC ends, over 10,000 genes have been placed on the B73 physical map by oligonucleotide (overgo) hybridization (Gardiner *et al*., 2004a), and their source sequences are suitable for marker development. These single-copy sequences are used to design PCR amplicons, which are then re-sequenced from the two mapping parents. A modest-size amplicon, of 300–500 bp almost invariably, reveals SNP or indel polymorphisms between parents of a maize mapping population. It is important to confirm the genetic location of the amplified sequences, for example through the use of maize–oat addition lines or by segregation analysis. The polymorphisms present in these amplicons may then be used to generate a genotype of the recombinant individuals and order the recombination breakpoints. Genotyping the informative progeny by direct DNA sequencing of amplicons will save time and effort. In maize, several steps of amplicon design and sequence genotyping of recombinant individuals usually allow the identification of a small interval containing only one or a few candidate genes. Recent cloning of a domestication gene, teosinte glume architecture, provides an interesting example of very fine resolution mapping (Wang *et al*., 2005).

SNP Markers in Duplicated Chromosome Regions and in Polyploids

As is the case with any genetic markers, the use of SNPs in polyploids presents special challenges. For example, in wheat, which is an allohexaploid, three different constituent genomes are closely related and a PCR or hybridization-based genetic marker is likely to 'pile up' information from three separate chromosomes. In many cases it is possible to identify sufficient DNA sequence differences between the A, B and D genomes at the loci of interest to enable the design of genome-specific markers (Zhang *et al*., 2003). In wheat, aneuploid stocks are available to verify the specificity of amplification.

One approach to the identification of genome-specific markers is to start with a primer set which amplifies from the three genomes. The product of this amplification may be cloned, 12–36 random clones sequenced and the three constituent sequences defined. Once the differences between the A-, B- and D-derived sequences are known it may be possible to develop genome-specific primers or genome-specific SNP assays (such as through a primer extension assay).

Maize is an ancient polyploid which genetically behaves as a diploid. However, many genes are duplicated and in some cases the two copies may be nearly identical (Emrich *et al*., 2007). Such duplications may have arisen recently as a result of the activity of helitrons which have been demonstrated to carry fragments of genes. In maize, a similar approach to the one described for wheat may be used, although it may be easier to separately sequence

chromosome-specific amplification products from maize–oat addition lines as long as the duplicated genes are on different chromosomes.

An even more complex situation arises in autopolyploids. Cultivated sugarcane is an octaploid, sometimes resulting from a cross between two different species. Therefore, some identical copies and some non-identical copies of a gene may exist in a cultivar. Developing single-dose markers may therefore be difficult. Pyrosequencing (Ronaghi *et al*., 1998; Ching and Rafalski, 2002) is well suited for the determination of gene copy number relative to an internal control. For example, a SNP present in two genomes of three in a hexaploid may be easily distinguished from one present in only one of the constituent genomes. In combination with genome-specific primers, pyrosequencing or similar quantitative methods may allow a more detailed characterization of such complex polyploids.

Outlook

While it is always risky to predict the future, some trends in the analysis and applications of SNPs are clear. DNA sequencing is becoming faster and cheaper, suggesting that within a few years it will be possible to rapidly re-sequence multiple inbred lines and discover all polymorphisms within the single-copy regions of the genome. It is already becoming possible to simultaneously re-sequence reduced representations of several inbreds or perhaps of pools of individuals representing phenotypic extremes. It will also be possible to sequence individuals of a segregating population identifying the location of recombination events with resolution limited only by the frequency of polymorphisms. High-density microarray technology and sequencing appear to be converging, and array-based whole genome high-density genotyping, currently possible for smaller genomes, will become a reality for larger genomes as well. This will in turn enable true whole genome scanning genetic association mapping in large collections of germplasm, in some cases enabling the identification of causative polymorphisms. It is becoming increasingly clear that a large fraction of functionally relevant alleles may affect change in gene expression, rather than amino acid sequence, and thus resolution may not be reduced to a single causal mutation (Doebley and Lukens, 1998; Clark *et al*., 2006).

As genotyping becomes increasingly routine, accurate phenotyping becomes rate-limiting, especially in high-resolution mapping or positional cloning exercises where success depends on the correct phenotype assignment of a single recombinant individual or where quantitative, multigenic traits are being considered.

In plant genotyping, rapid and automated extraction of DNA is still a challenge and developments in this area would be most welcome, as well as further advances in field sampling and encoding. Such technological advances will enable further reduction of cost per genotype.

Increases in the amount of data generated also challenge computer information systems, especially the ability to integrate different data types

and enable a rapid selection of relevant information such as the phenotypic value of genome segments present in a breeding population on the basis of current and historical data (Peleman and van der Voort, 2003).

Acknowledgements

We would like to thank Howie Smith and Stan Luck for sharing unpublished data.

References

Ananiev, E.V. *et al.* (1997) Oat–maize chromosome addition lines: a new system for mapping the maize genome. *Proceedings of the National Academy of Sciences of the USA* 94, 3524–3529.

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.

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 Physiology* 132, 84–91.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register III, J.C., Tingey, S.V. and Rafalski, A. (2002) Insertion–deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers. *Plant Molecular Biology* 48, 539–547.

Chen, X. and Sullivan, P.F. (2003) Single nucleotide polymorphism genotyping: biochemistry, protocol, cost and throughput. *Journal of Pharmacogenomics* 3, 77–96.

Chen, X., Cho, Y.G. and McCouch, S.R. (2002) Sequence divergence of rice microsatellites in *Oryza* and other plant species. *Molecular Genetics and Genomics* 268, 331–343.

Ching, A. and Rafalski, A. (2002) Rapid genetic mapping of ESTs using SNP pyrosequencing and indel analysis. *Cellular and Molecular Biology Letters* 7, 803–810.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M. and Rafalski, A.J. (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. *BMC Genetics* 3, 19.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H. and Nielsen, R. (2005) Ascertainment bias in studies of human genome-wide polymorphism. *Genome Research* 15, 1496–1502.

Clark, R.M., Linton, E., Messing, J. and Doebley, J.F. (2004) Pattern of diversity in the genomic region near the maize domestication gene tb1. *Proceedings of the National Academy of Sciences of the USA* 101, 700–707.

Clark, R.M., Wagler, T.N., Quijada, P. and Doebley, J. (2006) A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. *Nature Genetics* 38, 594–597.

Coe, E., Cone, K., McMullen, M., Chen, S.S., Davis, G., Gardiner, J., Liscum, E., Polacco, M., Paterson, A., Sanchez-Villeda, H., Soderlund, C. and Wing, R. (2002) Access to the maize genome: an integrated physical and genetic map. *Plant Physiology* 128, 9–12.

Cone, K.C., McMullen, M.D., Bi, I.V., Davis, G.L., Yim, Y.S., Gardiner, J.M., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Schroeder, S.G., Havermann, S.A., Bowers, J.E., Paterson, A.H., Soderlund, C.A., Engler, F.W., Wing, R.A. and Coe, E.H.J. (2002) Genetic, physical, and informatics resources for maize. On the road to an integrated map. *Plant Physiology* 130, 1598–1605.

Doebley, J. and Lukens, L. (1998) Transcriptional regulators and the evolution of plant form. *Plant Cell* 10, 1075–1082.

Emrich, S., Li, L., Wen, T.J., Yandeau-Nelson, M.D., Fu, Y., Guo, L., Chou, H.H., Aluru, S., Ashlock, D.A. and Schnable, P.S. (2007) Nearly identical paralogs: implications for maize (*Zea mays* L.) Genome evolution. *Genetics* 175, 429–439.

Fan, J.B., Gunderson, K.L., Bibikova, M., Yeakley, J.M., Chen, J., Wickham Garcia, E., Lebruska, L.L., Laurent, M., Shen, R. and Barker, D. (2006) Illumina universal bead arrays. *Methods in Enzymology* 410, 57–73.

Flint-Garcia, S.A., Thornsberry, J.M. and Buckler, E.S.I. (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54, 357–374.

Gardiner, J., Schroeder, S., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Morgante, M., Landewe, T., Fengler, K., Useche, F., Hanafey, M., Tingey, S., Chou, H., Wing, R., Soderlund, C. and Coe, E.H., Jr (2004a) Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. *Plant Physiology* 134, 1317–1326.

Gardiner, J., Schroeder, S., Polacco, M.L., Sanchez-Villeda, H., Fang, Z., Morgante, M., Landewe, T., Fengler, K., Useche, F., Hanafey, M., Tingey, S., Chou, H., Wing, R., Soderlund, C. and Coe, E.H.J. (2004b) Anchoring 9,371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimensional overgo hybridization. *Plant Physiology* 134, 1317–1326.

Gaut, B.S. and Long, A.D. (2003) The lowdown on linkage disequilibrium. *The Plant Cell* 15, 1502–1506.

Gibbs, J.R. and Singleton, A. (2006) Application of genome-wide single nucleotide polymorphism typing: simple association and beyond. *PLoS Genetics* 2, e150.

Gray, I.C., Campbell, D.A. and Spurr, N.K. (2000) Single nucleotide polymorphisms as tools in human genetics. *Human Molecular Genetics* 9, 2403–2408.

Grotewold, E., Drummond, B.J., Bowen, B. and Peterson, T. (1994) The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell* 76, 543–553.

Gut, I.G. (2001) Automation in genotyping single nucleotide polymorphisms. *Human Mutation* 17, 475–492.

Hsia, A.P., Wen, T.J., Chen, H.D., Liu, Z., Yandeau-Nelson, M.D., Wei, Y., Guo, L. and Schnable, P. (2005) Temperature gradient capillary electrophoresis (TGCE) – a tool for the high-throughput discovery and mapping of SNPs and IDPs. *Theoretical and Applied Genetics* 111, 218–225.

Huang, X., Barbee, K., Chen, Y.J. and Roller, E. (2006) Towards a $1000 human genome. *Nanomedicine* 2, 271–272.

Hyten, D.L., Choi, I.Y., Song, Q., Shoemaker, R.C., Nelson, R.L., Costa, J.M., Specht, J.E. and Cregan, P.B. (2007) Highly variable patterns of linkage disequilibrium in multiple soybean populations. *Genetics* 175, 1937–1944.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437, 1299–1320.

Jung, M., Ching, A., Bhattramakki, D., Dolan, M., Tingey, S., Morgante, M. and Rafalski, A. (2004) Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. *Theoretical and Applied Genetics* 109, 681–689.

Konieczny, A. and Ausubel, F.M. (1993) A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. *The Plant Journal* 4, 403–410.

Kwok, P.Y. (2001) Methods for genotyping single nucleotide polymorphisms. *Annual Review of Genomics and Human Genetics* 2, 235–258.

Kynast, R.G., Riera-Lizarazu, O., Vales, M.I., Okagaki, R.J., Maquieira, S.B., Chen, G., Ananiev, E.V., Odland, W.E., Russell, C.D., Stec, A.O., Livingston, S.M., Zaia, H.A., Rines, H.W. and Phillips, R.L. (2001) A complete set of maize individual chromosome additions to the oat genome. *Plant Physiology* 125, 1216–1227.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y. and Gish, W.R. (1999) A general approach to single-nucleotide polymorphism discovery. *Nature Genetics* 23, 452–456.

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J., Hagenblad, J., Kreitman, M., Maloof, J.N., Noyes, T., Oefner, P.J., Stahl, E.A. and Weigel, D. (2002) The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nature Genetics* 30, 190–193.

Olivier, M. (2005) The invader assay for SNP genotyping. *Mutation Research* 573, 102–110.

Palaisa, K., Morgante, M., Williams, M. and Rafalski, A. (2003a) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *The Plant Cell* 15, 1795–1806.

Palaisa, K.A., Morgante, M., Williams, M. and Rafalski, A. (2003b) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. *Plant Cell* 15, 1795–1806.

Palaisa, K., Morgante, M., Tingey, S. and Rafalski, A. (2004) Long-range patterns of diversity and linkage disequilibrium surrounding the maize Y1 gene are indicative of an assymetric selective sweep. *Proceedings of the National Academy of Sciences of the USA* 101, 9885–9890.

Peakall, R., Gilmore, S., Keys, W., Morgante, M. and Rafalski, A. (1998) Cross-species amplification of soybean (*Glycine max*) simple sequence repeat (SSRs) within the genus and other legume genera: implications for the transferability of SSRs in plants. *Molecular Biology and Evolution* 15, 1275–1287.

Peleman, J. and van der Voort, J.R. (2003) Breeding by design. *Trends in Plant Science* 8, 330–334.

Powell, W., Morgante, M., Andre, C., Hanafey, M., Vogel, J., Tingey, S. and Rafalski, A. (1996) The comparison of RFLP, RAPD, AFLP and SSR (Microsatellite) markers for germplasm analysis. *Molecular Breeding* 2, 225–238.

Remington, D.L., Thornsberry, J.M., Matsuoka, Y., Wilson, L.M., Whitt, S.R., Doebley, J., Kresovich, S., Goodman, M.M. and Buckler, E.S.T. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the USA* 98, 11479–11484.

Ronaghi, M., Uhlen, M. and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate. *Science* 281, 363–365.

Tenaillon, M.I., Sawkins, M.C., Long, A.D., Gaut, R.L., Doebley, J.F. and Gaut, B.S. (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proceedings of the National Academy of Sciences of the USA* 98, 9161–9166.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E.S.I. (2001) Dwarf8 polymorphisms associate with variation in flowering time. *Nature Genetics* 28, 286–289.

Useche, F., Gao, G., Hanafey, M. and Rafalski, A. (2001) High-throughput identification, database storage and analysis of SNPs in EST sequences. *Genome Informatics* 12, 194–203.

Vigouroux, Y., Jaqueth, J.S., Matsuoka, Y., Smith, O.S., Beavis, W.D., Smith, J.S. and Doebley, J. (2002) Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution* 19, 1251–1260.

Wang, H., Nussbaum-Wagler, T., Li, B., Zhao, Q., Vigouroux, Y., Faller, M., Bomblies, K., Lukens, L. and Doebley, J.F. (2005) The origin of the naked grains of maize. *Nature* 436, 714–719.

Wang, R.L., Stec, A., Hey, J., Lukens, L. and Doebley, J. (1999) The limits of selection during maize domestication. *Nature* 398, 236–239.

Yu, J., Pressoir, G., Briggs, W.H., Vroh, B.I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* 38, 203–208.

Zhang, W., Gianibelli, M.C., Ma, W., Rampling, L. and Gale, K.R. (2003) Identification of SNPs and development of allele-specific PCR markers for gamma-gliadin alleles in *Triticum aestivum*. *Theoretical and Applied Genetics* 107, 130–138.

3 Rare SNP Discovery with Endonucleases

M.J. Cross

Introduction

The discovery and use of genetic variation is essential for answering many biological questions. Single nucleotide polymorphisms (SNPs) represent the most common form of genetic variation in both plants and animals and play a key role in revealing the molecular mechanisms underlying traits. SNPs may be classed as either common or rare – depending on the frequency in which they occur in a population. Common SNPs find use as markers in mapping and association studies, ultimately enabling the positioning and cloning of genes responsible for phenotypes of interest. Rare SNPs also play a role in elucidating gene function, such as in the context of chemical mutagenesis where induced single-base changes offer knockout and missense mutations for potentially any gene of interest.

The identification of such subtle DNA variations has proven to be a challenging task in the past; however, the postgenomics era is offering new opportunities for the discovery of SNPs. The discovery of common SNPs, for example, has become a relatively straightforward process, since the simple comparison of a small number of samples against a reference sequence, such as the Columbia wild-type genome sequence of *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), can reveal a multitude of polymorphisms (Jander *et al*., 2002; Kwok and Chen, 2003). Such comparisons may be performed one locus at a time by standard re-sequencing of PCR amplicons. However, the advent of new-age technologies, such as total genomic array hybridization (Borevitz *et al*., 2003) and new sequencing inventions (Margulies *et al*., 2005), will greatly facilitate the global discovery of common SNPs, via the capacity to compare large stretches of sequence (multiple loci) between individuals. Such technologies, although highly parallel, are limited with regards to the number of individuals that can be analysed in a cost-effective manner. This feature makes these technologies unsuitable for the discovery

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

of less common polymorphisms within candidate genes, since rare mutations are unlikely to be present when only a limited number of individuals are screened. Detection strategies which instead enable the pooling of multiple samples into a single assay are more suitable for the discovery of rare SNPs in candidate genes of interest.

Nucleic acid annealing provides the foundation for a broad range of mutation detection strategies. Such techniques are based on the physical characteristics (including structural conformations and melting properties) that can distinguish wild-type from mutant (or mutant-containing) DNA. For separation-based methods (using technologies such as chromatography and electrophoresis) where mutation detection is achieved by resolving variant DNA species from wild-type controls, both mutant and wild-type samples must be present within an individual assay. This requirement for sample pooling is an important feature, since the screening of multiple samples in a single assay helps to facilitate throughput. A further advantage of annealing-based strategies is that unknown genetic variation may be discovered without the need for detailed sequence information on each sample under investigation.

Both double- and single-stranded DNA may be exploited for the purpose of mutation detection by nucleic acid annealing techniques. Although wild-type and mutant duplexes may possess subtle sequence variations, such as a SNP (Fig. 3.1A), the associated melting properties can often be specific enough to enable a means of separating the two species. If, for example, a duplex possesses strengthened hydrogen bonding due to the presence of a mutated sequence, then the local melting domain will have an increased tendency to retain its double-stranded form compared to that of its wild-type counterpart. Additionally, if this variation occurs within the lower-melting domain of the dsDNA (i.e. the region, usually the end, which melts before the remaining portions of the duplex due to a lower GC content), then this region of the duplex will branch out into single-stranded conformations which are specific to each sequence type, as conditions favouring denaturation are approached (Fig. 3.1B). These molecules may be thought of as taking on a 'Y-shaped' structure, with a clamped duplexed region in the high-melting domain and an open conformation in the low-melting domain. Thus, by subjecting the samples to near-denaturing conditions, the specific properties of each Y-conformation may be exploited for the purpose of mutation detection. This structural feature of nucleic acids was first utilized for the purpose of mutation detection by Fischer and Lerman (1983) (see below).

While minor melting temperature differences between homoduplex DNA species may enable a means of mutation detection, a more flexible and robust strategy involves the distinct duplex species formed when mutant strands re-hybridize with their wild-type reverse complements. Irregularity in the duplex is formed wherever bases are mismatched. Re-hybridization of sample pools is easily achieved by heat denaturation into single-stranded DNA (Fig. 3.1C), followed by slow cooling so that heteroduplexes may form whenever both mutant and wild-type alleles are present. Four distinct double-stranded species are thus formed following re-hybridization, comprising of two newly created heteroduplex species, in addition to the original two

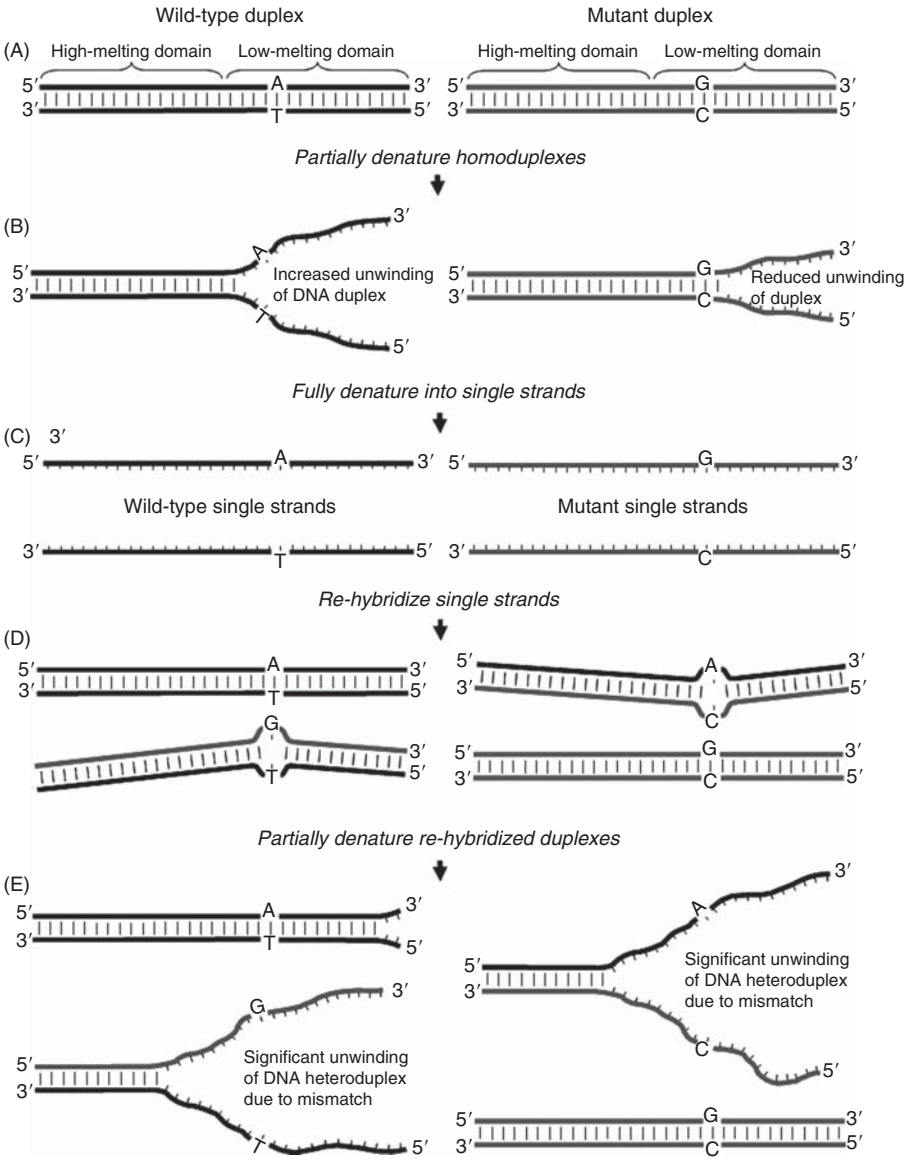


Fig. 3.1. Various DNA structures exploited for mutation detection. (A) Schematic illustration of wild-type and mutant duplexes in their native state. The polymorphism shown is located within the low melting domain of the sequence. (B) Strengthened annealing due to the G:C hydrogen bond in the mutant results in reduced unwinding of the helix under partial-denaturing conditions. The differences in helical unwinding are significant because the polymorphism is located within the low melting domain. (C) Increased denaturation results in complete unwinding of each helix so that single strands result. (D) Random re-annealing of the single-strand mix allows heteroduplexes to form between wild-type strands and their mutant reverse complements. (E) Heteroduplexes, as a set, have more significantly weakened hydrogen bonding compared to their homoduplex counterparts, with their helices thus being unwound to a greater extent as denaturing conditions are increased.

homoduplex species (Fig. 3.1D). Heteroduplexes possess distinct structural features which enable a means for their detection. The destabilized nature of a non-Watson/Crick base pair is believed to facilitate modification of the site by specific chemicals and/or enzymes, due to enhanced physical access of such factors to the exposed nucleotide bases. Such modifications may offer a means of heteroduplex detection via mechanisms such as binding of the mismatch to create a duplex with altered properties, and/or cleavage of the DNA followed by subsequent size fractionation. Additionally, nucleotide mismatches are known to create radically kinked duplex structures, which are implicated in the recognition process of certain endonucleases. Furthermore, melting temperature may also offer a means of separation, as it does for wild-type and mutant homoduplex species (see above). However, in contrast to the minor differences in melting properties between wild-type and mutant homoduplexes, both heteroduplex species have significantly lower melting points than their homoduplex founders. Thus, when relying on Y-conformations as a means of separating different DNA species, heteroduplexes, as a set, have an enhanced ability to be resolved from their homoduplex counterparts due to their increased rate of denaturation (Fig. 3.1E). Since heteroduplexes contain mutated DNA strands, their separation from homoduplex species is a means of mutation detection.

In addition to duplex DNA, a further type of structural feature exploited for mutation detection involves the conformations generated when individual DNA strands fold back and self-anneal on to themselves (Fig. 3.2). Here sequence variations can alter the resultant 'loop' structures formed, also allowing mutant strands to be distinguished from their wild-type counterparts.

A variety of tools have been developed for the detection of mutations based on such structural characteristics. These include two broad classes – those that separate the variants based purely on their specific mobilities under a range of physical conditions and those that separate the variants following structure-specific cleavage (or binding) events. The former class includes a range of highly effective techniques such as denaturing gradient gel electrophoresis (DGGE), denaturing high-performance liquid chromatography (DHPLC), constant denaturing gel electrophoresis (CDGE) and single-strand conformation polymorphism (SSCP).

DGGE is a mutation scanning technique first described by Fischer and Lerman (1983). DGGE relies on the specific mobilities generated when DNA helices begin to melt into a single-stranded state as they are electrophoresed through a gradient of increasing denaturing conditions. As denaturation of the duplex commences, the strands of each helix will unwind into a Y-shaped structure which, if located within the low-melting domain of the duplex, will result in greatly retarded electrophoretic mobility (see above). The specific melting properties of wild-type and mutant DNA thus enable a means of separation, since one species will tend to melt sooner than the other and thus have a retarded mobility. Gradients are usually generated by combining a higher denaturation gel mix (containing stronger concentrations of the chemical denaturant – typically urea and formamide) with a lower denaturation mix, such that the two flow into each other. Although the minor melting-property differences between

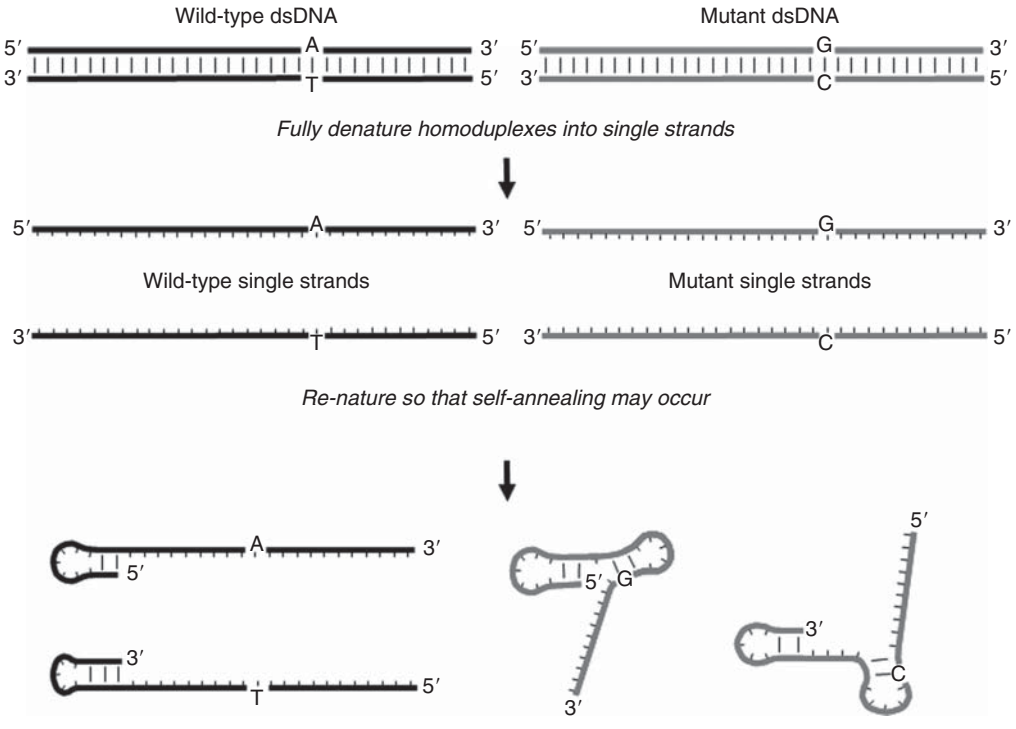


Fig. 3.2. Single-stranded conformations useful for mutation detection. Following denaturation, single strands are isolated from each other to prevent re-hybridization into duplexes and subjected to native conditions to promote self-annealing. The genetic sequence is specific to each strand, resulting in unique annealing properties. Sequence-specific annealing results in four unique single-stranded structures which may subsequently be exploited for the purpose of mutation detection.

homoduplex-wild-type and homoduplex-mutant DNA are often sufficient to achieve resolution, a more effective approach to DGGE involves the formation of both homoduplex and heteroduplex DNA species via a prior re-hybridization step (Fig. 3.1D). Heteroduplexes have a more significantly reduced melting temperature and are thus resolved from homoduplex species to a greater extent than direct comparisons between homoduplex wild-type and homoduplex mutant samples (Fig. 3.1E). This enhanced ability to resolve mutant-containing DNA (i.e. heteroduplexes) from wild-type was first reported in 1985 by Myers *et al*. (1985d). Although the generation of denaturing gradients is impractical for capillary electrophoresis (CE), a similar effect may be achieved using temperature as a means of controlling the level of denaturation. Temperature gradient capillary electrophoresis (TGCE) is thus a simple modification of DGGE, where varying temperatures are applied along the length of an electrophoretic capillary to help resolve wild-type from mutant (or mutant-containing) DNA species.

CDGE is a closely related electrophoretic method which relies on the specific mobilities of homoduplex and heteroduplex DNA at near-denaturing conditions, following a prior re-hybridization step. As with DGGE, those

DNA species with lower melting points (such as heteroduplexes) will more readily shift into a partially melted state and thus be greatly retarded as they move through the electrophoretic medium. A variation of CDGE is conformation-sensitive capillary electrophoresis (CSCE). Capillary DNA analysers are based on the principle of laser-induced fluorescence (LIF) and are noted for their excellent resolution and superior sensitivity. The minor mobility shifts associated with the presence of DNA variants are thus readily detectable by CSCE, even when mutant alleles are significantly diluted among a pool of predominantly wild-type individuals. Furthermore, the resolution, resulting from a capillary-based platform, often allows the separation of all four DNA duplex species present among the re-hybridized pool (Esteban-Cardenosa *et al*., 2004), a feature that is not readily achievable by other heteroduplex detection methods. While CSCE is a very similar technique to TGCE, the lack of requirement for a temperature (denaturing) gradient makes it a far simpler method to perform. When such practicalities are combined with other advantages of CE, such as dye multiplexing and the capacity for automation, CSCE becomes one of the most robust and high-throughput options available for the discovery of rare SNPs in candidate genes of interest.

DHPLC was first applied to mutation detection by Underhill *et al*. (1997). DHPLC relies on the different binding affinities of heteroduplex and homoduplex molecules (formed by a prior re-hybridization step) as they migrate through a chromatographic column under partial denaturing conditions. Phosphate groups of the duplex samples are able to form ionic bonds with the positively charged groups of the column matrix. As the run continues, the attraction of each duplex to the column is weakened by increasing concentrations of acetonitrile, until eventually the DNA will separate from the matrix and be eluted. Heteroduplex DNA species have a slightly weaker affinity for the matrix and will thus separate and elute from the column before their homoduplex counterparts. Mutant-containing pools will therefore produce two main peak groups – one resulting from the elution of both heteroduplex species and one resulting from the elution of both homoduplex species. In addition, DHPLC separation is often powerful enough to resolve both duplex types within each group, so that all four duplex species are resolved on the resultant chromatogram.

Finally, SSCP is a further electrophoretic method for the discovery of simple mutations, first described by Orita *et al*. (1989). Separation is based on the specific three-dimensional structures (Fig. 3.2) that are formed when individual nucleic acid strands are allowed to self-anneal under non-denaturing conditions. Prior to electrophoresis, the DNA samples are kept in a single-stranded state by storage in a denaturing loading buffer such as formamide. The samples are then electrophoresed through a non-denaturing gel matrix. Once the strands have entered the electrophoretic medium, re-hybridization into a duplex form is inhibited by the surrounding gel matrix. As the samples migrate away from the denaturant and into increasingly native conditions, each strand will begin to self-anneal. The resulting single-stranded conformations are determined by the sequence context of each molecule in a specific

manner, allowing subtle variations to be resolved from each other (Fig. 3.2). As with DGGE/TGCE and CDGE/CSCE, individual conformations with more branched structures will have reduced mobilities in the electrophoretic medium. Since non-denaturing polymer matrices are supported by many capillary instruments, SSCP is well suited to such platforms and thus benefits from the greater resolution and sensitivity associated with CE, as well as the capacity for automation and dye multiplexing.

Although the above-mentioned techniques are excellent tools for the discovery of unknown genetic variation, their application is limited to scanning short DNA sequences only. The structural differences that may separate variant DNA species are localized within the domain of the nucleic acid where the polymorphic sequence resides. As the length of the sample to be scanned increases, the annealing properties of other domains begin to play a more significant role in determining the overall mobility of each molecule. Eventually a limit is reached in which wild-type and mutant (or mutant-containing) DNA may no longer be reliably resolved from each other due to the diminished effect of their variant domains on electrophoretic mobility. This sequence-length limitation is most prominent when trying to separate homoduplex wild-type from homoduplex mutant DNA (Fig. 3.1B), since their respective melting properties are so similar. However, it should be noted that the physical separation of heteroduplexes from their homoduplex counterparts is still a considerable challenge for SNP detection, since mismatched bases (unlike loop structures) maintain a hydrogen bond attraction to each other, so that overall melting temperature differences are still quite minor. For the basic application of techniques such as DGGE/TGGE, CDGE/CSCE and SSCP, the scan region limit is approximately 200–500 bp (depending on the sequence context of the region to be scanned). However, a specialized innovation was later developed for DGGE, wherein one end of the duplex was appended with a sequence of higher melting temperature (GC clamp) as a means of ensuring that the polymorphic region is located within the low-melting domain of each sample (Myers *et al*., 1985a,b). Here mutant identification may be possible in scanning regions of up to 1000 bp in length for DGGE and other techniques that rely on the physical separation of partially denatured duplexes. Scans of around 1 kb may also be achieved by DHPLC, although this technology is somewhat limited compared to electrophoretic platforms, due to the lack of multichannel arrays for increased sample processing.

Since mutation detection strategies that are based purely on the physical separation of variant DNA structures are limited with regards to the length of the region that can be scanned, techniques which directly target the structural differences associated with the presence of a mutation may offer significant advantages. Both chemical and enzymatic methods are available for the targeting of heteroduplexes directly at the point of mismatch. Such methods include binding of the enzyme (or chemical) to mismatched sites as a means of radically altering the properties of the duplex. The protein MutS, for example, will attach to mismatched DNA so that the bound heteroduplexes will have greatly reduced electrophoretic mobilities under non-denaturing

conditions. Subsequent size fractionation will thus identify mutant-containing pools due to the presence of additional bands that are easily resolved from their wild-type (unbound) counterparts. In addition, MutS may also be used to inhibit exonuclease digestion so that bound heteroduplexes will resist degradation at mismatched regions, whereas their homoduplex counterparts will be completely digested (Ellis *et al*., 1994). Mutant-containing pools are thus easily identified by the presence of bands. Similar results may be obtained when MutS binding is used as a means of terminating primer extension so that shorter bands are present when bound heteroduplexes are used as template for PCR. Chemical alternatives to heteroduplex binding are also available. Carbodiimide will attach to mismatched guanine (G) and thymine (T) bases so that mutant (or mutant-containing) samples may be identified by any of the three above-mentioned techniques (reduced electrophoretic mobility, inhibited exonuclease digestion or termination of PCR primer extension). Additionally, the carbodiimide/mismatch complex also acts as substrate for an ABC excinuclease system of proteins, enabling cleavage of the heteroduplex at the site of mismatch. Hence, when the resultant digestion products are subsequently size-fractionated, heteroduplex-containing samples are identified by the presence of two readily resolvable sub-fragments, the size of which will sum approximately to give the full fragment length. Thus, by using techniques which directly target heteroduplex molecules at the site of the mismatch, the requirement for fine resolution of wild-type and variant DNA species is often eliminated. Similar results may also be obtained when other structural differences associated with the presence of a SNP, such as single-stranded conformations, are directly targeted. Significantly, the lack of requirement for fine resolution is associated with an enhanced ability to scan longer stretches of nucleotide sequence, thus providing two key advantages over purely physical separation methods for SNP detection.

The cleavage of mismatched DNA (Fig. 3.1D) provides a platform for structure-specific mutation scanning with some distinct advantages. Typically, re-hybridization is a simple matter of heat denaturation followed by slow cooling to enhance the formation of heteroduplexes in pools where both mutant and wild-type alleles are present (see above). Following the cleavage step, size fractionation of the resultant products is performed, typically by electrophoresis. Since incision will produce two sub-fragments, the size of which will sum to produce the full length of the sequence under study, the need for fine resolution between similarly sized nucleic acid fragments is eliminated. As a consequence, high-resolution instrumentation for the separation of nucleic acid fragments is not a necessary requirement for mismatch assays, with many successful applications being described for slab gel systems.

Mismatch cleavage techniques comprise the vast majority of tools that enable the separation of subtle DNA variants following structure-specific cleavage events. Mismatch cleavage tools include both enzymatic and chemical techniques. In addition to the carbodiimide/ABC excinuclease system mentioned above, other tools include: the use of S1 nuclease; ribonuclease A; chemical cleavage of mismatch (CCM); DNA N-glycosylases; T4 endonuclease VII (T4E7); the MutH/MutL/MutS (MutHLS) system of

proteins; CEL nuclease; and endonuclease V. A further platform for mutation scanning is that provided by the cleavage of single-stranded conformations (Fig. 3.2). Here, cleavage represents the sole tool for such an approach to mutation scanning.

S1 nuclease

S1 nuclease from *Aspergillus oryzae* was first reported as having the potential to cleave heteroduplex DNA at single-base mismatches by Shenk in 1975 (Shenk *et al*., 1975). Shenk's work involved the successful detection of heteroduplexed DNA from Simian Virus 40, the temperature-sensitive variants of which were assumed to be the result of single-point mutations. S1 and its orthologues, P1 from *Penicillum citrinum* and Mung Bean nuclease, are members of the single-strand-specific (SSS) family of nucleases. The high specificity of such nucleases for single strands only is well supported by a range of evidence. X-ray crystal structures of S1 and P1 nucleases, for example, have revealed that their active sites are located within a protein cleft that is too narrow to accommodate double-stranded DNA or RNA (Suck, 1997). S1 and Mung Bean nucleases have also been shown to cleave single-stranded overhangs (sticky ends) in a quantitative fashion, so that the double-stranded portion of the DNA will remain undigested in its entirety (Ghangas and Wu, 1975). Although SSS nucleases are capable of endonucleolytic cleavage of double-stranded DNA, this is known to be favoured under conditions which disrupt base pairing of the helix, so that individual strands may separate to allow recognition by the enzyme. Such helical disruptions may be generated by intercalating agents (Fuchs, 1975) or conditions which favour denaturation or other conformational transitions (Johnson and Laskowski, 1970).

The highly specific requirement of S1 nuclease and its orthologues to access single strands presents problems for the purpose of mutation scanning. Single-base mismatches will often create only minor strand separation, typically allowing flanking nucleotides to remain strongly bound to their reverse complements. Under these circumstances of minimal duplex separation, SSS nucleases are incapable of mismatch cleavage. As a consequence SSS nucleases have been found to cleave only a small proportion of mismatches, allowing most heteroduplexes to go undetected. Modification of digestion conditions to promote 'breathing' of the duplex may increase the proportion of single-base mismatches detectable by SSS nucleases. The addition of dioxane, for example, has been shown to enable cleavage of mutations which were otherwise undetectable; however, this approach also resulted in end-hydrolysis of the DNA samples under investigation, since fraying of the duplex termini into single strands allowed their subsequent digestion (Howard *et al*., 1999). Similarly, increased enzyme concentration may also enhance the ability to cleave at a site of mismatch, at the sacrifice of generating significant background due to non-specific cleavage – occurring principally in AT-rich regions (Johnson and Laskowski, 1970). S1 and its orthologues are generally active on mismatches of two nucleotides or greater (Dodgson and Wells, 1977), hydro-

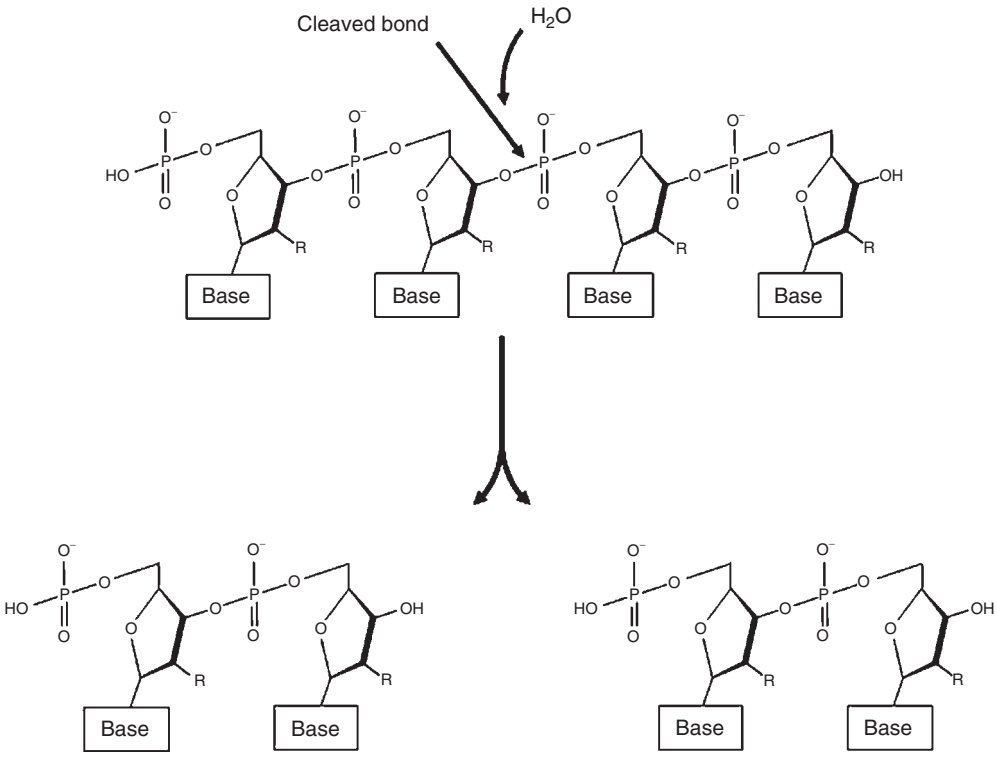


Fig. 3.3A. Cleavage of DNA and RNA by S1 nuclease and its orthologues. Nucleic acid scission occurs by a hydrolysis reaction involving nucleophilic attack of the phosphate by H_2O. The scissile bond is between the 3' oxygen atom and the phosphorous. Reaction products are a 3' hydroxyl and a 5' phosphoryl group. Single-strandedness is a requirement for cleavage. R = H for DNA and OH for RNA.

lysing nucleic acids to form 3' hydroxyl and 5' phosphoryl groups (Fig. 3.3A). The lack of preference for single-base mismatches thus severely limits the application of SSS nucleases to SNP detection.

Ribonuclease A

Ribonuclease A was first applied to the detection of nucleic acid mismatches in 1981 by Freeman and Huang (1981), although in this case RNA heteroduplexes were used as substrate for enzyme cleavage. Based on Shenk's previous work (see above), these authors surmised that RNA duplexes distorted by a mismatch, may be sensitive to cleavage by either S1 nuclease or other enzymes specific for RNA single strands, such as ribonuclease A. Freeman and Huang's study relied on the ability of their viral system to replicate the single-stranded molecule of anti-sense RNA which comprises the entire genome of Vesicular Stomatitis Virus. Labelled messenger RNA (mRNA)

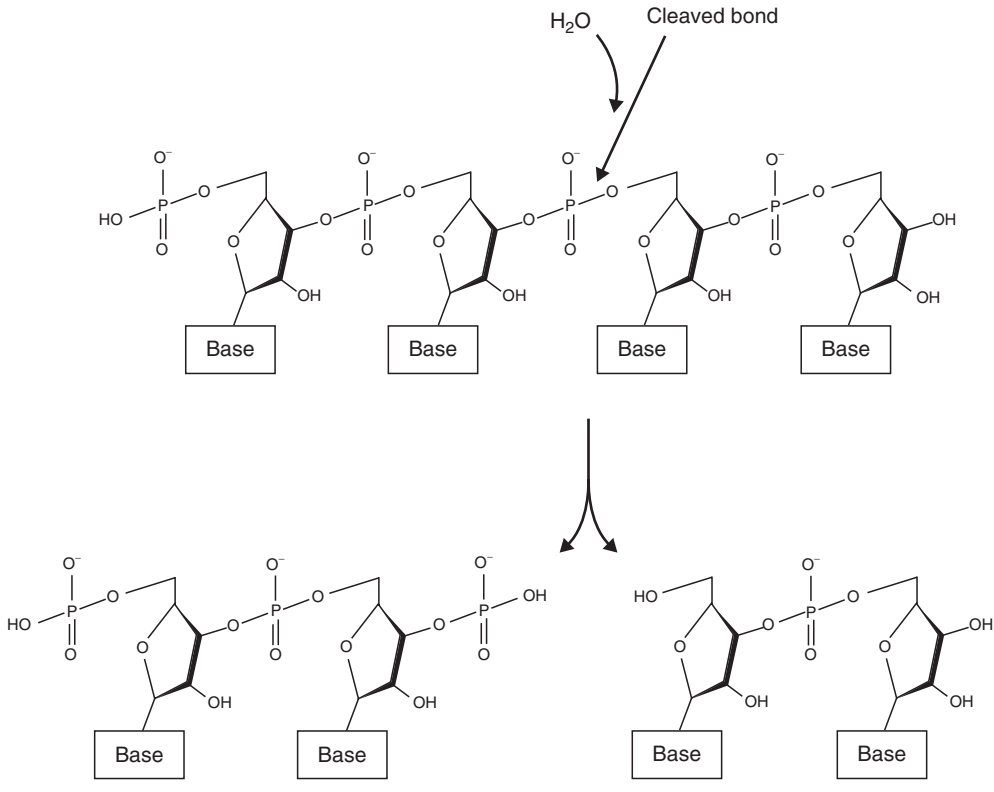


Fig. 3.3B. Cleavage of RNA by ribonuclease A and its orthologues. RNA scission occurs by a hydrolysis reaction. The scissile bond is between the 5' oxygen atom and the phosphorous. Reaction products are a 3' phosphoryl and a 5' hydroxyl group. Single-strandedness is a requirement for cleavage.

was isolated and hybridized with whole genomic RNA and the resultant duplexes digested with ribonuclease A. Similar to Shenk's work, Freeman and Huang used temperature-sensitive mutants to generate their positive control substrates, with the resulting heteroduplexes being successfully cleaved by the ribonuclease. Later advances enabling labelled mRNA to be more readily synthesized via transcription vectors (Melton *et al*., 1984) prompted further innovations of the method. Successful adaptations were subsequently made to mammalian systems using both RNA:RNA (Winter *et al*., 1985) and RNA:DNA (Myers *et al*., 1985c) hybrids, with many of the heteroduplexes tested being cleaved with great efficiency.

Although ribonuclease A demonstrated an improved capacity for the detection of SNPs relative to S1 nuclease, the method was found to be subject to some significant limitations. Ribonuclease A and its orthologues hydrolyse RNA to form 3' phosphoryl and 5' hydroxyl groups (Fig. 3.3B). Ribonuclease A has a strong preference for incising after a pyrimidine residue, often making purine-containing mismatches refractory to incision. Of the 12 possible mismatch types in a RNA:DNA duplex, only four, C:A, C:C, C:T and U:T (RNA:

DNA), were found to be cleaved with high efficiency (Myers *et al*., 1985c). Hence, there is a one in three chance of detecting all mismatch types when only the sense RNA strand is labelled as a probe. Since the same probability would exist if an anti-sense RNA probe was to be constructed, the maximum chance of detecting all mismatch types in an RNA:DNA heteroduplex is about two-thirds. This estimate is considered a minimum, since additional mismatches to the above-mentioned may also be cleaved with lesser efficiency (Myers *et al*., 1985c).

Alternatively, when working with RNA heteroduplexes, if both the sense and anti-sense strands are simultaneously labelled, potentially all mismatch types can be detected, since at least one pyrimidine will be present in each lesion (Cotton, 1997). Although this outcome was impractical upon first development of the method (Winter *et al*., 1985), subsequent technologies have greatly enhanced RNA:RNA mismatch screening. The production of synthetic RNA has benefited greatly from the development of the polymerase chain reaction (PCR) (Saiki *et al*., 1985). PCR-based transcription templates (containing T7 and SP6 promoter sequences), for example, have eliminated the requirement for laborious cloning into transcription vectors. In addition, the treatment of RNA heteroduplexes with intercalating agents (ethidium bromide) has been found to produce substrates that are recognized by two ribonuclease A orthologues, ribonuclease 1 and ribonuclease T1, with high mismatch specificity. Such a modified assay has been described and named non-isotopic ribonuclease cleavage analysis (NIRCA) (Goldrick *et al*., 1996). Ribonucleases 1 and T1 incise both strands at the site of a mismatch to produce blunt-ended cleavage products that are largely refractory to further hydrolysis. The double-stranded nature of the digestion products and the use of ethidium bromide as the intercalating agent are of crucial importance since the capacity to visualize the RNA fragments under UV light is enabled, thus eliminating the need for direct strand labelling. Unlike ribonuclease A, the combination of enzymes used with NIRCA can cleave all intercalated mismatches, and do so with greater efficiency and specificity (Goldrick, 2001). Many NIRCA studies have detected 100% of the mismatches screened and in some cases mutations have been identified that were missed by direct sequencing. However, as with all intercalating approaches to nucleic acid visualization, the signal to noise ratio is limited due to background from non-intercalated dyes. Although few studies have investigated the level of sensitivity of NIRCA, the successful identification of mutant alleles at concentrations as low as 2–4% has been reported (Prescott *et al*., 1999). Such results are exceptional and a higher proportion of mutant alleles (~10%) would be recommended for routine mutation scanning. Thus, the pooling of five diploid samples may be possible, assuming that a single mutant allele within a heterozygous individual can be detected among the remaining nine wild-type alleles. Unfortunately adaptation of NIRCA to a CE platform creates limitations, since the presence of ethidium bromide and residual ribonucleases has been reported to interfere with the capillary injection process (Goldrick, 2001). The emission maxima of ethidium bromide are not compatible with most CE instrumentation, and hence direct labelling of the RNA

with an additional dye is a likely requirement. This, however, would introduce further complications such as high background from residual ethidium bromide and potential energy transfer interactions between dyes. Thus, NIRCA cannot readily benefit from the increased sensitivity and capacity for automation offered by CE platforms.

The refractory nature of single-point mismatches to S1 nuclease, ribonuclease A and their orthologues highlights a significant feature of non-Watson/Crick base pairings. Mismatched nucleotides can still maintain a minor hydrogen bond attraction to each other, such that the local sequence will largely retain its duplexed state. Mismatched bases will only show significant substrate properties when intercalating agents are added to facilitate strand separation. Methods, such as the above, which target mismatches based solely on single-strandedness are therefore fundamentally limited. Fortunately however, mismatched duplexes possess additional properties which may be exploited for the purpose of SNP scanning.

Chemical cleavage of mismatch

Chemical cleavage of mismatch (CCM) is another heteroduplex-based technique which was first applied to mutation scanning in 1988 by Cotton *et al*. (1988). Previous studies had revealed the potential for chemicals to react with the non-paired nitrogenous bases of tRNA (Cramer, 1971). Furthermore, the DNA sequencing protocol of Maxam and Gilbert had reported on the ability of certain chemical treatments to break pyrimidine rings, making the sugar/phosphate backbone susceptible to cleavage at the modified site with alkali (Maxam and Gilbert, 1977). Upon testing a variety of chemicals, Cotton and colleagues discovered that osmium tetroxide and hydroxylamine react with mismatched T and cytosine (C), respectively. The DNA strand was then cleaved at the modified base by subsequent treatment with the alkali, piperidine. Similar to ribonuclease A, the CCM chemistry is highly specific to pyrimidines only. Destabilization of the pyrimidines caused by non-Watson/Crick base pairings provides the mechanism for reactivity, with the chemicals having access to the bases in their non-protected state (Fig. 3.4). The high specificity of CCM for pyrimidines was demonstrated in the authors' original paper since all single-point heteroduplexes tested (13 T and 21 C mismatches) were successfully cleaved. In a follow-up paper, it was discovered that cleavage resulting from purine-only mismatches was possible and in fact quite common. This was presumably due to the transfer of destabilizing forces from the purine mismatch to neighbouring pyrimidines (Cotton and Campbell, 1989). So although mismatches that contain adenine (A) and G only do not react with osmium tetroxide/hydroxylamine, they may distort the duplex sufficiently such that nearby (matched) C and T bases may lose their stability and become reactive. Such effects are obviously related to the sequence context surrounding each mismatch. This 'proximal mismatch' reactivity enhances the specificity of CCM further by increasing the capacity for cleavage. Although cleavage resulting from purine-containing mismatches is not

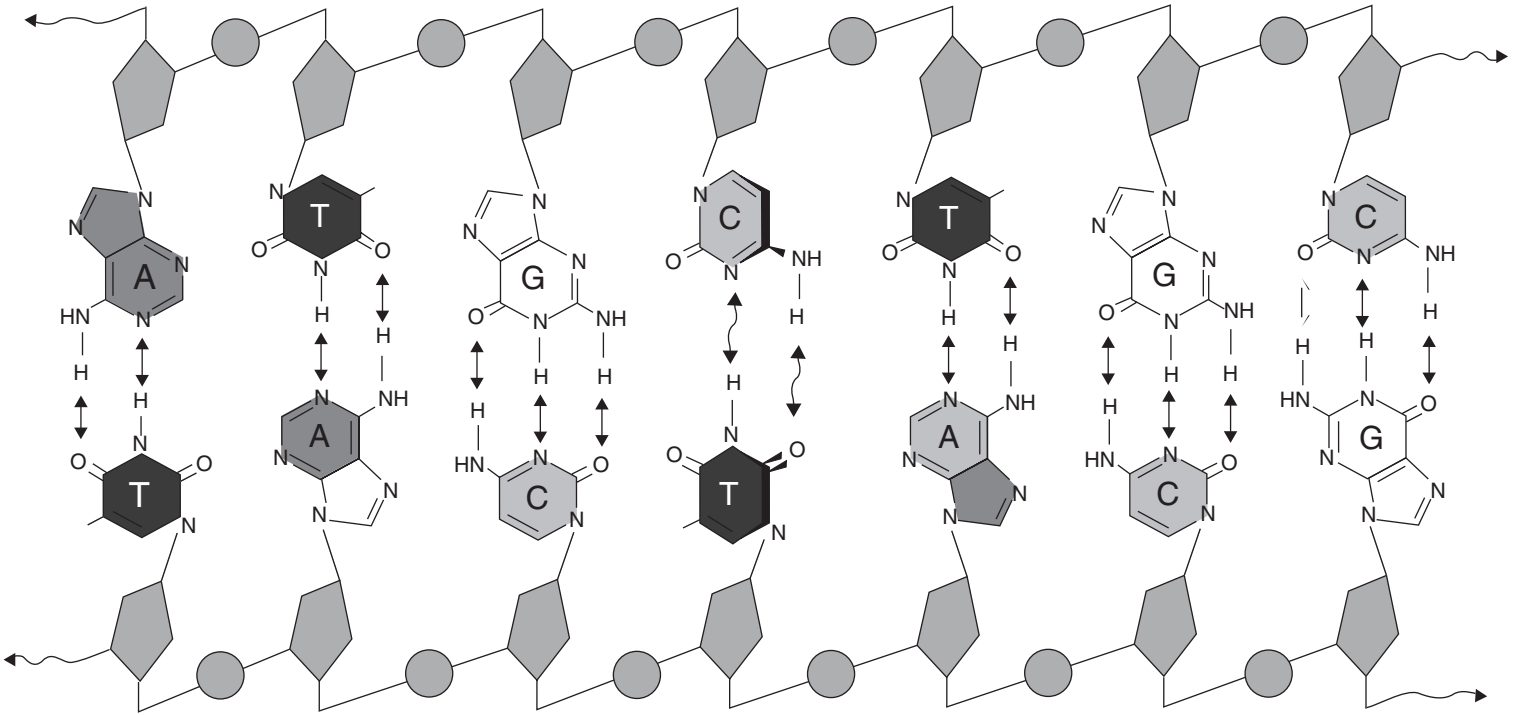


Fig. 3.4. Schematic representation of a single nucleotide mismatch and surrounding matched sequence. The Watson/Crick base pairs of A:T and C:G are stably locked together by two and three hydrogen bonds, respectively. In contrast, the C:T mismatch shown is characterized by reduced hydrogen bonding attractions and unstable geometry. Since nucleic acid strand separation is minimal, single nucleotide mismatches are largely refractory to SSS nucleases. The destabilized nature of a single mismatch may make it susceptible to chemical or enzymatic activity.

favoured by CCM, labelling of the anti-sense strand was proposed as a means of detecting all mutations (Cotton *et al*., 1988). Thus, when a sense probe fails to detect mismatches that contain A and/or G purines, an anti-sense probe will produce complementary T- and/or C-containing mismatches and be susceptible to cleavage. This method works in an identical fashion to that described for ribonuclease treatment of RNA:RNA hybrids and has been successfully applied to CCM (Forrest *et al*., 1991).

The original CCM protocol relied on the generation of sample DNA by vector cloning. Thus, as with many molecular biology applications, the development of PCR technology (Saiki *et al*., 1985) has helped to simplify the method greatly. The CCM method was originally reported as detecting all pyrimidine mismatches with the same level of sensitivity and was thus a significant improvement on the original ribonuclease A technique. The excellent specificity and sensitivity of CCM was quickly exploited finding use with a range of mutation scanning applications (Smooker and Cotton, 1993). Key innovations included the use of RNA:DNA heteroduplexes (Cotton and Wright, 1989), labelling of probes with ^{35}S for improved signal to background (Saleeba and Cotton, 1991) and non-radioactive visualization of the DNA by silver staining (Saleeba *et al*., 1992). The most significant innovation of the CCM technique was the use of fluorescently labelled DNA probes (Haris *et al*., 1994; Verpy *et al*., 1994). Labelling of nucleic acids with fluorescent dyes offers many advantages over other DNA visualization techniques. As mentioned above, the sensitivity of intercalating techniques is limited by high background. Although silver staining and radio-labelling offer an improved signal to noise ratio over intercalating methods, none of these three techniques is compatible with automated DNA analysis such as that offered by many slab gel and capillary instruments. Direct labelling using techniques such as the 5'-modification of PCR primers with fluorescent dyes provides a significantly improved signal to noise ratio and, in addition, benefits from the capacity for multiplexing and automation offered by many DNA analysers. An approach known as fluorescence-assisted mismatch analysis (FAMA) involves the tagging of both sense and anti-sense strands with different fluorescent dyes, so that the strand undergoing cleavage may be identified (Verpy *et al*., 1994). For pyrimidine/purine mismatches, FAMA does not have the capacity to positively identify a cleaved strand by summation of two sub-fragments (to produce the full fragment length), since only the cleavage fragment containing the dye-labelled terminus is detectable. This differs from the original CCM protocol used by Cotton and colleagues in which each probe was internally labelled with ^{32}P. However, if both bases involved in a mismatch are pyrimidines then such summation may be allowed, in a 5'-terminally labelled system, since both strands are cleaved and sub-fragments on both sides of the lesion may be visualized. Alternatively internally labelled fluorescent probes may be generated by PCR, using dye-labelled dNTP substrates, to help positively link any resultant sub-fragments to a mismatch cleavage event (Rowley *et al*., 1995).

The strength of CCM has been demonstrated by its extensive use for the characterization of genetic variation within human disease loci (Ellis *et al*.,

1998). Comparative studies have revealed CCM to have superior specificity to SSCP and CDGE (Condie *et al*., 1993), detecting 100% of mismatches under study. In addition, the superior sensitivity of CCM for detecting low levels of mutant alleles, compared to SSCP and direct sequencing, has been demonstrated (Weghorst *et al*., 1994), with the FAMA technique easily detecting variants when present at as low as 10% in a sample pool (Verpy *et al*., 1994). Such sensitivity should allow pooling of diploid samples up to at least fivefold. CCM also has a greatly enhanced capacity to scan longer stretches of DNA, consistently detecting mismatches in fragments of up to 1000 bp or more. The combined effects of 100% specificity, superior sensitivity and longer scanning reads make CCM an extremely effective SNP discovery method. Although originally CCM was reported to detect all pyrimidine-containing mismatches with equal sensitivity (Cotton, 1989), subsequent results revealed that some T:G mismatches were refractory to cleavage, depending on their surrounding sequence context (Forrest *et al*., 1991; Cotton *et al*., 1993). Various studies have shown this mismatch to be particularly stable, thus explaining its unreactive nature. The refractory nature of the T:G mismatch emphasizes the importance of incorporating an extra step so that the anti-sense strand is labelled in addition to the sense strand. The key drawback of the CCM method is thus the relatively involved protocol, including the preparation of both sense and anti-sense probes and two separate treatment steps with hazardous chemicals.

DNA glycosylases

The use of DNA glycosylases for mismatch cleavage was first reported in 1992 by Lu and Hsu (1992). The base excision repair pathway is an organism's primary tool used for the detection and repair of mutations. The initial step of this pathway involves the excision of the mutated base by a DNA glycosylase. Excision occurs via hydrolysis of the glycosidic bond which links the nucleobase to its sugar residue within the DNA backbone. The resulting abasic site is referred to as apurinic/apyrimidinic (AP). AP sites are scissile, often being hydrolysed by a separate enzyme with AP endonuclease activity. However, certain glycosylases, such as MutY from *Escherichia coli*, possess an additional lyase activity in which the AP sugar is subsequently removed (Fig. 3.5). Lu and Hsu made use of the dual glycosylase/AP lyase activity of MutY as a means of cleaving DNA at specific sites of mismatch. MutY activity is highly specific for A:G (and to a lesser extent, A:C) mismatches, cleaving the A base to trigger subsequent scission of the resultant AP strand. Sample heteroduplexes were generated by PCR, radio-labelled and digested by the glycosylase. The excellent specificity of MutY for its A:G substrate and absence of background cleavage at matched sites allowed for an extremely sensitive assay in which mutant alleles were confidently detected when present at concentrations as low as 1–2%. Lu and Hsu named their method mismatch repair enzyme cleavage (MREC). The detection rate of MREC is limited. The following list comprises all possible sets of single-base pair mismatches – A:C, T:C,

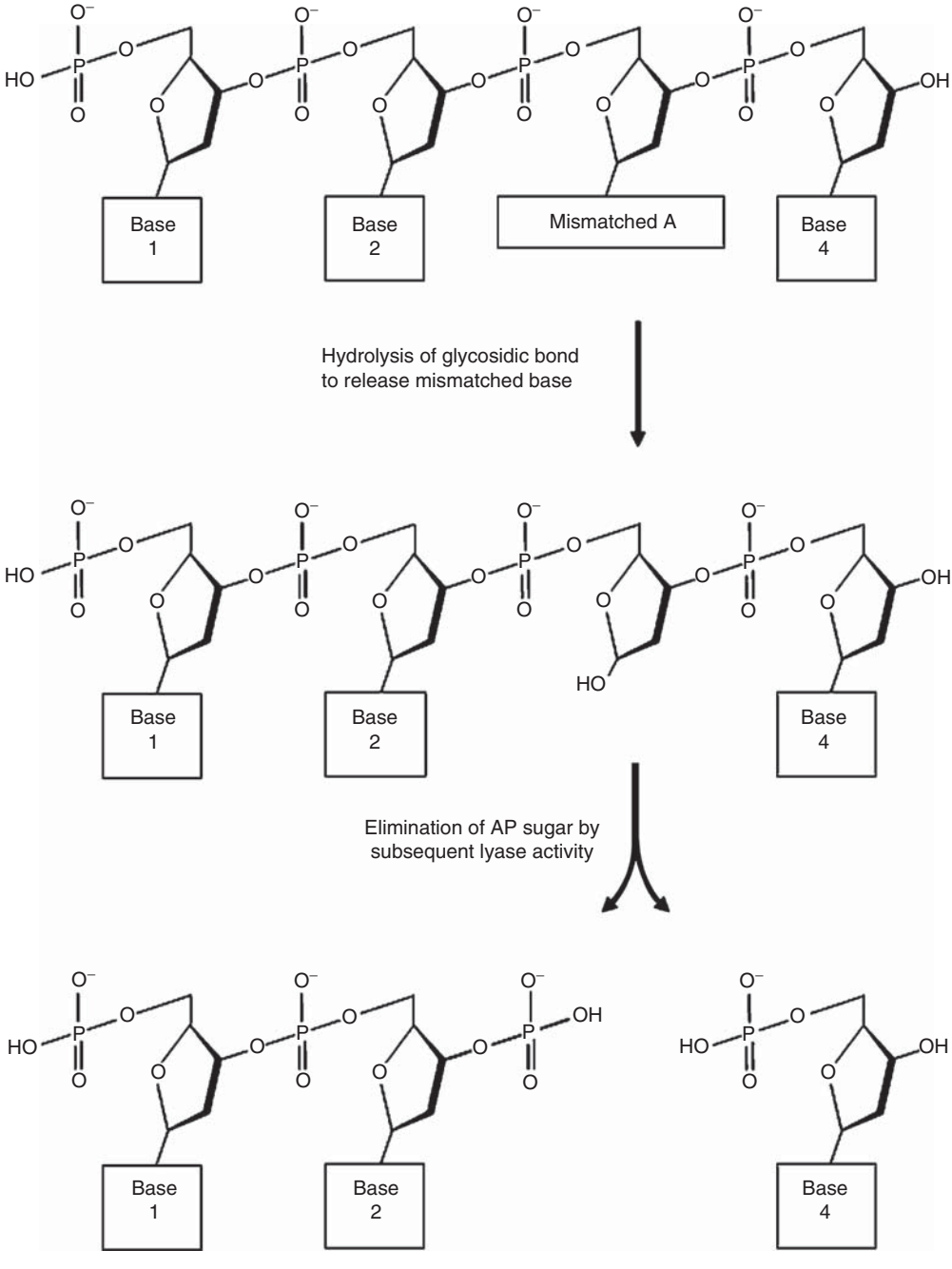


Fig. 3.5. Cleavage of mismatches by MutY. The mismatched adenine is recognized and excised specifically by MutY. The AP site is then cleaved on both sides by a separate lyase activity of MutY to produce a single nucleotide gap with 3' and 5' phosphate ends.

A:G, T:G, A:A, T:T, G:G and C:C (Cotton, 1997). It can be seen that MutY is capable of detecting one-quarter of these mismatches when a sense strand is considered and an additional one-quarter when the anti-sense strand is also screened (T:C and T:G being detected as A:C and A:G). Thus, only half of all mismatches may be detected by MutY. This fact has contributed towards MREC having only limited application to the discovery of unknown mutations; however, the use of alternative DNA glycosylases (Wiebauer and Jiricny, 1990) to broaden the mismatch cleavage potential of the technique was suggested. The lack of availability of a commercial source of MutY has also contributed to its rare usage since the original publication.

T4 endonuclease VII

T4 endonuclease VII was first proposed as a tool for mutation detection in 1993 by Youil *et al*. (1993). Here the authors reported on a pilot experiment in which examples from each of the four possible mismatch types (G:A/C:T, C:C/G:G, A:A/T:T and C:A/G:T) were successfully cleaved upon treatment with the endonuclease. This original study was later expanded to report that 17 of 18 single-base mutations were detectable by the enzyme cleavage method (Youil *et al*., 1995). They named their assay enzyme mismatch cleavage (EMC). Similar findings were reported soon afterwards using T4 endonuclease VII in conjunction with another bacteriophage resolvase, T7 endonuclease I (Mashal *et al*., 1995).

T4 endonuclease VII has been shown to cleave a range of branched DNA structures including Holliday junctions, cruciforms, forks, strand overhangs and, most significantly, a range of mismatches. The enzyme scans DNA until an aberrant site is recognized. A nick is produced in one strand followed by a counter-nick in the complementary strand, so that two duplex subfragments are typically produced. The two incisions are hydrolysis reactions (Fig. 3.3A) occurring up to 3 bp 3' of the aberrant region in each strand so that a ~2–6-nucleotide overhang results on each cleaved sub-fragment. When the local nucleotide sequence surrounding the aberrant region is fixed, the enzyme shows little discrimination between different classes of branched DNA, cleaving overhangs, forks and Y-structures with equal efficiency (Pottmeyer and Kemper, 1992). The nature of the nucleotide sequence at the point of branching is more significant, often resulting in varied cleavage efficiencies. X-ray structures of T4 endonuclease VII have suggested that duplexes containing a local perturbation may be capable of bending so as to bind appropriately to the enzyme to enable cleavage (Suck, 1997). This model explains the broad specificity of the enzyme for a range of 'bent' substrates, with certain sequences at the point of branching capable of bending more than others.

A clear demonstration of the specificity and sensitivity of EMC was demonstrated in follow-up work by Youil and colleagues in which 81 of 81 known point mutations were detected (Youil *et al*., 1996). The original EMC method involved the use of radio-labelled wild-type DNA, hybridized to an excess of

unlabelled test DNA. These methods, however, resulted in strong autoradiography background due to cleavage of matched, labelled DNA (much of which is caused by homoduplex wild-type species). The non-specific cleavage is attributed to sequence-induced duplex bends that act as substrate for the enzyme. To reduce background problems, the test PCR products were biotin-labelled and the wild-type PCR products radio-labelled. Hence, following re-hybridization, heteroduplexes of test- and wild-type strands could be separated from the background-enhancing wild-type homoduplexes by streptavidin and the resulting background was significantly reduced (Babon *et al*., 1995).

Significant enhancement in the sensitivity of T4 endonuclease scanning was later achieved by the use of fluorescently labelled PCR primers and LIF electrophoresis, in a technique named enzymatic mutation detection (EMD) (Del Tito *et al*., 1998). Unlike other techniques, T4 endonuclease VII is readily adaptable to the LIF electrophoresis platform. For EMD, the sensitivity of detection of fluorescein-labelled amplicon strands by an ABI 377 DNA analyzer is successfully exploited, readily detecting mutant alleles when diluted in a pool as low as 5%. Furthermore, the number of steps involved in EMD is greatly reduced relative to radio-labelling, involving the simple amplification of sample DNA with 5'-modified primers followed by re-hybridization and CE. The strength of the EMD assay has been attested by a number of successful applications, particularly to disease research (Inganas *et al*., 2000; Otway *et al*., 2000), with excellent levels of sensitivity being consistently featured. The inherent problem of non-specific background has limited the EMD's specificity to around ~80–90% of mismatches in a number of studies. One study in particular deemed EMD insufficient for the screening of p53 (Norberg *et al*., 2001).

CEL nuclease

CEL nuclease was first applied to the detection of mutations in 1998 by Oleykowski *et al*. (1998). Here, the authors reported the discovery of a novel nuclease from celery that hydrolyses duplex DNA at single-base mismatches and other destabilized or single-stranded regions. The nuclease was designated CEL I. Enzymes of similar function were also found in other plant species and the general term CEL nuclease has since been coined to describe all CEL I orthologues. Incision by CEL nuclease was found to occur primarily at the phosphodiester bond immediately 3' of the mismatch, resulting in hydrolysis products identical to those produced by SSS nucleases (Fig. 3.3A). CEL nuclease was found to possess a number of distinct properties useful for the purpose of mutation detection. In contrast to other mismatch-specific enzymes, CEL nuclease was found to cleave all mismatch types with excellent efficiency, thus providing a 100% detection rate. Non-specific scission of matched nucleotides was insignificant and mismatch cleavage was unaffected by sequence context, even when flanked by GC-rich regions. CEL nuclease shows no particular preference for nicking one side of the mismatch

over the other, so that both strands are incised at equal rates. Furthermore, the activity of the enzyme may be controlled such that only one strand is cut per heteroduplex molecule. This feature is especially pertinent in the context of the mismatch detection assay reported by the authors. Here, PCR-amplified samples are generated so that a different fluorescein label is attached to each 5' terminus. As with typical mismatch protocols, multi-sample pools are re-hybridized before enzyme treatment to enhance the formation of heteroduplexes. With conditions set to favour the cutting of just one side of the mismatch, and the selection of this side being random for each cleavage event, CEL nuclease digestion will result in two labelled fragments, the size of which will sum to produce the full fragment length. This capacity for two independent cleavage events that complement each other to confirm the presence of a mutation (Oleykowski *et al*., 1998) is in contrast to other techniques such as ribonuclease A digestion of RNA:RNA hybrids or the FAMA innovation of CCM, in cases where only one pyrimidine is located in a mismatch and hence only one strand is capable of being cleaved.

The strength of the CEL nuclease mutation scanning assay has been demonstrated by its extensive use as the screening method of choice for the powerful reverse genetics technology of targeting induced local lesions in genomes (TILLING) (Colbert *et al*., 2001; Till *et al*., 2006). Here, Oleykowski's labelling method of 5'-modification with fluoresceins has been substituted with infrared dyes to reliably enable the detection of one heterozygous mutant in a total pool size of eight diploid individuals (6.25%). Although few innovations have been made since its first inception, adaptation of the protocol to include dye-labelled universal primers that may hybridize to 5' end tails of otherwise gene-specific primers have been employed to reduce the costs associated with direct labelling of each locus (Wienholds *et al*., 2003; Till *et al*., 2006). CEL nuclease's 100% specificity for mismatches, excellent sensitivity for low levels of substrate and simplicity of application, rate it among the most effective of all mismatch cleavage methods. However, limitations of the use of this enzyme have been reported. A significant problem associated with the use of CEL nuclease became apparent when simplified protocols were developed for ethidium detection systems. Normally, the sensitivity of fluorescent electrophoresis is many fold superior to that of ethidium-intercalated/agarose detection systems; however, the maximum pooling capacities resulting from the former (Wienholds *et al*., 2003; Qiu *et al*., 2004; Till *et al*., 2006) and latter (Greber *et al*., 2005; Sato *et al*., 2006; Raghavan *et al*., 2007) are noted as being similar. The loss of signal from 5'-labelled amplicons has been hypothesized to explain the relatively disappointing sensitivity of this means of DNA detection (Yeung *et al*., 2005). Upon further investigation, our group reported data which suggested an exonucleolytic activity of CEL nuclease wherein 5' signal is rapidly degraded from homoduplex (matched) controls (Cross *et al*., 2008). Furthermore, it was also noted that internally labelled DNA probes will largely retain their signal strength under conditions where 5' end signal is degraded. With a proof-of-concept demonstrated, our group developed an alternative protocol for mismatch screening with CEL nuclease which was termed endonucleolytic mutation analysis by internal labelling (EMAIL)

(Cross *et al*., 2008). EMAIL exploits the ability of DNA polymerases to incorporate fluorescently labelled nucleotide substrates into PCR extension products. The resultant DNA probes contain random substitutions of the labelled nucleotide analogue for site normally occupied by the natural nucleotide. Since the internal signals are unaffected by nuclease digestion, EMAIL provides an increased signal strength relative to that obtained by the standard 5'-labelled approach. As an added feature, the background associated with internal labelling is significantly improved compared to the 5'-approach. With more robust signal and reduced background, EMAIL is expected to facilitate further improvement of mismatch screening. Furthermore, EMAIL is expected to find application as a beneficial labelling method for other mismatch endonucleases, many of which are reported as being limited by their degradation of 5' end signal.

References

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815.

Babon, J.J., Youil, R. and Cotton, R.G. (1995) Improved strategy for mutation detection – a modification to the enzyme mismatch cleavage method. *Nucleic Acids Research* 23, 5082–5084.

Borevitz, J.O., Liang, D., Plouffe, D., Chang, H.S., Zhu, T., Weigel, D., Berry, C.C., Winzeler, E. and Chory, J. (2003) Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Research* 13, 513–523.

Colbert, T., Till, B.J., Tompa, R., Reynolds, S., Steine, M.N., Yeung, A.T., McCallum, C.M., Comai, L. and Henikoff, S. (2001) High-throughput screening for induced point mutations. *Plant Physiology* 126, 480–484.

Condie, A., Eeles, R., Borresen, A.L., Coles, C., Cooper, C. and Prosser, J. (1993) Detection of point mutations in the p53 gene: comparison of single-strand conformation polymorphism, constant denaturant gel electrophoresis, and hydroxylamine and osmium tetroxide techniques. *Human Mutation* 2, 58–66.

Cotton, R.G. (1989) Detection of single base changes in nucleic acids. *Journal of Biochemistry* 263, 1–10.

Cotton, R.G.H. (1997) *Mutation Detection*. Oxford University Press, Oxford.

Cotton, R.G. and Campbell, R.D. (1989) Chemical reactivity of matched cytosine and thymine bases near mismatched and unmatched bases in a heteroduplex between DNA strands with multiple differences. *Nucleic Acids Research* 17, 4223–4233.

Cotton, R.G. and Wright, P.J. (1989) Rapid chemical mapping of dengue virus variability using RNA isolated directly from cells. *Journal of Virological Methods* 26, 67–76.

Cotton, R.G., Rodrigues, N.R. and Campbell, R.D. (1988) Reactivity of cytosine and thymine in single-base-pair mismatches with hydroxylamine and osmium tetroxide and its application to the study of mutations. *Proceedings of the National Academy of Sciences of the USA* 85, 4397–4401.

Cotton, R.G., Dahl, H.H., Forrest, S., Howells, D.W., Ramus, S.J., Bishop, R.E., Dianzani, I., Saleeba, J.A., Palombo, E., Anderson, M.J. *et al.* (1993) Analysis of sequence contexts flanking T.G mismatches leads to predictions about reactivity of the mismatched T to osmium tetroxide. *DNA Cell Biology* 12, 945–949.

Cramer, F. (1971) Three-dimensional structure of tRNA. *Progress in Nucleic Acid Research and Molecular Biology* 11, 391–421.

Cross, M.J., Waters, D.L.E., Lee, L.S. and Henry, R.H. (2008) Endonucleolytic mutation analysis by internal labeling (EMAIL). *Electrophoresis* (in press).

Del Tito, B.J., Jr, Poff, H.E. III, Novotny, M.A., Cartledge, D.M., Walker, R.I. II, Earl, C.D. and Bailey, A.L. (1998) Automated fluorescent analysis procedure for enzymatic mutation detection. *Clinical Chemistry* 44, 731–739.

Dodgson, J.B. and Wells, R.D. (1977) Action of single-strand specific nucleases on model DNA heteroduplexes of defined size and sequence. *Biochemistry* 16, 2374–2379.

Ellis, L.A., Taylor, G.R., Banks, R. and Baumberg, S. (1994) MutS binding protects heteroduplex DNA from exonuclease digestion *in vitro*: a simple method for detecting mutations. *Nucleic Acids Research* 22, 2710–2711.

Ellis, T.P., Humphrey, K.E., Smith, M.J. and Cotton, R.G. (1998) Chemical cleavage of mismatch: a new look at an established method. *Human Mutation* 11, 345–353.

Esteban-Cardenosa, E., Duran, M., Infante, M., Velasco, E. and Miner, C. (2004) High-throughput mutation detection method to scan BRCA1 and BRCA2 based on heteroduplex analysis by capillary array electrophoresis. *Clinical Chemistry* 50, 313–320.

Fischer, S.G. and Lerman, L.S. (1983) DNA fragments differing by single base-pair substitutions are separated in denaturing gradient gels: correspondence with melting theory. *Proceedings of the National Academy of Sciences of the USA* 80, 1579–1583.

Forrest, S.M., Dahl, H.H., Howells, D.W., Dianzani, I. and Cotton, R.G. (1991) Mutation detection in phenylketonuria by using chemical cleavage of mismatch: importance of using probes from both normal and patient samples. *American Journal of Human Genetics* 49, 175–183.

Freeman, G.J. and Huang, A.S. (1981) Mapping temperature-sensitive mutants of vesicular stomatitis virus by RNA heteroduplex formation. *Journal of Genetic Virology* 57, 103–117.

Fuchs, R.P. (1975) *In vitro* recognition of carcinogen-induced local denaturation sites native DNA by S1 endonuclease from *Aspergillus oryzae*. *Nature* 257, 151–152.

Ghangas, G.S. and Wu, R. (1975) Specific hydrolysis of the cohesive ends of bacteriophage lambda DNA by three single strand-specific nucleases. *Journal of Biological Chemistry* 250, 4601–4606.

Goldrick, M.M. (2001) RNase cleavage-based methods for mutation/SNP detection, past and present. *Human Mutation* 18, 190–204.

Goldrick, M.M., Kimball, G.R., Liu, Q., Martin, L.A., Sommer, S.S. and Tseng, J.Y. (1996) NIRCA: a rapid robust method for screening for unknown point mutations. *BioTechniques* 21, 106–112.

Greber, B., Tandara, H., Lehrach, H. and Himmelbauer, H. (2005) Comparison of PCR-based mutation detection methods and application for identification of mouse Sult1a1 mutant embryonic stem cell clones using pooled templates. *Human Mutation* 25, 483–490.

Haris, I.I., Green, P.M., Bentley, D.R. and Giannelli, F. (1994) Mutation detection by fluorescent chemical cleavage: application to hemophilia B. *PCR Methods and Application* 3, 268–271.

Howard, J.T., Ward, J., Watson, J.N. and Roux, K.H. (1999) Heteroduplex cleavage analysis using S1 nuclease. *BioTechniques* 27, 18–19.

Inganas, M., Byding, S., Eckersten, A., Eriksson, S., Hultman, T., Jorsback, A., Lofman, E., Sabounchi, F., Kressner, U., Lindmark, G. *et al*. (2000) Enzymatic mutation detection in the P53 gene. *Clinical Chemistry* 46, 1562–1573.

Jander, G., Norris, S.R., Rounsley, S.D., Bush, D.F., Levin, I.M. and Last, R.L. (2002) Arabidopsis map-based cloning in the post-genome era. *Plant Physiology* 129, 440–450.

Johnson, P.H. and Laskowski, M., Sr (1970) Mung bean nuclease I. II. Resistance of double stranded deoxyribonucleic acid and susceptibility of regions rich in adenosine and thymidine to enzymatic hydrolysis. *Journal of Biological Chemistry* 245, 891–898.

Kwok, P.Y. and Chen, X. (2003) Detection of single nucleotide polymorphisms. *Current Issues in Molecular Biology* 5, 43–60.

Lu, A.-L. and Hsu, I.-C. (1992) Detection of single DNA base mutations with mismatch repair enzymes. *Genomics* 14, 249–255.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z.T. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.

Mashal, R.D., Koontz, J. and Sklar, J. (1995) Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. *Nature Genetics* 9, 177–183.

Maxam, A.M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the USA* 74, 560–564.

Melton, D.A., Krieg, P.A., Rebagliati, M.R., Maniatis, T., Zinn, K. and Green, M.R. (1984) Efficient *in vitro* synthesis of biologically active RNA and RNA hybridization probes from plasmids containing a bacteriophage SP6 promoter. *Nucleic Acids Research* 12, 7035–7056.

Myers, R.M., Fischer, S.G., Lerman, L.S. and Maniatis, T. (1985a) Nearly all single base substitutions in DNA fragments joined to a GC-clamp can be detected by denaturing gradient gel electrophoresis. *Nucleic Acids Research* 13, 3131–3145.

Myers, R.M., Fischer, S.G., Maniatis, T. and Lerman, L.S. (1985b) Modification of the melting properties of duplex DNA by attachment of a GC-rich DNA sequence as determined by denaturing gradient gel electrophoresis. *Nucleic Acids Research* 13, 3111–3129.

Myers, R.M., Larin, Z. and Maniatis, T. (1985c) Detection of single base substitutions by ribonuclease cleavage at mismatches in RNA:DNA duplexes. *Science* 230, 1242–1246.

Myers, R.M., Lumelsky, N., Lerman, L.S. and Maniatis, T. (1985d) Detection of single base substitutions in total genomic DNA. *Nature* 313, 495–498.

Norberg, T., Klaar, S., Lindqvist, L., Lindahl, T., Ahlgren, J. and Bergh, J. (2001) Enzymatic mutation detection method evaluated for detection of p53 mutations in cDNA from breast cancers. *Clinical Chemistry* 47, 821–828.

Oleykowski, C.A., Mullins, C.R.B., Godwin, A.K. and Yeung, A.T. (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Research* 26, 4597–4602.

Orita, M., Suzuki, Y., Sekiya, T. and Hayashi, K. (1989) Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. *Genomics* 5, 874–879.

Otway, R., Tetlow, N., Hornby, J. and Kohonen-Corish, M. (2000) Evaluation of enzymatic mutation detection trade mark in hereditary nonpolyposis colorectal cancer. *Human Mutation* 16, 61–67.

Pottmeyer, S. and Kemper, B. (1992) T4 endonuclease VII resolves cruciform DNA with nick and counter-nick and its activity is directed by local nucleotide sequence. *Journal of Molecular Biology* 223, 607–615.

Prescott, J., Patel, H., Tillman, S., McHugh, T. and Ralph, D. (1999) Cleavage of double-stranded copy RNA by RNase 1 and RNase T1 provides a robust means to detect p53 gene mutations in clinical specimens. *Electrophoresis* 20, 1149–1161.

Qiu, P., Shandilya, H., D'Alessio, J.M., O'Connor, K., Durocher, J. and Gerard, G.F. (2004) Mutation detection using surveyor (TM) nuclease. *BioTechniques* 36, 702–707.

Raghavan, C., Naredo, M.E.B., Wang, H.H., Atienza, G., Liu, B., Qiu, F.L., McNally, K.L. and Leung, H. (2007) Rapid method for detecting SNPs on agarose gels and its application in candidate gene mapping. *Molecular Breeding* 19, 87–101.

Rowley, G., Saad, S., Giannelli, F. and Green, P.M. (1995) Ultrarapid mutation detection by multiplex, solid-phase chemical cleavage. *Genomics* 30, 574–582.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. and Arnheim, N. (1985) Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* 230, 1350–1354.

Saleeba, J.A. and Cotton, R.G. (1991) 35S-labelled probes improve detection of mismatched base pairs by chemical cleavage. *Nucleic Acids Research* 19, 1712.

Saleeba, J.A., Ramus, S.J. and Cotton, R.G. (1992) Complete mutation detection using unlabeled chemical cleavage. *Human Mutation* 1, 63–69.

Sato, Y., Shirasawa, K., Takahashi, Y., Nishimura, M. and Nishio, T. (2006) Mutant selection from progeny of gamma-ray-irradiated rice by DNA heteroduplex cleavage using *Brassica* petiole extract. *Breeding Science* 56, 179–183.

Shenk, T.E., Rhodes, C., Rigby, P.W.J. and Berg, P. (1975) Biochemical method for mapping mutational alterations in DNA with S1 nuclease: the location of deletions and temperature-sensitive mutations in simian virus 40. *Proceedings of the National Academy of Sciences of the USA* 72, 989–993.

Smooker, P.M. and Cotton, R.G. (1993) The use of chemical reagents in the detection of DNA mutations. *Mutation Research* 288, 65–77.

Suck, D. (1997) DNA recognition by structure-selective nucleases. *Biopolymers* 44, 405–421.

Till, B.J., Zerr, T., Bowers, E., Greene, E.A., Comai, L. and Henikoff, S. (2006) High-throughput discovery of rare human nucleotide polymorphisms by ecotilling. *Nucleic Acids Research* 34, 5352–5352.

Underhill, P.A., Jin, L., Lin, A.A., Mehdi, S.Q., Jenkins, T., Vollrath, D., Davis, R.W., Cavalli-Sforza, L.L. and Oefner, P.J. (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Research* 7, 996–1005.

Verpy, E., Biasotto, M., Meo, T. and Tosi, M. (1994) Efficient detection of point mutations on color-coded strands of target DNA. *Proceedings of the National Academy of Sciences USA* 91, 1873–1877.

Weghorst, C.M., Dragnev, K.H., Buzard, G.S., Thorne, K.L., Vandeborne, G.F., Vincent, K.A. and Rice, J.M. (1994) Low incidence of point mutations detected in the P53 tumor-suppressor gene from chemically-induced rat renal mesenchymal tumors. *Cancer Research* 54, 215–219.

Wiebauer, K. and Jiricny, J. (1990) Mismatch-specific thymine DNA glycosylase and DNA polymerase beta mediate the correction of G.T mispairs in nuclear extracts from human cells. *Proceedings of the National Academy of Sciences USA* 87, 5842–5845.

Wienholds, E., van Eeden, F., Kosters, M., Mudde, J., Plasterk, R.H.A. and Cuppen, E. (2003) Efficient target-selected mutagenesis in zebrafish. *Genome Research* 13, 2700–2707.

Winter, E., Yamamoto, F., Almoguera, C. and Perucho, M. (1985) A method to detect and characterize point mutations in transcribed genes: amplification and overexpression of the mutant c-Ki-ras allele in human tumor cells. *Proceedings of the National Academy of Sciences of the USA* 82, 7575–7579.

Yeung, A., Hattangadi, D., Blakesley, L. and Nicolas, E. (2005) Enzymatic mutation detection technologies. *BioTechniques* 38, 749–758.

Youil, R., Kemper, B.W. and Cotton, R.G.H. (1993) Screening for mutations by enzyme cleavage of mismatch using T4 Endonuclease VII. *American Journal of Human Genetics* 53, 1257–1257.

Youil, R., Kemper, B.W. and Cotton, R.G. (1995) Screening for mutations by enzyme mismatch cleavage with T4 endonuclease VII. *Proceedings of the National Academy of Sciences USA* 92, 87–91.

Youil, R., Kemper, B. and Cotton, R.G. (1996) Detection of 81 of 81 known mouse beta-globin promoter mutations with T4 endonuclease VII–the EMC method. *Genomics* 32, 431–435.

4 Sequence Polymorphisms in the Flanking Regions of Microsatellite Markers

G. Ablett and R.J. Henry

Introduction

SNP markers may be discovered using many strategies. In this chapter we describe the approach of finding SNP markers in the flanking sequences surrounding microsatellites (simple sequence repeats (SSRs)) (Ablett *et al*., 2006). A major advantage of this strategy is that SSR markers have been widely studied and mapped in plant genomes, thus providing a readily available source of genetic loci that have already been mapped and that are readily convertible to SNP markers.

Genetic markers are used in modern breeding programmes for variety identification (Plaschke *et al*., 1995) and marker-assisted selection (Harker *et al*., 2001) and in evolutionary studies (Matsuoka *et al*., 2002). A genetic marker is a known sequence difference (or potential difference) between genomic sources such as varieties or species. A marker is not necessarily within a gene of interest but is close enough to it for the marker and the gene to be inherited together and very rarely to be separated by events such as crossovers. Genetic analysis takes advantage of the known sequence differences between genomic sources. A relationship between these differences and phenotypic outcomes is valued when breeding for specific outcomes. For a marker to become useful it needs to be specifically associated with a variety or with a trait. Once a marker has been discovered it is associated with traits or given a chromosomal location (mapped) in relation to other chromosomal markers and traits. This has been done in many species including wheat (Dubcovsky *et al*., 1996; Roder *et al*., 1998; Somers *et al*., 2004) and barley (Ablett *et al*., 2003). SSRs have been markers of choice because they are highly informative, easy to use and many are now well characterized.

SSR Analysis

SSRs are sequences of 1–6 bases which repeat sequentially several times. To be developed as markers these repeats and the bases surrounding them

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

(flanking regions) are sequenced. Roder *et al*. (1995) used a lambda phage library. They found a frequency of one microsatellite per 270 kb for GA plus GT repeats. The sequences in the flanking regions are used to develop PCR primers to amplify a region around the repeat region. The reason SSRs are useful is because they are 'hyper-variable'. They mutate at a high rate by gaining and losing repeat units by 'DNA replication slippage' (Dieringer and Schlotterer, 2003). The rate of mutation depends on several factors including number of repeats, class of repeat (di-, tri-, etc.), GC content, chromosomal location or location in relation to a gene. The repeat region is much more variable than the flanking region. Thus, while primers designed to the flanking regions may amplify this section of DNA in variously related germplasm, the product size may vary due to changes to the number of repeats and therefore re-sequencing is not necessary in order to find polymorphisms. The product size polymorphism, which is usually caused by differences in repeat numbers, may at times be caused by indels in the region flanking the repeat region (Curtu *et al*., 2004, who investigated European Oak). This is especially so since 'microsats are often located in highly repetitive regions that are subject to high rates of mutation' but will be more common when more divergent varieties or species are studied. For this reason size homoplasy is more likely to pose a problem in conservation biology than in agronomic studies (Estoup *et al*., 2002). 'Size homoplasy' is the occurrence of alleles appearing monomorphic by marker size when sequencing would show sequence polymorphism due to indels. Polymorphisms including indels have even been found in the flanking regions of chloroplast SSRs in wheat (Matsuoka *et al*., 2005).

Why SNP Markers

The two main reasons for using SNP markers are the ability of new technology to analyse many markers at the same time (highly parallel) and the high frequency of SNPs.

By sequencing, polymorphisms may be found between varieties which may be exploited. The sequence variation may be due to a base change (an 'SNP') or an inclusion or deletion of one or more bases (an 'indel'). SNPs are found by comparing genetic sequences. This can be done by comparing published sequences, *de novo* sequencing or re-sequencing (when a sequence that is known in one germplasm is compared by re-sequencing in another). Discovering a SNP requires sequencing which can be costly because of the preparative processes of developing primers, PCR conditions, sequencing and analysing. An alternative is comparing published sequences such as the many thousands of ESTs for common grains (Somers *et al*., 2003). Somers *et al*. found in wheat that SNPs between genotypes occurred at one per 540 bp between genomes in the 12 varieties that they tested. Validation is required, however, especially for *in silico* SNP discovery because differences may be caused by sequencing errors. Validation means developing an analytical method to test the SNP (Sommers *et al*., 2003 and other chapters, this volume). *In silico* SNP discovery in wheat is made more complicated because of

the very large hexaploid genome and therefore polymorphisms between the genomes within a variety are common and must be distinguished from those between varieties. The two types of intra-genomic SNP are homoeologous SNPs between the three genomes that make up the hexaploid of wheat and paralogous SNPs between gene families of a single genome (Fig. 4.1). Finally, for the marker to be useful it must be characterized. Characterizing the marker means making the marker useful by associating the marker with a variety, a trait or by mapping the marker to a specific chromosome location (Ablett *et al*., 2003).

Why Use SSR Markers to Find SNPs

There are many publicly available published SSR markers. These have primers and established PCR methods. They have also been characterized by mapping or trait association. This means that a large part of the work has already been done. If SNPs can be found in these sequences, well-characterized SNP markers can be easily developed. There was some evidence that the flanking regions contain sequence polymorphisms. Null alleles may be due to sequence polymorphisms within the primer binding sites (Curtu *et al*., 2004). Also analysis of PCR product size polymorphism often reveals single base pair differences.

The presence of SNPs in barley microsatellite markers was explored in our laboratory. When good sequence was available for several barley varieties, SNPs were commonly found (P. Bundock, Lismore, 2006, unpublished data). Unfortunately good sequencing data was not always found in the short PCR products of the microsatellites. Even though barley is diploid with a fairly small genome, few of the markers yielded adequate results. The problem was that good-quality sequencing did not start until about 30 bases in from the sequencing primer. Good quality was often lost after the repeat region. This left only a small series of bases in which to look for polymorphisms. Frequent polymorphisms in the flanking regions of SSR were, however, confirmed. Rossetto *et al*. (2002) used SNPs in the flanking regions of SSR to examine relationships in

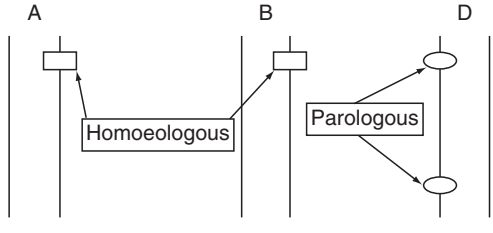


Fig. 4.1. The hexaploid nature of wheat complicates the search for a SNP in ESTs because of the presence of homoeologous genes, which are similar genes found on the same location of two or all three of the genomes. Paralogous genes are found on different locations of the same genome.

wild relatives of grape (Vitaceae). In this study the flanking regions were either directly sequenced using di-deoxy sequencing or sequenced following cloning when evidence of apparent heterozygosity of sequence was detected in attempts at direct sequencing. These SNPs were generally more phylogenetically informative than the repeat lengths of the SSR. The flanks were a good source of both SNPs and indels for evolutionary studies.

Blankenship *et al*. (2002) compared flanking region polymorphism with repeat number polymorphism in salmon. Primers were developed from the original clonal sequence in which the SSR was discovered. Sequencing was done by subcloning the PCR products and sequencing these. In 300 bases of flanking region, a SNP and an indel were found at three positions. Several haplotypes were found. The sequence polymorphism results reveal information about the evolutionary stability of this fish population in addition to the SSR size data alone. This is because of the difference in the frequency of point mutations in the order of 10^{-9}–10^{-10} as compared with SSR mutation in the order of 7.7×10^{-4} mutations per generation. Vigouroux *et al*. (2002) also used cloning to sequence SSR flanking regions in maize. They showed that although most mutations in SSR markers were in the repeat regions, mutations in the flanking regions were significant. Some were directly sequenced using the PCR primer as the sequencing primer. Mogg *et al*. (2002) found frequent sequence polymorphisms in the flanking regions of *Zea mays* microsatellites. Most of the 54 microsatellites studied showed multiple sequence polymorphisms in the 11 maize lines studied.

SNPs in Wheat SSR Markers

SNPs in ESTs: More than 400,000 EST sequences for the 16 Gb wheat genome have been published and aligned (http://wheat.pw.usda.gov/ITMI/2002/WheatSNP.html). These have been successfully used to find SNPs (Somers *et al*., 2003). The sequences were aligned so that near identical sequences from different varieties could be compared. The process is more complicated for EST sequences which are duplicated on the genome (paralogous sequences). In wheat this was more complex again because it is made up of three genomes, so nearly all ESTs will have homoeologous sequences. When searching for polymorphisms in wheat one must differentiate the desired heterologous polymorphisms from the paralogous and homoeologous polymorphisms. Somers *et al*. (2003) estimated one SNP every 540 bp in 12 varieties of wheat tested.

An advantage of finding SNPs in the published SSR markers of wheat is that they are usually single locus. Gupta *et al*. (2002) found only 12 multiple locus primers from 58 primer pairs and Roder *et al*. (1998) found that 80% of 230 primer sets were genome-specific. This means that the problem of polymorphisms between the three genomes in a single variety is reduced (see Fig. 4.1). That still leaves the problem of sequencing the short sequence between primers and the repeat region.

SNP Discovery Methods

DNA: The material from which the DNA is extracted needs to come from a reliable source. We obtained seeds of the wheat we wanted to survey from the Australian Winter Cereal Collection, Tamworth. Good-quality DNA was obtained from the shoots using a DNeasy extraction kit (Qiagen) (Ablett *et al*., 2006). Many publicly available markers like the Wheat Microsatellite Consortium markers can be accessed from sites such as the Graingenes site (http://wheat.pw.usda.gov/ggpages/SSR/WMC/wmcIndex.HTM), where full sequences of the source of the marker can be obtained. It also has the PCR primers and suggested PCR conditions. In order to take advantage of the cost savings by using the 'universal biotinylated primer' technique (Pacey-Miller and Henry, 2003), the primers were ordered with a 5' end tagged with 11 bases (GCCCCCGCCCG) and an untagged set. Each PCR was done in duplicate with either the forward- or reverse-tagged primer. In some cases minor changes to PCR conditions were required, possibly because of effects of the tag. After PCR of the wheat varieties the products were tested on agarose gel and good PCR products were used for sequencing.

Pyrophosphate Sequencing

Pyrophosphate sequencing is ideally suited to the problem of short sequences. Pyrophosphate sequencing works by single base additions starting at the first base after the primer (Pacey-Miller and Henry, 2003). Briefly an aliquot of the first round product was further amplified in a second round of PCR using a biotinylated GCCCCCGCCCG oligonucleotide to replace the tagged primer (Ablett *et al*., 2006). This was optimized to 25 cycles of 94°C, 50°C and 72°C for 30 s each preceded by denaturation at 94°C for 5 min to denature the Platinum taq. The PCR products were attached to magnetic beads and then denatured to create single strands. PCR primers were used as sequencing primers but because the binding is done at 28°C some of these may not be ideal sequencing primers. By sequentially probing bases in the order of sequence expected by the published sequence we were able to extend the readable sequence.

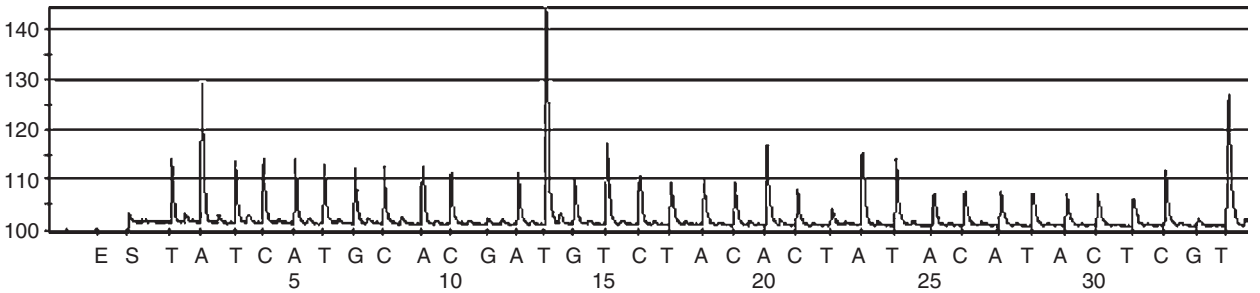
The pyrograms of Fig. 4.2 show how the polymorphisms can be displayed. By comparing them with the published sequences (Table 4.1 and Figs 4.3 and 4.4), oligonucleotides or other validation techniques can be designed to detect these polymorphisms (refer to Chapter 6, this volume).

This technique is ideally suited to finding sequence polymorphisms in published SSR markers. Plant species which have received a lot of scientific attention usually have large numbers of mapped SSR available. The new information can add to the information already made available from the SSR size data, especially when looking at a longer evolutionary scale (Blankenship *et al*., 2002). The repeat regions of these markers are highly variable and may evolve to either larger or smaller sizes. For this reason there is some homoplasia (markers of the same size but different evolutionary background). This can cause some problems when microsatellites are used in evolutionary studies. Because single nucleotide changes are more rare and irreversible, the SNP in flanking regions may reveal extra information for evolutionary studies.

Entry: Wmc 11F

Sample: Mercia

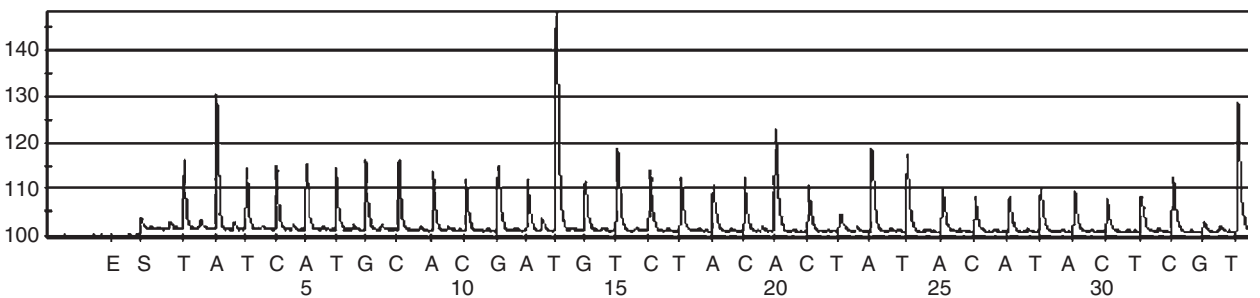
Result: TAATCATGCACATTTTTGTT CTACAACAAT TACA



Entry: Wmc 11F

Sample: Sumei

Result: TAATCATGCACGATTTTTGT TCTACAACAA TTACATA



In Sumei, a ‘G’ was found inserted at the 11th base position as compared with Mercia

Fig. 4.2. Pyrograms showing polymorphisms in the Wmc 11F marker.

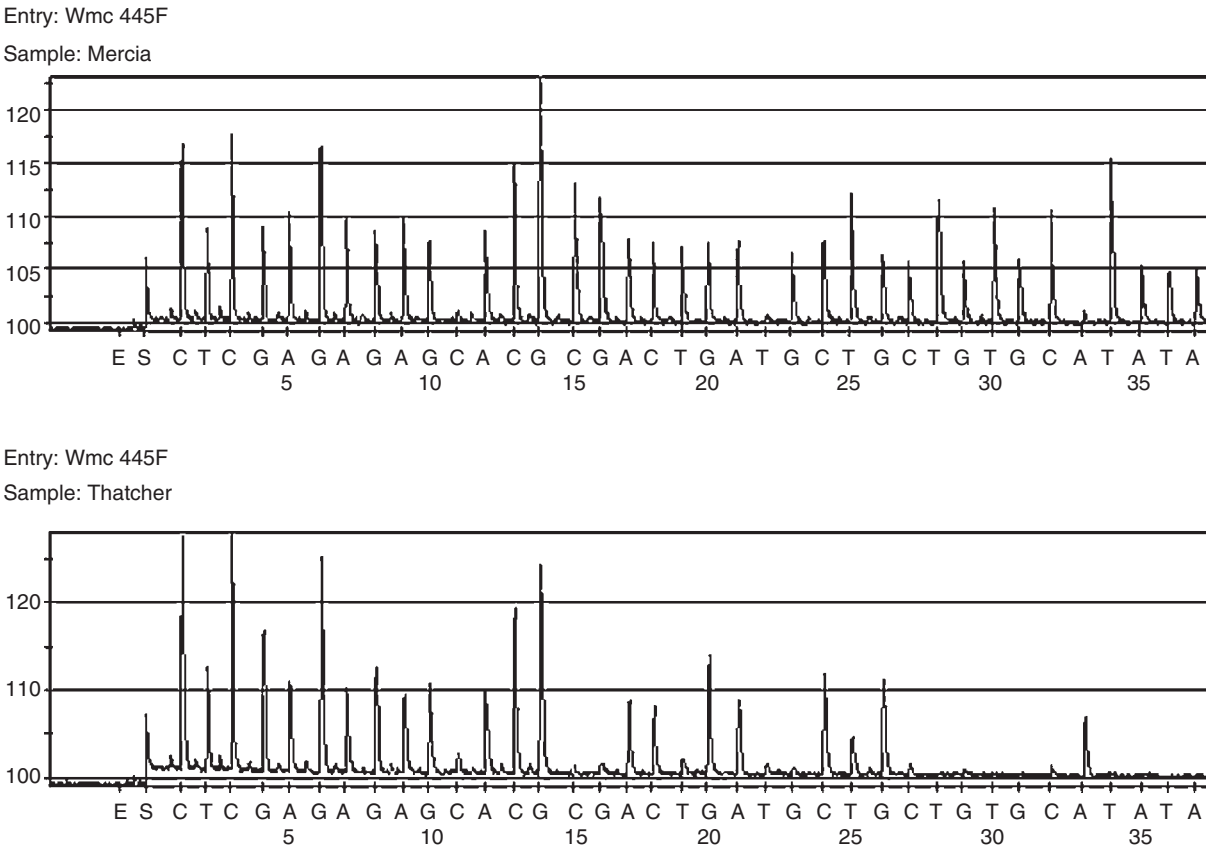


Fig. 4.3. Pyrograms showing polymorphisms in the Wmc 445F marker.

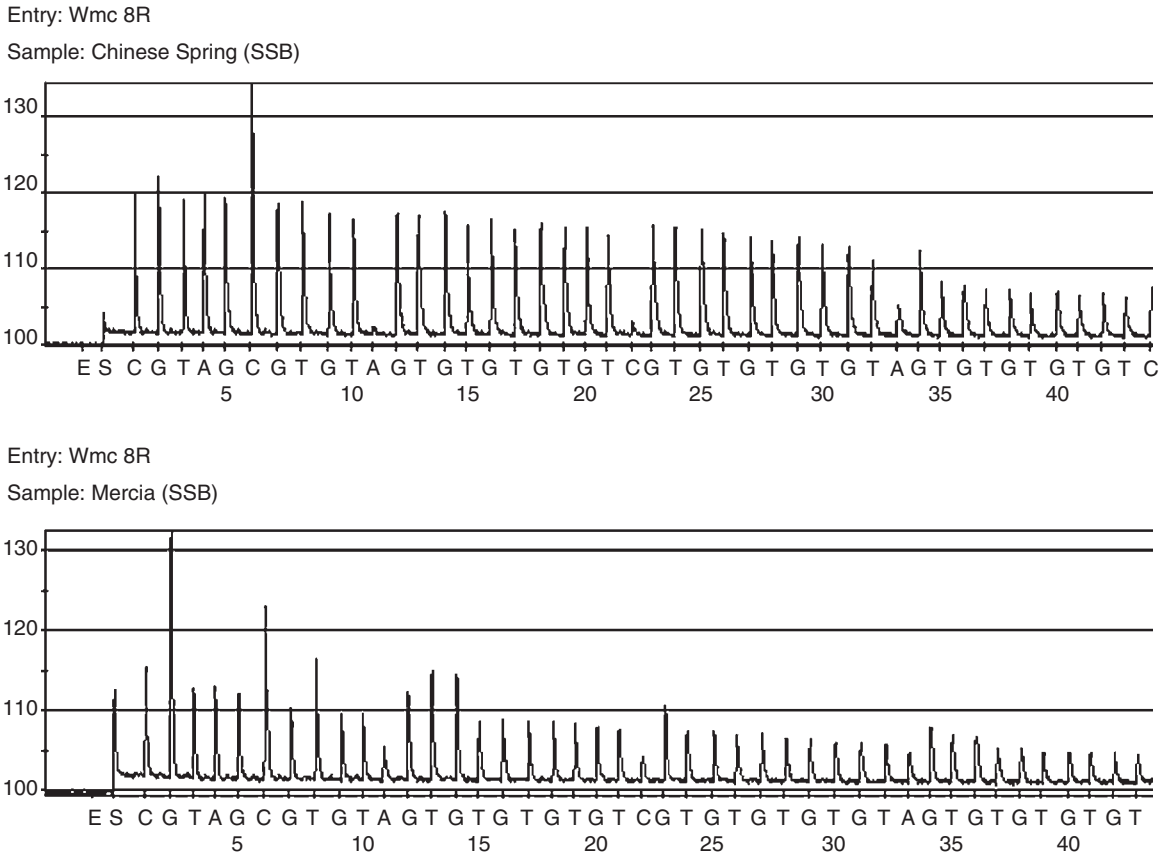


Fig. 4.4. Pyrograms showing polymorphisms in the Wmc 8R marker.

Table 4.1. Markers found to have sequence polymorphisms.

			SNP found											
			gwm	261	F	wmc	8	R	wmc	219	F	wmc	428	F
gwm	33	F	gwm	291	F	wmc	10	F	wmc	219	R	wmc	429	F
gwm	33	R	ISC	1	R	wmc	10	R	wmc	270	F	wmc	432	F
gwm	44	F	psr	6465	F	wmc	11	F	wmc	314	F	wmc	445	F
gwm	44	R	psr	6465	R	wmc	20	F	wmc	314	R	wmc	222	R
gwm	136	F	psr	6512	R	wmc	84	F	wmc	326	F			
gwm	161	F	wmc	3	F	wmc	121	F	wmc	334	F			
gwm	164	F	wmc	3	R	wmc	167	F	wmc	402	R			
gwm	219	R	wmc	4	R	wmc	170	R	wmc	406	R			

References

Ablett, G., Hill, H. and Henry, R.J. (2006) Sequence polymorphism discovery in wheat microsatellite flanking regions using pyrophosphate sequencing. *Molecular Breeding* 17, 281–289.

Ablett, G.A., Karakousis, A., Banbury, L., Cakir, M., Holton, T.A., Langridge, P. and Henry, R.J. (2003) Application of SSR markers in the construction of Australian barley genetic maps. *Australian Journal of Agricultural Research* 54, 1187–1195.

Blankenship, S.M., May, B. and Hedgecock, D. (2002) Evolution of a perfect simple sequence repeat locus in the context of its flanking sequence. *Molecular Biology and Evolution* 19, 1943–1951.

Curtu, A.-L., Finkeldey, R. and Gailing, O. (2004) Comparative sequencing of a microsatellite locus reveals size homoplasy within and between European oak species (*Quercus* spp.). *Plant Molecular Biology Reporter* 22, 339–346.

Dieringer, D. and Schlotterer, C. (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Research* 13, 2242–2251.

Dubcovsky, J., Luo, M.-C., Zhong, G.-Y., Bransteitter, R., Desai, A., Kilian, A., Kleinhofs, A. and Dvorak, J. (1996) Genetic map of diploid wheat, *Triticum monococcum* L., and its comparison with maps of *Hordeum vulgare* L. *Genetics Society of America* 143, 983–999.

Estoup, A., Jarne, P. and Cornuet, J.-M. (2002) Homoplasy and mutation model at microsatellite loci and their consequences population genetics analysis. *Molecular Ecology* 11, 1591–1604.

Gupta, P.K., Balyan, H.S., Edwards, K.J., Isaac, P., Korzun, V., Roder, M.S., Gautier, M.-F., Joudrier, P., Schlatter, A.R., Dubcovsky, J., De la Pena, R.C., Khairallah, M., Penner, G., Hayden, M.J., Sharp, P., Keller, B., Wang, R.C.C., Hardouin, J.P. and Jack, P. (2002) Genetic mapping of 66 new microsatellite (SSR) loci in bread wheat. *Theoretical and Applied Genetics* 105, 413–422.

Harker, N., Rampling, L., Shariflou, M.R., Hayden, M.J., Holton, T.A., Morell, M., Sharp, P., Henry, R.J. and Edwards, K.J. (2001) Microsatellites as markers for Australian wheat improvement. *Australian Journal of Agricultural Research* 52, 1–10.

Matsuoka, Y., Mitchell, S.E., Kresovich, S., Goodman, M. and Doebley, J. (2002) Microsatellites in Zea – variability, patterns of mutations, and use for evolutionary studies. *Theoretical and Applied Genetics* 104, 436–450.

Matsuoka, Y., Mori, N. and Kawahara, T. (2005) Genealogical use of chloroplast DNA variation for intraspecific studies of *Aegilops tauschii* Coss. *Theoretical and Applied Genetics* 111, 265–271.

Mogg, R., Batley, J., Hanley, S., Edwards, D., O'Sullivan, H. and Edwards, K.J. (2002) Characterisation of the flanking regions of *Zea mays* microsatellites reveals a large number of useful sequence polymorphisms. *Theoretical and Applied Genetics* 105, 532–543.

Pacey-Miller, T. and Henry, R.J. (2003) SNP detection in plants using a single stranded pyrosequencing protocol with a universal biotinylated primer. *Analytical Biochemistry* 317, 165–170.

Plaschke, J., Ganal, M.W. and Roder, M.S. (1995) Detection of genetic diversity in closely related bread wheat using microsatellite markers. *Theoretical and Applied Genetics* 91, 1001–1007.

Roder, M.S., Plaschke, J., Konig, S.U., Borner, A., Sorrells, M.E., Tanksley, S.D. and Ganal, M.W. (1995) Abundance, variability and chromosomal location of microsatellites in wheat. *Molecular and General Genetics* 246, 327–333.

Roder, M.S., Korzun, V., Wendehake, K., Plaschke, J., Tixier, M.-H., Leroy, P. and Ganal, M.W. (1998) A microsatellite map of wheat. *Genetics Society of America* 149, 2007–2023.

Rossetto, M., McNally, J. and Henry, R.J. (2002) Evaluating the potential of SSR flanking regions for examining relationships in the Vitaceae. *Theoretical and Applied Genetics* 104, 61–66.

Somers, D.J., Kirkpatrick, R., Moniwa, M. and Walsh, A. (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome* 46, 431–437.

Somers, D.J., Isaac, P. and Edwards, K.J. (2004) A high-density microsatellite consensus map for bread wheat (*Triticum aestivum* L.). *Theoretical and Applied Genetics* 109, 1105–1114.

Vigouroux, Y., Jaqueth, J.S., Matsuoka, Y., Smith, O.S., Beavis, W.D., Smith, J.S.C. and Doebley, J. (2002) Rate and pattern of mutation at microsatellite loci in maize. *Molecular Biology and Evolution* 19, 1251–1260.

5 SNP Discovery by Ecotilling Using Capillary Electrophoresis

F. Eliott, G. Cordeiro, P.C. Bundock and R.J. Henry

Introduction

Ecotilling (derived from targeting-induced local lesions in genomes (TILLING)) is a high-throughput method of detecting naturally occurring DNA polymorphisms in specific gene sequences. We have utilized sugarcane, a complex polyploid species, as a model to develop and test new protocols for high-throughput ecotilling using capillary electrophoresis (CE) systems.

TILLING

TILLING is a technique that was developed to identify single nucleotide polymorphism (SNP) mutants in specific sequences of DNA (McCallum *et al*., 2000; Colbert *et al*., 2001; Till *et al*., 2003b). The method was originally designed to detect single base pair changes resulting from chemical mutagenesis. The mutations are detected using dual fluorescent-labelled, site-specific amplicons from DNA pooled from mutant lines. The fragments are denatured and annealed to form heteroduplexes between polymorphic DNA strands, then digested with CEL 1 endonuclease that cleaves single base mismatch positions on the heteroduplexed DNA strands. The cleaved products are visualized by fluorescence detection using denaturing polyacrylamide gel electrophoresis. For a useful, revised protocol for TILLING and ecotilling, refer to the recently published paper by Till *et al*. (2006b).

TILLING has proven to be a powerful, high-throughput and cost-effective method of SNP discovery and following its original application to *Arabidopsis thaliana* (McCallum *et al*., 2000; Till *et al*., 2003b) and *Drosophila melanogaster* (Bentley *et al*., 2000), it has been applied to a wide range of organisms. In plants, TILLING has been used for discovery of induced mutations in maize (Till *et al*., 2004a), lotus (Perry *et al*., 2003), barley (Caldwell *et al*., 2004), wheat

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

(Slade *et al*., 2005) and rice (Till *et al*., 2007). TILLING studies in animal species include Drosophila (Winkler *et al*., 2005), zebrafish (Wienholds *et al*., 2003) and rats (Smits *et al*., 2004).

Ecotilling

Ecotilling is a variation on the TILLING method, and although the protocol is essentially the same, ecotilling targets the identification of natural variation in populations without the use of mutagenesis (Comai *et al*., 2004). Refer to Fig. 5.1 for a diagrammatic representation of the steps involved in the ecotilling process.

The ecotilling approach is particularly well suited to screening large populations for the discovery of rare alleles (Comai and Henikoff, 2006). DNA polymorphisms which can be detected through ecotilling include SNPs, small insertions and deletions (indels) and variation in microsatellite (SSR) repeat number (Comai *et al*., 2004; Till *et al*., 2006b).

Ecotilling has increased in popularity as the pooling of samples makes it a cost-effective strategy for identifying DNA polymorphisms and haplotypes and the need for conventional sequencing is limited to the fraction of individuals representing unique haplotypes. Another advantage is that

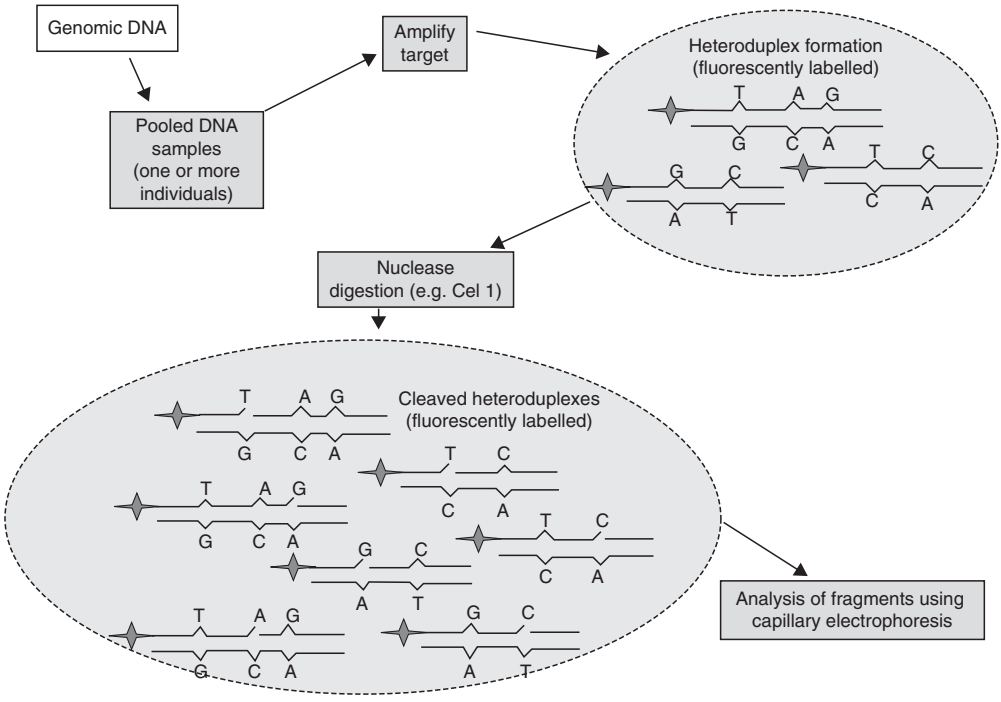


Fig. 5.1. TILLING protocol.

the whole genome sequence for the target species is not required. Provided there is sufficient sequence information to design primers that amplify the region of interest, sequences from related species or from EST databases can be utilized.

Originally developed on *Arabidopsis* (Comai *et al*., 2004), the ecotilling strategy has recently been utilized in a number of SNP discovery projects. Examples include screening for genetic variation in wild populations of *Populus trichocarpa* (Gilchrist *et al*., 2006) and natural populations of *Brassica rapa* (Wu *et al*., 2005) and SNP discovery in the *alk* gene in *Oryza sativa* L. (Kadaru *et al*., 2006). Allelic variation has been identified in genes associated with virus susceptibility in *Cucumis* spp. (Nieto *et al*., 2007), powdery mildew resistance in barley (Mejlhede *et al*., 2006) and herbicide resistance in *Monochoria vaginalis* (Wang *et al*., 2007). Ecotilling has also proved useful for the discovery of rare alleles in the human genome (Till *et al*., 2006a).

In sugarcane, an ecotilling approach has been used by Hermann *et al*. (2006) to investigate the molecular diversity of four resistance gene analogue families (SoRPID, SoPTO, SoXa21 and SoHs1pro-1) and also for mapping SNPs in the sucrose phosphate synthase (*SPS*) gene (McIntyre *et al*., 2006).

Ecotilling Sugarcane Using Capillary Electrophoresis

Sugarcane (*Saccharum* sp.) cultivars are highly polyploid and aneuploid and typically contain 100–130 chromosomes (Price, 1957). Derived from interspecific hybridization, cultivars inherit 80% of their chromosomes from *S. officinarum* ($2n = 80$, $x = 10$), 10–15% from *S. spontaneum* ($2n = 40$–124, $x = 8$) and the remaining 5–10% are recombinant chromosomes (D'Hont *et al*., 1996). Sugarcane cultivars are estimated to contain 8–14 copies of every chromosome (Rossi *et al*., 2003; Aitken *et al*., 2004). The high ploidy level combined with typical mapping population sizes of ~300 progeny allows only single-dose (single copy present at a locus) or double-dose (two copies) markers to be mapped in sugarcane (Wu *et al*., 1992). Consequently, the complexity of the sugarcane genome with its large and variable number of chromosomes presents a challenge for mapping projects in this species.

Ecotilling and TILLING methods as published rely largely on electrophoretic gel systems, such as the LI-COR 4300 DNA Analysis System, to separate and visualize the products. The protocol using gel-based systems did not have the sensitivity required to detect the single-dose SNPs of interest for gene mapping in sugarcane. We have utilized sugarcane as a 'pre-pooled' model to develop and test new protocols for high-throughput ecotilling using CE systems (Cordeiro *et al*., 2006b) which have faster sample processing times and are also reported to be more sensitive than gel electrophoresis systems (Mardis, 1999). This protocol should prove useful for the many laboratories not equipped with a LI-COR system but possessing one of the various CE systems.

The modifications are based on the protocol as provided with the SURVEYOR Mutation Detection Kit by Transgenomic (www.transgenomic.com). Our protocol includes modifications to the PCR conditions, nuclease

digestion times and termination and digestion product precipitation. We also attempted to reduce genotyping costs by exploring a universal primer-labelling strategy and by multiplexing samples.

To optimize the various parameters we used the control DNA provided with the SURVEYOR kit and two sugarcane genotypes, $Q165^{A}$ ($2n \sim 115$), a hybrid cane variety, and the *S. officinarum* clone IJ76-514 ($2n = 80$). Over 15 replicates of the optimized protocol were carried out on the test genotypes. To confirm the presence of segregating SNPs, we screened a population of 190 Q165 × IJ76-514 progeny.

PCR Amplification

The sugarcane samples were amplified with PCR template primers specific to each of four members of the SPS family of genes which were 5'-end labelled with the ABI dyes for detection using the Applied Biosystems 3730 DNA Analyzer. Allele sizes were scored in reference to the ABI GS500 LIZ internal size standard using the computer program GENEMAPPER version 3.7.

PCR amplification was carried out using either the high-fidelity, proofreading Optimase Polymerase (Transgenomic, Omaha, Nebraska, USA) or Platinum Taq polymerase (Invitrogen – Life technologies) or a mixture of both polymerases (Mixed Taq). Amplified products were visualized on agarose gels and overall, the greatest product yield was obtained with either Platinum Taq or the Mixed Taq. After ecotilling, products amplified using the Optimase PCR buffer with Mixed Taq produced the cleanest electropherograms.

CEL 1 Nuclease Digestion and Precipitation

We compared two sources of the CEL I endonuclease enzyme, a purified recombinant protein supplied with the SURVEYOR Mutation Detection Kit (Transgenomic) and a celery extract obtained using the method described by Till (Till *et al*., 2003a). As sugarcane is highly polymorphic, we expected numerous SNPs in each PCR fragment. Therefore, we needed to achieve partial digestion in order to capture all SNPs. We found the optimal conditions for a partial digest in sugarcane samples was 5 min at 42°C using 1 µl of CEL I enzyme in a 10 µl reaction. The signal strength of peaks at all SNP loci decreased with increasing digestion times, with no product remaining in samples digested for longer than 10 min. We also found that using the CEL 1 enzyme at the higher concentration of 2 µl per reaction completely digested the entire product at all digestion times tested. There was no discernible difference between samples digested with the two sources of CEL 1 and the omission of the SURVEYOR Enhancer F did not appear to alter the efficiency of the CEL 1 activity.

Several methods were tested for terminating the CEL 1 digest (the SURVEYOR Stop Solution, SDS and EDTA). While SDS and EDTA produced similar results, the electropherograms for samples stopped with EDTA

showed a lower noise level. However, distinctly lower signal strength was evident in reactions terminated with the SURVEYOR Stop Solution.

Without exception, reactions which were ethanol precipitated and concentrated prior to CE produced the best results with a higher signal to noise ratio and therefore greater ease and confidence in fragment scoring. We found that signal strength was increased further by including a wash step in the ethanol precipitation.

Universal Primer Labelling

The use of universal primer labelling strategies has been reported in a number of ecotilling and TILLING studies using gel-based systems to visualize digested products (Wienholds *et al*., 2003; Winkler *et al*., 2005; Wu *et al*., 2005; Till *et al*., 2006a).

We followed the method described by Schuelke (2000) to label products for separation on the ABI 3730 CE system. PCR fragments were labelled using three primers in a single reaction. Each reaction contained a universal FAM-labelled M13 (−21) primer, and the SPS PCR primers described above. The forward SPS primer had the M13 (−21) sequence added to the 5' end. Equimolar amounts of the universal and reverse SPS primers were used and the forward SPS primer was tested at concentrations ranging from 1/4 to 1/80 of the labelled M13 and reverse primers. While the results were encouraging, they were not consistent. Spurious peaks would sometimes appear and SNP peaks were not always reproducible. Further optimization is required for this strategy to be applied to ecotilling on CE systems.

Multiplexing

Samples with the ABI dyes FAM, PET, VIC or NED were multiplexed after the CEL 1 digestion step and prior to ethanol precipitation. Concentrations of each product needed to be adjusted to allow for the varying intensity of these dyes. The multiplexing worked well although the signal strength decreased for each dye compared to single plex samples. We achieved good results multiplexing three products and, with optimization, higher plex levels would be attainable which would significantly reduce genotyping costs.

Conclusion

Segregating SNPs were easily identified in the progeny of our sugarcane population using the ABI 3730 CE system. In the SPS Gene Family 1, a cluster of SNP peaks formed a distinctly different pattern in each of the parents Q165 and IJ76-514, and these two patterns clearly segregate in the progeny (Fig. 5.2).

In ecotilling or TILLING using gel-based systems, polymorphisms are typically localized to within ±10 bases (Henikoff *et al*., 2004). In our analyses

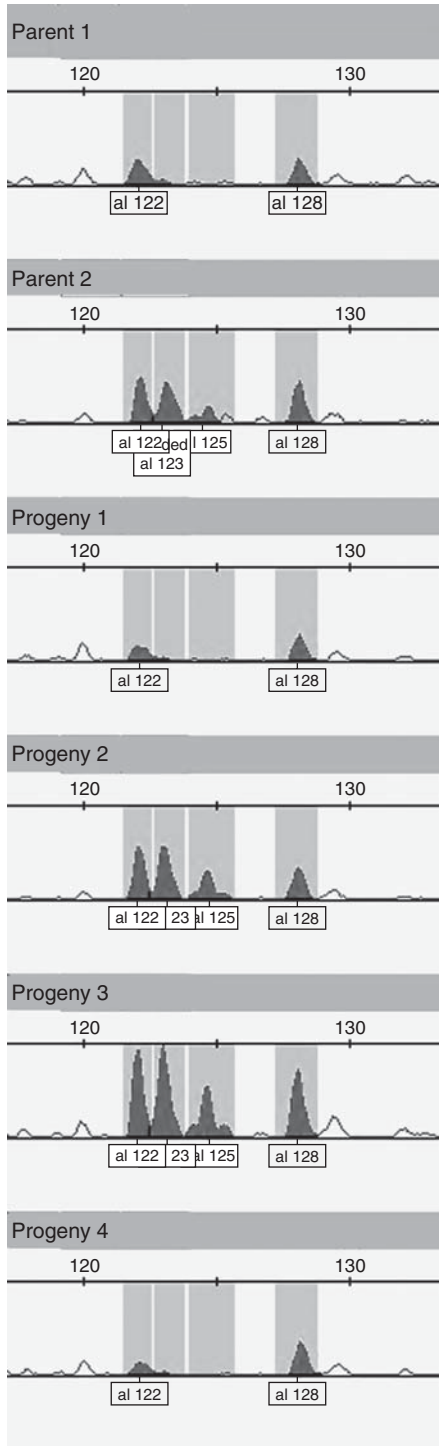


Fig. 5.2. Analysis of progeny of a sugarcane cross using ecotilling.

using CE, SNPs were typically localized to within ±3 bases of that determined through sequence information (Cordeiro *et al*., 2006a). Thus, a higher level of precision in the assignment of fragment sizes based on the size standards is achieved with the sensitivity of a CE system compared with gel electrophoresis systems.

In the two sugarcane genotypes, Q165 and IJ76-514, SNP loci in SPS Gene Family III, identified through ecotilling, correspond precisely with SNPs identified in and through the respective alignment of the 58 and 86 sequences described by McIntyre (McIntyre *et al*., 2006) amplified with the same set of primers used in ecotilling (Cordeiro *et al*., 2006a).

Background noise is common in ecotilling and may result from incomplete polymerase extension during PCR and sequence-specific background bands resulting from endonucleolytic cleavage (Colbert *et al*., 2001; Till *et al*., 2004b). By cleaning and concentrating the digested product through ethanol precipitation we obtained clear electropherograms with significantly reduced noise from incomplete polymerase extension, although bands from sequence-specific endonucleolytic cleavage still remained. Where ecotilling is to be used for SNP discovery, labelling both forward and reverse primers should assist in the confirmation of identified SNPs (Comai *et al*., 2004). In this case, costs can be reduced by using different dye labels for multiplexing. In situations where the SNP location(s) is/are known, only one primer needs to be labelled, in which case selecting the forward primer will assist in the ease of scoring.

An ecotilling strategy utilizing the sensitivity of a CE system has been shown as a robust means for genetic mapping in sugarcane. In a number of instances, multiple single-dose SNPs belonging to homoeologous genes were detected and mapped to separate locations on the Q165 map (L. McIntyre, Brisbane, 2007, unpublished).

Summary of the Optimized Protocol for Ecotilling Using CE

Note: Due to the light-sensitive nature of the fluorescent dyes used, where possible, sample tubes were covered with aluminium foil.

Step 1 – PCR amplification

Create a PCR mix with a final volume of 35 µl containing 25 ng template DNA, 0.2 µM of forward and reverse primers, 0.2 mM of each dNTP, 1 X Optimase Taq buffer and 0.5 U each of Optimase Polymerase and Platinum Taq.

PCR cycle commences with 2 min at 94°C followed by 35 cycles of 10 s at 94°C, 30 s at the appropriate annealing temperature, 30 s at 75°C and a final extension step of 3 min at 75°C. Hold at 15°C.

Confirm successful amplification by visualizing 4 µl of product on a 2% agarose gel.

Step 2 – DNA hybridization

In a thermocycler, denature and reanneal remaining amplified products with the following programme: 95°C for 2 min; 95°C to 85°C at −2°C/s, 85°C to 25°C at −0.1°C/s, hold at 4°C.

Step 3 – CEL I endonuclease digestion

Transfer 10 µl of the hybridized DNA to a new PCR plate on ice. Add 1 µl of either SURVEYOR CEL I or celery extract CEL I to each reaction and incubate at 42°C for 5 min. Immediately return to ice and stop the reaction with 1/10 volume 0.25 mM EDTA (pH7.0).

Step 4 – Ethanol precipitation

Ethanol precipitates directly in the PCR plate by adding 2.5 volumes of ice cold 100% ethanol. Seal the plate and invert several times to mix and place at −20°C for 30 min. Spin the plate at 3000 × g for 15 min in an appropriate centrifuge. Gently invert plate and tip out ethanol. Place the inverted plate on a kimwipe and spin briefly for 10 s at 180 × g to remove remaining ethanol. To wash, add 70 µl of 70% ethanol and repeat the spin cycles described above. Air-dry for 15–30 min.

Step 5 – Capillary electrophoresis

Resuspend DNA pellet with 18 µl ABI Hi-Di Formamide and 0.5 µl ABI GS500 LIZ size standard, heat at 95°C for 2 min and electrophorese.

References

Aitken, K., Jackson, P., Piperidis, G. and McIntyre, L. (2004) QTL identified for yield components in a cross between a sugarcane (Saccharum spp.) cultivar Q165[A] and a *S. officinarum* clone IJ76-514. *Proceedings for the 4th International Crop Science Congress, Brisbane, Australia, 26 September–1 October 2004*. Available at: www.cropscience.org.au

Bentley, A., MacLennan, B., Calvo, J. and Dearolf, C.R. (2000) Targeted recovery of mutations in drosophila. *Genetics* 156, 1169–1173.

Caldwell, D.G., McCallum, N., Shaw, P., Muehlbauer, G.J., Marshall, D.F. and Waugh, R. (2004) A structured mutant population for forward and reverse genetics in barley (*Hordeum vulgare* L.). *Plant Journal* 40, 143–150.

Colbert, T., Till, B.J., Tompa, R., Reynolds, S., Steine, M.N., Yeung, A.T., McCallum, C.M., Comai, L. and Henikoff, S. (2001) High-throughput screening for induced point mutations. *Plant Physiology* 126, 480–484.

Comai, L. and Henikoff, S. (2006) TILLING: practical single-nucleotide mutation discovery. *Plant Journal* 45, 684–694.

Comai, L., Young, K., Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R. and Henikoff, S. (2004) Efficient discovery of DNA polymorphisms in natural populations by Ecotilling. *Plant Journal* 37, 778–786.

Cordeiro, G., Eliott, F.G. and Henry, R.J. (2006a) An optimized ecotilling protocol for polyploids or pooled samples using a capillary electrophoresis system. *Analytical Biochemistry* 355, 145–147.

Cordeiro, G.M., Eliott, F. and Henry, R.J. (2006b) An optimised ecotilling protocol for polyploids or pooled samples using a capillary electrophoresis system. *Analytical Biochemistry* 355, 145–147.

D'Hont, A., Grivet, L., Feldmann, P., Rao, S., Berding, N. and Glaszmann, J.C. (1996) Characterisation of the double genome structure of modern sugarcane cultivars (*Saccharum* spp.) by molecular cytogenetics. *Molecular and General Genetics* 250, 405–413.

Gilchrist, E.J., Haughn, G.W., Ying, C.C., Otto, S.P., Zhuang, J., Cheung, D., Hamberger, B., Aboutorabi, F., Kalynyak, T., Johnson, L., Bohlmann, J., Ellis, B.E., Douglas, C.J. and Cronk, Q.C.B. (2006) Use of ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology* 15, 1367–1378.

Henikoff, S., Till, B.J. and Comai, L. (2004) TILLING. Traditional mutagenesis meets functional genomics. *Plant Physiology* 135, 630–636.

Hermann, S., Brumbley, S. and McIntyre, C.L. (2006) Analysing diversity in sugarcane resistance gene analogues. *Australasian Plant Pathology* 35, 631–641.

Kadaru, S.B., Yadav, A.S., Fjellstrom, R.G. and Oard, J.H. (2006) Alternative ecotilling protocol for rapid, cost-effective single-nucleotide polymorphism discovery and genotyping in rice (*Oryza sativa* L.). *Plant Molecular Biology Reporter* 24, 3–22.

Mardis, E. (1999) Capillary electrophoresis platforms for DNA sequence analysis. *Journal of Biomolecular Techniques* 10, 137–143.

McCallum, C.M., Comai, L., Greene, E.A. and Henikoff, S. (2000) Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiology* 123, 439–442.

McIntyre, C.L., Jackson, M., Cordeiro, G.M., Amouyal, O., Hermann, S., Aitken, K.S., Eliott, F., Henry, R.J., Casu, R.E. and Bonnett, G.D. (2006) The identification and characterisation of alleles of sucrose phosphate synthase gene family III in sugarcane. *Molecular Breeding* 18, 39–50.

Mejlhede, N., Kyjovska, Z., Backes, G., Burhenne, K., Rasmussen, S.K. and Jahoor, A. (2006) EcoTILLING for the identification of allelic variation in the powdery mildew resistance genes mlo and Mla of barley. *Plant Breeding* 125, 461–467.

Nieto, C., Piron, F., Dalmais, M., Marco, C.F., Moriones, E., Gomez-Guillamon, M.L., Truniger, V., Gomez, P., Garcia-Mas, J., Aranda, M.A. and Bendahmane, A. (2007) EcoTILLING for the identification of allelic variants of melon *eIF4E*, a factor that controls virus susceptibility. *BMC Plant Biology* 7.

Perry, J.A., Wang, T.L., Welham, T.J., Gardner, S., Pike, J.M., Yoshida, S. and Parniske, M. (2003) A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. *Plant Physiology* 131, 866–871.

Price, S. (1957) Cytological studies in saccharum and allied genera lll. Chromosome numbers in interspecific hybrids. *Botanical Gazzette (Chicago lll)* 118, 146–159.

Rossi, M., Araujo, P.G., Paulet, F., Garsmeur, O., Dias, V.M., Chen, H., Van Sluys, M.A. and D'Hont, A. (2003) Genomic distribution and characterization of EST-derived resistance gene analogs (RGAs) in sugarcane. *Molecular Genetics and Genomics* 269, 406–419.

Schuelke, M. (2000) An economic method for the fluorescent labeling of PCR fragments. *Nature Biotechnology* 18, 233–234.

Slade, A.J., Fuerstenberg, S.I., Loeffler, D., Steine, M.N. and Facciotti, D. (2005) A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. *Nature Biotechnology* 23, 75–81.

Smits, B.M.G., D'Souza, U.M., Berezikov, E., Cuppen, E. and Sluyter, F. (2004) Identifying polymorphisms in the *Rattus norvegicus* D-3 dopamine receptor gene and regulatory region. *Genes Brain and Behavior* 3, 138–148.

Till, B.J., Colbert, T., Tompa, R., Enns, L.C., Codomo, C.A., Johnson, J.E., Reynolds, S.H., Henikoff, J.G., Greene, E.A., Steine, M.N., Comai, L. and Henikoff, S. (2003a) *Plant Functional Genomics: Methods and Protocols*. Humana Press, Totowa, New Jersey, pp. 205–220.

Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., Young, K., Taylor, N.E., Henikoff, J.G., Comai, L. and Henikoff, S. (2003b) Large-scale discovery of induced point mutations with high-throughput TILLING. *Genome Research* 13, 524–530.

Till, B., Reynolds, S., Weil, C., Springer, N., Burtner, C., Young, K., Bowers, E., Codomo, C., Enns, L., Odden, A., Greene, E., Comai, L. and Henikoff, S. (2004a) Discovery of induced point mutations in maize genes by TILLING. *BMC Plant Biology* 4, 12.

Till, B.J., Burtner, C., Comai, L. and Henikoff, S. (2004b) Mismatch cleavage by single-strand specific nucleases. *Nucleic Acids Research* 32, 2632–2641.

Till, B.J., Zerr, T., Bowers, E., Greene, E.A., Comai, L. and Henikoff, S. (2006a) High-throughput discovery of rare human nucleotide polymorphisms by ecotilling – art. no. e99. *Nucleic Acids Research* 34, E99.

Till, B.J., Zerr, T., Comai, L. and Henikoff, S. (2006b) A protocol for TILLING and ecotilling in plants and animals. *Nature Protocols* 1, 2465–2477.

Till, B.J., Cooper, J., Tai, T.H., Colowit, P., Greene, E.A., Henikoff, S. and Comai, L. (2007) Discovery of chemically induced mutations in rice by TILLING – art. no. 19. *BMC Plant Biology* 7, 19–19.

Wang, G.X., Tan, M.K., Suj, A.R.C., Saitoh, H., Terauchi, R., Imaizumi, T., Ohsako, T. and Tominaga, T. (2007) Discovery of single-nucleotide mutations in acetolactate synthase genes by ecotilling. *Pesticide Biochemistry and Physiology* 88, 143–148.

Wienholds, E., van Eeden, F., Kosters, M., Mudde, J., Plasterk, R.H.A. and Cuppen, E. (2003) Efficient target-selected mutagenesis in zebrafish. *Genome Research* 13, 2700–2707.

Winkler, S., Schwabedissen, A., Backasch, D., Bokel, C., Seidel, C., Bonisch, S., Furthauer, M., Kuhrs, A., Cobreros, L., Brand, M. and Gonzalez-Gaitan, M. (2005) Target-selected mutant screen by TILLING in drosophila. *Genome Research* 15, 718–723.

Wu, J., Sun, R., Zhang, Y. and Wang, X. (2005) Establishment of ecotilling for discovery of DNA polymorphisms in *Brassica rapa* natural population. *Agricultural Sciences in China* 4, 654–659.

Wu, K.K., Burnquist, W., Sorrells, M.E., Tew, T.L., Moore, P.H. and Tanksley, S.D. (1992) The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theoretical and Applied Genetics* 83, 294–300.

6 Genotyping by Allele-specific PCR

D.L.E. Waters, P.C. Bundock and R.J. Henry

Introduction

Single nucleotide polymorphisms (SNPs) represent the most common type of sequence polymorphism found in plant and animal genomes, with well over a million SNPs detected and catalogued for the human genome alone (Sachidanandam *et al*., 2001). SNPs are the basis for many polymorphisms that are detected using systems such as restriction fragment length polymorphisms (RFLPs), randomly amplified polymorphic DNAs (RAPDs) and amplified fragment length polymorphisms (AFLPs) (Schork *et al*., 2000). With recent advances in DNA sequencing technology and hence output, it is now cost-effective to detect SNPs directly. For species which have publicly available expressed sequence tag (EST) databases or extensive genomic sequence data derived from more than one individual or cultivar, putative SNPs can be identified *in silico*.

SNPs are a common currency that can be transferred between laboratories and many methods have been developed for SNP genotyping. However, the assays are commonly not transferable due to the diversity of assay technologies available and utilized. High-throughput automated systems have recently become available (Gupta *et al*., 2001; Gut, 2001), largely in response to the demand for personalized medical applications which have been made possible by the availability of whole genome sequence. Although these methods are high throughput, they are often high cost because of their reliance on expensive equipment. Simple cost-effective methods for SNP marker genotyping still have a place in the world of high-throughput, high-cost genotyping, particularly in the agricultural context. There are active plant breeding programmes in all parts of the world, many of which do not have the resources available which would allow them to exploit the opportunities presented by sophisticated high-throughput genotyping platforms.

Cleaved amplified polymorphic site (CAPS or PCRRFLP) markers have been adopted by a number of groups to enable mapping of SNP markers

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

identified in ESTs (Graner *et al*., 2004; Sato *et al*., 2004). Because this system relies on each SNP being associated with a restriction enzyme site, only a proportion of SNPs are amenable to CAPS. In addition, the enzyme digest step is both time-consuming and often unreliable. An alternative method is allele-specific PCR (AS-PCR). The original report of two primer, single product AS-PCR (Wu *et al*., 1989), was confined to the analysis of one, albeit important, SNP and has likely been used for many assays; however, this type of assay has a reputation for a low rate of success in producing robust markers. The reliability of AS-PCR can be increased by the addition of destabilizing mismatches within the allele-specific primer (ASP), reducing the false positive rate, and a parallel positive control PCR reducing false negatives (Newton *et al*., 1989). The positive control PCR takes place in the same tube as the diagnostic PCR, competing with it for access to polymerase and nucleotides and this too may contribute to a reduction in the false positive rate. Other approaches which include four competing primers in one assay and enable heterozygote identification are, in chronological order of publication date, tetra primer PCR (Ye *et al*., 1992), Bi-PASA (Liu *et al*., 1997), CTTP (Hamajima *et al*., 2000) and tetra primer ARMS-PCR (Ye *et al*., 2001). Although all of these approaches use four primers, each has unique features that differentiate it from the other methods.

The method of Ye *et al*. (1992) was applied to one SNP only and used flanking primers which had a higher annealing temperature than the two inner ASPs, each of which annealed to the alternate SNP alleles in opposite directions. The allele-specific nucleotide was placed in the centre of the primers. The flanking product was first generated by ten rounds of PCR at a higher (63°C) annealing temperature followed by 20 cycles at a lower (46°C) temperature which allowed generation of the AS-PCR product. The technique was sensitive enough to detect one molecule in 40 although the authors did not state whether one or both alleles were detected in pooled samples.

Bidirectional PCR amplification of specific alleles (Bi-PASA; Liu *et al*., 1997) also utilizes two primers which flank the SNP and has the following distinguishing features. The ASPs carry 5' G + C-rich tails with the mismatch nucleotides being placed at or near the 3' end. The annealing temperature and concentration of the outer primers match those of the ASPs. Perhaps the key feature of the method is the presence of the 5' G + C tails, the purpose of which is to increase the efficiency of amplification of allele-specific products during later rounds of PCR. Guidelines for primer design were developed using two human G to A transitions, one in a high (63%) G + C region the other in a low (40%) G + C region. In both cases, only one allele-specific mismatch was tested in either direction, A in the 'forward' direction and C in 'reverse' for both SNPs. The composition of the 10 bp tag and the length of the target-specific region of the primer were varied while the position of the allele-specific nucleotide ranged from the terminal nucleotide to the third base from the terminal nucleotide. The outcomes of these experiments allowed the formation of guidelines which were successfully used to design assays for three additional SNPs. Briefly, the guidelines were as follows: (i) the melting temperature (Tm) of the positive control primers should be 20–25°C lower than that of the

300–1000 bp PCR product itself; (ii) the Tm of the inner PCR products should be ~35°C lower than the positive control PCR product; and (iii) the annealing temperature should be 20°C below the Tm of the positive control PCR product. The recommended optimization pathway involved alterations to both positive control and ASP concentrations.

Hamajima *et al*. (2000) reported assays for two human SNPs. The allele-specific nucleotides were placed at the terminal 3' base and the flanking primers were placed asymmetrically, allowing heterozygote detection. However, no optimization data were presented.

Tetra primer ARMS-PCR as described by Ye *et al*. (2001) was successfully applied to three SNPs, G/A, G/C and A/C. They developed and utilized primer design software, which they made publicly available. The ASPs contained a deliberate mismatch 2 bp from the 3' terminal base and were applied in a concentration ratio of 10:1 relative to the two flanking primers. Allele specificity was achieved by placing the allele-specific nucleotide at the 3' terminal position. The suggested order of primer design was allele-specific before positive control primers.

Soleimani *et al*. (2003) simplified these approaches by using three primers only in combination with a hot start enzyme and a touchdown PCR protocol. The technique was compared with dideoxy primer extension (SNaPshot™ Multiplex Kit). The authors suggested this was an efficient and very cost-effective technique relative to the SNaPshot™ Multiplex Kit although it is not clear how many of the 214 loci that were identified by direct sequencing and alignment of ESTs were validated with either AS-PCR or dideoxy primer extension.

The preceding examples were validated on at most two or three SNPs. Because there was a lack of available data which indicated what the likelihood would be of converting any known SNP to an AS-PCR assay, Bundock *et al*. (2006) examined the success rate of SNP marker development based on the nested three primer PCR approach of Soleimani *et al*. (2003).

Assays were designed for SNPs that had been found using public domain barley (*Hordeum vulgare*) EST sequences, most of which had been used as RFLP probes and mapped in barley. SNPs have been shown to occur at high frequency in expressed barley sequences (Kota *et al*., 2001; Kanazin *et al*., 2002; Bundock *et al*., 2003). Their high frequency makes them a more attractive option for mapping expressed genes than simple sequence repeats (SSRs) which occur in only a small proportion of ESTs (Kota *et al*., 2001; Holton *et al*., 2002; Varshney *et al*., 2002).

Likely SNP sites flanked by reliable sequence became the focus of primer design for three primer AS-PCR (Fig. 6.1). At potential SNP sites, two primers (forward and reverse orientations) that flanked the SNP-containing region were designed using MacVector™ 6.5 (Accelrys, San Diego, California, USA). ASPs were designed to interrogate 49 SNPs in 31 sequence clusters, using the nested three primer system. Of these 49 SNP sites, robust AS-PCR markers were developed for 36 located in 28 EST clusters.

Two approaches were taken which reflect two potentially different motives for developing markers. The first approach was a 'saturation'

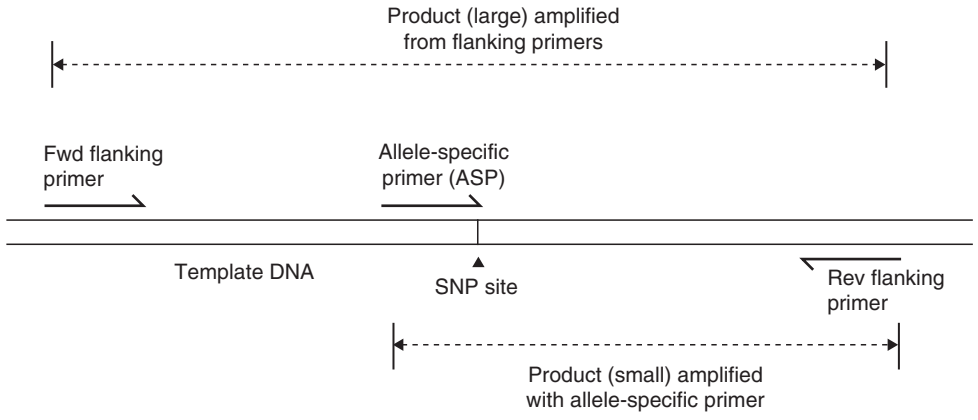


Fig. 6.1. Diagram illustrating the position and orientation of PCR primers for the three primer allele-specific PCR relative to the matching template DNA.

approach designed to provide a high probability of developing an assay for each SNP. In this approach only a single SNP within each SNP containing EST cluster (contig) which was considered to be highly reliable was targeted. For these SNPs, eight ASPs were synthesized in two sets of four. In the first set there were two primers (one for each SNP allele at the terminal base) in each orientation (forward and reverse), i.e. two orientations for each allele. The second set of primers was identical to the first except for the inclusion of a mismatch 3 bp from the 3' terminal base (Kwok *et al*., 1990; Zhang *et al*., 2003). The mismatch with a base of the same identity (i.e. A–A, T–T, C–C and G–G) was expected to increase primer specificity. The only parameter considered in the design of the ASP was Tm, which was as close as possible to the Tm of the competing flanking primers.

The second approach involved selecting a minimum number of the 'best' ASPs to enable a greater coverage of SNP sites for a number of primers. Candidate ASPs were removed from consideration for synthesis based on warnings displayed in MacVector™. The remaining candidate primers were chosen based on closeness of the Tm to the competing flanking primer. There were 6.2 (111/18) primers per SNP site marker for the second approach compared with 10.6 (190/18) for the first approach.

PCR amplifications were carried out with an annealing temperature gradient of 1°C difference in temperature between adjacent wells across a 12-well block from 59°C to 70°C. By doing this the optimum annealing temperature and an assessment of how robust each allele-specific reaction was determined. A non-3'–5' proofreading 'hot start' Taq polymerase was utilized. Proofreading ability has the potential to interfere with allele-specific amplifications by removing mismatch nucleotides. All AS-PCR amplifications were run as nested PCRs with the two flanking primers and the internal ASP present in the reaction mix.

A high proportion of the ASPs led to the development of a robust SNP assay. As expected, the saturation approach produced a higher success rate

for development of one or more markers at each SNP site (18 SNP sites out of 19 targeted or 95%) compared with the second approach (18 from 30, or 60%). Importantly, the saturation approach generated markers for both SNP alleles at 16 out of 19 sites. However, the cost of oligonucleotide primers for the saturation approach per SNP site developed was on average nearly twice the cost of the second approach and there were a larger number of tests to determine useful primer combinations. Although no assay was developed for many individual SNP sites, the second method of selecting a small number of the 'best' primers was more cost-effective in terms of providing the maximum number of SNP assays for minimum primer cost. Only two flanking primers failed to amplify – a rate of 0.03 (2/66). In contrast, the failure rate of amplification from the ASPs was 0.2 (39/191). A total of 20 primers did not display allele-specific amplification. From these there were 15 targeting nine SNP sites that were verified as SNPs by other primers.

Amplification failure is a problem with PCR in general but there are several reasons why amplification failure from internal primers might be higher for the type of assay described here. First, the sequence of internal ASP is entirely dictated by the position of the SNP site, while the best choice is made from many possible flanking primers. In addition, the flanking primers were designed as a pair, whereas the ASPs were designed separately but were required to pair with an independently designed flanking primer for amplification, reducing the likelihood of success for the ASPs relative to the flanking primers. The alternative approach where flanking primers are designed to pair with the respective ASP would probably increase the frequency of ASPs amplifying but reduce the likelihood of flanking primers amplifying a competing product. This strategy may perform no better than the strategy used here since amplification from the flanking primers provides a competitive product for allele absent templates (genotypes). An indirect cause of failure of ASPs may be introns. In some cases it appears that introns may have prevented the development of allele-specific assays where the product amplified from flanking primers was larger than one would predict from the cDNA sequence due to the presence of an intron.

The mismatch primers (comprising half the primers used in the first approach) provided a total of 38 (out of 76) primers giving allele-specific amplifications compared with successful allele-specific amplifications for 41 (out of 76) of the corresponding primers without this mismatch. These primers were not as robust as the non-mismatch primers since the average operational temperature range for mismatch primers allele-specific amplification was approximately 4°C compared with 6°C for the primers without this mismatch. This, in combination with the finding that failure of ASP amplification was the major cause of assay failure and a mismatch destabilizes primers, suggests that these primers do not on average confer any advantage over those without a mismatch.

A combination of the two approaches described here would be an efficient means of developing SNP markers for projects that are on a modest budget without access to sophisticated equipment and where there is sufficient knowledge of DNA sequence around the SNP sites. An initial round of

primer design using the more efficient 'best' primer strategy, followed by a second round using a modified saturation approach which excluded the mismatch primers and finally followed by the addition of mismatch primers where non-allele-specific amplification was evident should yield markers for around 80% of target SNPs for minimum cost.

Given the capacity of AS-PCR to deliver low-cost, simple, accurate, reliable and semi-high-throughput assays, we applied the technique to two important rice quality traits, fragrance and starch gelatinization temperature (GT). Rice is the staple cereal in many countries, many of which have low per capita incomes. Because these are low-cost assays, they should have utility within the breeding programmes of many of these countries.

Consumers are willing to pay a premium price for fragrant rice varieties. The flavour and fragrance of Basmati and Jasmine style rice have been associated with increased levels of 2-acetyl-1-pyrroline (2AP) (Buttery *et al*., 1983; Lorieux *et al*., 1996; Widjaja *et al*., 1996; Yoshihashi, 2002). A number of subjective sensory methods have been utilized to assist breeders in selecting fragrant rice but they have limitations when processing large numbers of samples. An objective method of 2AP identification using gas chromatography is available but the assay requires large tissue samples and is time-consuming (Lorieux *et al*., 1996; Widjaja *et al*., 1996). More recently molecular markers genetically linked to fragrance have been developed for the selection of fragrant rice (Cordeiro *et al*., 2002). Although these markers had the advantage of being inexpensive, simple, rapid and only requiring small amounts of tissue, they were only linked with fragrance and therefore did not allow prediction of the fragrant status of any one rice sample with 100% accuracy.

With the finding that an 8 bp deletion and three SNPs in the gene encoding a betaine aldehyde dehydrogenase 2 (BAD2) homologue is the likely cause of fragrance in Jasmine and Basmati style rice (Bradbury *et al*., 2005), an opportunity was clearly available for the construction of a perfect marker for fragrance in rice. A four primer AS-PCR assay was designed and then validated on a diverse collection of 14 fragrant and 74 non-fragrant varieties in addition to a population of 168 field-grown F2 individuals segregating for fragrance. The PCR relied on two flanking primers which annealed to both fragrant and non-fragrant genotypes and two ASPs which were in opposite orientations which annealed to each of the alternative alleles. The flanking primers bound asymmetrically around the deletion, so each allele generated products of differing sizes. As expected the flanking primers generated a PCR product of approximately 580 bp that was present in every sample while fragrant individuals had a second product of 257 bp and non-fragrant individuals were identified by a product of 355 bp. Heterozygotes could be identified by the presence of all three PCR products.

Starch is a polymer of glucose which presents as a mixture of two forms, amylose and amylopectin. Amylose is principally composed of a linear polymer of α (1–4)-linked glucose with some α (1–6) linkages while amylopectin is a more complex mixture of both α (1–4)-linked glucose extensively branched by α (1–6) linkages. In its native state, rice starch has a semi-crystalline structure which is disrupted by cooking, transforming the starch into a softer edible

gel-like material. Because it is associated with the cooking time and texture of cooked rice, and cool cooked rice, the temperature at which rice starch gelatinizes is an important component of rice eating quality (Maningat and Juliano, 1978).

A link between the GT of rice starch and the enzymes of starch biosynthesis was made with the finding that a major gene that controls rice GT, as determined by the indirect measure of alkali spreading, genetically maps to a region of chromosome six that is clearly different from the *waxy* locus (Umemoto *et al*., 2002). This gene co-segregated with soluble starch synthase IIa (SSIIa) and a gene that affects amylopectin structure (Umemoto *et al*., 2002) suggesting the cause and effect or genotype to phenotype hierarchy of GT is as follows: SSIIa, amylopectin structure, GT.

Given there was strong evidence of different alleles of the SSIIa encoding gene governing rice starch GT, there was benefit in uncovering the sequence differences which were associated with each GT class. Umemoto *et al*. (2004) initially reported four haplotypes of the gene that codes for SSIIa, based on four combinations of three functional SNPs. However, GT and chain length distribution were not unique to all haplotypes in that study.

By including an additional 2bp polymorphism that affects the SSIIa coding sequence and which was previously reported by Gao *et al*. (2003), two GT rice cultivar classes were evident (Umemoto and Aoki, 2005). Waters *et al*. (2006) working on the sequence encoding the SSIIa gene in a set of genotypes different to that utilized by Umemoto *et al*. (2002) found these same two polymorphisms within exon 8 of the SSIIa encoding gene, A/G and GC/TT, allowed the differentiation of the rice varieties examined into two discrete GT classes which differed by 8°C. The high-GT class was found to have the haplotype G, GC while the low-GT class was either haplotype G–TT or A–GC.

Following the guidelines developed by Bundock *et al*. (2006), robust dominant assays were developed. However, this first pass assay design could detect the presence of the A and TT polymorphisms, but not each of the alternative G and GC. Depending on the target market either high- or low-GT starches are desirable, and so it is important to be able to select for or against each allele at this locus within a breeding programme. Because of this, a series of primers which were either longer or shorter than the initial set was designed. Primers which were non-discriminatory due to the generation of false positives in the first pass were destabilized by the removal of nucleotides from the 5' end while primers which generated false negatives were stabilized by the addition of nucleotides to the 5' end. By doing this it was found that a primer two bases longer than the initial primer (which did not have a mismatch three nucleotides from the terminal base) was stabilized and allowed discrimination of the A/G SNP by detecting the presence of 'G'. In contrast, a primer one base shorter than the original design and with a mismatch was sufficiently destabilized such that it amplified in the presence of 'GC' but not 'TT'. Each polymorphism associated with GT class could then be identified.

Competitive AS-PCR has proven to be a robust, low-cost flexible technique. If a panel of SNPs and indels are already available, low-cost assays

can be generated for many of these polymorphisms which can then be used for a range of applications including mapping, QTL analysis and fingerprinting. In addition, if the polymorphism which affects a trait is known, there is a high probability of generating a functional assay for such a trait.

References

Bradbury, L.M.T., Henry, R.J., Jin, Q. and Waters, D.L.E. (2005) A perfect marker for fragrance genotyping in rice. *Molecular Breeding* 16, 279–283.

Bundock, P., Christopher, J., Eggler, P., Ablett, G., Henry, R. and Holton, T. (2003) Single nucleotide polymorphisms in cytochrome P450 genes from barley. *Theoretical and Applied Genetics* 106, 676–682.

Bundock, P.C., Cross, M.J., Shapter, F.M. and Henry, R.J. (2006) Robust allele-specific PCR markers developed for SNP's in expressed barley sequences. *Theoretical and Applied Genetics* 112, 358–365.

Buttery, R.G., Ling, L.C., Juliano, B.O. and Turnbaugh, J.G. (1983) Cooked rice aroma and 2-acetyl-1-pyrroline. *Journal of Agricultural and Food Chemistry* 31, 823–826.

Cordeiro, G.M., Christopher, M.J., Henry, R.J. and Reinke, R.F. (2002) Identification of microsatellite markers for fragrance in rice by analysis of the rice genome sequence. *Molecular Breeding* 9, 245–250.

Gao, Z., Zeng, D., Cui, X., Zhou, Y., Yan, M., Huang, D., Li, J. and Qian, Q. (2003) Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. *Science in China* 46, 661–668.

Graner, A., Kota, R., Perovic, D., Potokina, E., Prasad, M., Scholz, U., Stein, N., Thiel, T., Varshney, R. and Zhang, H. (2004) Molecular mapping: shifting from the structural to the functional level. Oral presentations. *Proceedings 9th International Barley Genetics Symposium*, Brno, Czech Republic, pp. 49.

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. *Current Science* 80, 524–535.

Gut, I.G. (2001) Automation in genotyping of single nucleotide polymorphisms. *Human Mutation* 17, 475–492.

Hamajima, N., Saito, T., Matsuo, K., Kozaki, K., Takahashi, T. and Tajima, K. (2000) Polymerase chain reaction with confronting two-pair primers for polymorphism genotyping. *Japanese Journal of Cancer Research* 91, 865–868.

Holton, T.A., Christopher, J.T., McClure, L., Harker, N. and Henry, R.J. (2002) Identification and mapping of polymorphic SSR markers from expressed gene sequences of barley and wheat. *Molecular Breeding* 9, 63–71.

Kanazin, V., Talbert, H., See, D., DeCamp, P., Nevo, E. and Blake, T. (2002) Discovery and assay of single-nucleotide polymorphisms in barley (*Hordeum vulgare*). *Plant Molecular Biology* 48, 529–537.

Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J. and Graner, A. (2001) Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). *Hereditas* 135, 145–151.

Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Goda, L., Levenson, C. and Sninsky, J.J. (1990) Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Research* 18, 999–1005.

Liu, Q., Thorland, E.C., Heit, J.A. and Sommer, S.S. (1997) Overlapping PCR for bidirectional PCR amplification of specific alleles: a rapid one-tube method for simultaneously differentiating homozygotes and heterozygotes. *Genome Research* 7, 389–398.

Lorieux, M., Petrov, M., Huang, N., Guiderdoni, E. and Ghesquiere, A. (1996) Aroma in rice: genetic analysis of a quantitative trait. *Theoretical and Applied Genetics* 93, 1145–1151.

Maningat, C.C. and Juliano, B.O. (1978) Properties of lintnerized starch granules from rices of different amylose content and gelatinization temperature. *International Rice Research Newsletter* 3, 7–8.

Newton, C.R., Graham, A., Heptinstall, L.E., Powell, S.J., Summers, C., Kalsheker, N., Smith, J.C. and Markham, A.F. (1989) Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research* 17, 2503–2516.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Marth, G., Sherry, S., Kwok, P.-Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Etten, W.J.V., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S. and Altshuler, D. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933.

Sato, K., Nankaku, N., Motoi, Y. and Takeda, K. (2004) A large scale mapping of ESTs on the barley genome. *Proceedings 9th International Barley Genetics Symposium*. Brno, Czech Republic, pp. 79.

Schork, N.J., Fallin, D. and Lanchbury, J.S. (2000) Single nucleotide polymorphisms and the future of genetic epidemiology. *Clinical Genetics* 58, 250–264.

Soleimani, V.D., Baum, B.R. and Johnson, D.A. (2003) Efficient validation of single nucleotide polymorphisms in plants by allele-specific PCR, with an example from barley. *Plant Molecular Biology Reporter* 21, 281–288.

Umemoto, T. and Aoki, N. (2005) Single nucleotide polymorphisms in rice starch synthase IIa that alter starch gelatinisation temperature and starch association of the enzyme. *Functional Plant Biology* 32, 763–768.

Umemoto, T., Yano, M., Satoh, H., Shomura, A. and Nakamura, Y. (2002) Mapping of a gene responsible for the difference in amylopectin structure between japonica-type and indica-type rice varieties. *Theoretical and Applied Genetics* 104, 1–8.

Umemoto, T., Aoki, N., Lin, H.X., Nakamura, Y., Inouchi, N., Sato, Y., Yano, M., Hirabayashi, H. and Maruyama, S. (2004) Natural variation in rice starch synthase IIa affects enzyme and starch properties. *Functional Plant Biology* 31, 671–684.

Varshney, R.K., Thiel, T., Stein, N., Langridge, P. and Graner, A. (2002) *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. *Cellular and Molecular Biology Letters* 7, 537–546.

Waters, D.L.E., Henry, R.J., Reinke, R.F. and Fitzgerald, M.A. (2006) Gelatinisation temperature of rice explained by polymorphisms in starch synthase. *Plant Biotechnology Journal* 4, 115–122.

Widjaja, R., Craske, J.D. and Wootton, M. (1996) Comparative studies on volatile components of non-fragrant and fragrant rices. *Journal of the Science of Food and Agriculture* 70, 151–161.

Wu, D.Y., Ugozzoli, L., Pal, B.K. and Wallace, R.B. (1989) Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anaemia. *Proceedings of the National Academy of Sciences of the USA* 86, 2757–2760.

Ye, S., Humphries, S. and Green, F. (1992) Allele specific amplification by tetra-primer PCR. *Nucleic Acids Research* 20, 1152.

Ye, S., Dillon, S., Ke, X.Y., Collins, A.R. and Day, I.N.M. (2001) An efficient procedure for genotyping single nucleotide polymorphisms. *Nucleic Acids Research* 29, e88.

Yoshihashi, T. (2002) Quantitative analysis on 2-acetyl-1-pyrroline of an aromatic rice by stable isotope dilution method and model studies on its formation during cooking. *Journal of Food Science* 67, 619–622.

Zhang, W., Gianibelli, M.C., Ma, W., Rampling, L. and Gale, K.R. (2003) Identification of SNPs and development of allele-specific PCR markers for c-gliadin alleles in *Triticum aestivum*. *Theoretical and Applied Genetics* 107, 130–138.

7 The MassARRAY System for Plant Genomics

D. Irwin

Introduction

Plant characterization is the specific detection and identification of a genus or species as well as the classification of potential strains present in the population – an important aspect for the recognition of plant origin, commercial value and monitoring of populations for genetic drift and infiltration from alternate strains. In addition, monitoring of quantitative traits and resistance to diseases and pests can have a significant impact on the commercial return on production crops.

Dominant systems for plant detection and characterization are based on either phenotypic or genotypic identification and characterization. Phenotypic characterization systems are based upon the detection of the organism itself or its metabolic products and can therefore be quite variable and dependent on environmental conditions like temperature, soil quality, water availability and the overall sample preparation.

Systems based on genetic characteristics (the DNA or RNA constitution of the organism), on the other hand, have the advantage that DNA sequence characteristics are stable and less influenced by external factors. They are useful tools and reliable alternatives to phenotypic methods.

Laboratories have begun to utilize nucleic acid amplification-based methods of informative genus or species-specific genomic marker regions. Polymerase chain reaction (PCR)-based procedures for plant genus/species identification and characterization have been demonstrated as have association studies to discover quality trait loci (Werner *et al*., 2005; Yu *et al*., 2006). Specific amplification by PCR is followed by analysis of the amplification product via gel electrophoresis, fluorescence detection, sequencing or mass spectrometry, as presented in this chapter. PCR provides a means of producing large numbers of DNA copies from a relatively small amount of starting material and has enabled rapid identification of all varieties of organisms.

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

The underlying genomic marker regions are chosen based on typability and discriminatory power, the ability to obtain an unambiguous result from all types within the species or genus of interest and the ability to differentiate between closely related strains. Therefore, different genomic sequences enable different levels of identification to the genus, species, subspecies or strain-specific level.

The following chapter introduces the MassARRAY system as an emerging analysis tool translating genomic capability into superior solutions for rapid and effective plant identification and characterization. The application of the system to plant identification and characterization benefits from 10 years of technological development and expertise in high-throughput DNA analysis in human genomics and in the arena of the Human Genome project.

The MassARRAY System

SEQUENOM launched the MassARRAY system in 2000 as an automated genotyping platform for the detection of genetic variations (single nucleotide polymorphisms (SNPs)) within the human genome. The mass spectrometry-based system has become one of the leading technologies for high-throughput analysis and high-fidelity measurements of nucleic acid sequence variations (CorbachoCorbacho *et al*., 1999; Jurinke *et al*., 2002, 2004). It has since been applied successfully to large-scale genome-wide genetic association studies revealing disease susceptibility genes (Bansal *et al*., 2002; Downes *et al*., 2004; Kammerer *et al*., 2004, 2005) and to a gamut of additional applications such as quantitative gene expression, comparative sequencing, haplotyping and epigenetic gene regulatory mechanisms via methylation analysis (Beaulieu *et al*., 2003; Honisch *et al*., 2004; Stanssens *et al*., 2004; Tang *et al*., 2004; Ehrich *et al*., 2005).

The MassARRAY system components include liquid handling robotics for automated sample processing and a MALDI-TOF mass spectrometer optimized for nucleic acid analysis along with automated computer algorithms for data acquisition and analysis (Fig. 7.1). Reagents and protocols are standardized and the sample preparation is automated and simple.

All system applications are based on a target-specific amplification by PCR in 5µl volumes using a standard 384-microtitre plate format (Fig. 7.1, Step 1). Adaptations to 96-well formats are practicable. Resulting PCR products are subject to post-PCR processing and two kit-based biochemistries – MassEXTEND or MassCLEAVE – as described in more detail later in this chapter (Fig. 7.1, Step 2). Post-PCR products are conditioned and desalted *in situ* via ion exchange resin and automatically transferred one sample to one pad on a 384 SpectroCHIP (Fig. 7.1, Step 3). Chips are placed two at a time into the MassARRAY analyser for data acquisition (Fig. 7.1, Step 4). The instrument is automatically calibrated with defined control analytes prior to each run and results are obtained with real-time control of data quality.

This scheme allows for parallel processing in 384-well formats in combination with miniaturized sample preparation on matrix pads on the surface of

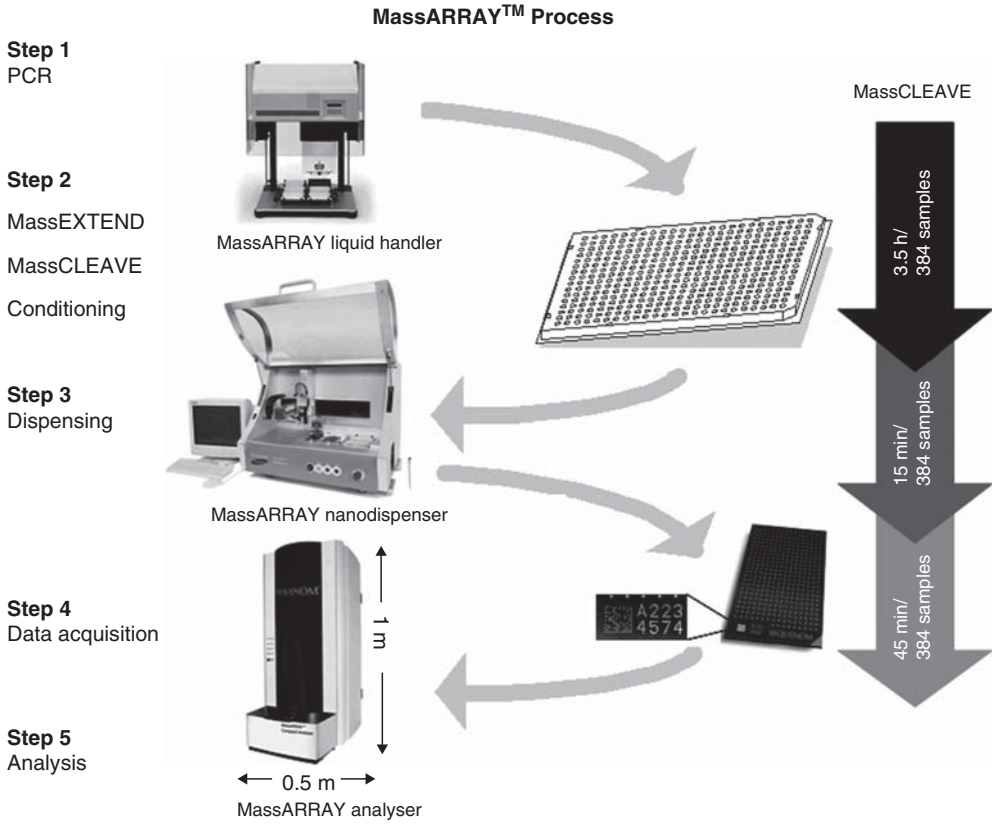


Fig. 7.1. MassARRAY process. (Step 1) Liquid-handling robotics for amplification of the target region by PCR and (Step 2) for post-PCR processing by MassEXTEND for plant identification and typing or by MassCLEAVE for sequence variant detection. Assay formats are homogeneous and no purification is required. (Step 3) 1:1 transfer of the reaction products from the 384-microtitre plate to a SpectroCHIP for analysis. (Step 4) Automated data acquisition by MALDI-TOF MS and databases storage of analysis results. (Step 5) Data analysis by application-specific software packages. Less than 5 h turn-around time to result for identification, typing and subtyping of 384 samples.

modified silicon chip carriers, 384 SpectroCHIPs. The format is flexible and allows for analysis of up to 384 samples with a minimum of one assay or a minimum of 384 assays and one sample. Results are obtained in less than 5 h per 384 plate and automatically loaded into a database, which allows for convenient data analysis and data comparability (Fig. 7.1, Step 5). The MassARRAY analyser, a bench-top instrument, has greatly reduced analysis time and has created a cost-effective analytical tool that is extremely attractive to a non-experienced operator (Fig. 7.2).

The instrument utilizes matrix-assisted laser desorption ionization time-of-flight (TOF) mass spectrometry (MALDI-TOF MS) for accurate, high-throughput analysis of DNA and RNA. Profound advantages are the speed

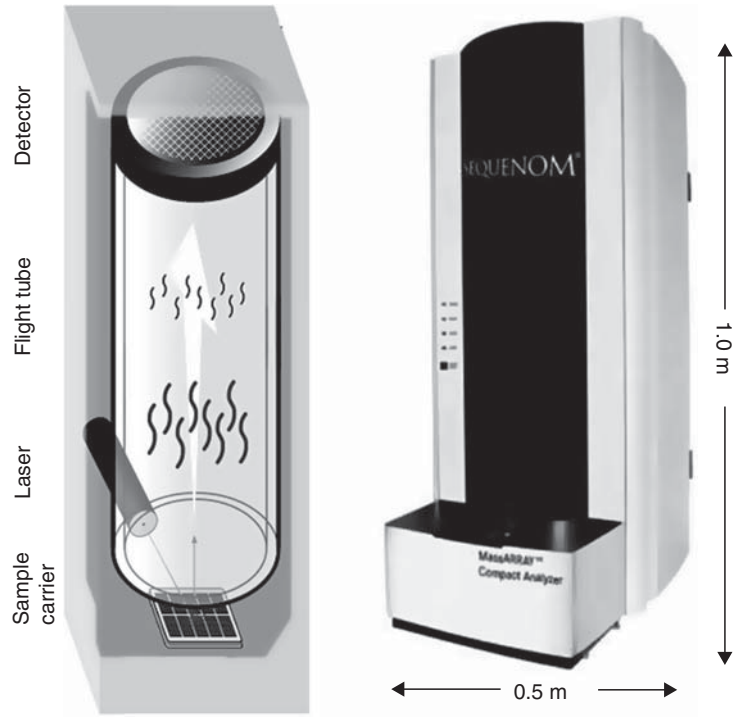


Fig. 7.2. MALDI-TOF MS and the MassARRAY compact analyser. In MALDI-TOF MS a laser hits an analyte patch on the sample carrier. It creates a plume of nucleic acid fragment ions that are separated in the flight tube and delivered to the detector in the order of their mass. Small molecular weights travel fastest. MassARRAY compact analyser is a bench-top MALDI-TOF MS. Instrument footprint: 1.07 m × 0.67 m × 0.51 m, dynamic range: 109 molecules.

of analysis, with a mass spectrum being generated in a matter of milliseconds, and the direct detection of an intrinsic analyte property, its molecular mass. Labelling with expensive fluorophores or potentially hazardous radioactive isotopes can thus be avoided.

The basic components of a mass spectrometer consist of an ionization source (e.g. a UV laser), an analyser and a detector (Fig. 7.2). For analysis, the biological sample (e.g. a mixture of nucleic acids) is mixed with matrix material on pads of a silicon chip surface. The MALDI-TOF process is then initiated by laser-induced desorption of the matrix–analyte mix (Karas and Hillenkamp, 1988). Nucleic acid compomer masses are calculated from TOF measures, which reflect the time a laser-ionized and accelerated compomer requires to drift through the flight tube (1–2 m long) of the TOF analyser and reach the detector of the instrument. In the detector, the ionized compomers generate an electrical signal that gets recorded by a data system and is finally converted into a mass spectrum.

The TOF increases proportionally to the mass to charge ratio (m/z) of the ionized molecule. MALDI-TOF MS predominantly generates single-charged,

non-fragmented ions, so that DNA or RNA compomers reach the detector according to their nucleotide composition and length, those with small molecular weights travelling the fastest. The output is a mass spectral profile determining the molecular masses of the ions in the matrix–analyte mix.

Mass signals in the spectrum are interpreted as compomer masses in Daltons. Nucleotide compomers in a range between 1000 Da (~3 nucleotides) and 10,000 Da (~30 nucleotides) can be resolved with a resolution of ~400–800 (mass in Da/peak width at half peak height). This allows for a distinction down to a single nucleotide difference between compomers.

MassARRAY and Plant Genotyping

The following paragraphs elaborate on agricultural applications, specifically the detection of plants using the genotyping and multiplexing capabilities of the MassARRAY system, and introduce quantitative nucleic acid analysis using MALDI-TOF MS. Known marker regions can be utilized to detect gene sequence variations and quantify them within mixtures.

This MassARRAY process combines three consecutive steps (Fig. 7.3):

1. Marker region-specific PCR;
2. MassEXTEND assay; and
3. MALDI-TOF MS analysis.

The biochemistry is simple and homogeneous. PCR, primer extension reaction and sample conditioning are performed in a homogenous assay format within the same reaction vial (Storm *et al*., 2003). Reaction conditions are universal with no purification required prior to analysis (Fig. 7.1).

The assay design is simple and flexible. An assay design software defines PCR and extension primers so that resulting combinations of mass signals are resolvable in the spectrum and provide unique, unambiguous information for each variant in the mixture. The specificity of the MassEXTEND detection assay is determined by two successive steps, each exhibiting locus-specific primer annealing. This scheme allows for the design of customized detection assays that can be used for plant identification, routine monitoring and disease detection.

Performing multiple reactions in a single vial is a way to increase the analysis throughput and reduce costs. The Whitehead Institute Center for Genome Research validated the system performance in a five-plex study on 598,466 genotypes with a data accuracy of ~99.6% (Gabriel *et al*., 2002). Current developments in SNP genotyping on human genomic DNA enable plexing levels of up to 40 reactions per vial for qualitative analysis (Fig. 7.3).

In addition to the qualitative genotyping call, frequencies of the variant products are obtained as the proportion of either peak area relative to the total area for both expected peaks. This provides a semi-quantitative result of the variant ratios within the mixture. Multiple studies have compared the quantitative ability of various platforms in the context of allele frequencies in populations of nucleic acid molecules. MALDI-TOF MS measurements of

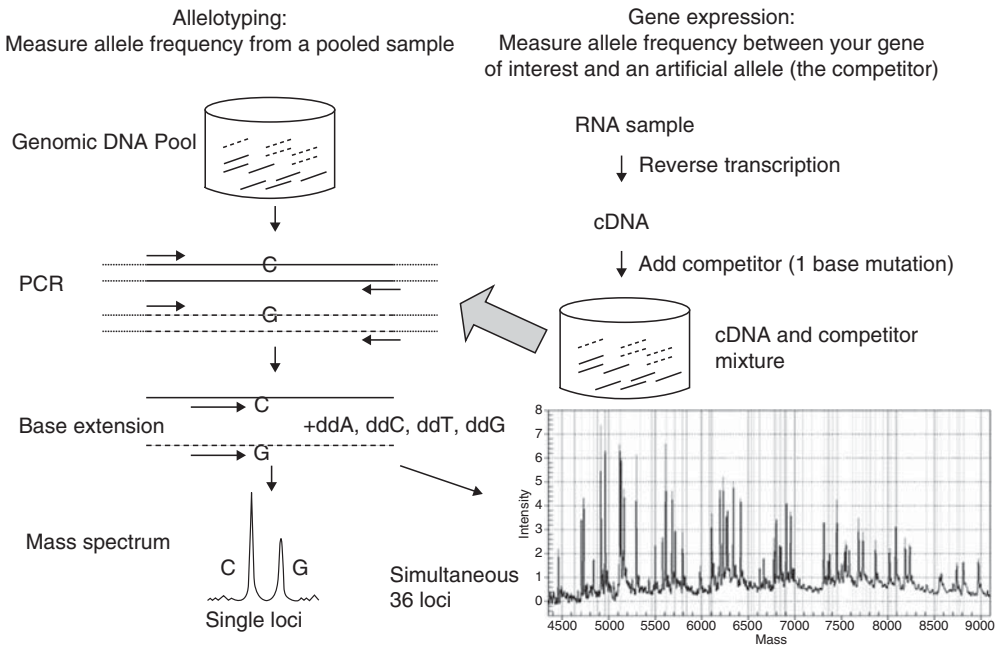


Fig. 7.3. MassEXTEND process. Amplification of the target region by PCR and post-PCR processing by MassEXTEND for plant identification and typing. Assay formats are homogeneous and no purification is required. Semi-quantitative mass spectrum identifies variants present as individual mass peaks and frequency of the alleles in the starting sample. Incorporation of an artificial allele or 'competitor' provides an internal control for absolute quantitation of the number of starting copies of alleles present. Multiplexing provides simultaneous qualitative analysis of up to 36 loci per reaction.

primer extension reaction have been shown to be as accurate, sensitive and reproducible as all available technologies (Le Hellard *et al*., 2002; Sham *et al*., 2002; Shifman *et al*., 2002) with a limit of detection (LOD) of 2% and a limit of quantitation (LOQ) of 5% for the minor allele or variant. For quantitative applications, 4–5 plexes have successfully been applied (Andell *et al*., 2004) with up to 20 reactions per vial currently being developed.

Sugarcane varieties are complex polyploid, aneuploid interspecific hybrids between the domesticated *Saccharum officinarum* ($x = 8$) and the wild relative *S. spontaneum* ($x = 10$). Cultivar chromosome numbers range from 100 to 130 with approximately 10% contributed by *S. spontaneum*. Genetic analysis of sugarcane is made difficult by the presence of 8–16 gene homologues and, as such, reliable mapping and quantifying gene copy number is yet to be mastered.

To perform genetic analysis of sugarcane, polymorphisms must first be identified in the genes of interest (Fig. 7.4) by clonal sequencing or alternative

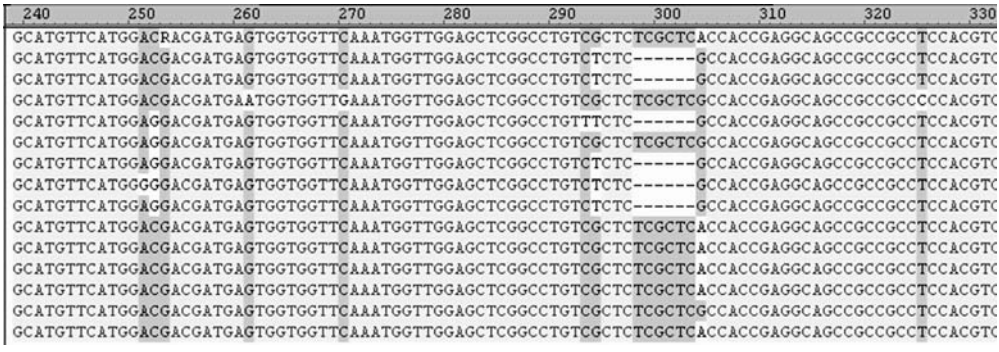


Fig. 7.4. Sequence alignment of a representative gene (90 bases). SNPs that discriminate for a single homologue. Panels of polymorphisms can be developed that in combination discriminate for single copies (shaded). Here a single polymorphism can be used to quantify a group of gene copies and analysis of further polymorphisms can determine the copy that has been duplicated or is absent.

techniques. Ideally polymorphisms will be identified that distinguish a single homologue from the background of up to 15 copies. This presents the first challenge for genetic analysis platforms, detecting the minimum allele ratio of 1:15 (6%) required for gene copy mapping. Quantitative genotyping was performed to determine the presence or absence of the homologue of interest in single plants. The superior signal to noise ratio of the Sequenom Mass Spectrometry-based platform provided clear evidence of the presence of the homologue (Fig. 7.5). Cluster analysis demonstrates high confidence and reproducibility in the results. In addition, the quantitative capabilities and the 1% standard deviation provided the required sensitivity to determine if a gene duplication event had occurred which would result in a change from a 1:15 peak area ratio (6%) to 2:15 (12%).

Due to the high sequence homology between gene copy sequences and limited sequence information available, not all homologues had a polymorphism discriminatory for a single copy. In these instances it is necessary to develop panels of polymorphisms that in combination discriminate for single copies (Fig. 7.4). Here a single polymorphism can be used to quantify a group of gene copies and analysis of further polymorphisms can determine the individual copy that has been duplicated or is absent.

Absolute Quantitation and Gene Expression

PCR inhibitors and the low abundance of the gene copy of interest in some polyploid organisms or within population mixtures are challenges in PCR-based plant detection methods. False negatives may be the result if an organism-specific PCR is inhibited or if the organism/gene copy of interest is present in very low copy numbers. With the MassARRAY system, these issues can be addressed by the addition of a competitive template (Fig. 7.3),

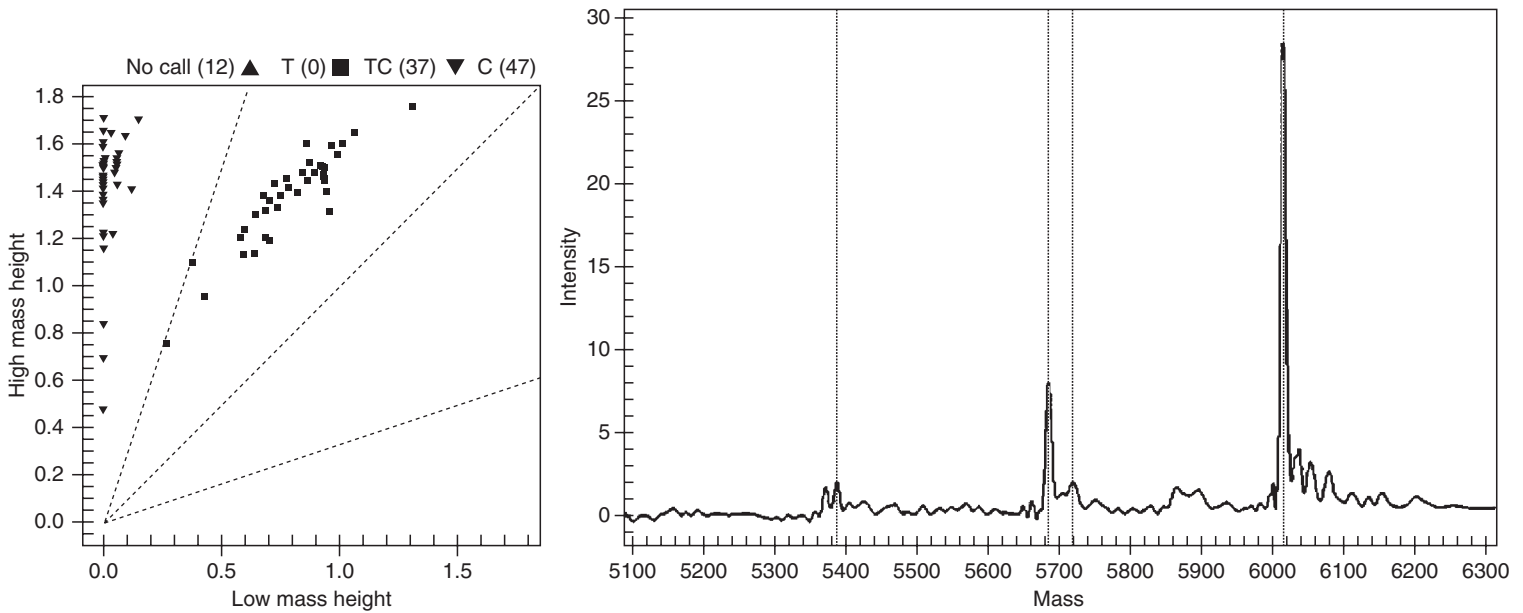


Fig. 7.5. Analysis of a single locus that distinguishes the homologue of interest from the majority allele with up to 15 copies, detecting the minimum allele ratio of 1:15 (6.25%) required for gene copy mapping. A gene duplication event would result in a change from a 1:15 peak area ratio (6.25%) to 2:15 (11.8%).

which serves as an internal standard for inhibition, sensitivity and performance control of the assay and sample. Oligonucleotides or artificial plasmids can be designed to match the actual target sequence except for a single base mutation. This single base change can then be used as a discriminator from the target during MassEXTEND and MALDI-TOF MS.

This approach allows for absolute quantitation if the concentration of the internal standard is known. Unknown analyte concentrations can be determined prior to analysis by titrating the internal standard over a range of concentrations, so that the unknown analyte peak area and the internal standard peak area are detected at an approximate 1:1 ratio (Fig. 7.6). Exact molar concentrations can then be assigned to the unknown analyte via regression analysis of the plotted peak area ratios versus the concentration of the internal standard.

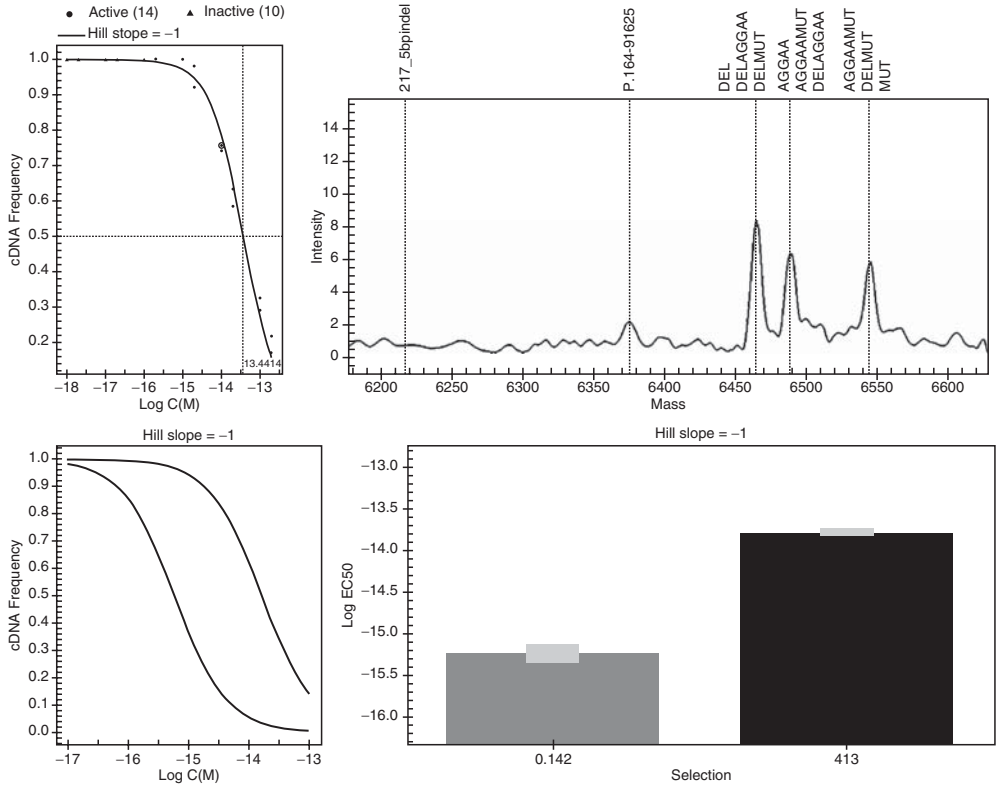


Fig. 7.6. Titration of the internal standard over a range of concentrations, so that the unknown analyte peak area and the internal standard peak area are detected at an approximate 1:1 ratio (top). Absolute quantitation of a genomic sequence polymorphism (bottom). There are clearly more copies of the insertion than the deletion in the sample. Quantitation of the number of copies present, contrasted with the number of genome copies present, can delineate the absolute number of copies of each variant in the sample, for example six copies of the insertion sequence versus five copies of the deletion sequence in the genomic DNA.

If the unknown analyte is a genomic sequence, the numbers of copies can be delineated. Utilizing the power of multiplexing, additional 'housekeeper' sequences of known copy numbers can be incorporated to clearly differentiate between duplication and deletion events. In the above example of sugarcane genotyping, the observed differences in allele frequencies of 1:15 peak area ratio (6%) versus 2:15 (12%) were extrapolated to indicate gene duplication. Similarly this could have been caused by deletion events such that the allele ratio was 1:8 (12%). It was concluded that this event was sufficiently unlikely to not warrant further interrogation.

In studies of less complex plant genomes or more complex copy number variations it may be advantageous to make distinction between duplication of the gene of interest and deletion of the alternate copies. This technique is yet to be applied to analysis of plants; however, it has been extensively used in other organisms. Figure 7.6 provides an example of absolute quantitation of a genomic sequence polymorphism. There are clearly more copies of the insertion than the deletion in the sample. Quantitation of the number of copies present, contrasted with the number of genome copies present, can delineate the absolute number of copies of each variant in the sample, for example six copies of the insertion sequence versus five copies of the deletion sequence.

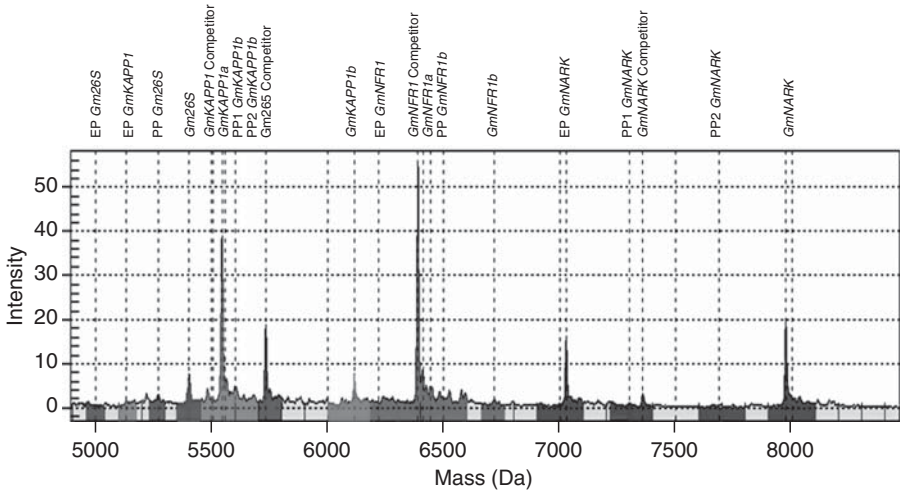
This competitive control approach has been extensively used in quantitative gene expression analysis (Jurinke *et al*., 2003; Ding and Cantor, 2004; Ding *et al*., 2004; Tang *et al*., 2004) with absolute determination of the number of transcripts. Therefore, highly quantitative genotyping of gene copy number can be followed up with accurate and absolute transcriptional profiling of expression using a single technology. Gene expression studies in plants are often hindered by the amount of available material particularly where isolation of sub-tissue samples is required to avoid signal dilution. Additionally, measurement of gene expression at low transcript levels, as found for many transcription factors, regulatory RNAs and receptor kinases, is one of the key components in almost any functional genomics project.

Differential transcription and expression due to allelic variance in the coding region of genes provides an important link between individual genetic variation and disease in simple diploid human systems (Yan *et al*., 2002; Lo *et al*., 2003). Considering the complex polyploidy of plant genomes, it is likely that allele-specific expression biases will provide an important link between genotype and phenotype penetrance. Therefore, monitoring of individual expression of paralogous genes at ultra-low expression levels and/or in extremely small tissue samples is considered highly desirable in plant genomic studies.

Soybean is an ancestral paleopolyploid ($2n = 4x = 40$) that at the DNA level results in genomic duplications of various sizes. Most of the known soybean genes have transcribed homologues (usually genetically unlinked) elsewhere in the genome (Pagel *et al*., 2004). These duplicated sequences often differ by only a handful of SNPs, but their function could differ dramatically. An excellent example is the *GmNARK* gene (which, if mutated, results in the loss of autoregulation of nodulation (AON) and subsequent supernodulation) (Men *et al*., 2002; Searle *et al*., 2003). The soybean genome

contains a homologous and genetically unlinked partner of *GmNARK*, namely *GmCLV1A*, which fails to complement recessive *GmNARK* mutations though it is about 90% identical at the nucleotide level. Likewise the nodulation factor receptor genes *GmNFR1α* and *GmNFR1β*, again unlinked and 96% identical at the DNA level with identical intron–exon structure, are functionally differentiated so that *GmNFR1β* does not complement *Gmnfr1α* recessive mutations.

Assays were designed to interrogate SNPs in a hexaplex format to differentiate between two paralogous genes, with a 'third allele' for the synthetic 'competitor oligonucleotide' serving as an internal reference to estimate the target transcript concentration (Fig. 7.7). We also observed good concordance in absolute copy numbers of each gene regardless of the level of multiplexing or the level of expression. It is noteworthy that the levels of expression in these gene paralogues were low, representing very small copy numbers of starting mRNA (Fig. 7.7). Assays with triple allele design produced consistent resolution for *GmKAPP1a/b* and *GmNFR1α/β* and also showed a good correlation with RT PCR expression values for *GmNARK* normalized against *Gm26S* (Nontachaiyapoom *et al*., 2007).



Gene	*Gm26S*	*GmNARK*	*GmKAPP1a*	*GmKAPP1b*	*GmNFR1a*	*GmNFR1b*
Concentration	0.25 ± 0.02 fM	1.1 ± 0.21 aM	1.0 ± 0.13 aM	1.2 ± 0.11 aM	0.6 ± 0.07 aM	0.2 ± 0.07 aM

Fig. 7.7. MALDI-TOF mass spectra for a multiplex expression assay with six amplicons: *Gm26S*, *GmNARK*, *GmKAPP1a*, *GmKAPP1b*, *GmNFR1α*, *GmNFR1β*. The absolute concentrations of expressed genes in soybean leaf cDNA samples per 5 ng of total RNA are provided in the table. fM and aM are femto-mole (10–15) and atto-mole (10–18) amounts of the original mRNA transcripts, respectively. Standard deviation values are based on four independent reactions.

In order to detect the precise range of concentration where all target DNA molecules were equally represented, it was initially necessary to generate titration curves by using a constant amount of each cDNA sample and decreasing amounts of corresponding competitor sequences. Then the differences in 'area-under-mass-peak' between competitor and target amplicon(s) were calculated for each reaction where the peak areas were approximately 1:1. Figure 7.8 represents an example of such titration for *GmNARK* versus *GmNARK* competitor and both alleles of *GmKAPP1* versus *GmKAPP1* competitor in a hexaplex experiment.

This experiment demonstrates that MALDI-TOF MS readily distinguished between nearly identical gene transcripts and accurately measured differential levels in a single assay using minute amounts of plant total RNA (5ng). The novelty of the method is in its use of a single SNP as a target marker that differs between two paralogous genes, to design the 'third allele' for the synthetic 'competitor oligonucleotide' serving as an internal reference to estimate the target transcript concentration. It is noteworthy that the degree of multiplexing of the genes assayed with the MALDI-TOF platform is higher than for Real Time PCR methods, where at present a maximum of four fluorescent dyes can be resolved (Wittwer *et al*., 2001).

As many agriculturally important plant species manifest high levels of polyploidy (e.g. lucerne, banana, potato and wheat), this methodology could be very effective to monitor paralogous gene expression as well as differential transcription of allelic variants of the same gene. Even simple diploid genomes like *Arabidopsis* have long stretches of highly homologous DNA and duplicated genes (Bevan *et al*., 2001) and therefore would benefit from this analysis. We also believe that this analytical approach will be particularly useful in the analysis of wild-type transgene performance in a mutant background during complementation or overexpression studies.

Conclusions

The MassARRAY system provides a mature application portfolio for plant genomic analysis. Biochemical assays have been adapted to suitable process automation and the analytical speed of mass spectrometry. Results are obtained with real-time control of data quality. Software algorithms translate signal information automatically into plant genomic information, delivering qualitative as well as quantitative analysis results.

Because small mass differences can be detected, MassARRAY can conclusively distinguish highly concordant gene paralogues as the case in many complex plant genomes. In addition, incorporation of an internal standard provides highly quantitative data on absolute copy number. Using this technique it is possible to analyse gene rearrangement, duplication and deletion events as well as expression from closely related genes, for example from those occurring as multi-gene families or in the case of transgene/resident gene expression comparisons.

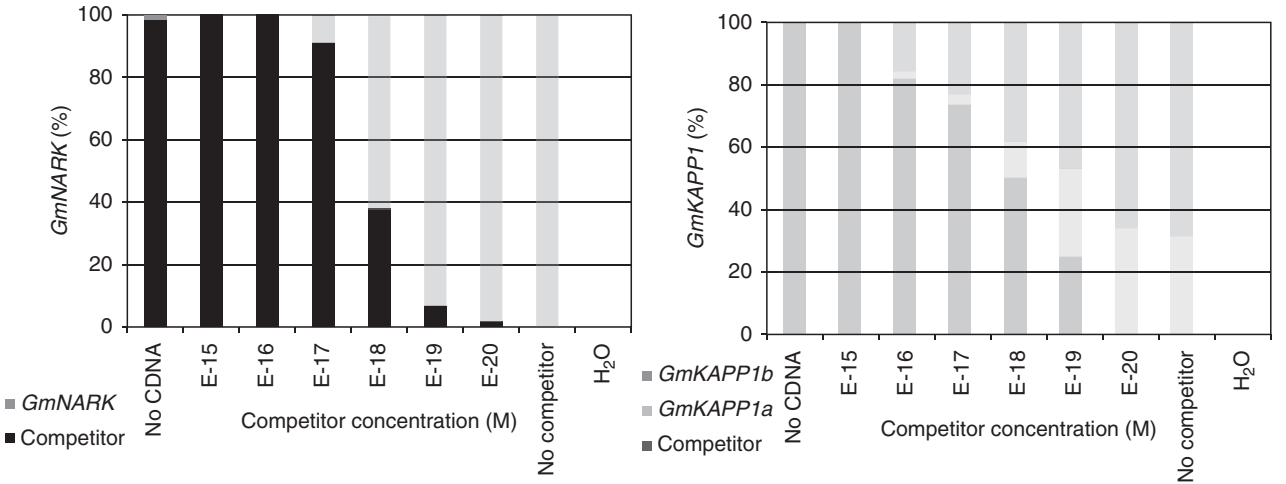


Fig. 7.8. Serial dilution and titration for 'competitor' molecule in a hexaplex PCR for *GmNARK* and *GmKAPP1a/b* transcripts. *X*-axis represents molar amount of the competitor. *Y*-axis shows the proportion of amplified product. The most reliable data point is reached at the 1:1 molar ratio between the competitor peak and the transcript peak (in case of *GmNARK* ~10–18 moles), or at 1:1:1 ratio between the competitor and two homologous transcripts (in case of *GmKAPP1* ~10–19 moles).

Performing multiple reactions in a single vial in addition to parallel processing in 384-well formats in combination with miniaturized sample preparation is a way to increase the analysis throughput and reduce costs. For these reasons MassARRAY shows promise to become a standard method for identification and typing of plants as well as measurement of gene expression at low transcript levels – a key component in almost any plant functional genomics project.

Acknowledgements

Thanks to Artem Men, David Hawkes and Rust Turulakov for their contribution to the figures and content of this chapter. Thanks also to Christiane Honish, Dirk van den Boom and Paul Oeth for background content and figures on MassARRAY and Mass Spectrometry.

References

Andell, Y., Correll, D., Oeth, P. and Jurinke, C. (2004) Multiplexed gene expression analysis using competitive PCR and MassARRAY. *Sequenom Application Note*. Available at: www.sequenom.com

Bansal, A., van den Boom, D., Kammerer, S., Honisch, C., Adam, G., Cantor, C.R., Kleyn, P. and Braun, A. (2002) Association testing by DNA pooling: an effective initial screen. *Proceedings of the National Academy of Sciences of the USA* 99, 16871–16874.

Beaulieu, M. and Kosman, D. (2003) Molecular haplotyping combining a novel AS-PCR technique with base-specific cleavage and mass spectrometry. In: *SNPs, Haplotypes and Cancer: Application in Molecular Epidemiology*. Conference Proceedings, Key Biscayne, Florida, September 13–17, A60.

Bevan, M., Mayer, K., White, O., Eisen, J.A., Preuss, D., Bureau, T., Salzberg, S.L. and Mewes, H.-W. (2001) Sequence and analysis of the *Arabidopsis* genome. *Current Opinions in Plant Biology* 4, 105–110.

CorbachoCorbacho, J.L., SuarezSuarez, J.C., MolledaMolleda, F., Graber, J.H., Smith, C.L. and Cantor, C.R. (1999) Differential sequencing with mass spectrometry. *Genetic Analysis: Biomolecular Engineering* 14, 215–219.

Ding, C. and Cantor, C.R. (2004) Quantitative analysis of nucleic acids – the last few years of progress. *Journal of Biochemistry and Molecular Biology* 37, 1–10.

Ding, C., Maier, E., Roscher, A.A., Braun, A., Cantor, C.R. (2004) Simultaneous quantitative and allele-specific expression analysis with real competitive PCR. *BMC Genetics* 5, 8.

Downes, K., Barratt, B.J., Akan, P., Bumpstead, S.J., Taylor, S.D., Clayton, D.G. and Deloukas, P. (2004) SNP allele frequency estimation in DNA pools and variance components analysis. *Biotechniques* 36, 840–845.

Ehrich, M., Nelson, M., Stanssens, P., Zabeau, M., Liloglou, T., Xinarianos, G., Cantor, C.R., Field, J.K. and van den Boom, D. (2005) Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. *Proceedings of the National Academy of Sciences of the USA* 102, 15785–15790.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J. and Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science* 296, 2225–2229.

Honisch, C., Raghunathan, A., Cantor, C.R., Palsson, B.O. and van den Boom, D. (2004) High-throughput mutation detection underlying adaptive evolution of *Escherichia coli*-K12. *Genome Research* 14, 2495–2502.

Jurinke, C., van den Boom, D., Cantor, C.R., Köster, H. and Hoheisel, J. (2002) The use of MassARRAY technology for high throughput genotyping. *Advances in Biochemical Engineering/Biotechnology* 77, 57–74.

Jurinke, C., Oeth, P. and Correll, D. (2003) Gene expression analysis using MassARRAY. *Sequenom Application Note*. Available at: www.sequenom.com

Jurinke, C., Oeth, P. and van den Boom, D. (2004) MALDI-TOF mass spectrometry: a versatile tool for high-performance DNA analysis. *Molecular Biotechnology* 26, 147–164.

Kammerer, S., Roth, R.B., Reneland, R., Marnellos, G., Hoyal, C.R., Markward, N.J., Ebner, F., Kiechle, M., Schwarz-Boeger, U., Griffiths, L.R., Ulbrich, C., Chrobok, K., Forster, G., Praetorius, G.M., Meyer, P., Rehbock, J., Cantor, C.R., Nelson, M.R. and Braun, A. (2004) Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus. *Cancer Research* 64, 8906–8910.

Kammerer, S., Roth, R.B., Hoyal, C.R., Reneland, R., Marnellos, G., Kiechle, M., Schwarz-Boeger, U., Griffiths, L.R., Ebner, F., Rehbock, J., Cantor, C.R., Nelson, M.R. and Braun, A. (2005) Association of the *NuMA* region on chromosome 11q13 with breast cancer susceptibility. *Proceedings of the National Academy of Sciences of the USA* 102, 2004–2009.

Karas, M. and Hillenkamp, F. (1988) Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry* 60, 2299–2301.

Le Hellard, S., Ballereau, S.J., Visscher, P.M., Torrance, H.S., Pinson, J., Morris, S.W., Thomson, M.L., Semple, C.A.M., Muir, W.J., Blackwood, D.H.R., Porteous, D.J. and Evans, K.L. (2002) SNP genotyping on pooled DNAs: comparison of genotyping technologies and a semi automated method for data storage and analysis. *Nucleic Acids Research* 30, e74.

Lo, H.S., Wang, Z., Hu, Y., Yang, H.H., Gere, S., Buetow, K.H. and Lee, M.P. (2003) Allelic variation in gene expression is common in the human genome. *Genome Research* 13, 1855–1862.

Men, A.E., Laniya, T.S., Searle, I.R., Iturbe-Ormaetxe, I., Gresshoff, I., Jiang, Q., Carroll, B.J., and Gresshoff, P.M. (2002) Fast neutron mutagenesis of soybean (*Glycine soja* L.) produces a supernodulating mutant containing a large deletion in the linkage group H. *Genome Biology* 1, 147–155.

Nontachaiyapoom, S., Scott, P.T., Men, A.E., Kinkema, M., Schenk, P.M. and Gresshoff, P.M. (2007) Promoters of orthologous soybean and *Lotus japonicus* nodulation autoregulation genes interchangeably drive phloem-specific expression in transgenic plants. *Molecular Plant–Microbe Interactions* 20, 769–780.

Pagel, J., Walling, J.G., Young, N.D., Shoemaker, R.C. and Jackson, S.A. (2004) Segmental duplications within the glycine max genome revealed by fluorescence *in situ* hybridization of bacterial artificial chromosomes. *Genome* 47, 764–768.

Searle, I.R., Men, A.E., Laniya, T.S., Buzas, D.M., Iturbe-Ormaetxe, I., Carroll, B.J. and Gresshoff, P.M. (2003) Long-distance signaling in nodulation directed by a CLAVATA1-like receptor kinase. *Science* 299, 109–112.

Sham, P., Bader, J.S., Craig, I., O'Donovan, M. and Owen, M. (2002) DNA Pooling: a tool for large-scale association studies. *Nature Review Genetics* 3, 862–871.

Shifman, S., Pisante-Shalom, A., Yakir, B. and Darvasi, A. (2002) Quantitative technologies for allele frequency estimation of SNPs in DNA pools. *Molecular Cell Probes* 16, 429–434.

Stanssens, P., Zabeau, M., Meersseman, G., Remes, G., Gansemans, Y., Storm, N., Hartmer, R., Honisch, C., Rodi, C.P., Bocker, S. and van den Boom, D. (2004) High-throughput MALDI-TOF discovery of genomic sequence polymorphisms. *Genome Research* 14, 126–133.

Storm, N., Darnhofer-Patel, B., van den Boom, D. and Rodi, C.P. (2003) MALDI-TOF mass spectrometry-based SNP genotyping. *Methods in Molecular Biology* 212, 241–262.

Tang, K., Oeth, P., Kammerer, S., Dennisenko, M.F., Ekblom, J., Jurinke, C., van den Boom, D., Braun, A. and Cantor, C.R. (2004) Mining disease susceptibility genes through SNP analyses and expression profiling using MALDI-TOF mass spectrometry. *Journal of Proteome Research* 3, 218–227.

Werner, J.D., Borevitz, J.O., Warthmann, N., Trainer, G.T., Ecker, J.R., Chory, J. and Weigel, D. (2005) Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proceedings of the National Academy of Sciences of the USA* 102, 2460–2465.

Wittwer, C.T., Herrmann, M.G., Gundry, C.N. and Elenitoba-Johnson, K.S. (2001) Real-time multiplex PCR assays. *Methods* 25, 430–442.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., Kresovich, S. and Buckler, E.S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Natural Genetics* 38, 203–208.

Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2002) Allelic variation in human gene expression. *Science* 297, 1143.

8 Mutation Screening

L. Izquierdo

Spontaneous Mutations

Naturally, a mutation is a spontaneous change occurring within a gene, potentially altering it permanently and subsequently enabling it to be passed on from generation to generation. The ability of a mutation to produce a phenotypical change is determined by the degree of dominance, the cell type affected and the stage of the organism's life cycle. Its inheritability in multicellular organisms implies that the change must occur within cells committed to produce gametes (or germ line cells). These mutated cells must be able to compete, continue to persist during the ontogenesis of the plant, participate in the sporogenesis phase and consequently generate chimeric individuals in the next reproductive event (Micke and Donini, 1993). In nature, mutation is the main evolutionary driving force that generates variation and it is essential in producing the variability that allows organisms to adapt to new environments. The frequency associated with spontaneous or natural mutations is low (~10^{-6}–10^{-8} per locus/haploid genome/plant generation). Spontaneous changes in DNA can occur through diverse processes including replication errors, polymerase accuracy failure, free radicals of oxygen (oxidation), base degradation (deamination and depurination) and the addition of alkyl groups such as methyl, ethyl and occasionally propyl to the bases (Montelone, 1998).

Several classification schemes for mutations have been proposed (Van Harten, 1998). A broad classification is shown in Table 8.1. In general, mutations have been classified as deletions, insertions, substitutions, duplications, translocations and inversions. Their effect on the functionality of the final product (protein) varies from having no effect to not functioning at all. Among other things, this depends upon the size of the mutation, the location within the genome, which chemical process originated it (oxidation being the most harmful) and how efficiently the DNA-repaired mechanisms work (Van Harten, 1998; Montelone, 1998).

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

Table 8.1. Types of mutations occurring in genes and chromosomes. (From Van Harten, 1998.)

Gene	Base substitution (single point)	Missense	Alteration of a codon resulting in a different amino acid in the peptide
		Nonsense	Change of a codon that specifies for an amino acid to a stop codon resulting in a truncate protein
		Silent	The change in the DNA does not alter the final product
		Splice site	Alteration of the sequence in the intron region generating a wrong splicing and producing an incorrect protein
	Insertions and deletions		One or two base pairs are added or removed from the DNA sequence. It might cause the reading frame shift to lead to a useless protein
Chromosome	Structure	Duplications	Some genes are duplicated
		Translocations	Transference of DNA segments between non-homologous chromosomes. It is generally a reciprocal event between chromosome fragments
		Inversions	A region of the chromosome flips its orientation, opposite to the rest of the chromosome
		Deletions	Loss of a chromosome section (several genes lost)
		Fragmentations and fusions	
	Number		Failure to separate the opposite pole of the chromosome during cell division resulting in one cell daughter with an extra chromosome and the other with one less

Plant Breeding

Plant breeding is the production of plants that are genetically improved so as to express certain traits required to suit human needs. Plant breeding has been extensively used in crop domestication. Plants with an altered phenotype, as a result of spontaneous mutations, were probably selected and used by early humans from the moment they began to utilize plants for food, clothing and housing. For crop domestication, it all started around 10,000 years ago, with the selection of plants with desirable traits such as producing a high yield, seed recovery, resistance to pests and diseases and the ability to grow in man-made environments. However, by the end of the 19th century farmers had accumulated some basic information regarding hybridization, segregation

and recombination procedures, and they used that knowledge to experiment with fertilization processes. The manipulation of crops has not stopped since, but the methods applied have changed (Micke and Donini, 1993).

Induced Mutations

In the last century we have seen an enormous development of agricultural systems and have gained a great amount of knowledge of the genetic material, DNA. The green revolution was in some way possible due to our better understanding of genetics and the subsequent use of innovative molecular approaches such as induced mutations. It all started with the discovery of the X-ray's ability to induce mutations in *Drosophila* by Hermann Muller in 1927. As a result of this finding a completely new approach towards genetic analysis was born (Snustad and Simmons, 2006). Induced mutations are caused by external agents, or mutagens, and are randomly generated within the genome. They provide simple and relatively efficient means of altering genetic make-up and obtaining desired genotypes from already well-adapted ones (FAO/IAEA, 1998, 2004).

Induced mutation technology has increased plant breeding efficiency for those traits without obvious phenotypes, enhanced natural genetic resources and improved cereal, fruit and other crop cultivars (Lee *et al*., 2002). It has also played an important role in helping us to understand the changes that take place in DNA at a molecular level; fingerprinting genotypes precisely; improving the yield and genetic background of specific cultivars; increasing resistance to pests, toxicity and diseases; developing genetic maps; and generating new attributes with a market value to suit human needs (Bacha *et al*., 2002). Since 1966, over 2300 mutant plants, from both seed and propagated crops (http:/www-mvd.iaea.org/), have been officially released from 59 countries (Jain, 2005).

The fact that many critical steps in relevant biosynthetic pathways – as well as important characteristics distinguishing cultivated and domesticated plants from their wild relatives – are controlled by one or a few major genes makes feasible the use of breeding and mutation technologies (FAO/IAEA, 2004). Superior cultivars from important crops such as maize, soybean, cotton, tomato and wheat have been developed through mutation technologies (FAO/IAEA, 2003; Liu *et al*., 2004; Jain, 2005).

One advantage of induced mutations is that they occur at a higher rate than spontaneous mutations (10^3 times higher), thereby increasing the chances that more than one gene changes; yet many of them will go unnoticeable. It has been suggested that some regions within a chromosome (which are not evenly distributed) are gene-rich and may be more vulnerable to mutations (Sandhu and Gill, 2002). One example of this has been found in mutagenized rice, where genes associated with the life cycle and amylose content showed more mutations than those associated with gelatinization temperature (Bacha *et al*., 2002).

The number of induced mutations occurring depends upon several factors such as plant species, the amount of primary damage, genotype (Wu *et al*.,

2005), tissue and physiological condition of the cells (particularly the water content), ploidy level, efficiency of the DNA-repair mechanisms and the probability that the mutation produces an altered phenotype (Montelone, 1998; Wu *et al*., 2005). Kimball (1987), in his comprehensive review on the processes of DNA repair, commented on the complexity of the mechanisms developed by plant cells to protect their genome stability after being altered by mutagenic treatments.

There are two main categories of mutagenic agents: physical (or radiation) and chemical. A third category, biological techniques, has been suggested to include the use of callus culture (due to its potential to generate somaclonal variation) and transposable elements (Van Harten, 1998).

Physical Mutagens

Physical mutagens comprise corpuscular radiation (e.g. moving particles as protons, neutrons) and electromagnetic waves. The portion of the electromagnetic spectrum with wavelengths shorter, and of higher energy, than visible light is subdivided into ionizing radiation (X-, gamma and cosmic rays) and non-ionizing radiation (Snustad and Simmons, 2006). Ionizing radiation can penetrate tissue, and in doing so it collides with atoms, promoting the ejection of electrons (Lapins, 1983; Jain, 2005). The ionization of the atoms of plant tissue leads to the formation of free radicals that react with plant molecules, thus releasing more electrons and altering the structure of the DNA. Although these physical and radiochemical events occur very quickly, their effects take time to appear. Physical mutagens tend to produce large deletions from 17 bp to 20 cM in length (Naito *et al*., 2005) but they can also generate complex inversions and translocations. Small deletions probably result from two or three hits occurring very close to one another. Nevertheless, Sato *et al*. (2006) reported point mutations, transversions, transitions and 2–4 bp deletions in material irradiated by gamma rays.

Chemical Mutagens

Chemical mutagens have been used mainly in seed-propagated crops. They tend to produce point mutations, small alterations such as deletions up to 50 bp and high density of irreversible mutations. There are different types of chemical mutagens based on their chemical structure and mode of action (Van Harten, 1998; Jain, 2005).

The first report of a chemical mutagenic action was in 1942 by Charlotte Auerbach, who showed that mustard gas, used in the First and Second World Wars as a toxic weapon, was able to induce mutations in *Drosophila*. Soon after the Second World War, a Swedish group, led by Åke Gustafsson, applied the mustard gas to barley, *Hordeum vulgare* (Van Harten, 1998).

Since then many other mutagenic chemicals have been identified, characterized and applied to higher plants. Of specific interest are those that alter

DNA structure and the pairing properties of bases. In particular, chemical mutagens with alkyl groups (C–H molecules with unpaired electrons) react with plant tissues through an alkylation process (where an H atom in the DNA molecule is replaced by an alkyl radical such as ethyl, methyl, propyl and *n*-ethyl groups) interfering in the synthesis of the genetic material by yielding a baseless site or facilitating a mispairing between the bases (Lapins, 1983; Jain, 2005).

In the late 1950s, a considerable awareness of the potential of alkylating agents as efficient mutagens in higher plants was established. Some well-characterized and frequently used alkylation agents are: diethyl sulphate (DES), ethyl methane sulphonate (EMS), methyl methane sulphonate (MMS) and isopropyl methane sulphonate (iPMS). These agents are very reactive, even in water (FAO/IAEA, 1977).

EMS is the main chemical mutagen used. Its mutagenic effect is remarkably consistent on species diverging 1 billion years ago, as *Arabidopsis thaliana* and *Drosophila melanogaster*, or with very different genome sizes. EMS alkylates G bases, and consequently they mispair with T instead of C, generating mutations G/C to A/T transitions with heteroduplexes such as G/T or A/C mismatches. EMS is usually used for the mutagenesis of seeds in concentrations of between 20 and 100 mM for up to 10–20 h. Seeds are then washed many times with water and finally sown (FAO/IAEA, 1977; Greene *et al*., 2003; Henikoff and Comai, 2003).

When using EMS, the treatment should be long enough to allow complete penetration of the mutagen into the target tissue. Data have shown that the rate of uptake and level of saturation of the embryo depend upon the seed size, permeability of the seedcoat and contents of cell constituents. It is possible to reduce the time seeds spend in the EMS solution by soaking them in water before applying the mutagen. This pre-soaking treatment induces the germination process, causing the seeds to enter a meristematic stage. As a result, they are more sensitive to mutation induction but with less chance of suffering chromosomal damage (FAO/IAEA, 1977).

Mutagens and Plant Material

The way cells of higher plants respond to physical mutagens is influenced, to a varying degree, by a number of biological, environmental and chemical factors (Atak *et al*., 2004). In seed-propagated plants, seeds are the most convenient and practical starting material for most mutagenic experiments, as they are easily stored and handled. However, it is important to keep in mind that there is a genotypical response, and that the water content of seeds could modify the effectiveness (mutation per unit dose) and efficiency (ratio of mutation) of the treatment. If seeds are used, they should be soaked prior to mutagenic treatment so as to facilitate penetration of the mutagen (Van Harten, 1998). A large plant population must be available for treatment (usually, thousands of seeds are required) and post-treatment regeneration to prevent losses due to sterility and lethality.

When selecting a mutagen it is important to consider its effectiveness, efficiency, specificity and a balance between frequency of generated mutations per

Table 8.2. Advantages and disadvantages of chemical mutagens. (From Montelone, 1998; McCallum *et al.*, 2000; Jain, 2005.)

Advantages	Disadvantages
Generate mainly point mutations	Difficult penetration in multicellular plant tissues
Less chromosomal damage occurred	Uneven penetration
Generate a different mutation spectra	Low reproducibility
High mutation frequency	Involves more safety precautions
Due to its reliability, the probability of recovering a knockout can be calculated in advance	Not precise dosimetry
A range of useful missense mutations that can be very useful in genetic studies	Persistence of the mutagen in the treated material
	Health risks-associated handling

exposure unit of the mutagen and frequency of unwanted effects, e.g. chromosome aberrations, sterility and lethality (Lapins, 1983). Among all, EMS and gamma rays are the most popular mutagens used for inducing mutations (Lapins, 1983; McCallum *et al*., 2000; Jain, 2005). Table 8.2 describes the main advantages/disadvantages of chemical mutagens over physical ones. To achieve a high mutation rate it is important to use an optimal dose of the mutagen. A dose can be defined as a particular quantity for a definitive period of time at a particular temperature. For physical mutagens doses are expressed as roentgens or rads/min or Gy (Grays); 1 Gy = 100 R (Van Harten, 1998). If relevant data are not available for the specific plant material a preliminary experiment applying different doses (dosimetry experiment) must be performed. Very high doses can lead to sterility or lethality while doses too low would result in lower mutation density. The ideal dose would be one that induces the highest number of mutations while keeping the plant alive and able to reproduce and generate a mutant population (Wu *et al*., 2005). In general, a dose around what is called the lethal dose (LD_{50}) is applied; at this value 50% of the irradiated material dies as a result of the treatment (Lapins, 1983).

The first generation after the mutagenic treatment is M1 and it is a chimera population containing undetectable mutation which in many cases will not be inherited by the next generation, M2. The second generation of plants will be the first that can be screened for mutations using technologies, such as PCR. Any possible mutant will be in a heterozygous state (Henikoff and Comai, 2003). A successful outcome of a screening project relies on the method chosen. It must be efficient, highly sensitive and easy to apply.

Reverse Genetics

The development of functional genomics as a discipline has been possible due to the completion of DNA sequencing for a series of genomes, bioinformatics

advances and introduction of new high-throughput technologies that can be applied to any organism, as long as the available sequence information is adequate (Larson *et al*., 2000; Tang *et al*., 2004). Reverse genetics aims to identify the function of a gene with known sequences by phenotype analysis of cells and individuals in which the function is impaired (Perry *et al*., 2003). Technologies, such as transposon-tagging, T-DNA and physical mutagenesis; RNA-mediated gene silencing or RNA interference and TILLING (targeted induced local lesions in genomes) (http://en.wikipedia.org/wiki/Reverse_genetic) are commonly used to study gene functions. Reverse genetics also involves the process of genotyping or analysis of existing polymorphism in an individual DNA sample (Comai *et al*., 2004; Gilchrist *et al*., 2006).

In the past, crop domestication and improvement was possible through the wider range of visual markers available for increased yield, adaptation and quality traits. However, those markers have an important limitation: the requirement of a visual phenotype. The advances in DNA research have led to the development of technologies (e.g. precise fingerprinting, genetic maps, molecular markers, mutation technology) that, used in combination, have increased the efficiency of plant breeding applications for those traits that do not have obvious phenotypes. Additionally, the recent development of robots for DNA extraction, multiplex analysis, use of capillary electrophoresis (CE), high-throughput technique platforms and software have made a huge impact on the discovery of polymorphisms in many organisms.

High-throughput Techniques

High-throughput methods involve the analysis of large number of samples in parallel by using diverse tools. In the context of mutation detection, they must be able to discriminate the variation, be inexpensive and simple to perform. The idea to uncover mutations by high-throughput DNA screening was first applied on nematode by Liu *et al*. (1999) and it consisted in using a pair of primers to target a specific gene to screen DNA pools of putative mutants (Wu *et al*., 2005).

A popular high-throughput technique is CE. It was first introduced in 1989 by Beckman instruments (Lin and Barringer, 2004). This type of electrophoresis is performed by highly flexible capillaries made of fused silica, which are 50 µm in the inner diameter, 150–350 µm in the outer diameter and 30 cm in length. The capillaries are filled with a polymer that has sieving properties, which allows separation of DNA fragments into their respective sizes following migration in an electric field produced within the capillary (Dovichi and Zhang, 2000). A very high voltage (~15 kV) is applied to generate an electrical field that is easily dissipated from the capillary. Movement of molecules in CE is complex; it depends not only on their mass and net charge but also on the electric-osmotic flow (Le *et al*., 1997).

CE offers some advantages over slab gels: automated sample loading, higher resolution and sensitivity, high analysis speed (up to several hundred assays per day); small reagent consumption, single primer labelling allowing the use of

various colours to label several fragments; and the ability to discriminate small size differences along with automated genotype calls (Kozlowski and Krzyzosiak, 2001). CE has made easier the detection of unknown mutations.

The ABI 3110/3730 genetic analysers (Applied Biosystems) are instruments used in conjunction with CE to detect variants in a population. They are fast, precise, sensitive (able to detect differences as small as 1 bp) and high-throughput platforms. These analysers are built for loading 16, 48 or 96 samples, which are injected automatically during the run. The instrument has a series of automatic features that minimizes the intervention of an operator and decreases the risk of human errors for sample handling and polymer delivery (Le *et al*., 1997). The analyser has a tray from which samples are injected into the capillaries at the anode end; then the capillaries are moved to an anode buffer chamber where a high electrical field (~15 kV) is applied. The cathode end of the instrument contains a buffer chamber and a pump to supply fresh polymer for each run (Andersen and Larsen, 2004). The separated DNA fragments are detected at the cathode end by a laser-induced fluorescence detector that registers the fluorescence signal in four different spectral channels. The separation of samples occurs at ambient temperature through the use of a heating plate that controls the temperature of the entire capillary. The ABI 3730 is capable of analysing samples serially and it uses an internal algorithm to calibrate samples against an internal control. Having four spectral channels offers the possibility of multiplexing through the use of four dyes in combination with several size fragments (up to 430 bp). ABI 3730 is also able to analyse fragments between 450 and 800 bp although the chemistry of the platform at the present time does not work very efficiently. Additionally, the sensitivity of the platform slows down when it tries to analyse data at the end of the fragment. All signals are collected and stored in a designated computer. Results are obtained as raw data that can be displayed graphically (electropherograms) showing fluorescence intensity in the ordinate and data points (or bp) in the abscissa. The raw data are normalized by the computer, using the peak position of a DNA standard, before running the analysis with specific software. DNA variants or mutants cause a mobility shift in the electrophoresis and appear as electropherogram alterations when compared to profiles of control samples.

The level of sensitivity of ABI platforms allows PCR multiplexing and pooling of DNA samples (Slade and Knauf, 2005). PCR multiplexing requires some time for optimization of concentration of PCR reagents such as $MgCl_2$, buffer, primers, template and cycling conditions but it is worthwhile if thousands of samples are going to be screened (Lachtermacher *et al*., 2000). Regarding DNA pooling, it should be properly tuned; too much pooling would reduce sensitivity. For example, a mutation rate of 10^{-5}/nucleotide and a fragment of 1 kb mean that a mutation will be present on average once per 100 fragments. In this case an 8× pooling will require approximately 20 assays (100/8), including the first pool screening plus the individual assays in the second round. Only those samples that come out positive for variation after individual screenings will be sequenced (Henikoff and Comai, 2003).

TILLING

TILLING (McCallum *et al*., 2000) is a very efficient, sensitive, cost-effective and high-throughput reverse genetics technology for screening variants in a mutant population (Slade *et al*., 2005). It combines chemical mutagenesis (specifically EMS) of seeds or pollen, targeted PCRs and an enzymatic DNA cleavage (Taylor, 1999). It has been used to generate allelic series for specific genes including missense and knockout mutations which can be very useful in gene function studies (Perry *et al*., 2003). However, its main application has been in detecting single natural or induced nucleotide polymorphisms (SNPs), which are changes in a single base at specific positions in the genome (Gut, 2004; Sood *et al*., 2006). TILLING was originally developed to fit denaturing high performance liquid chromatography (HPLC) requirements. However, it has been successfully adapted to agarose and sequencing gels; and to high-throughput technologies such as CE (Greene *et al*., 2003; Henikoff and Comai, 2003; Sood *et al*., 2006).

Once seeds have been treated with EMS, they are sown and the resulting plants (M1) are taken to the next generation (M2) to allow any potential mutation to be stably inherited. The next step is to extract DNA from leaves of M2 plants and store the seeds from that generation in order to create a stock material or TILLING population that can be a resource over time (Fig. 8.1A). Building a TILLING population can take up to 2 years depending on the species involved. The second stage in TILLING technology is the use of PCR technology. Primers targeting specific genes are designed, and at least one in a pair is dye-labelled and used to amplify PCR fragments from pooled DNA samples from several individuals. PCR products are then denatured and let to reanneal at room temperature. This will allow the formation of mismatched base pairs (Fig. 8.1B) which destabilize the DNA helix. These mismatches or heteroduplexes represent natural or induced (caused by mutagens) SNPs. After heteroduplexes are formed, the next step involves the use of an endonuclease, CEL I that is able to recognize these destabilized regions in DNA helices and cut them at the 3' end of the mismatch. CEL I is a single-stranded specific nuclease that attacks DNA and RNA; it is present in several plant tissues and active between pH 6 and 9. It was originally isolated from celery (Oleykowski *et al*., 1998), but other CEL-like enzymes have been found in lucerne, sprout, asparagus and tomato. Sato *et al*. (2006) used the extract of CEL I-like nuclease from petioles of *Brassica rapa* and from celery and found no differences in activities. All these nucleases recognize DNA conformational changes due to mismatches and cleavage in them on one of the two strands. It has been reported that CEL-like enzymes require the presence of TAQ polymerase to be functional through a mechanism which is still unknown (Oleykowski *et al*., 1998).

TILLING is an updated version of mutation breeding, a technology that has been practised for decades by plant breeders. It differs from mutation breeding in the method used to detect mutations. While traditional mutation breeding has been used primarily for readily observable phenotypes,

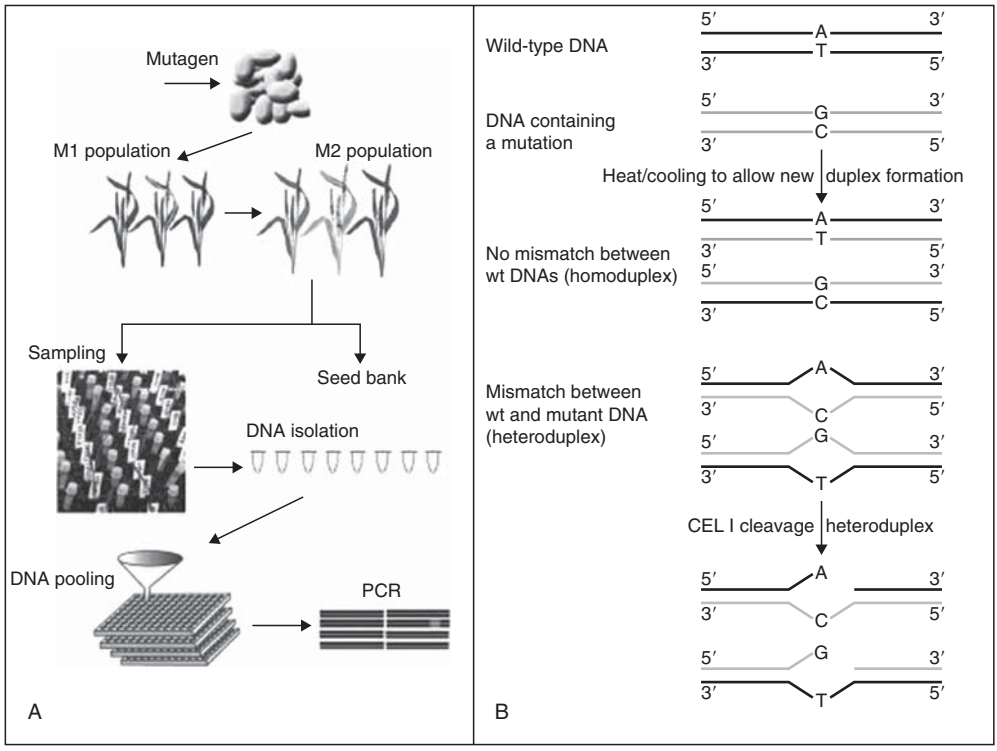


Fig. 8.1. TILLING process (A) and detail of duplex formation and CEL I activity (B).

such as plant height or disease resistance (Slade and Knauf, 2005), TILLING detects the mutation directly in the DNA sequence of the gene of interest. TILLING is a high-throughput technology, sensitive enough to detect a mutation in a pool of 16, cost-effective and fast (Henikoff and Comai, 2003; Till *et al*., 2003).

A similar protocol to detect mutants can be used when applying X- or gamma rays. However, in this case there is no cleavage step. PCR samples are directly analysed, in pools or individually, by CE in an ABI analyser. DNA fragments containing mutations will generate different DNA profiles compared to the wild type and must be sequenced to confirm the presence of a mutation (Lachtermacher *et al*., 2000).

Starch as an example of application of mutation technology

Starch is the most abundant glucose polymer in plants and a primary source of dietary carbohydrates. It is also used extensively for various industrial purposes. In cereals, the endosperm tissue is an important storage of starch. These carbohydrate reserves develop in a membrane-bound organelle, the plastid (amyloplast), which is lost at maturity and is accumulated within the

endosperm as granules embedded in a matrix of proteins. Starch is an association of two main polysaccharides arranged radially, forming concentric layers of amylopectin (usually covers 70–80% w/w of the grain) with amylose molecules overlaying it. Starch is also associated with other minor components such as lipids and proteins either inside or on the surface (Buleon *et al*., 1998). In contrast to tubers and legume starches, cereal starches are characterized by the presence of monoacyl lipids (free fatty acids (FFAs)) and lysophospholipids (LPLs) in amounts positively correlated to amylose content (Stone, 1996).

Amylopectin is a large and usually highly branched polymer, forming sometimes a helical structure that holds diverse molecules, such as fatty acids and fatty alcohols. The extensive branching of amylopectin provides an open structure that is more accessible to digestive enzymes. Instead, amylose is a linear polymer, smaller and more soluble than amylopectin. Its glucosyl monomers are joined via α-1,4 linkages. Amylose seems to play an important role on gel firmness, being the chief material in forming gel networks during water-binding of starch. The behaviour of starch in water at different temperatures determines its potential end-use properties in the food industry (Beta *et al*., 2000).

Main enzymes involved

In the biosynthesis of starch the most important enzymes are:

- ADP-glucose pyrophosphorylase, which catalyses the first reaction in starch synthesis, rending ADP-glucose from glucose 1-PO_4 and ATP.
- Soluble starch synthase (SSS), which is involved in the elongating of linear chains by the formation of α-1,4 linkages (Anglani, 1998). Four isoform synthases, SSI, SSIIa, SSIIb and SSIII, have been reported and they seem to be implicated only in the synthesis of amylopectin.
- Granule-bound starch synthase (GBSS), which is encoded by the *waxy* (*Wx*) locus and is in charge of synthesis and elongation of amylose. *Waxy* mutants typically show a remarkable reduction in amylose. However, amylose content does not only depend on the *Wx* dosage.
- Two debranching enzyme (DBE) families exist in plants, isoamylase and pullulanase types. Both hydrolyse α-1,6 linkages but differ in substrate specificity.
- Finally, the starch branching enzyme (SBE), which is responsible for the branching processes through the generation of α-1,6 linkages (Repellin *et al*., 2001; James *et al*., 2003; Mutisya *et al*., 2003).

Starch genes show conservative domains among cereal crops, although, multiple isoforms of SBE have been reported for barley, wheat, maize, rice, pea, potato and *Arabidopsis* (Opsahl-Ferstad and Rudi, 2000). They are classified as I (B) or II (A) depending on their substrate specificity and gene expression patterns. The SBEII family can be further divided into IIa and IIb isoforms. They are encoded by different genes (*sbeIIa* and *sbeIIb*) and exhibit differen-

tial expression patterns during plant development. For example, in barley the *sbeIIb* gene expression is restricted to the endosperm while in maize it is also detected in the embryo. It has been recently reported that the specificity of *sbeIIb* to endosperm is mediated by the BbI element of the large second intron of the gene. BbI is a *cis*-element that seems to be involved in recruiting a repressor in non-expressing tissues. A *sbeIIb* gene has been isolated from sorghum but it lacks a *cis*-element. The clone was 7.8 kb and a BLAST search showed that it shared the highest degree of nucleotide identity with *sbeIIb* in maize, followed by rice, barley then wheat (Mutisya *et al*., 2003).

Although starch is an important carbohydrate and used in many ways, its biosynthesis has not completely been elucidated. Several mutants associated with quantitative and qualitative starch characteristics have been identified and characterized at a molecular level. They have made an important contribution towards understanding the properties and biosynthesis of starch (Opsahl-Ferstad and Rudi, 2000).

Amylose Starch Mutants

Sato *et al*. (2006) were able to screen a series of *waxy* alleles (point mutations) from a gamma-irradiated rice population using heteroduplex cleavage. Their results suggested that the rate of point mutations by gamma rays was lower than with EMS, but the rate of knockout mutations was higher. Verhoeven *et al*. (2004) characterized three novel starch mutants of sodium azide mutagenized oats. Two of the mutants were *waxy* type (amylose-free), and the third one contained phytoglycogen, a highly branched polysaccharide that had only been reported in mutant plants that lacked or had reduced isoamylase activity (Zeeman *et al*., 1998). These authors studied the degree of involvement of isoamylase in determining the amylopectin structure in *Arabidopsis* through the analysis of an isoamylase-deficient mutant (obtained from EMS- and X-ray-induced mutagenized populations). Interestingly, this mutant was also able to highly accumulate phytoglycogen. The analysis of this mutant not only provided confirmation that debranching enzymes participate in amylose synthesis but that amylose is not required, at least in *Arabidopsis*, to initiate starch degradation. Ball *et al*. (1991) unveiled the requirements necessary to obtain a substantial starch synthesis in higher plants through the study of a low-starch mutant obtained after X-ray mutagenesis treatment of *Chlamydomonas reinhardtii* cells.

Resistant Starch and Health

There has been a recent health-related interest in high amylose, *waxy* or resistant starches (RSs) in simple or processed food associated with the desire to reduce digestible calories, increase the fibre content and provide energy over extended periods (Anglani, 1998). Consumption of RS is very low in countries such as Australia and the USA where the risk of colorectal cancer is high

(Baghurst *et al*., 1996). Furthermore, RSs seem to play a role in decreasing the amount of glucose that is released as the result of digestive action which consequently could have an impact on the treatment of diabetes (Brown, 1996). Large bowel health could be improved through greater consumption of RS foods (Champ, 2004). However, given that there are likely to be barriers to substantial dietary changes at the individual level, enrichment of foods with RS as an ingredient is a more feasible option. At present, there is a high-amylose maize which is being used in the manufacturing of a number of food products in order to raise their RS content (Brown *et al*., 1995). High-amylose (up to 85%) starches with a dietary fibre content of 33% have also been obtained from other cereal crops using traditional breeding, mutagenesis and genetic engineering (Quintero-Fuentes *et al*., 1999; Fujita *et al*., 2001; Pedersen *et al*., 2005). However, within the cereal group only *waxy* maize is produced commercially. Some examples of *waxy* mutants in cereals are mentioned in the following paragraphs.

Barley (*Hordeum vulgare* L.)

A potential new cultivar of barley with low-starch and high-amylose content has been identified. The mutant contains a single point mutation in the gene encoding for starch synthase IIa (Bird *et al*., 2004) and it was generated by chemical mutagenesis of the barley cultivar 'Himalaya 292'. Another high-amylose barley (Barleyplus) has been developed in Australia (CSIRO, 2000) by chemical mutagenesis.

Wheat (*Triticum* sp.)

All species under the common name of wheat (*Triticum* sp.) comprise one of the most important crops found in the human diet. However, few mutants have been developed for the commercially grown wheat species mainly due to its polyploidy level. Common (hexaploid) wheat has three *waxy* genes on three genomes producing three kinds of waxy proteins. To create a polyploidy *waxy* mutant in hexaploid wheat, it would be necessary to mutate various loci in more than two genomes. Luckily, a hexaploid *waxy* mutant has been produced by EMS treatment of 'partial *waxy* mutants' lacking two of the three waxy proteins (Yasui *et al*., 1997). Fujita *et al*. (2001) chose *Triticum monococcum*, a diploid species as a model system to study starch metabolism in *Triticum* sp. They treated seeds with 1% EMS and generated M1, M2 and M3 populations. During the screening of M3 population they isolated a *waxy* mutant that exhibited properties that could help to elucidate starch metabolism in wheat. The genotype of this mutant was controlled by a single *waxy* allele that not only lacked amylose, 59-kDa waxy protein and GBSS I activity in the endosperm, but also showed a gene dosage effect in terms of amylose content, the amount of waxy protein and the GBSS activity.

Mutant Resources and Databases

More than 2000 mutant plants have been generated since the 1960s through mutation technology. The FAO in conjunction with the International Atomic Energy Agency (IAEA) created the first database of mutant plants officially released (http://www-mvd.iaea.org/) into breeding programmes of approximately 60 countries; it was called the Mutant Varieties Database (MTD). Most of the mutants reported in this database correspond to crop improvement initiatives and were developed directly from physical mutagenesis (FAO/IAEA, 2003, 2004).

The advances in bioinformatics, the completion in 2000 of the *Arabidopsis* sequencing project and the progress towards the conclusion of other model plant genomes opened the door to functional genomics – the study of the function of thousands of genes. The use of *Arabidopsis* in the analysis of the relationship between gene identification and function has limited applications for specialized traits shared by a specific group of plants. For example, the study of genes associated with N_2 fixation in legumes and characteristics of the seeds (e.g. grain texture, protein and lipid composition) in cereals (where a considerable homology exists within the group but not with *Arabidopsis*) requires the development of alternative genetic resources. Many mutant populations for important crops have been generated using diverse technologies, but few of them have been organized into databases. Some databases and web resources are mentioned below.

The goal for *Arabidopsis* now, promoted by the National Science Foundation, is to determine and understand the function of the 25,000 *Arabidopsis* genes by 2010 (http://www.nsf.gov/od/lpa/news/press/99/pr9958.htm). To accomplish this project multiple technologies are required, TILLING being one of them. The *Arabidopsis* TILLING Project, ATP (Till *et al*., 2003), has been created and the information is publicly available on the *Arabidopsis* Information Resource (http://www.arabidopsis.org) and ATP (http://tilling.fhcrc.org:9366) web sites. The data for ATP come from TILLING populations created and analysed for the model plant by several groups (McCallum *et al*., 2000; Greene *et al*., 2003; Till *et al*., 2003, 2006).

Rice was the second genome (2004) to be completed for a plant species. It is also the model plant for cereal crops. The International Rice Research Institute (IRRI) (http://www.irri.org/) has built a pool of genetic resources including a mutant collection generated using different mutagens. Also, a TILLING population has been generated (Till *et al*., 2007). This group is building a TILLING service for the rice community (http://tilling.ucdavis.edu), promising it will be soon accessible. Other research groups have also developed TILLING and gamma-irradiated populations for rice (Wu *et al*., 2005; Sato *et al*., 2006; http://www.shigen.nig.ac.jp/rice/oryzabase/top/top.jsp).

Maize is another important cereal crop and many resources exist for this species. In particular, a TILLING project (http://genome.purdue.edu/maizetilling/) has been developed to complement their mutant collection. This TILLING initiative is a collaborative effort to develop and screen TILLING populations for maize research community. For some of the populations they used pollen as a starting material.

A couple of initiatives are under development to generate sorghum gamma-irradiated and TILLING mutant populations. One is a TILLING project and it is being developed by the US Department of Agriculture (Xin *et al*., 2007) with the aim of generating a library of annotated, individually pedigreed, mutagenized sorghum (AIMS) lines that may serve as a valuable genetic resource for sorghum genomics and for elucidation of the genetic mechanisms underpinning important agronomic traits.

The other sorghum initiative is being carried out at the Centre for Plant Conservation Genetics (http://www.scu.edu.au/research/cpcg/sxn3/staff.php?id=15) in Australia, where physical (gamma rays) and chemical (EMS) treatments have been applied to generate sorghum DNA variants for genes associated with starch biosynthesis. At this stage, M2 plants are being screened.

For the gamma-irradiated project, 300 Gy was used as the optimal dose (LD_{50}) after a dosimetry experiment was performed using eight doses in combination with dried and wet seed treatments. Our preliminary results (Izquierdo *et al*., 2004) showed differences in germination, mortality and growth rates between dried and wet seeds, and doses. However, germination was not affected by dose. The screening of the first 500 M2 plants by a 3730 ABI genetic analyser showed that 10% of the mutants exhibited variations in the DNA profiles compared to the wild types (L. Izquierdo, Lismore, 2006, unpublished data).

Two TILLING populations (Slade *et al*., 2005) have been developed for wheat, one for hexaploid bread wheat and another for tetraploid pasta wheat with the aim to identify an allelic series of variants in the granule-bound starch synthase I (GBSSI) gene and develop a *waxy* wheat that could be grown commercially. Slade *et al*. (2005) identified an extensive allelic series (250 novel alleles) of the *waxy* genes for both bread and pasta wheat. In addition, they have applied bioinformatics tools to predict a diverse range of effects on *waxy* traits (from severe to none). One interesting result is that 94 of those 250 alleles caused changes within the coding region of the gene but the mutants were phenotypically indistinguishable from the wild types.

Gottwald *et al*. (2007) are developing an EMS-mutagenized TILLING population of the barley cultivar Barke, an elite malting barley. The project is part of the genome programme GABI2 (http://www.gabi-till.de/) conformed by German institutions. Also, two large-scale EMS mutant populations for the cultivar Optic have been developed to promote both forward and reverse genetics in this crop (Caldwell *et al*., 2004).

Lotus japonicus is the model plant for the legume group. A mutant database called *Lotus japonicus* mutant finder (http://data.jic.bbsrc.ac.uk/cgi-bin/lotusjaponicus/) has been developed at John Innes Centre (Perry *et al*., 2003). The database is the result of a 4-year programme and contains information for *c*.4000 M2 families. It mainly focuses on starch and symbiotic characteristics.

The Legume Genetics and Ecophysiology Research Unit (URLEG, Dijon, France) and the National Institute for Agricultural Research (INRA) generated a publicly available TILLING resource for *Medicago truncatula* in 2002. In

2006, the programme extended to pea (*Pisum sativum*). Both initiatives are part of the European project 'Grain Legumes' (http://www.dijon.inra.fr/urleg/web-urgap-gb/ressourcesgenetiques-gb/ressourcesgenetiques-gb.htm).

The International Centre for Tropical Agriculture (CIAT; http://www.ciat.cgiar.org/beans/index.htm) is in the early stages of developing a TILLING population for the common bean, *Phaseolus vulgaris*. They have a collection of 1000 fertile M1 mutant plants. They are interested in genes associated with agronomic traits, drought tolerance, N-fixation and Al tolerance.

Remarks

Induced mutation technology has increased breeding efficiency for many traits in important crops. It has also provided a tool to understand gene functions. At present the combination of mutagenesis and high-throughput technologies is playing a key role in developing new cultivars to suit new market demands in the food, health and environmental sectors. Mutation technology is a more friendly technology in the public domain than the transgenic approach, which is very important in times when consumers have a say. The challenges are vast. Although, some web-genetic resources have been developed for important crops, the information is still far from organized and centralized. More collaborative work should be done in order to avoid duplication efforts. We hope advances in bioinformatics will make it possible to generate virtual allelic series for important genes and predict potential functions.

Acknowledgement

The author is grateful to Dr Diego Holmgren for critical reading of the manuscript and useful suggestions.

References

Andersen, P.S. and Larsen, L.A. (2004) High-throughput mutation screening. In: Rapley, R. and Harbron, S. (eds) *An Overview of Genotyping and Single Nucleotide Polymorphisms (SNPs)*. Wiley, New York, pp. 71–98.

Anglani, C. (1998) Sorghum carbohydrates – A review. *Plant Foods for Human Nutrition* 52, 77–83.

Atak, C., Alikamanoglu, S., Acik, L. and Canbolat, Y. (2004) Induced of plastid mutations in soybean plant (*Glycine max* L. Merrill) with gamma radiation and determination with RAPD. *Mutation Research* 556, 35–44.

Bacha, R.E., Yokoyama, S. and Ishiy, T. (2002) Induced mutations as a method of obtaining iron toxicity resistant and high quality rice cultivars. In: Maluszynski, M. and Kasha, K.J. (eds) *Mutations,* In Vitro *and Molecular Techniques for Environmental Sustainable Crop Improvement*. Kluwer Academic Publisher, Dordrecht, The Netherlands.

Baghurst, P.A., Baghurst, K.I. and Record, S.J. (1996) Dietary fibre, non-starch polysaccharides and resistant starch – A review. *Food Australian* 48, S3–S35.

Ball, S., Marianne, T., Dirick, L., Fresnoy, M., Delrue, B. and Decq, A. (1991) A *Chlamydomonas reinhardtii* low-starch mutant is defective for 3-phosphoglycerate activation and orthophosphate inhibition of ADP-glucose pyrophosphorylase. *Planta* 185, 17–26.

Beta, T., Corke, H. and Taylor, J.N.R. (2000) Starch properties of Barnard Red, a South African red sorghum variety of significance in traditional African brewing. *Starch* 52, 467–470.

Bird, A.R., Flory, C., Davies, D.A., Usher, S. and Topping, D.L. (2004) A novel barley cultivar (Himalaya 292) with a specific gene mutation in starch synthase IIa raises large bowel starch and short-chain fatty acids in rats. *Journal of Nutrition* 134, 831–835.

Brown, I. (1996) Recent advances in the utilisation of starch for use in the human diet. *3rd Australian Sorghum Conference*. Tamworth, New South Wales, Australia.

Brown, I.L., McNaught, K.J. and Moloney, E. (1995) Hi-maizeTM: new directions in starch technology and nutrition. *Food Australian* 47, 272–275.

Buleon, A., Colonna, P., Planchot, V. and Ball, S. (1998) Starch granules: structure and biosynthesis. *International Journal of Biological Macromolecules* 23, 85–112.

Caldwell, D.G., McCallum, N., Shaw, P., Muehlbauer, G.J., Marshall, D.F. and Waugh, R. (2004) A structure mutant population for forward and reverse genetics in barley (*Hordeum vulgare* L.). *The Plant Journal* 40, 143–150.

Champ, M. (2004) Resistant starch. In: Eliasson, A. (ed.) *Starch in Food, Structure, Function and Applications*. Woodhead Publishing, Cambridge, pp. 560–574.

Comai, L., Young, K., Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R. and Henikoff, S.N. (2004) Efficient discovery of DNA polymorphisms in natural populations by ecotilling. *Plant Journal* 37, 778–786.

CSIRO (2000) BarleyplusTM. *Annual Report*. CSIRO, Australia, pp. 192.

Dovichi, N.J. and Zhang, J.Z. (2000) How capillary electrophoresis sequenced the human genome. *Angewandte Chemie International Edition* 39, 4463–4468.

FAO/IAEA (1977) Manual on mutation breeding. International Atomic Energy Agency, Vienna, Austria.

FAO/IAEA (1998) Application of DNA based marker mutations for improvement of cereals and other sexually reproduced crop plants. International Atomic Energy Agency, Vienna, Austria.

FAO/IAEA (2003) Improvement of new and traditional industrial crops by induced mutations and related biotechnology. International Atomic Energy Agency, Vienna, Austria, 164 pp.

FAO/IAEA (2004) Genetic improvement of under-utilized and neglected crops in low income food deficit countries through irradiation and related techniques. International Atomic Energy Agency, Vienna, Austria.

Fujita, N., Hasegawa, H. and Taira, T. (2001) The isolation and characterization of a waxy mutant of diploid wheat (*Triticum monococcum* L.). *Plant Sciences* 160, 595–602.

Gilchrist, E.J., Haughn, J.W., Ying, C.C., Otto, S.P., Zhuang, J., Cheung, D., Hamberger, B., Aboutorabi, F., Kalynyak, T., Johnson, L., Bohlmann, J., Ellis, B.L., Douglas, C.J. and Cronk, Q.C.B. (2006) Use of ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Molecular Ecology* 15, 1367–1378.

Gottwald, S., Bauer, P., Altschmied, L. and Stein, N. (2007) A TILLING population for functional genomics in Barley cv.'Barke'. *Plant and Animal Genomics XV Conference*. San Diego, California.

Greene, E.A., Codomo, C.A., Taylor, N.E., Henikoff, J.G., Till, B.J., Reynolds, S.H., Enns, L.C., Burtner, C., Johnson, J.E., Odden, A.R., Comai, L. and Henikoff, S. (2003) Spectrum of chemically induced mutations from a large-scale reverse-genetic screen in *Arabidopsis*. *Genetics* 164, 731–740.

Gut, I.G. (2004) An overview of genotyping and single nucleotide polymorphisms (SNPs). In: Rapley, R. and Harbron, S. (eds) *Molecular Analysis and Genome Discovery*. Wiley, Chichester, UK, pp. 43–67.

Henikoff, S. and Comai, L. (2003) Single-nucleotide mutations for plant functional genomics. *Annual Review of Plant Biology* 54, 375–401.

Izquierdo, L., Lee, S., Watson, L. and Henry, R.J. (2004) Preliminary study to establish the best gamma irradiation dose to induce mutation in three important crop seeds. *5th Australasian Workshop on Mutation Detection*. Queenstown, New Zealand.

Jain, S.M. (2005) Major mutation-assisted of plant breeding programs supported by FAO/IAEA. *Plant Cell, Tissue and Organ Culture* 82, 113–123.

James, M.G., Denyer, K. and Myers, A.M. (2003) Starch synthesis in the cereal endosperm. *Current Opinions in Plant Biology* 6, 215–222.

Kimball, R.F. (1987) The development of ideas about the effects of DNA repairs on the induction of gene mutations and chromosomal aberrations by radiation and chemicals. *Mutation Research* 186, 1–34.

Kozlowski, P. and Krzyzosiak, W.J. (2001) Combined SSCP/duplex analysis by capillary electrophoresis for more efficient mutation detection. *Nucleic Acids Research* 29, 2–14.

Lachtermacher, M.B.R., Seuanez, H.N., Moser, H.W. and Smith, K.D. (2000) One-step multiplex PCR strategy for identification of mutations by SSCP and DNA sequencing. *Biotechniques* 29, 234.

Lapins, K.O. (1983) Mutation breeding. In: Moore, J.N. and Janick, J. (eds) *Methods in Fruit Breeding*. Purdue University Press, West Lafayette, Indiana, pp. 74–99.

Larson, S.R., Rutger, J.N., Young, K.A. and Raboy, V. (2000) Isolation and genetic mapping of a non-lethal rice (*Oryza sativa* L.) *low phytic acid 1* mutation. *Crop Science* 40, 1397–1405.

Le, H., Fung, D. and Trent, R.J. (1997) Applications of capillary electrophoresis in DNA mutation analysis of genetic disorders. *Molecular Pathology* 50, 261–265.

Lee, Y.I., Lee, I.S. and Lim, Y.P. (2002) Variations in sweet potato regenerates from gamma-ray irradiated embryogenic callus. *Journal of Plant Biotechnology* 4, 163–170.

Lin, S. and Barringer, G.E. (2004) Method for detection of molecular species in a crude sample using capillary electrophoresis. Groton Biosystems, Massachusetts. Available at: http://www.patentdebate.com/PATAPP/20040259269

Liu, L.X., Spoerke, J.M., Mulligan, E.L., Chen, J., Reardon, B., Westlund, B., Sun, L., Abel, K., Armstrong, B., Hardiman, G., King, J., McCague, L., Basson, M., Clover, R. and Johnson, C. (1999) High-throughput isolation of *Caenorhabditis elegans* deletion mutants. *Genome Research* 9, 859–867.

Liu, L., Zanten, L.V., Shu, Q.Y. and Maluszynski, M. (2004) Officially released mutant varieties in China. *Mutation Breeding Review*. FAO/IAEA, Vienna, Austria.

McCallum, C.M., Comai, L., Greene, E.A. and Henikoff, S. (2000) Targeted screening for induced mutations. *Nature Biotechnology* 18, 455–457.

Micke, A. and Donini, B. (1993) Induced mutations. In: Hayward, M.D., Bosemark, N.O. and Romagosa, I. (eds) *Plant Breeding: Principles and Prospects*. Chapman & Hall, London, pp. 52–62.

Montelone, B.A. (1998) Mutation, mutagens, and DNA repair outline. Available at: http://www-personal.k-state.edu/~bethmont/mutdes.html

Mutisya, J., Sathish, P., Sun, C.X., Andersson, L., Ahlandsberg, S., Baguma, Y., Palmqvist, S., Odhiambo, B., Aman, P. and Jansson, C. (2003) Starch branching enzymes in sorghum (*Sorghum bicolor*) and barley (*Hordeum vulgare*): comparative analyses of enzyme structure and gene expression. *Journal of Plant Physiology* 160, 921–930.

Naito, K., Kusaba, M., Shikazono, N., Takano, T., Tanaka, A., Tanisaka, T. and Nishimura, M. (2005) Transmissible and nontransmissible mutations induced by irradiating *Arabidopsis thaliana* pollen with gamma-rays and carbon ions. *Genetics* 169, 881–889.

Oleykowski, C.A., Mullins, C.R.B., Godwin, A.K. and Yeung, A.T. (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Research* 26, 4597–4602.

Opsahl-Ferstad, H.G. and Rudi, H. (2000) Endosperm development, gene regulation and starch synthesis in cereals. *Curent Topics in Plant Biology* 2, 83–92.

Pedersen, J.F., Bean, S.R., Graybosh, R.A., Park, S.H. and Tilley, M. (2005) Characterization of waxy grain sorghum lines in relation to granule-bond starch synthase. *Euphytica* 151–156.

Perry, J.A., Wang, T.L., Welham, T.J., Gardner, S., Pike, J.M., Yoshida, S. and Parniske, M. (2003) A tilling reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. *Plant Physiology* 131, 866–871.

Quintero-Fuentes, X., McDonough, C.M., Rooney, L.W. and Almeida-Dominguez, H. (1999) Functionality of rice and sorghum flours in baked tortilla and corn chips. *Cereal Chemistry* 76, 705–710.

Repellin, A., Baga, M. and Chibbar, R.N. (2001) Characterization of a cDNA encoding a type I starch branching enzyme produced in developing wheat (*Triticum aestivum* L.) kernels. *Journal of Plant Physiology* 158, 91–100.

Sandhu, D. and Gill, K.S. (2002) Gene-containing regions of wheat and other grass genomes. *Plant Physiology* 128, 803–811.

Sato, Y., Shirasawa, K., Takahashi, Y., Nishimura, M. and Nishio, T. (2006) Mutant selection from progeny of gamma-ray-irradiated rice by DNA heteroduplex cleavage using *Brassica* petiole extract. *Breeding Science* 56, 179–183.

Slade, A.J. and Knauf, V.C. (2005) TILLING moves beyond functional genomics into crop improvement. *Transgenic Research* 14, 109–115.

Slade, A.J., Fuerstenberg, S.I., Loeffler, D., Steine, M. and Facciotti, D. (2005) A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. *Natural Biotechnology* 23, 75–81.

Snustad, D.P. and Simmons, M.J. (2006) *Principles of Genetics*. Wiley, New York, 866 pp.

Sood, R., English, M.A., Jones, M., Mullikin, J., Wang, D.M., Anderson, M., Wu, D., Chandrasekharappa, S.C., Yu, J., Zhang, J. and Liu, P.P. (2006) Methods for reverse genetic screening in zebrafish by resequencing and TILLING. *Methods* 39, 220–227.

Stone, B.A. (1996) Cereal grain carbohydrates. In: Henry, R.J. and Kettlewell, P.S. (eds) *Cereal Grain Quality*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 251–288.

Tang, T., Huang, J.Z., Zhong, Y. and Shi, S.H. (2004) High-throughput S-SAP by fluorescent multiplex PCR and capillary electrophoresis in plants. *Journal of Biotechnology* 114, 59–68.

Taylor, G.R. (1999) Enzymatic and chemical cleavage methods. *Electrophoresis* 20, 1125–1130.

Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.C., Johnson, J.E., Burtner, C., Odden, A.R., Young, K., Taylor, N.E., Henikoff, J.G., Comai, L. and Henikoff, S. (2003) Large-scale discovery of induced point mutations with high-throughput TILLING. *Genome Research* 13, 524–530.

Till, B.J., Colbert, T., Codomo, C.A., Enns, L.C., Johnson, J., Reynolds, S., Henikoff, J.G., Greene, E.A., Steine, M.N. and Comai, L. (2006) High-throughput TILLING for *Arabidopsis*. *Methods in Molecular Biology* 323, 2632–2641.

Till, B.J., Cooper, J., Tai, T.H., Colowit, P., Greene, E.A., Henikoff, S. and Comai, L. (2007) Discovery of chemical induced mutations in rice by TILLING. *BMC Plant Biology* 7.

Van Harten, A.M. (1998) *Mutation Breeding*. Cambridge University Press, Cambridge, 353 pp.

Verhoeven, T., Fahy, B., Leggett, M., Moates, G. and Denyer, K. (2004) Isolation and characterization of novel starch mutants of oats. *Journal of Cereal Science* 40, 69–79.

Wu, J., Wu, C., Lei, C., Baraoidan, M., Bordeos, A., Madamba, M.R.S., Ramos-Pamplona, M., Mauleon, R., Portugal, A., Ulat, V.J., Bruskiewich, R., Wang, G., Leach, J., Khush, G. and Leung, H. (2005) Chemical and irradiation-induced mutation of indica rice IR64 for forward and reverse genetics. *Plant Molecular Biology* 59, 85–97.

Xin, Z., Wang, M., Barkley, N., Franks, C., Burow, G., Pederson, G. and Burke, J. (2007) Development of a TILLING population for sorghum functional genomics. *Plant and Animal Genomes XV Conference*, 13–17 January 2007, San Diego, California.

Yasui, T., Sasaki, T., Matsuki, J. and Yamamori, M. (1997) Waxy endosperm mutants of bread wheat (*Triticum aestivum* L) and their starch properties. *Breeding Science* 47, 161–163.

Zeeman, S.C., Unemoto, T., Lue, W., Au-Yeung, P., Martin, C., Smith, A.M. and Chen, J. (1998) A mutant of *Arabidopsis* lacking a chloroplastic isoamylase accumulates both starch and Phytoglycogen. *The Plant Cell* 10, 1699–1711.

9 Nanotechnology: The Future of Cost-effective Plant Genotyping

J.A. Pattemore, M. Trau and R.J. Henry

Introduction

The exponential growth in biotechnology and nanotechnology industries today along with associated decreasing costs can be compared with that of the silicon chip revolution. Gordon Moore, a co-founder of Intel, is often quoted as having said in 1965 that computer processing power would double every 18 months, an observation now known as Moore's law. Stated correctly, however, Moore said that the complexity for minimum component costs increased at a rate of roughly a factor of two per year. That is, cost per unit could be reduced by increasing the number of features per unit area in addition to decreasing unit size. Moore's observations in 1965 led him to speculate quite accurately on the impact of inexpensive computing power on the way we go about our daily lives today.

So what has the silicon revolution got to do with genotyping molecular markers in plants? While production and commercialization of transgenic plants has been relatively successful, molecular marker technology for plant breeding has not been adopted as swiftly (Gupta *et al*., 2001). Factors affecting whether and how molecular markers should be used in plant breeding programmes include cost of the technology, cost of measuring the trait and sample turnaround time (Lamkey and Lee, 1993). Large segregating populations are difficult to manage and cost-effective, high-throughput (HTP) approaches to date have not been available, thus restricting wider implementation of molecular marker technology in plant breeding (Gupta *et al*., 2001). Ultimately, in order for molecular marker technology to become more widely adopted, reducing the cost of interrogating a data point for the purposes of identifying an allele, trait or individual remains the challenge.

The scope of this chapter includes a review of first and second generation molecular marker genotyping techniques with the focus on single nucleotide polymorphism (SNP) detection. Using barley (*Hordeum vulgare* L.) as a

 Plant Genotyping II: SNP Technology (ed. R.J. Henry)

case study, we identify the useful application of these genotyping techniques and discuss the limitations of these to HTP genotyping applications. Nanotechnology and a number of nanoparticle-based assays are subsequently explored and the potential cost-effective implications of this third generation genotyping technology on multiplexing and HTP genotyping are discussed.

Genotyping Single Nucleotide Polymorphisms in Barley

Barley is an important crop used for both food and beer-making. As a species, barley is extremely variable and is cultivated in a wide range of environments (Kanazin *et al*., 2002). In plant breeding terms, the capacity to generate genome-wide molecular markers rapidly and easily would constitute a significant improvement to quantitative trait loci (QTL) analysis, marker-assisted selection (MAS) and variety identification (Kanazin *et al*., 2002). Correct identification of barley varieties is imperative for controlling the quality of goods requiring different grain attributes, for example, beer-making and animal feed, and to ensure the best use of agronomic genotypes available (McCausland and Wrigley, 1977). Verification of genotype may also be necessary to ensure that royalties on improved varieties are paid to plant breeders, thus providing more resources for further plant improvement. Thirty years ago, molecular marker identification of barley varieties involved starch gel electrophoresis and isoelectric focusing, in combination with the phenotypic marker aleurone colour (McCausland and Wrigley, 1977). In the three decades since, significant quantities of molecular marker data have been generated for barley using a variety of methodologies including PCR (Chiapparino *et al*., 2004; Bundock *et al*., 2006), sequencing (Kanazin *et al*., 2002; Bundock *et al*., 2003; Bundock and Henry, 2004; Russell *et al*., 2004), high-performance liquid chromatography (HPLC; Kota *et al*., 2001b), matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS; Paris *et al*., 2002, 2003) and *in silico* SNP discovery from expressed sequence tag (EST) databases (Kota *et al*., 2001a, 2003). SNPs are loci where the sequence of DNA differs by a single base and are ideal for genotyping as they are the most abundant, stable marker in both animal and plant genomes. Estimates of the frequency of barley SNP markers vary from around 1/130 to 1/200 bases (Kanazin *et al*., 2002; Bundock *et al*., 2003; Rostoks *et al*., 2005b), likely due to differences in the region targeted for analysis. Regardless, these estimates are significantly more frequent than the estimated average for plant genomes of two SNPs per kilobase (Paris *et al*., 2003) and one SNP per kilobase in humans (Wang *et al*., 1998).

SNPs are found in both coding and non-coding regions of the genome and are rapidly becoming the molecular marker of choice for individual genotyping applications. Not all SNPs are useful markers; however, those found in cDNA and promoter region sequence alignments can directly affect gene function and are therefore considered 'perfect' molecular markers (Paris *et al*., 2003). In theory a SNP site could involve four possible alleles, although in

practice SNPs are biallelic, i.e. only two are generally observed at a specific site in a population (Gupta *et al*., 2001). Many methods are available for SNP genotyping, and the choice depends mainly on the scale of the study and the scientific question that is being addressed, i.e. whether large numbers of SNPs are required from a small number of individuals or vice versa (Gut, 2001).

All current methods of SNP genotyping involve generation of an allele-specific product followed by allele interrogation (Tost and Gut, 2002). Broadly speaking, SNP genotyping can be divided into gel-based and non-gel-based methods. The following sections briefly outline the most widely used allele identification methods. However, it is by no means exhaustive, and is intended to highlight the pros and cons of various SNP genotyping methodologies used in plant genotyping, past and present, in order to set the scene for introduction to potential applications of nanotechnology in plant genotyping. For more comprehensive information on SNPs in genotyping and epidemiology see Schork *et al*. (2000); for SNP genotyping in plant systems, see Gupta *et al*. (2001).

First Generation Assays (Gel-based)

First generation molecular markers can be divided into probe-based and PCR-based assays. Probe-based assays include restriction fragment length polymorphisms (RFLPs) and minisatellites. RFLP patterns are generated by cutting genomic DNA with restriction enzymes and polymorphisms occur by the gain or loss of a restriction site. Apart from being time-consuming, the main disadvantage of RFLPs is that the marker must be found in a restriction cutting site. Minisatellites are tandem arrays of short repeated sequences which are found throughout the genome (Jones *et al*., 1997). Polymorphic loci are generated by variable numbers of repeat units and multi-allelic forms (Jones *et al*., 1997). Probe-based assays have generally been time-consuming and also require radioactive labelling.

PCR amplification techniques are varied and include whole genome or targeted allele approaches. Random amplified polymorphic DNA (RAPD) (Williams *et al*., 1990) uses short arbitrary oligomers to prime random sites within the genome producing amplicons between 200 and 2000 kb long (Jones *et al*., 1997). Sequence polymorphisms at priming sites in the genome generate variation between banding patterns. RAPDs are inexpensive and easy to produce; however, they have limitations which have restricted their use including amplification of impurities, marker anonymity and the sensitivity of marker reproducibility to reaction conditions (Bachman, 1994).

Amplified fragment length polymorphism (AFLP) (Vos *et al*., 1995) involves digesting genomic DNA, ligation of the fragments to oligonucleotide adapters and selective PCR amplification of the fragments followed by gel analysis. Denaturing polyacrylamide gels provides high resolution of up to 100 restriction fragments, providing a powerful DNA fingerprinting technique (Vos *et al*., 1995). However, the methodology is time-consuming and requires the use of acrylamide, which is a dangerous neurological toxin.

Allele-specific PCR (AS-PCR) is an inexpensive, targeted approach where the identity of the amplified locus is generally predetermined by sequence analysis. AS-PCR specificity is influenced by the nature of the polymorphism and the flanking sequences. Guanine cytosine (GC)-rich loci make AS-PCR assays and multiplexing difficult to optimize (Shi *et al*., 1999).

Single-stranded conformation polymorphism (SSCP) is an inexpensive, sensitive method for screening PCR amplicons for novel polymorphisms. Under appropriate conditions, single-stranded DNA folds into sequence-specific structures which affects electrophoresis migration rate. A single base change can alter the structure, thus affecting mobility rate through the gel. The sensitivity of the assay is inversely proportional to the fragment size, and thus smaller fragments are preferable (Sunnucks *et al*., 2000). This method is useful for SNP discovery; however, it does not provide any other information except presence or absence of sequence variation between two samples. Once a putative polymorphic locus is found, it must be sequenced to confirm the presence and identity of the allele.

First generation marker technologies have had an enormous impact on the field of molecular biology and gel-based assays are generally facile for small numbers of samples; however, probe and PCR-based methods are not amenable to HTP applications because the assays are time-consuming, labour-intensive and expensive (Bassam *et al*., 1996). As a result, non-gel-based genotyping technologies have emerged as the dominant genotyping platforms for HTP, particularly in the field of pharmacogenetics (Shi, 2001).

Second Generation Assays (Non-gel-based)

Real-time PCR and fluorescent SNP genotyping

Real-time or quantitative PCR (qPCR) monitors the accumulation of target amplicons via fluorescently labelled probes during the reaction. Hence, while only the end point of a standard PCR is assessed via electrophoresis, qPCR enables the user to monitor lag, exponential and plateau phases as they occur. Quantitative PCR and fluorescent SNP genotyping generally involves simultaneous PCR amplification of the target and detection via one or more fluorescently tagged oligonucleotide probes specific to the allele of interest. Probe/target hybridization events generated during each PCR cycle result in an exponential increase in fluorescence. Prior to hybridization with the target, probes are rendered non-fluorescent via contact quenching or fluorescence resonance energy transfer (FRET) quenching. Contact quenching occurs when the reporter fluorophore is brought into direct contact with the quenching fluorophore by secondary folding of the oligo, for example, molecular beacons and Scorpion probes. Upon hybridization to the target, probes undergo conformational change resulting in spatial separation of the reporter and quencher producing fluorescence. FRET quenching occurs when the emission of the reporter fluorophore is altered by the absorption spectrum of a quenching fluorophore in close proximity on the same oligonucleotide, for

example TaqMan probes. On hybridization to the target, a TaqMan probe is cleaved by 5'-exonuclease activity of Taq DNA polymerase, resulting in the irreversible separation of reporter and quencher producing fluorescence.

Molecular beacons are capable of distinguishing targets that differ by a single nucleotide, are more specific than linear oligonucleotide probes (Bonnet *et al*., 1999) and have been successfully exploited for detecting SNPs and transgenes in barley (Kota *et al*., 1999). TaqMan assays have been used extensively in barley pathogen studies, in particular detection and quantification of toxin-producing *Fusarium* spp. (Strausbaugh *et al*., 2005; Leisova *et al*., 2006; Sarlin *et al*., 2006). Fluorescent assays are not suitable for large-scale multiplexing but are amenable to HTP when used to detect a small number of targets from a large number of samples. Fluorescent oligonucleotide probes may be expensive per data point, depending on the amount of oligo modification and while TaqMan chemistry is supported by dedicated software, a significant amount of optimization is required for each of these assays. The major advantage of qPCR over standard PCR includes the homogeneous or 'closed-tube' format, reducing the chance of cross-contamination, elimination of electrophoresis, dual specificity of primers and probes and, significantly, the ability to quantify the target.

Sequencing

Sequencing remains the most reliable method for detecting sequence mutations. The cost of sequencing has decreased over the past decade due to a number of factors including the replacement of gel-based sequencing with automated fluorescent capillary sequencers, improvements in sequencing chemistries and availability of lower-cost HTP sequencers.

The Sanger dideoxy chain terminator sequencing assay (Sanger *et al*., 1977) has been widely used for nearly three decades and capillary sequencers are capable of sequencing up to 700 bases from 96 DNA templates in a single 1-h run. To date, Sanger sequencing has been the main generator of sequence information; however, the requirement for cloning, amplification and purification of individual templates has been a limiting factor to cost-effective, HTP sequencing (Margulies *et al*., 2005).

Pyrosequencing (Ronaghi *et al*., 1996) relies on the conversion of pyrophosphates (PPi) to adenosine triphosphates (ATP) stimulating the light-producing luciferase in a base-dependent reaction. A primer is designed to hybridize adjacent to the 5'-side of the SNP position and all possible variants can be determined in a single-tube reaction as the region is sequenced (Fakhrai-Rad *et al*., 2002). Homozygous and heterozygous variants each give a unique pattern (pyrogram) compared to the wild-type (Ronaghi *et al*., 1996; Ahmadian *et al*., 2000). SNPs can be directly scored without the need for post-sequence editing and multiplexing can be applied to this technique (Fakhrai-Rad *et al*., 2002). Ninety-six-well and 384-well plate instruments are currently available facilitating HTP analysis of 5000–50,000 SNPs per day (Fakhrai-Rad *et al*., 2002). Pyrosequencing has been employed to identify cereal species

including barley (McIntosh *et al*., 2005) and to rapidly classify β-amylase allelic variation in barley (Polakova *et al*., 2003). Pyrosequencing is quickly emerging as an attractive platform for SNP analysis as the technology is time- and cost-competitive (approximately US$0.20 per sample); however, when compared to other genotyping platforms, the cost per data point is still a matter of concern, primarily due to the cost of biotinylated primers and specialized, dedicated instrumentation (Fakhrai-Rad *et al*., 2002).

Recently, a novel method of sequencing, '454 sequencing' (Margulies *et al*., 2005), was described which could potentially sequence 25 million bases in a 4-h run, generating a 100-fold increase in throughput over current Sanger sequencing technology. The 454 sequencing is PPi-based sequencing in picolitre-sized wells. Genomic DNA is isolated, mechanically sheared into small fragments and ligated to adapters. A single fragment is captured on to a bead and the DNA is clonally amplified on the bead, within a droplet in an emulsion (Margulies *et al*., 2005). Beads bearing single-stranded DNA clones are distributed into wells of a fibre-optic slide (PicoTitre-Plate) and the DNA is sequenced by pyrosequencing (Margulies *et al*., 2005). The method was shown to be extremely effective on compact microbial genomes (Margulies *et al*., 2005). However, due to short read lengths of 454 sequences, it was unknown how 454 sequencing would perform on large, highly repetitive genomes such as those of wheat and barley. The 454 sequencing was compared directly with Sanger sequencing on barley Bacterial Artificial Chromosome (BAC) clones, and while repetitive regions were problematic, deep 454 sequencing provided more even coverage of the BAC clones than did Sanger sequencing. A sequencing strategy incorporating some Sanger sequencing to partly compensate for the short read length of 454 sequences was developed by Wicker *et al*. (2006). The authors also concluded that due to difficulties encountered in repetitive regions, a whole-genome shotgun approach may not be practical for very large plant genomes; however, they suggested that small pools of BACs may be more practical for 454 sequencing (Wicker *et al*., 2006). Perhaps the greatest advantages of 454 sequencing over Sanger sequencing include direct sequencing from genomic template and reduced time and labour inputs, for example, 20 mb of sequences can be obtained in 4 h in a single 454-sequencing run (Margulies *et al*., 2005; Wicker *et al*., 2006). In combination with targeted Sanger sequencing, 454 sequencing promises rapid and cost-effective sequencing of coding sections of large, complex genomes (Wicker *et al*., 2006), providing unprecedented acceleration of SNP discovery and genotyping.

Sequencing provides a reliable genotyping method for barley as studies have shown that SNP haplotype is congruent with barley germplasm group (Kanazin *et al*., 2002). Bundock and Henry (2004) used sequencing to determine SNP haplotype diversity of the *Isa* gene of barley. The *Isa* gene is putatively involved in pathogen defence in the seed, and while cultivated barley varieties were less diverse at the *Isa* loci than wild barley (*H. spontaneum*), evidence of nine recombination events in cultivated varieties suggested a recombination hot spot driven by selection (Bundock and Henry, 2004). Genetic diversity and SNP haplotype were then determined from *Isa* sequences of eight wild barley populations from Israel (Cronin *et al*., 2007). Genetic diversity

was significantly correlated to key environmental water variables and genetic diversity at the *Isa* loci in wild accessions of barley from arid regions was significantly higher compared with those accessions from wetter climates. It was concluded that the higher diversity of *Isa* defence proteins observed may be a result of selection pressure by more diverse microbial populations present in arid environments (Cronin *et al*., 2007). Gene diversity linked to agronomically important traits such as plant defence mechanisms is of enormous potential and importance to plant breeding programmes. SNP discovery and genotyping by sequencing has clearly demonstrated capacity to identify potentially valuable loci in barley research programmes.

DNA chips/microarrays/high-density oligonucleotide arrays

The basic principle of 'DNA chips' or microarrays is a binding or hybridization assay (Bier and Kleinjung, 2001). 'DNA chip' refers to miniaturized arrays of nucleic acid oligomers (or probes) immobilized on a flat solid support, usually a glass slide (Hacia and Collins, 1999). Applications of microarrays range from SNP detection and scoring, gene expression to mutation analysis of large genes (Hacia and Collins, 1999; Bier and Kleinjung, 2001). The target is PCR amplified and fluorescently labelled before incubation on the array. After incubation, the array is read by exciting the fluorescent signal with a laser by scanning each spot or imaging the entire array. Thousands of probes can be either synthesized on to the chip during manufacture (Affymetrix, California) or, alternatively, spotted on to the chip by robot after synthesis.

The microarray technique has facilitated multiple measurements of many hybridization events to be performed in parallel, reducing labour and reagent costs, and is readily automated (Bier and Kleinjung, 2001). Oligonucleotide arrays were used extensively in large-scale identification and genotyping of SNPs in the human genome (Wang *et al*., 1998). In 2003, the Barley GeneChip (Affymetrix) was released as the culmination of worldwide collaboration among the barley research community (Close *et al*., 2004). Over 21,000 genes are represented on the array, generated from more than 84 libraries worldwide (Close *et al*., 2004). The chip has potential applications in analysis of malting properties, pest and disease control, abiotic stress tolerance, nutritional characteristics and reproductive development.

Some limitations to microarrays, however, need to be understood for their astute use, particularly in the field of SNP discrimination. Microarrays are primarily a screening tool and their major limitation is low sensitivity to rare events and sequences in trace populations (Hacia and Collins, 1999; Bunney *et al*., 2003). For a non-exhaustive SNP screen where an error rate of 5–10% is acceptable, microarray assays for heterozygous sequence changes are quite useful; however, for more specific applications it is crucial to define the sensitivity and specificity of DNA chip assays beforehand (Hacia and Collins, 1999).

Hybridization efficiency of microarrays is influenced by sample concentration, size and density of probes and melting temperature of each probe.

Each bound probe has its own sequence-dependent melting temperature; thus, it is difficult to ensure homogeneous hybridization conditions for all probes across an array (Bier and Kleinjung, 2001). In its current methodology, discrimination of SNPs at array level may require extremely stringent washes, which could result in the loss of much information (Bier and Kleinjung, 2001). A new approach to discriminate SNP mismatches on microarrays may be to observe the dissociation kinetics during the detection process. Dissociation is much more rapid when mismatches are present and kinetic analysis may be a more accurate method to discriminate mismatches from wild types (Bier and Kleinjung, 2001).

Microarrays have had an important impact on the capacity of DNA screening applications; however, due to their two-dimensional nature and limited pixel size, their use is limited to libraries of 5×10^5 compounds (Trau and Battersby, 2001; Battersby *et al*., 2002). The identity of each individual spot is determined by its position on the grid and high-cost devices are required to keep track of each position (Trau and Battersby, 2001). While microarrays are a highly automated, established platform, the cost per data point must become more economical (Battersby *et al*., 2001).

Minisequencing

Minisequencing is a microarray-based approach to screening for all possible alleles. Probes, designed to hybridize adjacent to a SNP site in the target, are immobilized on a surface at the 5' end leaving an exposed 3'-OH group. The hybridized target serves as template, while the probe acts as a primer for extension with labelled dideoxynucleotide triphosphates. The allele is thus identified by the incorporated ddNTP (Hacia and Collins, 1999).

MALDI-TOF Mass Spectrometry

MALDI-TOF MS is a powerful and reliable tool for HTP SNP genotyping (Tost and Gut, 2002). Paris *et al*. (2003) used MALDI-TOF MS to identify a barley genome SNP which is possibly responsible for barley resistance to powdery mildew. This technology has also facilitated selection for superior alleles of β-amylase, a key starch-degrading enzyme in the malting process, in barley during the seedling stage (Paris *et al*., 2002).

A complete SNP genotyping package including Assay Design software, iPLEX chemistry, liquid handling, chip nanodispensing, MALDI-TOF MS and data analysis is provided by Sequenom. The homogeneous assay consists of PCR amplification of the target, followed by incubation with shrimp alkaline phosphatase (SAP) to inactivate extraneous nucleotides. An 'extend' primer is added, extended through the SNP by a single ddNTP under standard thermocycling conditions, followed by reaction termination. The mass of the primer extension product is determined by laser ionization and desorption of the extension product to determine the sequence of the nucleotide at the SNP site.

MALDI-TOF analysis is highly suitable for HTP SNP analysis and multiplexing because it is fast – ionization and detection take milliseconds (Griffin and Smith, 2000; Ragoussis *et al*., 2006). It relies on direct mass determination of amplicons rather than indirect analysis by fluorescent or radioactive tagging (Tost and Gut, 2002) and is amenable to automation (Griffiths *et al*., 2000). It is compatible with 384-well amplification systems and is a homogeneous, automated system, minimizing sample handling. Currently, 3840 samples can be run at a time, and the iPLEX chemistry has been proven to be cost-efficient (around a 'few cents' per genotype) when multiplexes of 25–29 SNPs on 384 samples are processed (Ragoussis *et al*., 2006).

PCR amplification of the target still remains a bottleneck in most genotyping applications (Galvin, 2002) including MALDI-TOF-based genotyping. The high start-up cost and maintenance of instrumentation is further exacerbated by expensive, single-use chips and large numbers of oligos during assay design and validation. Thus, cost-effectiveness can only be achieved by high levels of multiplexing, which has been facilitated by recent advances in Assay Design software (Ragoussis *et al*., 2006). The potential to reduce the cost per assay lies in the sensitivity of the instrument. Only 15 nl of analyte is spotted on to the MassARRAY chip, despite 5–10 µl PCR amplification volumes; therefore, significant savings could result from developing miniaturized nanolitre-scale PCRs (Ragoussis *et al*., 2006). For more comprehensive discussion of DNA and SNP analysis by MALDI-TOF MS, see Griffin and Smith (2000), Ragoussis *et al*. (2006) and Tost and Gut (2006).

In silico SNP discovery

Many databases are publicly available online and *in silico* SNP discovery is a relatively inexpensive process. EST databases consist of single-pass sequences of cDNA libraries and are particularly attractive resources for SNP discovery. Of the 40 million ESTs currently lodged on GenBank, over 430,000 are from barley. Other useful barley EST databases include CR-EST (http://pgrc.ipk-gatersleben.de/est/index.php), HarvEST (http://harvest.ucr.edu), BarleyBase (http://www.plexdb.org/plex.php?database=Barley) and the Barley SNP database (http://bioinf.scri.ac.uk/barley_snpdb/).

While the quality of individual EST data is lower than some other sequence resources, the sheer volume and redundancy of data available makes EST databases attractive for SNP discovery. A large portion of the libraries is obtained from different individuals and assembly of overlapping sequences from the same loci in different individuals can enable the discovery of new SNPs (Picoult-Newberg *et al*., 1999). Unigenes assembled by aligning (or clustering) multiple ESTs from the same clone are more reliable than those from individual ESTs alone; however, the success of any SNP mining strategy depends on culling EST sequence errors by imposing stringent computational analysis of the data pool (Picoult-Newberg *et al*., 1999). In the past, sequence trace files were required to unambiguously filter sequencing errors; however, a number of software algorithms have become available which

automatically call bases depending on the level of stringency set by the user, for example, Sputnik (Rudd *et al*., 2003) and SNiPpER (Kota *et al*., 2003). *In silico* SNP discovery has thus become a relatively simple but accurate process. Once an adequate number of SNPs have been identified and verified by re-sequencing, a sensitive, robust and inexpensive detection method is subsequently required.

The number of second generation assays listed here along with applications directly related to barley genotyping demonstrates the importance of these assays in widening our understanding of plant genomics, and their ability to generate markers useful to plant breeding. With the exception of MALDI-TOF MS, second generation assays still lack the capacity to offer HTP and ultra-HTP screening for SNPs and other markers. This is primarily due to bottlenecks caused by the reliance on PCR to generate the target allele of interest, the cost of sequencing and limitations imposed by the two-dimensional nature of assays such as microarrays.

Third Generation Assays – Nanotechnology

In the near future, HTP and ultra-HTP SNP genotyping will be facilitated by: (i) generating novel materials by microfabrication to lower cost and increase capacity; (ii) reducing reaction volumes; (iii) increasing reaction rates; (iv) automating sample handling; and finally (v) developing single-platform genotyping, data analysis and storage instrumentation (Galvin, 2002). Advances in the field of nanotechnology are generating devices which are capable of meeting these requirements, thereby providing powerful new opportunities in molecular diagnostics and HTP screening (Jain, 2005). Currently these technologies are limited to frontier clinical applications; however, they will undoubtedly make their way into plant genotyping research, fast-tracking an abundance of opportunities to the plant breeding community and plant-based agriculture in general.

Nanotechnology concerns materials or structures with at least one dimension less than 100 nanometres (nm) (Lim, 2004), i.e. approximately 1/1000th the width of a human hair. Significantly, nano-sized compounds exhibit different properties to those of the same compound at macro- or micro-scale (D'Aquino *et al*., 2006). The size-dependent properties include electrical conductivity, magnetic coercivity, mechanics (hardness, strength and ductility), luminescent efficiency, transparency, catalytic properties and reaction rates (Lim, 2004).

Nanobiotechnology is the union of engineering and molecular biology with the goal of developing structures and devices of atomic, molecular or supramolecular size (Jain, 2003). The novel properties of these nanodevices have vast biological applications in diagnostics, drug delivery and therapeutics. Indeed the number of nano-industries emerging from nanobiotechnology continues to grow rapidly particularly in the field of medicine, promising 'to consign current technologies to obsolescence' (Seetharam, 2006). Nano-therapeutics are destined to attract the greatest attention of the nano-revolution as targeted drug

delivery and non-invasive medical procedures become more widely available in the fight against diseases such as cancer (Moos and Barry, 2006; Shelley, 2006). However, some of the most recent and successful applications of nanobiotechnology are in the field of molecular diagnostics (D'Aquino *et al*., 2006; Wispelwey, 2006). Following is a short review on nanoparticle-based assays with particular promise in molecular recognition and diagnostics.

Quantum Dots and FloDots

Quantum dots (QDs) are the most popular nanoparticles used in diagnostics (Jain, 2005). QDs are semiconductor nanocrystals, the most common being made of a CdSe core, often capped with ZnS to increase quantum yield (i.e. ratio of absorbed and emitted light) (Chan *et al*., 2002). Semiconductors are formed by adding conductive metal atoms on to the surface of insulators – a process known as 'doping'. The addition of conductive metal atoms to the neutral surface of an insulator changes the availability of electrons. Upon excitation, electrons are free to move into vacant orbitals and carry a current; thus, semiconductor QDs are capable of absorbing and emitting light energy as photons.

The emission properties of QDs are size- and composition-dependent. For example, a 3 nm CdSe nanoparticle will emit green light, whereas a 6 nm CdSe nanoparticle will emit red light (Xu *et al*., 2003); hence, emission wavelengths from blue to near infrared can be excited from a single wavelength (Chan *et al*., 2002), simplifying the instrumentation required for detection. QDs are an attractive alternative to traditional fluorescent dyes for a number of reasons. The QD emission spectrum is narrow and nearly symmetric, whereas its excitation profile is broad and continuous (Chan *et al*., 2002). In other words, spectral overlap between fluorescent signals is minimized and sources of excitation light can be simple and cheap. QDs can themselves be linked to biomolecules. However, they are most useful when embedded in microspheres where their optical properties can be multiplexed. In theory, a combination of QDs emitting six different colours at six different intensities could yield around 40,000 unique optical codes (Han *et al*., 2001; Ng and Liu, 2006). Thus, large screening libraries can be created by conjugating allele-specific oligonucleotide probes to microspheres encoded with unique optical addresses via specific QD composition (Han *et al*., 2001). The optical signal of the microsphere identifies the target, while the labelled target/probe complex indicates presence or absence and abundance of the target (Han *et al*., 2001). QD-encoded microbeads can be imaged on standard microscopes; however, HTP detection and screening via regular flow cytometers is possible at the rate of 1000 beads/s (Gao and Nie, 2004). The main disadvantage to QDs is that the toxic CdSe core must be coated to render them biologically inert; however, for clinical diagnostics and treatment, there is still concern that some of the toxic ion may escape (Hogan, 2006). In addition to toxicity problems, QDs are subject to poor solubility, low quantum yield and agglutination problems (Yao *et al*., 2006).

A more recent development in fluorescent nanoparticle technology is FloDots (Yao *et al*., 2006) (developed at, and named after, the University of Florida). In comparison to the size-dependent emissions of CdSe-cored QDs, FloDots consist of a silica matrix embedded with thousands of luminescent dye molecules. Dye-doped FloDots have strong emission signals when properly excited and similar photo stability to QDs. The silica surface can be modified to contain functional groups, used as a substrate for immobilization of biomolecules, and has the advantage of promoting dispersion in water, overcoming some of the poor solubility and agglutination problems of QDs (Yao *et al*., 2006).

Microspheres and Bead-based Assays

Polystyrene, latex and silica microspheres have been developed as solid platforms for attaching biological recognition compounds in what can be collectively called bead-based assays. Microspheres (2–500 μm in diameter) are functionalized to allow the attachment of multiple compounds on each bead (Battersby *et al*., 2000; Trau and Battersby, 2001). The spherical nature of the beads allows spatial homogeneity of probe attachment and subsequent target hybridization (Spiro *et al*., 2000). The small bead size allows them to stay suspended in solution for several hours without remixing, while the surface area of the beads in solution promotes close proximity of probe and target, generating near-fluid-phase reaction kinetics which is faster than that of microarrays (Fulton *et al*., 1997; Spiro *et al*., 2000). Microsphere-based assays have the potential to generate very large combinatorial libraries ($>10^{10}$ compounds) (Battersby *et al*., 2001) by varying the number and intensity of covalently attached dyes. For example, a library of 4^{16} probes can theoretically be encoded with just six fluorophores (Battersby *et al*., 2001).

Luminex

The Luminex analysis system (formerly known as FlowMetrix) is an early example of fluorescent microsphere technology. FlowMetrix polystyrene beads were impregnated with red and orange fluorophores in different ratios to produce a library of up to 100 bead types, distinguishable by relative intensities of red and orange fluorescence. Up to 2×10^6 capture probes complementary to the target of interest can be attached to the uniquely encoded microsphere (Fulton *et al*., 1997). Fluorescent wavelength and intensity were detected and analysed using computer-enhanced flow cytometry. Specific PCR targets were identified by the red-orange microsphere colour code and quantified by their green fluorescent label. Smith *et al*. (1998) used FlowMetrix for simultaneous detection of both RNA and DNA viruses. The assay was rapid, sensitive, and cross-hybridization was not observed except in the presence of extremely large amounts of PCR product ($>10^{10}$ copies). Spiro *et al*. (2000) hybridized nucleic acid sequences to the surface of polystyrene beads

(Luminex, Texas) to identify and quantify prokaryotic DNA sequences in heterogeneous environmental samples. The bead-based assay indicated superior specificity compared with microarrays and allowed accurate quantification of the target (Spiro *et al*., 2000). These assays were subsequently multiplexed successfully (Spiro and Lowe, 2002).

In a variation on the theme, the properties of complementary base pairing were used to alter the fluorescence of encoded polystyrene microspheres by FRET (Ihara *et al*., 2004). Target DNA was complementary to oligos covalently immobilized to the surface of a pair of different coloured microspheres. Thus, when hybridized to multiple targets, the microspheres aggregate, producing a FRET-induced change to particle fluorescence, measurable by fluorescent microscopy (Ihara *et al*., 2004). This technique requires some modification, however, for HTP applications and improvement in colour differentiation (Ihara *et al*., 2004).

Qbeads

Qbeads (Xu *et al*., 2003) are latex microspheres embedded with QDs. In some ways similar to the Luminex assay, Qbeads are colour-coded with QDs of varying colours and intensity; however, the assays differ primarily in the dyes used and the allele-specific target hybridization method. Optically encoded Qbeads are conjugated with allele-specific oligos and subsequently hybridized to biotinylated PCR targets. A further step is required to hybridize a streptavidin-PE-Cy5 label to the PCR target, followed by flow cytometric analysis and decoding (Xu *et al*., 2003). When validated by comparison with TaqMan PCR, the Qbead assay called homozygous and heterozygous SNP alleles with 100% accuracy compared with sequencing, which was 96.5% accurate, possibly due to the susceptibility of sequencing to poor template quality, even after PCR template clean-up with a commercial kit (Xu *et al*., 2003). As little as 0.2 ng of genomic DNA per ten SNP genotype calls was used, thereby saving genomic template, and given that only one multiplexed amplification step was performed, time and expense associated with PCR and commercial column purification products were reduced (Xu *et al*., 2003).

Silica microspheres

Silica microspheres are an attractive alternative to polystyrene beads as they are more stable in most solvent environments (Corrie *et al*., 2006). Silica microspheres are porous (Johnston *et al*., 2005), facilitate biomolecule attachment via a range of surface reactions (Corrie *et al*., 2006) and can be optically encoded by covalent attachment of fluorescent dyes (Johnston *et al*., 2005). The dyes may be incorporated in a combinatorial 'split and mix' manner, producing a diverse range of optical signals (Johnston *et al*., 2005). Oligonucleotides coupled to the surface of the microspheres fluoresce brightly upon hybridization to a perfectly matched, labelled target, whereas mismatched probes show significantly

less fluorescence (Johnston *et al*., 2005). Further refinement led to the development of thiol-functionalized organosilica microspheres with a uniformly narrow size distribution (Miller *et al*., 2005). The microsphere surface was functionalized with either thiol or amines and it was found that dye molecules covalently bound to free thiol groups were distributed uniformly throughout the microsphere interior, whereas amine functionalized microspheres became more impermeable, resulting in the dyes localized to the surface (Miller *et al*., 2005). The organosilica microsphere retained fluorescent intensity under phosphoramidite DNA synthesis conditions, whereas commercially available polystyrene-divinylbenzene microspheres with non-covalently bound dyes lost significant fluorescence (Miller *et al*., 2005). Fluorescence of an A/T mismatch was 20% less than that of a perfectly matched target (Miller *et al*., 2005), which, while quantifiable by flow cytometry, may be necessary to improve probe-target specificity for SNP detection.

Fluorescent microspheres can be detected and quantified either by fluorescent microscopy or flow cytometry. Flow cytometers were originally designed to automate optical and fluorescent measurement of cells or particles in solution and fluorescent microspheres have played an integral role in flow cytometry as calibration standards (Wedemeyer and Potter, 2001). Flow cytometry is capable of analysing multiple wavelengths of fluorescent light from thousands of particles per second; thus, when coupled to molecules of interest, hundreds of thousands of bead-based assays can be examined very quickly (Iannone, 2001). This capacity makes flow cytometry an attractive instrument for HTP detection and quantification of a wide range of molecular interactions (Nolan and Sklar, 1998).

Gold Nanoparticles

Gold nanoparticles are attractive labels for biosensors because a range of analytical techniques can be used to detect them, including fluorescence, optical absorption and electrical conductivity (Jain, 2003). Storhoff *et al*. (2004) used gold nanoparticle probes to detect unamplified bacterial DNA. When hybridized to the target and spotted on to a glass slide, gold-labelled oligonucleotide probes (GNP-DNA) scattered yellow-orange light when illuminated with white light from the side. Unhybridized probes scattered green light. Visual readout removes the need for complex detection instrumentation while the assay itself is rapid and inexpensive, without the need for a PCR amplification step (Storhoff *et al*., 2004).

In a different approach, GNP-DNA was used to discriminate SNPs in genes associated with thrombotic disorders, from unamplified human genomic DNA (Bao *et al*., 2005). Briefly, a capture probe specific to the allele of interest and a gene-specific signal probe were designed to 'sandwich' the target DNA. Capture probes were hybridized to a microarray and signal probes were modified with a gold nanoparticle (GNP-DNA). Genomic DNA, fragmented by ultrasonication into 500 kb lengths, was then hybridized to the microarray-bound allele-specific capture probes under stringent conditions.

Non-specific hybridizations were removed by washing, and a second hybridization step bound the GNP-DNA to the genomic DNA target, thus sandwiching the target between the capture and signal probes. Signal strength was enhanced by precipitating elemental silver on to the nanoparticles. The silver-amplified gold nanoparticles were excited with white light and the scattered light was captured on a photo sensor. While allele-specific identification required two hybridization steps and silver precipitation to enhance the signal, the authors were able to reliably genotype SNPs from 150,000 genome copies or 500 ng human genomic DNA in approximately 1 h.

Nanobarcodes

Nanobarcodes (Nicewarner-Pena *et al*., 2001) are microscopic metallic nanowires (Sha *et al*., 2006) which are 'barcoded' by sequentially electroplating metal ions of different reflectivity into narrow channels using a lithographic process. Complex barcodes are achieved by varying sequence and type of metal being deposited. Analytes bound to the particle by affinity-capture are detected by fluorescence, and the differential reflectivity of the barcode enables identification by conventional light microscopy (Nicewarner-Pena *et al*., 2001). Sha *et al*. (2006) were able to discriminate SNPs within the cytochrome P450 family of genes using nanobarcodes. Each uniquely coded nanowire is fixed to a different oligonucleotide probe and added to the reaction along with PCR target and fluorescently labelled probes. In the presence of a perfect match, the probes and PCR target hybridize and the probes are enzymatically ligated. The nanowires are imaged and analysed to determine which allele is present.

Nanotechnology on a Chip – NanoChips and NanoArrays

The BioForce Nano eNabler, formerly known as the NanoArray (BioForce Nanosciences, USA), is the next generation, ultra-miniaturized version of the microarray, capable of ultramicro- and nanoscale fluid delivery. Thousands of 1–20 μm spots of nucleic acids, antibodies and nanomaterials such as QDs, colloids and other material can be direct-patterned by the Nano eNabler via a micro-cantilever print head. Up to 400 nanospots can now be arrayed in the same area as the traditional microarray spot, vastly increasing the number of samples which can be analysed (Jain, 2003).

Conclusions

As experienced in the silicon revolution, miniaturization is the key to reducing plant genotyping costs and increasing throughput. Rapid advances in SNP discovery are providing vast amounts of data, ready for application in detection and diagnostics, pharmacogenetics and epidemiology. The priority

now is developing cost-effective platforms capable of rapidly and accurately identifying genetic polymorphisms (Shi, 2001) combined with multiplex and automated HTP detection capabilities. The challenge for HTP plant genotyping lies in continued SNP discovery, utilizing the available data more effectively, and improving the capacity and cost-effectiveness of screening thousands of polymorphisms from large numbers of individuals. The greatest obstacle is perhaps the affordability of testing and validating novel HTP genotyping assays. First and second generation genotyping methods discussed here have contributed enormously to the wealth of data available and to our current capacity for generating and detecting molecular markers; however, apart from MALDI-TOF MS, cost-effective, HTP, multiplex genotyping is still largely unavailable.

In future, plant genotyping, elimination of PCR amplification is desirable as, in addition to the extra steps required to generate target amplicons, cross-contamination and variability in target amplification efficiency are inherent flaws to PCR in HTP applications (Griffin and Smith, 2000). In order to remove the requirement for PCR target enrichment, detection sensitivity will become the priority (Galvin, 2002) particularly in large genomes such as barley, where the number of repetitive elements may weaken the allele-specific signal (Rostoks *et al*., 2005a). At present, gold nanoparticle assays are sensitive enough to detect non-amplified targets (Storhoff *et al*., 2004; Bao *et al*., 2005); however, they are array-based. Multiple washings, hybridizations and silver precipitation to enhance signal increase time and labour required and reduce HTP potential.

Bead-based assays have the advantage of near-fluid-phase reaction kinetics which are rapid compared with the relatively slower kinetics of microarray-based technologies (Xu *et al*., 2003). Conjugated to specific oligo recognition events, bead-based assays have the potential to realize HTP SNP genotyping due to the low cost, small reaction volumes, rapid near-fluid-phase reaction rates and amenity to automated sample handling. This combined with the single-platform genotyping, data analysis and storage available in flow cytometry may provide the plant sciences with the most powerful tool to date in the quest for cost-effective genotyping.

References

Ahmadian, A., Gharizadeh, B., Gustafsson, A.C., Sterky, F., Nyren, P., Uhlen, M. and Lundeberg, J. (2000) Single-nucleotide polymorphism analysis by pyrosequencing. *Analytical Biochemistry* 280, 103–110.

Bachman, K. (1994) Molecular markers in plant ecology. *New Phytologist* 126, 403–418.

Bao, Y.P., Huber, M., Wei, T.-F., Marla, S.S., Storhoff, J.J. and Muller, U.R. (2005) SNP identification in unamplified human genomic DNA with gold nanoparticle probes. *Nucleic Acids Research* 33, e15.

Bassam, B.J., Allen, T., Flood, S., Stevens, J., Wyatt, P. and Livak, K.J. (1996) Nucleic acid sequence detection systems: revolutionary automation for monitoring and reporting PCR products. *Australasian Biotechnology* 6, 285–294.

Battersby, B.J., Bryant, D., Meutermans, W., Matthews, D., Smythe, M.L. and Trau, M. (2000) Toward larger chemical libraries: encoding with fluorescent colloids in combinatorial chemistry. *Journal of the American Chemical Society* 122, 2138–2139.

Battersby, B.J., Lawrie, G.A. and Trau, M. (2001) Optical encoding of microbeads for gene screening: alternatives to microarrays. *Drug Discovery Today* 6, S19–S26.

Battersby, B.J., Lawrie, G.A., Johnston, A.P.R. and Trau, M. (2002) Optical barcoding of colloidal suspensions: applications in genomics, proteomics and drug discovery. *Chemical Communications* 14, 1435–1441.

Bier, F.F. and Kleinjung, F. (2001) Feature-size limitations of microarray technology – a critical review. *Fresenius Journal of Analytical Chemistry* 371, 151–156.

Bonnet, G., Tyagi, S., Libchaber, A. and Kramer, F.R. (1999) Thermodynamic basis of the enhanced specificity of structured DNA probes. *Proceedings of the National Academy of Sciences of the USA* 96, 6171–6176.

Bundock, P.C. and Henry, R.J. (2004) Single nucleotide polymorphism, haplotype diversity and recombination in the *Isa* gene of barley. *Theoretical Applied Genetics* 109, 543–551.

Bundock, P.C., Christopher, J.T., Eggler, P., Ablett, G., Henry, R.J. and Holton, T.A. (2003) Single nucleotide polymorphisms in cytochrome P450 genes from barley. *Theoretical Applied Genetics* 106, 676–682.

Bundock, P.C., Cross, M.J., Shapter, F.M. and Henry, R.J. (2006) Robust allele-specific polymerase chain reaction markers develop for single nucleotide polymorphisms in expressed barley sequences. *Theoretical Applied Genetics* 112, 358–365.

Bunney, W.E., Bunney, B.G., Vawter, M.P., Tomita, H., Li, J., Evans, S.J., Choudary, P.V., Myers, R.M., Jones, E.G., Watson, S.J. and Akil, H. (2003) Microarray technology: a review of new strategies to discover candidate vulnerability genes in psychiatric disorders. *American Journal of Psychiatry* 160, 657–666.

Chan, W.C.W., Maxwell, D., Gao, X., Bailey, R.E., Han, M. and Nie, S. (2002) Luminescent quantum dots for multiplexed biological detection and imaging. *Current Opinion in Biotechnology* 13, 40–46.

Chiapparino, E., Lee, D. and Donini, P. (2004) Genotyping single nucleotide polymorphisms in barley by tetra-primer ARMS-PCR. *Genome* 47, 414–420.

Close, T.J., Wanamaker, S.I., Caldo, R.A., Turner, S.M., SAshlock, D.A., Dicerson, J.A., Wing, R.A., Muehlbauer, G.J., Kleinhofs, A. and Wise, R.P. (2004) A new resource for cereal genomics: 22K Barley GeneChip comes of age. *Plant Physiology* 134, 960–968.

Corrie, S.R., Lawrie, G.A. and Trau, M. (2006) Quantitative analysis and characterization of biofunctionalized fluorescent silica particles. *Langmuir* 22, 2731–2737.

Cronin, J.K., Bundock, P.C., Henry, R.J. and Nevo, E. (2007) Adaptive climatic molecular evolution in wild barley at the *Isa* defense locus. *Proceedings of the National Academy of Sciences of the USA* 104, 2773–2778.

D'Aquino, R., Harper, T. and Roman Vas, C. (2006) Nanobiotechnology: fulfilling the promise of nanomedicine. *Chemical Engineering Progress* 102, 35–37.

Fakhrai-Rad, H., Pourmand, N. and Ronaghi, M. (2002) Pyrosequencing(TM): an accurate detection platform for single nucleotide polymorphisms. *Human Mutation* 19, 479–485.

Fulton, R.J., McDade, R.L., Smith, P.L., Kienker, L.J. and Kettman Jr., J.R. (1997) Advanced multiplexed analysis with the FlowMetrix(TM) system. *Clinical Chemistry* 43, 1749–1756.

Galvin, P. (2002) A nanobiotechnology roadmap for high-throughput single nucleotide polymorphism analysis. *Psychiatric Genetics* 12, 75–82.

Gao, X. and Nie, S. (2004) Quantum dot-encoded mesoporous beads with high brightness and uniformity: rapid readout using flow cytometry. *Analytical Chemistry* 76, 2406–2410.

Griffin, T.J. and Smith, L.M. (2000) Single-nucletide polymorphism analysis by MALDI-TOF mass spectrometry. *Trends in Biotechnology* 18, 77–84.

Griffiths, R.I., Whiteley, A.S., O'donnell, A.G. and Bailey, M.J. (2000) Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Applied and Environmental Microbiology* 66, 5488–5491.

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. *Current Science* 80, 524–535.

Gut, I.G. (2001) Automation in genotyping of single nucleotide polymorphisms. *Human Mutation* 17, 475–492.

Hacia, J.G. and Collins, F.S. (1999) Mutational analysis using oligonucleotide microarrays. *Journal of Medical Genetics* 36, 730–736.

Han, M., Gao, X., Su, J.Z. and Nie, S. (2001) Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. *Nature Biotechnology* 19, 631–635.

Hogan, H. (2006) Fluorescent probes with a small but bright future. *Biophotonics International* 13, 44–48.

Iannone, M.A. (2001) Microsphere-based molecular cytometry. *Clinics in Laboratory Medicine* 21, 731–742.

Ihara, T., Tanaka, S., Chikaura, Y. and Jyo, A. (2004) Preparation of DNA-modified nanoparticles and preliminary study for colorimetric SNP analysis using their selective aggregations. *Nucleic Acids Research* 32, e105.

Jain, K.K. (2003) Nanodiagnostics: application of nanotechnology in molecular diagnostics. *Expert Reviews in Molecular Diagnostics* 3, 153–161.

Jain, K.K. (2005) Nanotechnology in clinical laboratory diagnostics. *Clinica Chimica Acta* 358, 37–54.

Johnston, A.P.R., Battersby, B.J., Lawrie, G.A. and Trau, M. (2005) Porous functionalised silica particles: a potential platform for biomolecular screening. *Chemical Communications* 7, 848–850.

Jones, N., Ougham, H. and Thomas, H. (1997) Markers and mapping: we are all geneticists now. *New Phytologist* 137, 163–177.

Kanazin, V., Talbert, H., See, D., DeCamp, P., Nevo, E. and Blake, T. (2002) Discovery and assay of single-nucleotide polymorphisms in barley (*Hordeum vulgare*). *Plant Molecular Biology* 48, 529–537.

Kota, R., Holton, T.A. and Henry, R.J. (1999) Detection of transgenes in crop plants using molecular beacon assays. *Plant Molecular Biology Reporter* 19, 363–370.

Kota, R., Varshney, R.K., Thiel, T., Dehmer, K.J. and Graner, A. (2001a) Generation and comparison of EST-derived SSRs and SNPs in barley (*Hordeum vulgare* L.). *Hereditas* 135, 145–151.

Kota, R., Wolf, M., Michalek, W. and Graner, A. (2001b) Application of denaturing high-performance liquid chromatography for mapping of single nucleotide polymorphisms in barley (*Hordeum vulgare* L.). *Genome* 44, 523–528.

Kota, R., Rudd, S., Facius, A., Kolesov, G., Thiel, T., Zhang, H., Stein, N., Mayer, K. and Graner, A. (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Molecular and General Genomics* 270, 24–33.

Lamkey, K.R. and Lee, M. (1993) Quantitative genetics, molecular markers, and plant improvement. In: Imrie, B.C. and Hacker, J.B. (eds) *10th Australian Plant Breeding Conference. Focused Plant Improvement: Towards Responsible and Sustainable Agriculture*. Gold Coast, Australia, pp. 104–115.

Leisova, L., Kucera, L., Chrpova, J., Sykorova, S., Sip, V. and Ovensna, J. (2006) Quantification of *Fusarium culmorum* in wheat and barley tissues using real-time PCR in comparison with DON content. *Journal of Phytopathology* 154, 603–611.

Lim, H.A. (2004) Nanotechnology in diagnostics and drug delivery. *Medicinal Chemistry Research* 13, 401–413.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B.,

Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.

McCausland, J. and Wrigley, C.W. (1977) Identification of Australian barley cultivars by laboratory methods: gel electrophoresis and gel isoelectric focusing of the endosperm proteins. *Australian Journal of Experimental Agriculture and Animal Husbandry* 17, 1020–1027.

McIntosh, S.R., Pacey-Miller, T. and Henry, R.J. (2005) A universal protocol for identification of cereals. *Journal of Cereal Science* 41, 37–46.

Miller, C.R., Vogel, R., Surawski, P.P.T., Corrie, S.R., Ruhmann, A. and Trau, M. (2005) Biomolecular screening with novel organosilica microspheres. *Chemical Communications* 4783–4785.

Moos, W.H. and Barry, S. (2006) Nanobiotechnology: it's a small world after all. *Drug Development Research* 67, 1–3.

Ng, J.K.-K. and Liu, W.-T. (2006) Miniaturized platforms for the detection of single-nucleotide polymorphisms. *Analytical and Bioanalytical Chemistry* 386, 427–434.

Nicewarner-Pena, S.R., Freeman, R.G., Reiss, G.D., He, L., Pena, D.J., Walton, I.D., Cromer, R., Keating, C.D. and Natan, M.J. (2001) Submicrometer metallic barcodes. *Science* 294, 137–141.

Nolan, J.P. and Sklar, L.A. (1998) The emergence of flow cytometry for sensitive, real-time measurements of molecular interactions. *Nature Biotechnology* 16, 633–638.

Paris, M., Jones, M.G.K. and Eglinton, J.K. (2002) Genotyping single nucleotide polymorphisms for selection of barley b-amylase alleles. *Plant Molecular Biology Reporter* 20, 149–159.

Paris, M., Potter, R.H., Lance, R.C.M., Li, C.D. and Jones, M.G.K. (2003) Typing *Mlo* alleles for powdery mildew resistance in barley by single nucleotide polymorphism analysis using MALDI-ToF mass spectrometery. *Australian Journal of Agricultural Research* 54, 1343–1349.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A. and Boyce-Jacino, M. (1999) Milling SNPs from EST databases. *Genome Research* 9, 167–174.

Polakova, K., Laurie, D., Vaculova, K. and Ovesna, J. (2003) Characterization of b-amylase alleles in 79 barley varieties with pyrosequencing. *Plant Molecular Biology Reporter* 21, 439–447.

Ragoussis, J., Elvidge, G.P., Kaur, K. and Colella, S. (2006) Matrix-assisted laser desorption/ionisation, time-of-flight mass spectrometry in genomics research. *PLoS Genetics* 2, 920–929.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* 242, 84–89.

Rostoks, N., Borevitz, J.O., Hedley, P.E., Russell, J., Mudie, S., Morris, J., Cardle, L., Marshall, D.F. and Waugh, R. (2005a) Single-feature polymorphism discovery in the barley transcriptome. *Genome Biology* 6, R54.

Rostoks, N., Mudie, S., Cardle, L., Russell, J., Ramsay, L., Booth, A., Svensson, J.T., Wanamaker, S.I., Walia, H., Rodriguez, E.M., Hedley, P., Liu, H., Morris, J., Close, T.J., Marshall, D.F. and Waugh, R. (2005b) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Molecular and General Genetics* 274, 515–527.

Rudd, S., Mewes, H.-W. and Mayer, K.F.X. (2003) Sputnik: a database platform for comparative plant genomics. *Nucleic Acids Research* 31, 128–132.

Russell, R., Booth, A., Fuller, J., Harrower, B., Hedley, P., Machray, G. and Powell, W. (2004) A comparison of sequence-based polymorphism and haplotype content in transcribed and anonymous regions of the barley genome. *Genome* 47, 389–398.

Sanger, F., Nicklen, S. and Couson, A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA* 74, 5463–5467.

Sarlin, T., Yli-Mattila, T., Jestoi, M., Rizzo, A., Paavanen-Huhtala, S. and Haikara, A. (2006) Real-time PCR for quantification of toxigenic *Fusarium* species in barley and malt. *European Journal of Plant Pathology* 114, 371–380.

Schork, N.J., Fallin, D. and Lanchbury, S. (2000) Single nucleotide polymorphisms and the future of genetic epidemiology. *Clinical Genetics* 58, 240–264.

Seetharam, R.N. (2006) Nanomedicine – emerging area of nanobiotechnology research. *Current Science* 91, 260.

Sha, M.Y., Walton, I.D., Norton, S.M., Taylor, M., Yamanaka, M., Natan, M.J., Xu, C., Drmanac, S., Huang, S., Bordherding, A., Drmanac, R. and Penn, S.G. (2006) Multiplexed SNP genotyping using nanobarcode particle technology. *Analytical and Bioanalytical Chemistry* 384, 658–666.

Shelley, S.A. (2006) Nanobiotechnology: cancer's newest deadly foe. *Chemical Engineering Progress* 102, 43–47.

Shi, M.M. (2001) Enabling large-scale pharmacogenetic studies by high-throughput mutation detection and genotyping technologies. *Clinical Chemistry* 47, 164–172.

Shi, M.M., Bleavins, M.R. and de la Iglesia, F.A. (1999) Technologies for detecting genetic polymorphisms in pharmacogenomics. *Molecular Diagnosis* 4, 343–351.

Smith, P.L., WalkerPeach, C.R., Fulton, R.J. and Dubois, D.B. (1998) A rapid, sensitive, multiplexed assay for detection of viral nucleic acids using the FlowMetrix system. *Clinical Chemistry* 44, 2054–2056.

Spiro, A. and Lowe, M. (2002) Quantitation of DNA sequences in environmental PCR products by a multiplexed, bead-based method. *Applied and Environmental Microbiology* 68, 1010–1013.

Spiro, A., Lowe, M. and Brown, D. (2000) A bead-based method for multiplexed identification and quantitation of DNA sequences using flow cytometry. *Applied and Environmental Microbiology* 66, 4258–4265.

Storhoff, J.J., Lucas, A.D., Garimella, V., Bao, Y.P. and Muller, U.R. (2004) Homogeneous detection of unamplified genomic DNA sequences based on colorimetric scatter of gold nanoparticle probes. *Nature Biotechnology* 22, 883–887.

Strausbaugh, C.A., Overturf, K. and Koehn, A.C. (2005) Pathogenicity and real-time PCR detection of *Fusarium* spp. in wheat and barley roots. *Canadian Journal of Plant Pathology* 27, 430–438.

Sunnucks, P., Wilson, A.C.C., Beheregaray, L.B., Zenger, K., French, J. and Taylor, A.C. (2000) SSCP is not so difficult: the application and utility of single-stranded conformation polymorphism in evolutionary biology and molecular ecology. *Molecular Ecology* 9, 1699–1710.

Tost, J. and Gut, I.G. (2002) Genotyping single nucleotide polymorphisms by mass spectrometry. *Mass Spectrometry Reviews* 21, 388–418.

Tost, J. and Gut, I.G. (2006) DNA analysis by mass spectrometry – past, present and future. *Journal of Mass Spectrometry* 41, 981–995.

Trau, M. and Battersby, B.J. (2001) Novel colloidal materials for high-throughput screening applications in drug discovery and genomics. *Advanced Materials* 13, 975–979.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Friters, A., Pot, J., Paleman, J., Kuiper, M. and Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research* 23, 4407–4414.

Wang, D.G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M. and Lander, E.S. (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082.

Wedemeyer, N. and Potter, T. (2001) Flow cytometry: an 'old' tool for novel applications in medical genetics. *Clinical Genetics* 60, 1–8.

Wicker, R., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B. and Stein, N. (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* 7, 275.

Williams, J.G.K., Kubelic, A.R., Livak, K.J., Rafalsky, J.A. and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Research* 18, 6531–6535.

Wispelwey, J. (2006) Nanobiotechnology: the integration of nanoengineering and biotechnology to the benefit of both. *Chemical Engineering Progress* 102, 34.

Xu, H., Sha, M.Y., Wong, E.Y., Uphoff, J., Xu, Y., Treadway, J.A., Truong, A., O'Brien, E., Asquith, S., Stubbins, M., Spurr, N.K., Lai, E.H. and Mahoney, W. (2003) Multiplexed SNP genotyping using the Qbead(TM) system: a quantum dot-encoded microsphere-based assay. *Nucleic Acids Research* 31, e43.

Yao, G., Wang, L., Wu, Y., Smith, J., Xu, J., Zhao, W., Lee, E. and Tan, W. (2006) FloDots: luminescent nanoparticles. *Analytical and Bioanalytical Chemistry* 385, 518–524.

10 Functionally Associated Molecular Genetic Markers for Temperate Pasture Plant Improvement

J.W. Forster, N.O.I. Cogan, M.P. Dobrowolski, M.G. Francki, G.C. Spangenberg and K.F. Smith

Introduction

Perennial ryegrass and white clover are the most important grass and legume species, respectively, of temperate pasture agriculture. The obligate outbreeding reproductive habits of these taxa require the design of synthetic varietal production systems based on the use of multiple parental genotypes. Although conventional methods for molecular genetic marker development and validation have been previously employed, novel strategies for implementation of candidate gene-derived functionally associated genetic markers are required. Species-specific genomic resources have been established to support these objectives, and functional analysis has been supported by gene annotation, extrapolation of data from related model species (wheat, barley, barrel medic and others), transcriptomics analysis and transgenic modification. Highly efficient experimental systems have been developed for *in silico* and *in vitro* discovery of single nucleotide polymorphisms (SNPs). Validated SNPs have been used to determine levels of nucleotide diversity, haplotype structure and linkage disequilibrium (LD), and to design association genetics experiments for haplotype–phenotype correlation. Methods have also been devised for implementation of genic markers to select for superior allele content in commercial pasture breeding programmes. Target traits for germplasm improvement include herbage quality; disease resistance; and tolerance to environmental stresses such as drought, heat, cold, soil toxicities (e.g. salt and aluminium) and nutrient deficiencies. Functionally associated genetic markers are applicable to a broad range of temperate pasture species and potentially to warm-season apomictic forage grasses.

Agronomic Importance of Temperate Pasture Species

Perennial ryegrass (*Lolium perenne* L.) and white clover (*Trifolium repens* L.) are cultivated, generally in combination, in grassland-producing regions of

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

Northern Europe, the Pacific North-west of the USA, Japan, South-eastern Australia and New Zealand and provide high-quality forage with superior palatability and nutrient content (Forster *et al*., 2001a). The cultivation of these species supports dairy, sheep meat, beef and wool production industries. A mixture of species in grazing swards can provide complementary qualities and increase animal productivity. For example, dairy cows can consume over 30% more white clover than perennial ryegrass, resulting in 25% more milk production (Rogers *et al*., 1982). Higher digestibility of white clover herbage is associated with lower neutral detergent fibre (NDF) content (Doyle *et al*., 2000). White clover also provides higher nutritive value than grasses for most of the year. This quality is a function of the availability of rumen bypass protein, which is superior to that of pasture grasses. However, rumen fermentation may not be optimized on pure legume diets due to an imbalance between high protein concentrations and inadequate readily fermentable water-soluble carbohydrate (WSC) content (Carruthers and Neil, 1997). Complementarity with white clover based on the use of high WSC content grasses, such as the perennial ryegrass cultivar Aurora, can address this problem, as demonstrated for meat production (Munro *et al*., 1992).

White clover also provides benefits to pasture production through biological nitrogen fixation, with the capacity to contribute up to 500 kg N/ha/year in combination with perennial ryegrass (Eckard, 1998). Although a white clover content of *c*.35% is optimal for fixation, contributing 100–350 kg N/ha/year, the value may be less than 10% in pastures suffering summer moisture stress, such as those of south-western Victoria in Australia (Doyle *et al*., 2000). Development of white clover cultivars with increased persistence due to improved abiotic stress is a breeding priority for temperate Australian conditions (Lane *et al*., 2000).

The major breeding priorities for perennial ryegrass are: improved dry matter yield and nitrogen recovery; herbage nutritive quality; resistance to invertebrate pests; resistance to viral, bacterial and fungal diseases such as crown rust (*Puccinia coronata* Corda f.sp. *lolii* Brown); tolerance to abiotic stresses such as heat and drought; and enhanced seed production (Wilkins and Humphreys, 2003). Traits such as accumulation of stem carbohydrates, which are important for nutritional value in perennial ryegrass, are also significant for tolerance to abiotic stresses in intensively studied cereal species, providing potential for the use of comparative biology and genetics to develop common functionally associated genetic markers.

The equivalent priorities for white clover include: yield and persistence; tolerance to cold, drought and saline soils; resistance to invertebrate pests such as clover cyst nematode; resistance to fungal, bacterial and viral diseases such as alfalfa mosaic virus (AMV); improved symbiotic interactions; compatibility with companion grasses; animal nutrition and welfare, including bloat safety; seed production; and positive environment impacts (Abberton and Marshall, 2005). As for perennial ryegrass, comparative approaches based on dicotyledonous plant model species with complete genome sequences are highly applicable to clover improvement.

Biology of Temperate Pasture Species

Breeding systems

Both perennial ryegrass and white clover are obligate allogamous species, with gametophytic self-incompatibility (SI) systems. The perennial system is controlled by two loci (*S* and *Z*), and incompatible matings occur when the alleles at both loci in the male gametophyte (pollen grain) match one of the two alleles at each locus in the female sporophyte (Cornish *et al*., 1979). The white clover system is controlled by allelic variation at a single locus (*S*) (Attwood, 1940, 1941, 1942a). These mechanisms ensure low levels of self-fertilization, and also limit the level of fertility in crosses between closely related individuals such as full-sibs. Rare instances of self-compatibility have been reported in white clover (Attwood, 1942b; Yamada *et al*., 1989), presumably due to the presence of self-fertile (S_f) alleles at the *SI* locus.

Varietal development systems

Due to the obligate outbreeding natures of both perennial ryegrass and white clover, both natural and synthetic populations are highly genetically heterogeneous. Varietal development is typically based on the following process: evaluation of base populations containing 2000–5000 individuals, selection of *c*.200 potential parental clones and polycrossing between selected parental genotypes to generate a synthetic 1 (Syn1) population (Vogel and Pedersen, 1993). The number of foundation individuals may vary from as low as four for perennial ryegrass to 50–100 for polyploid species such as tall wheat grass and lucerne (Bray and Irwin, 1999).

Genetic architecture

Members of the *Lolium* genus are diploids with a fundamental chromosome number of 7 ($2n = 2x = 14$), although synthetic autotetraploid varieties of perennial ryegrass and Italian ryegrass (*L. multiflorum* Lam.) are produced for commercial production. The haploid genome size of perennial ryegrass has been estimated as *c*.1.6×10^9 bp(Hutchinson *et al*., 1979; Seal and Rees, 1982). In common with other Poaceae family members, the genomes of *Lolium* species contain large numbers of dispersed repetitive sequences, frequently belonging to major retroelement families (Jenkins *et al*., 2000).

White clover is an allotetraploid species with a fundamental chromosome number of 8 ($2n = 4x = 32$). Study of chloroplast DNA and nuclear ribosomal DNA variation has implicated the diploid taxa *T. occidentale* D.E. Coombe and *T. pallescens* Schreber as paternal and maternal progenitors, respectively (Ellison *et al*., 2006). Haploid genome size has been estimated as *c*.8×10^8 bp, implying an average sub-genome size close to 400 µb, comparable to that of the related model legume species barrel medic (*Medicago truncatula* Gaertn.).

Genomic Resources for Temperate Pasture Species

High-throughput gene discovery by expressed sequence tag (EST) sequencing has generated collections of 44,534 and 42,017 sequences (Sawbridge *et al*., 2003a,b), corresponding to 12,170 and 15,989 unigenes from perennial ryegrass and white clover, respectively (Spangenberg *et al*., 2005). Complementary large insert DNA libraries have been generated using bacterial artificial chromosome (BAC) vectors. The perennial ryegrass BAC library consists of 50,304 clones with an average insert size of 113 kb, corresponding to 3.4 genome equivalents. The white clover BAC library consists of 50,302 clones with an average insert size of 101 kb, corresponding to 6.3 genome equivalents (Spangenberg *et al*., 2005).

Molecular Genetic Marker Technology for Improvement of Temperate Pasture Species

Perennial ryegrass

Molecular genetic marker development, associated map construction and trait-dissection studies have been comprehensively reviewed (Forster *et al*., 2001a, 2004; Yamada and Forster, 2005; Yamada *et al*., 2005). Genomic DNA-derived simple sequence repeat (SSR) markers have been developed using genomic cloning and enrichment library technology (Kubik *et al*., 1999; Jones *et al*., 2001; Jensen *et al*., 2005a; Lauvergeat *et al*., 2005), while the EST collection has also been used for the development of a set of EST-SSR primer pairs (Faville *et al*., 2004). Gene-associated SNP markers have been developed through both *in vitro* and *in silico* strategies (Shinozuka *et al*., 2005; Spangenberg *et al*., 2005; Cogan *et al*., 2006b).

The first perennial ryegrass reference genetic map contained restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP) and SSR markers (Jones *et al*., 2002a,b), spanning 811 cM across seven linkage groups (LGs), and allowing the inference of comparative relationships between perennial ryegrass and other Poaceae species. A predominant relationship between each perennial ryegrass LG and one of the homoeologous groups of hexaploid wheat was observed, providing a standardized nomenclature system, although several major translocation results were subsequently observed (Sim *et al*., 2005). This analysis has supported translational genetics between perennial ryegrass and the Triticeae. A second generation reference genetic mapping family was developed based on the $F_1(NA_6 \times AU_6)$ two-way pseudo-testcross family, generating two parental genetic maps dominated by EST-RFLP and EST-SSR loci (Faville *et al*., 2004).

Trait-dissection for perennial ryegrass has been performed in multiple populations to allow quantitative trait locus (QTL) analysis for characters such as vegetative and reproductive morphogenesis, reproductive development, winter hardiness, herbage quality, mineral content, gametophytic SI, photosynthetic efficiency and crown rust resistance (Armstead *et al*., 2002,

2004; Thorogood *et al*., 2002; Dumsday *et al*., 2003; Forster *et al*., 2004; Warnke *et al*., 2004; Yamada *et al*., 2004; Cogan *et al*., 2005a,b; Muylle *et al*., 2005a,b; Jensen *et al*., 2005b).

White clover

Comprehensive sets of genomic DNA-derived SSR and EST-SSR primer pairs have been developed (Kölliker *et al*., 2001; Barrett *et al*., 2004) and used to construct reference genetic maps based on different population structures. The F_2(I.4R × I.5J) population was obtained from parental genotypes from fourth and fifth generation inbred lines descended from plants containing the rare S_f allele and was used to generate a map dominated by AFLP and genomic DNA-derived SSR loci (Jones *et al*., 2003). A higher-resolution genetic map largely based on EST-SSR markers was constructed using the F_1(Sustain 6525-2 × NRS 364-7) mapping family (Barrett *et al*., 2004), allowing detection of homoeologous locations between the ancestral genomes at high frequency, and providing the basis for a standardized chromosome nomenclature. A third genetic map, incorporating information derived from genotyping of both genomic DNA-derived SSR and EST-SSR markers, has been recently developed using the F_1(GA43 × SVRR) population (Zhang *et al*., 2007).

The F_2(I.4R × I.5J) genetic map was exploited for QTL analysis of a number of vegetative morphogenesis, reproductive morphogenesis and reproductive development traits (Cogan *et al*., 2006a). The F_1(Sustain 6525-2 × NRS 364-7) population has also been used for QTL analysis, specifically targeting seed production traits such as inflorescence density, yield per inflorescence and thousand-seed weight (Barrett *et al*., 2005).

Functionally Associated Molecular Genetic Markers

Significance for pasture plant molecular breeding

The majority of strategies for selection of favourable marker allele–trait gene associations have been developed and implemented in obligate or facultative inbreeding species such as tomato, soybean, wheat, barley, rice and maize. For crops such these, trait-dissection and marker-assisted selection (MAS) are effectively coupled through pair-cross architecture. Marker–trait gene linkages are typically determined by QTL analysis of F_2, doubled haploid (DH) or recombinant inbred line (RIL) populations descended from contrasted parental genotypes, which generally represent donor and recipient varieties for MAS. Potential schemes for implementation include selection of the linked marker allele during recurrent backcrossing of progeny individuals to the recipient parent, in combination with genome-wide selection for restored background genotype. Such donor–recipient introgression schemes are not readily adaptable for outcrossing species with complex breeding systems, except under specific artificial conditions. In particular, development

and use of marker–trait linkages is problematic for several reasons. In a heterogeneous background, favourable *cis*-linkages may be reversed by recombination by the marker locus and the trait gene in any cycle of crossing, and may lead to inadvertent counter-selection. This effect may also occur in inbreeding pedigrees, but the degree of heterogeneity is constantly diminished, rather than remaining constant. The use of closely linked flanking marker alleles will mitigate such effects, but implies the necessity for more intensive trait-dissection analysis. This requirement is also dictated by the use of multiple polycross parents during varietal development: even for restricted base populations, with small numbers of foundation individuals (Forster *et al*., 2001b; Guthridge *et al*., 2001), multiple rounds of independent pair-cross-based trait-dissection studies would be required to identify all relevant marker locus–trait gene linkage relationships. Finally, consideration of potential introgression-based schemes (Forster *et al*., 2001a) has revealed high and probably unacceptable levels of logistical complexity, except for traits with high heritability which are under simple genetic control.

The limitations of molecular breeding strategies for pasture plants based on linked markers may be addressed through the development of diagnostic genetic markers based on functionally associated variation in candidate genes (Forster *et al*., 2004; Spangenberg *et al*., 2005; Cogan *et al*., 2006b). Similar strategies have been proposed for other outbreeding species, such as conifers (Neale and Savolainen, 2004). The reproductive habit and presumptive population structure of outbreeding forage species would be expected to dispose towards limited LD, extending over relatively short molecular distances (Mackay, 2001; Flint-Garcia *et al*., 2003; Forster *et al*., 2004; Gupta *et al*., 2005; Caldwell *et al*., 2006). This is especially relevant for long-established populations derived from a large number of founding parents, as expected for ecotypes and long-established varieties, in which many rounds of recombination have occurred. These factors would tend to favour the use of candidate gene-based functionally associated marker systems (Andersen and Lübberstedt, 2003; Lübberstedt *et al*., 2005) rather than whole genome scans (Rafalski, 2002) for association genetics, although newly synthesized populations with small numbers of parents may prove suitable for limited genome-wide marker-based analysis. Nucleotide variation in verified candidate genes will in such cases be closely associated with the casual mutations (quantitative trait nucleotide (QTN)) responsible for key agronomic trait variability and may be diagnostic for such variation (Sorrells and Wilson, 1997). Successful correlation of gene haplotype structure and phenotypic variation will provide the basis for a new paradigm in pasture plant molecular breeding based on direct selection of superior allele content at target genetic loci, allowing highly effective exploitation of germplasm collections for identification of potential parental genotypes. The subsequent progeny selection process will permit accumulation towards fixation of favourable alleles. In order to minimize the effects of inbreeding depression, parental individuals will be selected based on genome-wide genetic diversity at otherwise unselected genetic locations, and variability at these regions will be maximized during synthetic variety development (Forster *et al*., 2004; Spangenberg *et al*., 2005).

Candidate gene identification

The candidate gene-based approach for association genetics provides high resolution, potentially to the level of the QTN, in those species in which LD decays at very rapid rates (Neale and Savolainen, 2004). However, such analysis requires the presence of a large number of intragenic polymorphisms and the confidence that the correct target gene is under analysis. Candidate genes may be identified among within-species sequence resources or through translational genomics from related model and agronomic species. A hierarchy for candidate gene classification may be constructed (Forster *et al*., 2004). Primary candidate genes are those which may be directly identified by sequence annotation based on known function in either the target species or other taxa (Prioul *et al*., 1999; Pflieger *et al*., 2001). For perennial ryegrass, this category includes genes for lignin biosynthetic enzymes such as CCR (cinnamoyl CoA reductase) and CAD (cinnamyl alcohol dehydrogenase) (Lynch *et al*., 2002; McInnes *et al*., 2002) and for oligosaccharide metabolism enzymes such as 1-SST (1-sucrose:sucrose fructosyltransferase) (Chalmers *et al*., 2003). In each case, the biochemical function is well understood and both transgenic modification and mutant studies in related species (Lübberstedt *et al*., 2005) support functional classification and importance for the target trait.

The secondary candidate class corresponds to executor genes for biochemical processes which are currently unknown or generically annotated such as to provide minimal functional information. Transcriptomics analysis provides a method for identification of such candidates, either through comparison of contrasted cell type-specific or environment-specific gene expression (Kathiresan *et al*., 2006; Mager *et al*., 2006), or through comparison with the expression profile of a primary candidate template gene.

The tertiary candidate class corresponds to gene classes such as those for transcription factors and signal transduction factors, which are located higher in regulatory cascades than executor genes, but may show limited modulation of gene expression in response to developmental phase or environmental stimulus. Reverse genetics approaches such as anti-sense transcription and interfering RNA (RNAi) provide methods for functional classification of tertiary candidates (Vance and Vaucheret, 2001; Qi and Hannon, 2005). Induced mutagenesis methods (Waugh *et al*., 2006) such as EMS treatment (McCallum *et al*., 2000) and fast neutron bombardment (Li *et al*., 2001) are also capable of generating information for study of secondary candidates.

A final class of quaternary candidates may be described for which no relevant biological information is available. However, the physical position within the genome of such candidates may be correlated with quantitative variation for phenotypic traits. Although the confidence intervals for QTLs often extend over large molecular distances, especially for low heritability traits measured in small mapping families, verified candidate genes have in some instances been shown to be located close to maximum likelihood positions (Price, 2006). If comparative genetic analysis has defined the level of macrosynteny (and, ideally, microsynteny) between the target genome and

an appropriate model, a minimal number of candidate genes may be identified for more detailed analysis.

In practice, the different classes of candidate genes may coincide through a combination of analytical techniques. Genetical genomics uses genetic mapping, QTL detection and transcriptomics approaches which may combine secondary, tertiary and quaternary candidate identification (Jansen and Nap, 2001; Kadarmideen *et al*., 2006). Accurate identification of candidate genes may also be enhanced by the use of proteomic and metabolomic data, based on changes in protein structure and accumulation of specific metabolites.

Translational genomics for candidate gene identification

Perennial ryegrass

Perennial ryegrass is a member of the Poaeae tribe of the Poodae super-tribe in the Pooideae subfamily of the grass and cereal family Poaceae (Soreng and Davis, 1998). The *Lolium* and *Festuca* genera are closely allied, and the most closely related major cereal species is cultivated oats (*Avena sativa* L.) within the Aveneae tribe of the Poodae. The Triticeae cereal tribe (wheat, barley and rye) is located within the Triticodae super-tribe of the Pooidae. Rice (*Oryza sativa* L.), by contrast, is located in the Poaceae subfamily Bambusoideae. Translation genomics from rice to perennial ryegrass based on whole genome DNA sequence data consequently traverses a significant phylogenetic distance, but the use of partial genomic sequence and EST data from wheat (*Triticum aestivum* L.; Powell and Langridge, 2004) and barley (*Hordeum vulgare* L.) exploits closer taxonomic affinities. Future whole genome sequencing of the model plant species *Brachypodium distachyon* (Draper *et al*., 2001; Vogel *et al*., 2006), which is more closely related to the Triticeae than the Oryzeae, is also likely to be valuable for perennial ryegrass.

Examples of translational genomics to the benefit of perennial ryegrass candidate gene selection have been identified in several areas. For pathogen resistance, comparative studies have been performed for detection and validation of resistance gene analogues (RGAs) associated with resistance to the crown rust pathogen. The leaf rust receptor kinase gene (*LrK10*) which is located within the *Lr10* locus on chromosome 1AS of wheat has been found to confer resistance to leaf rust within certain wheat cultivars (Feuillet *et al*., 1998). *LrK10*-like sequences are conserved throughout the Poaceae, with comparisons between wheat and oat revealing 76% sequence similarity. Orthologous sequences have been identified at known chromosomal locations of conserved synteny in rice, wheat and oat (*Avena sativa* L.), which is also host to a *formae speciales* of crown rust (Cheng *et al*., 2002). This result suggests that similar orthologous sequences should be present in perennial ryegrass and may be accessible to a similar PCR-based isolation strategy. Primers designed to conserved regions at the end of the extracellular domain of the wheat *LrK10* gene were used to amplify products from perennial ryegrass genomic DNA. Subsequent cloning, sequencing and analysis of a 1.6 kb fragment revealed a single copy amplicon with a BLASTX match of 10^{-110} to

the wheat sequence. *In vitro* SNP discovery based on the parents of the $F_1(NA_6 \times AU_6)$ genetic mapping family detected 36 predicted SNPs and 14 indels. Validated SNPs were assigned to the upper region of LG1 in a region of predicted conserved synteny with wheat (1AS) and oat (LG-A) (P.M. Dracatos, Melbourne, 2007, unpublished data). This observation suggests that RGA structure and location is conserved at the macrosyntenic, and possibly the microsyntenic level. The *LpLrk10* SNP locus also maps in the region of LG1 which has previously been shown to contain two QTLs for quantitative field resistance to crown rust (Forster *et al*., 2004).

Translational genomics has also been applied to the study of gametophytic SI. Systems based on allelic variation at *S* and *Z* genes are present in a wide range of outbreeding Poaceae species, and it is assumed that inbreeding species carry loss-of-function alleles at either locus, or possibly at loci epistatic to *S* and *Z*. Genetic studies of SI have been performed in the blue canary grass species *Phalaris coerulescens* and cereal rye (*Secale cereale L.*), among others (Baumann *et al*., 2000). Genetic analysis in perennial ryegrass has been performed using the RFLP- and AFLP-based reference genetic map derived from the p150/112 population (Jones *et al*., 2002a). The *S* and *Z* loci were genetically mapped and located to LGs 1 and 2 respectively (Thorogood *et al.*, 2002), in the equivalent genomic regions defined by conserved synteny with the locations of their ortholoci in other self-incompatible Poaceae species such as cereal rye (Fuong *et al*., 1993; Voylokov *et al*., 1998; Korzun *et al*., 2001). This study was enabled by the detailed comparative genetic map analysis that defined the extent of conserved synteny between perennial ryegrass, the Triticeae cereals, oat and rice (Jones *et al*., 2002a).

A thioredoxin gene (*Bm*2) that was isolated from a meiotic tissue-derived cDNA library from *Phalaris coerulescens* has been demonstrated to detect a genetic locus closely linked to the *S* gene in both *P. coerulescens* (*c*.2 cM) and cereal rye (*c*.8 cM) (Baumann *et al*., 2000). The sequence of the rye ortholocus of *Bm*2 (*ScTrx*) was used to design a sequence tagged site (STS) marker. The STS primers also cross-amplified from the perennial ryegrass ortholocus of *Bm*2 (*LpTrx*), detecting a size polymorphism probably attributable to length variation in an internal SSR array. This size polymorphism was used to map *LpTrx* on to the p150/112 framework map, in addition to a number of genomic DNA-derived SSR loci (M.P. Dupal, Melbourne, 2007, unpublished data). The *LpTrx* locus coincided with the heterologous RFLP locus *xcdo98*, which is 0.6 cM from *S* on LG1. In addition, the SSR loci *xlpssrk14c02*, *xlpssrk15h05*, *xlpssrk09g05*, *xlpssrk03a02* and *xlpssrh02h04* are located in the vicinity of the *S* locus (Jones *et al*., 2002b). This mapping information has been augmented through the conversion of heterologous RFLP markers mapping close to the *P. coerulescens S* locus (Bian *et al*., 2004) into perennial ryegrass SNP loci. An expanded sib-ship of the $F_1(NA_6 \times AU_6)$ genetic mapping family (*c*.1000 genotypes) has been generated for fine-mapping of markers closely linked to *S*, in order to identify proximal and distal recombinants for selective phenotypic analysis. In addition, BAC clones containing linked sequences such as *LpTrx* have been selected to initiate physical mapping and subsequent positional cloning of the *S* gene (H. Shinozuka, Melbourne, 2007, unpublished data).

Translational genomics from perennial ryegrass may also be of benefit to other crop species. Genes for herbage digestibility and nutritive value have been intensively studied in the pasture grasses (Gallagher and Pollock, 1998; Heath *et al*., 1998, 2002; Lidgett *et al*., 2002; Lynch *et al*., 2002; McInnes *et al*., 2002; Chalmers *et al*., 2003, 2005; Johnson *et al*., 2003; Gallagher *et al*., 2004; Yamada *et al*., 2005), but to a lesser extent in the major cereals. However, lignin biosynthesis in cereals is of potential importance for pest resistance, control of lodging and residue feeding quality, while stem carbohydrates such as fructans may contribute tolerance to osmotic stresses (e.g. drought). Wheat ESTs showing significant nucleotide similarity to lignin biosynthetic genes from perennial ryegrass and other plant species were identified through annotation criteria. In particular, putative orthologues of the *Lp*CAD2, *Lp*CCR1 and *Lp*OMT1 genes were identified. RFLP loci detected by each of these genes are located within a 0.9 cM interval in the lower central region of perennial ryegrass LG7 (Cogan *et al*., 2005b), co-locating with QTLs for herbage quality traits. ESTs related to each of the template genes were identified within adjacent deletion bins (Endo and Gill, 1996; Qi *et al*., 2003) at the end of chromosome 7DL, while putative homoeoloci were also located on the other homoeologous group 7L chromosomes. The homoeologous group 3L chromosomes of wheat also contained putative ortholoci for each perennial ryegrass gene in distal bins. The distal regions of wheat group 3L and 7L chromosomes are the syntenic counterparts of the equivalent regions on perennial ryegrass LGs 3 and 7, in which clusters of herbage quality QTLs are located (Cogan *et al*., 2005b), suggesting that other members of lignin biosynthesis gene families may be located on LG3. Detection of QTLs for the solid stem character and sawfly resistance on wheat 3BL (Cook *et al*., 2004) provides the opportunity to target lignin biosynthesis gene ortholoci as candidates for marker-based selection in wheat.

DNA sequence comparisons were also made with genes for oligosaccharide metabolism, especially the fructosyltransferases (FTs) which are responsible for fructan biosynthesis (Chalmers *et al*., 2005) in perennial ryegrass. The *Lp*FT1 (Lidgett *et al*., 2002) and *Lp*1-SST (Chalmers *et al*., 2003) genomic sequences were aligned with rice genomic sequences and wheat ESTs, followed by intron–exon structure determination for the putative wheat FTs (Francki *et al*., 2006). Putative wheat ortholoci were assigned to deletion bins in regions of conserved synteny with the location of perennial ryegrass FTs, such as the group 7S homoeologous chromosomes. As rice is not a fructan accumulator, homologous sequences are likely to correspond to invertase (INV)-like genes which have evolved as FT genes in the Pooideae subfamily lineage that led to contemporary wheat and perennial ryegrass. This process is compatible with gene structural rearrangements, along with cycles of intragenomic gene duplication. As fructan accumulation has been reported to be associated with cold and drought tolerance (Vijn and Smeekens, 1999), wheat orthologues of FTs are candidates for a functional role in stem carbohydrate accumulation as a drought-adaptive feature.

Although FT and INV gene family structure has provided information on putative origin from a common ancestor, further comparative analysis of

gene function may be obtained by integration of data on quantitative trait variation from perennial ryegrass and wheat. In recent studies, QTL analysis has identified genomic regions influencing WSC content in perennial ryegrass and wheat in conserved syntenic locations corresponding to wheat homoloeologous chromosomes 1, 2, 5 and 6 (Cogan *et al*., 2005b; Turner *et al*., 2006; G. Rebetzke, Canberra, 2007, personal communication). Several wheat FT and INV-like genes (Francki *et al*., 2006) were assigned to map locations in these regions, providing the basis for more detailed analysis of candidate gene-QTL co-location in both species. A WSC QTL located on chromosome 2H of barley (Teulat *et al*., 2001) is also located in conserved synteny with the genomic region on perennial ryegrass LG2 (Cogan *et al*., 2005b), further supporting the translational genomics approach.

White clover

White clover is a member of the Trifolieae tribe of the cool-season Galegoid clade in the Papilinoideae subfamily of the legume family Fabaceae (Doyle and Luckow, 2003). The most closely related genus is *Melilotus* (sweet clovers) and the genus *Medicago*, including lucerne, is also part of the Trifolieae. As a consequence, the model legume species *M. truncatula* shares a common ancestor relatively recently in evolutionary time with white clover. Translational genomics based on whole genome sequencing of *M. truncatula* (Young *et al*., 2005; Zhu *et al*., 2005) is consequently anticipated to be highly efficient for members of the *Trifolium* genus. The other model legume species, *Lotus japonicus* Gifu, is also a Galegoid legume located in a separate tribe, Loteae.

Translational genomics for the benefit of white clover molecular breeding will be based on data not only from the model legumes, as well as soybean (*Glycine max* L.), but also from the dicotyledonous plant model species *Arabidopsis thaliana*. Saline stress tolerance is a major breeding objective for white clover. Increased prevalence of salt in soil in dryland areas and in high-usage irrigation zones, such as certain areas of Victoria, Australia, has adversely affected perennial pasture production. Although white clover is a salt-sensitive species, plants that are insensitive to low or moderate concentrations of NaCl have been observed, suggesting that intracellular ion compartmentation involving Na^+ and Cl^- exclusion from chloroplasts and cytoplasm may occur (Rogers *et al*., 1993, 1997). Salt-tolerant genotypes have been selected from within the Israeli cultivar Haifa and a salt-tolerant line (LCL) has been generated (Rogers *et al*., 1993). Pair-crosses between LCL genotypes and salt-intolerant individuals have been generated (Cogan *et al*., 2007), permitting trait-dissection of the salinity tolerance character, candidate gene mapping and evaluation of co-location with QTLs.

Template genes for salt tolerance have been identified from studies in *A. thaliana*. The *HKT1* gene class encodes a Na^+–K^+ co-transporter, which is a source of resistance to highly saline environments and is highly expressed in roots (Rus *et al*., 2004; Pardo *et al*., 2006; Rodriguez-Navarro and Rubio, 2006). The *A. thaliana* gene sequence was used for TBLASTX-based interrogation of the white clover EST sequence database and it identified a single candidate sequence. The *SOS1* and *SOS5* genes encode proteins with roles in antiporter

activity and in cell surface adhesion, respectively (Shi *et al.*, 2000, 2003). Both genes were identified through screening of *A. thaliana* mutagenized populations. The gene sequences for both these genes were used in TBLASTX-based interrogation analysis of the white clover database to identify convincing orthologues. The *SOS1* gene detected a putatively orthologous sequence in the white clover EST database with high confidence ($E = 1.1 \times 10^{-62}$ based on TBLASTX). In contrast, the *SOS5* gene sequence failed to identify any known sequence from white clover (N.O.I. Cogan, Melbourne, 2007, unpublished data). Access to gene sequences of this type will require the design of degenerate primers based on data from orthologous sequences of other legume species for evaluation on white clover template DNA. A further highly promising template sequence is provided by a *Krüppel*-like transcription factor gene which has been identified in *Medicago* species, is induced in roots in response to salt stress and confers tolerance through initiation of a recovery process (Frugier *et al*., 2000; Merchan *et al*., 2003). The *Mt*ZPT2 sequence identified three white clover ESTs contigs as potential ortholocus candidates (N.O.I. Cogan, Melbourne, 2007, unpublished data).

White clover requires frequent irrigation, and seasonal conditions have a marked effect on persistence (Doyle *et al*., 2000). Large white clover plants fragment into much smaller plants during spring due to the death of old stolons. The resultant plantlets are much more vulnerable to stress than larger plants, and with the onset of higher temperatures and moisture stress in late spring–early summer, a reduction of white clover content in pastures is observed. CBF/DREB (C-repeat binding factor/dehydration-responsive element-binding) gene families are represented in a range of plant species. This multigene family encodes a class of DNA-binding transcription factor proteins, and gene expression is induced in response to dehydration. A number of DREB genes have been described from soybean (*Glycine max* L.) (Gao *et al*., 2005) providing templates for gene identification in white clover.

SNP Discovery Strategies for Temperate Pasture Species

The validity of the candidate gene-based molecular marker concept for outbreeding pasture species has been determined through development and implementation of large-scale methods for SNP discovery. This process has been based on a twin phase strategy. An *in silico* discovery component (Buetow *et al*., 1999; Picoult-Newberg *et al*., 1999) has exploited computational identification of predicted SNPs in EST sequence contigs derived from multiple heterogeneous individuals of the white clover variety cultivar Grasslands Huia (Sawbridge *et al*., 2003a) and the perennial ryegrass cultivar Grasslands Nui (Sawbridge *et al*., 2003b), respectively (Spangenberg *et al*., 2005). Computational analysis of pre-existing sequence data sets provides an effective low-cost strategy for SNP discovery. However, EST collections may be generated from a limited number of genotypes and hence fail to capture significant proportions of SNP variation, including diversity relevant to specific germplasm. SNP variation may also fail to be represented in specific candidate genes, requiring targeted *in vitro* discovery activities. None the less, *in*

silico discovery provides the most obvious method for *de novo* SNP identification in a species with relatively underdeveloped genetic systems, and has been applied to a number of plant species including *Arabidopsis thaliana* (Schmid *et al*., 2003), wheat (Somers *et al*., 2003), barley (Kota *et al*., 2003; Rostocks *et al*., 2005) and tomato (Yang *et al*., 2004; Labate and Baldo, 2005).

The *in vitro* discovery component has exploited a method based on cloning and sequencing of gene-specific amplicons from the heterozygous parents of two-way pseudo-testcross mapping families. The cloned amplicons are then aligned to predict SNP incidence within and between parental genotypes. Although this method is costly and time-consuming compared to direct sequencing of PCR products, it provides several important advantages. Ambiguities in sequence traces due to heterozygous indels may be readily distinguished, and haplotype structures within amplicons are directly determined (Zhang and Hewitt, 2003). In addition, paralogous sequences amplified by conserved primers may be discriminated, and used to distinguish between members of multigene families (Edwards *et al*., 2007). The putative SNPs from both discovery phases are then validated in the progeny set from the parental cross, and cross-validated in other sib-ships and diverse germplasm. The *in vitro* discovery method has been progressed in the form of three distinct intensity streams: the low-intensity stream involved short ESTs, providing single SNP loci for structured map enhancement; the medium-intensity stream involved full-length cDNAs, providing several SNP loci and partial haplotypic data; and the high-intensity stream involved full-length genes with intron–exon structure, providing multiple SNP loci and determination of complete haplotype structures (Cogan *et al*., 2006b).

Status of Functionally Associated Genetic Marker Development for Perennial Ryegrass

The majority of perennial ryegrass SNPs have to date been obtained by *in vitro* discovery from genes which were previously analysed as RFLP markers (Faville *et al*., 2004). 'Proof-of-concept' for the *in vitro* discovery process was obtained with *Lp*ASRa2, an abscisic acid (ABA) inducible gene which is believed to be associated with osmotic stress tolerance in rice and maize (Vaidyanathan *et al*., 1999; Jeanneau *et al*., 2002). SNP haplotypes from the parents of the $F_1(NA_6 \times AU_6)$ genetic mapping family were defined and correlated with amino acid changes which may affect the functionality of the derived protein. Alternatively, the characterized SNPs may be in LD with functionally significant changes in the transcriptional control regions (Paran and Zamir, 2003), given haplotype stability over gene-length distances. Validated SNPs were used to determine diversity and putative haplotype structure within diverse germplasm collections, along with preliminary determinations of LD (Cogan *et al.,* 2006b). Data obtained from *Lp*ASRa2, several full-length herbage quality (lignin biosynthesis, oligosaccharide metabolism) genes (Ponting *et al*., 2005), and carbohydrate metabolism and flowering time control genes (Skøt *et al*., 2005a) suggested that contiguous

LD blocks in perennial ryegrass genes are unlikely to extend further than 1–2 kb, and that LD declines to minimal values ($r^2 = 0.2$) over less than 10 kb. LD decays slightly less rapidly in cultivars and ecotypes as compared to highly diverse genotype collections, but the rate of decline is still comparable to the length of the transcriptional unit.

Over 150 genes were introduced into the *in vitro* SNP discovery process for perennial ryegrass, of which 100 were sequenced and aligned, with a total of 1592 putative SNPs across 82 genes. SNPs were validated using the single nucleotide primer extension (SNuPe) assay for 66 genes. Over a total of 87 kb of resequenced DNA, a relatively high SNP frequency (for four haplotypes) of 1 per 55 bp was observed, with higher incidence in intron compared to exon sequences, as anticipated (Cogan *et al*., 2006b). Failure to validate predicted SNPs was observed for a minority (*c*.25%) of loci, presumably due to residual alignment of paralogous sequence derived from multigene families. The *in vitro* discovery process has been augmented with limited *in silico* discovery, and continues with an emphasis on genes associated with resistance to plant diseases such as crown rust, including defence response (DR) genes (chitinases, catalases, superoxide dismutases, β-glucanases, etc.) and resistance (R) genes (NBS-LRR class, receptor kinases, etc.). Validated gene-associated SNP loci have been assigned to the perennial ryegrass p150/112 (Jones *et al*., 2002a,b) and $F_1(NA_6 \times AU_6)$ (Faville *et al*., 2004) genetic maps and the parental maps obtained from a *L. perenne* × *L. multiflorum* interspecific cross ($F_1[Andrea_{1246} \times Lincoln_{1133}]$), and in many cases show co-location with pre-existing EST-RFLP loci (Cogan *et al*., 2006b). Co-location of SNP loci with QTLs for disease resistance, mineral content, flowering time and vegetative morphogenesis traits has been demonstrated (Cogan *et al*., 2005a; Shinozuka *et al*., 2005; Yamada *et al*., 2005).

Status of Functionally Associated Genetic Marker Development for White Clover

The complementary white clover SNP discovery process has been dominated by the *in silico* strategy, due to the presence of a relatively smaller number of well-characterized candidate genes. A sample of 236 EST contigs was selected for validation using the parents and progeny of the two-way pseudo-testcross mapping families $F_1(Haifa_2 \times LCL_2)$ and $F_1(S184_6 \times LCL_6)$. A total of 106 of the clusters (45%) were initially identified, and more detailed analysis confirmed 58 clusters as containing validated segregating SNPs. The SNP-containing genes belong to a range of predicted functional categories, including ribosomal proteins, heat-shock proteins, calmodulins, and lipid and organic acid biosynthesis enzymes (Cogan *et al*., 2007). The *in silico*-derived SNPs are currently being deployed along with genomic DNA-derived SSR and EST-SSR markers to generate second generation trait-specific genetic maps for white clover. Polymorphic SNP- and SSR-containing ESTs have also been used for comparative genomics analysis by comparison with whole genome data from the model legume species *M. truncatula* and *L. japonicus*, as well as *Arabidopsis*

thaliana. The homoeologous groups of white clover were aligned with the model species genomes, revealing a predominant correspondence between each group and one of the chromosomes of *M. truncatula* (George *et al*., 2006a), although evidence for evolutionary translocations and paralogous relationships was also obtained. Similar results were obtained from empirical mapping studies based on *M. truncatula*-derived SSR markers (Zhang *et al*., 2007).

The *in vitro* SNP discovery process for white clover is expected to be influenced by the allopolyploid genetic constitution of this species. *In vitro* SNP discovery from allopolyploids has so far been confined to inbreeding plant species such as wheat (Bryan *et al*., 1999; Caldwell *et al*., 2004; Lu *et al*., 2005). Multiple sequence haplotypes obtained from single homozygous genotypes may be confidently attributed to non-homologous genes, and use of aneuploids such as wheat nullisomic-tetrasomic (NT) substitution lines can assign different haplotypes to specific chromosomes (Caldwell *et al*., 2004). A similar strategy has been implemented for *in silico*-predicted wheat SNPs (Somers *et al*., 2003; Ogihara *et al*., 1994), permitting identification of putative homoeologous sequence variants (HSVs). In contrast, high levels of intrapopulation genetic diversity (George *et al*., 2006b) and intragenotype heterozygosity (Kölliker *et al*., 2001; Jones *et al*., 2003) have been observed in white clover. In addition, aneuploid lines for one-step assignment to chromosomes, such as those of hexaploid wheat, are not available. The combination of variability between paralogous, homoeologous and homologous sequences is likely to complicate *in vitro* SNP discovery and identification of both HSVs and paralogous sequence variants (PSVs).

A total of 43 white clover cDNAs were selected from public databases, including the cyanogenesis-associated linamarase gene *Tr*LIN (Oxtoby *et al*., 2004) and from the unigene resource of Sawbridge *et al*. (2003a), including genes for flavonoid biosynthesis (relevant to bloat safety) and organic acid biosynthesis (relevant to aluminium tolerance and phosphorus acquisition). The F_1(Haifa$_2$ × LCL$_2$) and F1(S184$_6$ × LCL$_6$) families provided the parental DNA templates for *in vitro* SNP discovery. DNA sequence suitable for SNP discovery was obtained from multiple amplicons of 35 genes, corresponding to total of 29.4 kb of consensus resequenced genomic DNA, at an average of 840 bp per template gene. High levels of haplotypic complexity were observed (Lawless *et al*., 2006), for the majority of template genes, at levels in excess of prediction for homologous sequence amplification. Elevated rates of failed validation for predicted SNPs were also observed (82% compared to 25% in perennial ryegrass, across comparable sample sizes) (Lawless *et al*., 2006; Edwards *et al*., 2007). None the less, individual SNP loci suitable for genetic mapping were validated for selected genes. In some instances, as for the *Tr*PK protein kinase gene, allelic haplotypes identified through SNP validation were highly differentiated at the sequence level through large intron-located indels. Although higher levels of intragenomic paralogy could account for the lower efficiency of *in vitro* SNP discovery in white clover compared to perennial ryegrass, homoeologous gene amplification (Cronn and Wendel, 1998) is a more likely explanation.

Refined methods are clearly required to improve the efficiency of *in vitro* SNP discovery, and will require discrimination of genome- and gene-specific

sequences. Putative paralogous sequences may be subtracted from haplotype sequence alignments, but this approach depends on the ability to identify domains associated with specific gene family members. The identification of putative progenitor genomes (Ellison *et al*., 2006) may allow a similar approach to homoeologous haplotype discrimination, based on generation and alignment of amplicons from contemporary *T. occidentale* and *T. pallescens* genotypes and subtraction of related haplotypes. However, significant recombination and sequence divergence since white clover speciation would confound this approach, and the possibility remains that other currently uncharacterized taxa may have been the genuine diploid progenitors. Gene-specific sequences may also be identified from BAC libraries rather than full-length cDNA or EST collections. Comparison of sequences from independent BACs selected with template genes will allow directed primer design in the direction of locus-specific features.

Association Genetics for Outbreeding Pasture Species

The feasibility of whole genome scans for association mapping in outbreeding forage species was assessed by AFLP profiling of perennial ryegrass (Skøt *et al*., 2002, 2005b). Natural populations from Europe were selected on the basis of pre-existing genecological data (Sackville-Hamilton *et al*., 2002) for traits such as cold tolerance and flowering time variation. A small number of AFLP loci were identified as potentially showing association with genes controlling the target traits, especially for flowering time, for which identified loci were shown to map to LG7 in the vicinity of a major QTL for heading date variation (Armstead *et al*., 2004). However, significant LD was also observed between unlinked loci, indicative of residual population structure effects. The focus of association mapping studies has subsequently shifted to the candidate gene-based approach.

Large-scale SNP development for perennial ryegrass and white clover has permitted the design of crucial 'proof-of-concept' experiments in association genetics (Thornsberry *et al*., 2001; Flint-Garcia *et al*., 2003). The current emphasis is on perennial ryegrass herbage quality genes (Lidgett *et al*., 2002; Lynch *et al*., 2002; McInnes *et al*., 2002; Chalmers *et al*., 2003), including those subjected to high-intensity SNP discovery such as the lignin biosynthesis genes previously identified as co-locating with QTLs for herbage digestibility on LG7 (Cogan *et al*., 2005a). An association mapping panel (AMP) of 384 genotypes has been established, including 192 individuals of globally diverse origin and 64 individuals from each of three ecotypes. Population structure was determined through genotyping with *c*.120 SSR markers, revealing minimal population differentiation, apart from a Swiss ecotype from the Zürich Uplands (M.P. Dobrowolski, Melbourne, 2007, unpublished data). SNP data was obtained for 144 loci across 21 cell wall biosynthesis and oligosaccharide metabolism genes (N.O.I. Cogan, Melbourne, 2007, unpublished data). In parallel, phenotypic analysis for lignin and WSC content is being performed through the use of calibrated near infrared reflectance spectroscopy (NIRS) analysis. Successful identification of correlations

between gene haplotype structure and phenotypic variation will validate the future use of this approach for other target agronomic traits, and generate diagnostic markers for implementation in practical breeding programmes (Dobrowolski and Forster, 2007). A similar approach has been used in a parallel programme for SNP loci within the *Lp*AlkInv alkaline INV gene, but significant associations with WSC content were not detected (Skøt *et al*., 2005a).

Future Prospects and Activities

Candidate gene verification

The concept of functionally associated markers as the optimal systems for molecular breeding of temperate pasture plants has been supported by results to date. However, there are substantial further requirements for the transition from 'proof-of-concept' to implementation in commercially relevant breeding programmes. In the first instance, enhanced methods for selection of candidate genes are required. The majority of perennial ryegrass and white clover candidates so far studied belong to the primary class, which will be continually expanded by further developments in translational genomics. Within-species functional genomics will also identify new target genes. Transcriptomics studies for perennial ryegrass and white clover have been based on high-density spotted cDNA microarrays containing approximately 15,000 unique genes for each species. The next generation of microarrays will be based on oligonucleotide arrays customized and synthesized *in situ* using the CombiMatrix (www.combimatrix.com) CustomArray synthesizer system. Prototype arrays of 60-mer oligonucleotides corresponding to the perennial ryegrass and white clover unigene sets have already been generated (R. Chapman, Melbourne, 2007, unpublished data) and are available for detection of secondary candidates modulated in response to developmental stage and environmental variation.

Direct verification of candidate gene function is possible through nullification of function. Gene silencing systems based on sense and anti-sense RNA expression, virus-induced gene silencing (VIGS) and small inhibitory RNAs (siRNAs) (Hammond *et al*., 2001; Vance and Vaucheret, 2001; Holzberg *et al*., 2002) have been developed for key forage species, as well as efficient methods for biolistic (perennial ryegrass) and T-DNA-mediated transformation (perennial ryegrass, white clover) and plant regeneration.

Mutagenesis approaches based on transposon insertion or treatment with alkylating agents and ionizing radiation are also of potential value. Targeted induced local lesions in genomes (TILLING) is a method for identification of candidate genes based on saturating EMS mutagenesis followed by highly sensitive detection of induced point mutations in pooled DNA samples (Till *et al*., 2003). Identification of candidate genes prioritizes TILLING targets, and DNA pools of varying depth are composed from M_2 individuals (selfed progeny of original heterozygous M_1 plants from an inbred line source), while M_3 seed from each M_2 individual is retained. Mutations in a target gene are

indicated through formation of heteroduplexes, which are detected by methods such as denaturing HPLC (dHPLC) (McCallum *et al*., 2000) or cleavage with the heteroduplex-recognizing enzyme CEL I (Oleykowski *et al*., 1998). DNA pools containing mutant alleles are decomposed, the mutant and wild type alleles are identified and the archived M_3 family seed can then be accessed for assessment of phenotype (Waugh *et al*., 2006). As a spectrum of allelic mutations is generated by EMS treatment, both loss-of-function and gain-of-function mutations may be generated, with relevance both for functional classification and potentially as sources of new variability (mutation breeding). As a complement to the reverse genetics strategy based on TILLING, fast neutron bombardment (FNB) provides opportunities for both forward and reverse genetics approaches to functional analysis. FNB typically generates chromosomal deletions of moderate size (1–10 kb), similar to the size range of typical gene or gene clusters (Li *et al*., 2001). The disadvantage of FNB for mutation breeding is that the majority of mutations lead to loss-of-function through gene ablation. However, FNB mutagenesis provides the capability for extremely deep pooling for mutation detection using methods such as critical extension-PCR (CE-PCR) (Jansen *et al*., 1997). FNB-mutagenized populations consequently provide a valuable complement to TILLING populations for reverse genetics. For forward genetics, maintenance of mutagenized populations is logistically challenging, but if novel target trait-related phenotypes can be identified, FNB-mutagenized populations are well suited for individual genotype screening using oligonucleotide arrays to identify the ablated gene sequence. Phenotypic pre-screening has also been proposed as a general approach for enrichment during reverse genetics screens (Starker *et al*., 2006).

As methods for analysis of loss-of-function mutations generated by transposon or physiochemical mutagenesis generally involve fixation of the mutation in a homozygous state, such approaches are expected to be limited for obligate outbreeding species such as perennial ryegrass and white clover. SI will complicate the fixation of alleles that are identical by descent, unless it is achieved by line breeding through different parallel lineages. It is possible to overcome the SI systems to some extent, either through the use of environmental conditions such as elevated temperature in perennial ryegrass (Wilkins and Thorogood, 1992) or through the use of self-fertile mutations such as the S_f mutation of white clover (Michaelson-Yeates *et al*., 1997). Advanced generation inbred lines have been produced for perennial ryegrass (Posselt, 1993). However, any significant inbreeding based on selection of a particular genomic region is also vulnerable to the effects of fixation of *cis*-linked deleterious alleles producing confounding phenotypic effects.

Two types of mutations should be immediately accessible to functional analysis. The first is a loss-of-function mutation derived from an effective allele that is brought into combination with another pre-existing recessive allele. This would correspond to a heterozygous combination at the allelic sequence level, but would produce a 'knockout' phenotype. The second is the generation of a dominant gain-of-function mutation, and again, the phenotypic effect should be apparent irrespective of the allelic combination. These two classes of mutations will also be of importance for applications in mutation

breeding. A further complication for functional analysis arises due to gene function redundancy. This is not unique to the outbreeding pasture species, as gene duplication and polyploidy are ubiquitous in the higher plants. However, the presence of two closely related diploid genomes in the allotetraploid taxon white clover (Ellison *et al*., 2006) may potentially complicate the analysis of loss-of-function mutations if the gene function disrupted in one genome is compensated by the corresponding locus in the other genome.

For functional genomics analysis, an alternative to the use of the main target species would be the development of mutagenized populations based on closely related inbreeding diploid species. For perennial ryegrass, the obvious choice would be darnel (*Lolium temulentum* L.). Lines of this species such as the IGER (Aberystwyth) line Ba3081 have been used extensively by grass physiologists to study developmental processes with minimal genotypic variation (Evans, 1999). However, certain aspects of perennial ryegrass biology (such as perenniality itself) are not available to this model system. For white clover, the putative progenitor species *T. occidentale*, as an autogamous diploid, provides an obvious model system. An EMS mutagenized population of 1075 M2 lines (consisting of *c*.8000 plants) has been generated, and phenotypically screened for mutations affecting condensed tannin (CT) biosynthesis (W.M. Williams, Palmerston North, 2007, personal communication).

Implementation in breeding programmes

Examination of pre-commercial and commercial pasture plant breeding programmes, especially for perennial ryegrass, allows definition of the optimum stages for gene-associated SNP haplotype marker implementation. Perennial ryegrass germplasm improvement at the Institute of Grassland and Environmental Research (IGER), Aberystwyth, UK, has typically been based on recurrent selection within a single germplasm pool, while the breeding programme of Pyne Gould Guinness (PGG) Seeds, New Zealand, at least prior to the recent merger with Agricom and Wrightsons, was characterized by the exploitation of highly divergent initial parental materials, often of exotic provenance. The breeding programme of Agriseeds New Zealand is intermediate in nature between these extremes, being based on initial crosses between parental plants from different adapted germplasm pools in order to introduce novel variation. More divergent material has been incorporated as required, but broad adaptation of the commercial germplasm pool to its target markets in Australasia has been maintained (F. Wilson, Christchurch, 2006, personal communication). Within this generic framework, the most obvious stages for potential intervention using gene-specific haplotype-based markers have been identified (Table 10.1) as in Year 1 (optimized selection of mother plants in the initial intervarietal crosses) and in Year 3 (optimized selection of parents of synthetic populations). The logistics of marker implementation differ between these two stages, with smaller numbers of intervarietal parental individuals requiring intensive genotyping (associated with multiple target loci) in Year 1, as compared to larger numbers of recombinant individuals

Table 10.1. Generic structure of the Agriseeds New Zealand perennial ryegrass breeding programme.

Year	Process	Logistics	Potential for haplotype-based selection
1	F_1 crosses between paired selected germplasm pools	*c.*40 crosses made per year, with *c.*20 genotypes per variety (*c.*800 parental plants)	Yes
2	Production of F_2 seed by cross-pollination within F_1 cross-derived families	Further advantageous recombination may be obtained by progressing to a F_3 generation, but this requires time in the programme and is not generally performed	
3–4	Establishment of progeny plants in spaced plant nursery and grazing-based selection	Grazing by cattle and sheep identifies persistence and regrowth over *c.*90,000 plants	
5	Selected plants established in clonal rows for visual evaluation	Clonal replicates of ten plants established for up to 2000 genotypes, in order to minimize environmental variation	Yes
6	Selected parental plants polycrossed to produce synthetic seeds	Number of synthetic parents variable, typically 4–12 (cv. Tolosa derived from eight parents). Generally 50 synthetics generated	
7–9	Multi-location small plot trials for specific synthetics	Conducted under conditions of grazing by cattle and sheep	
9–11	Selected synthetics from small plot trials established as large plot National Forage Variety Trials (NFVTs)	Conducted under conditions of grazing by cattle and sheep	
12	Potential variety release in concert with farmer trials	Conducted under conditions of grazing by cattle and sheep, typically six experimental cultivars obtained	

(10^3–10^4) requiring specific genotyping for permutation of the allelic variants detected in the initial crosses. Although accurate haplotype-based selection would be anticipated to be most effective in the long-term through deployment in Year 1, the relatively long lead time to the market also suggests that strategies for deployment at the Year 5 stage for current varietal programmes would be effective and timely. The effectiveness of precise strategies for haplotype-based marker deployment may be modelled and simulated in

order to calibrate initial models based on hypothetical gene variant effects with a defined range of effects on trait variation.

Cost-effective methods for low-cost genotypic analysis are of critical importance for implementation in forage plant breeding, due to the low profit margins associated with seed sales and extended renovation cycles characteristic of perennial pastures. Automated methods for genotyping at selected SNP loci over large numbers of target genotypes will be achieved through the use of technology platforms such as those based on micro-bead linked PCR-assays, with or without optical fibre array structure (Shen *et al*., 2005; Dunbar, 2006). Individually addressed micro-beads, such as the commercial digital hologram-based CyVera system (http://www.illumina.com/technology/cyvera_overview.ilmn), or optically bar-coded colloidal particles (Battersby *et al*., 2002), will provide maximum flexibility for such analysis through deployment at different multiplex ratios for different applications. Alternatively, moderate multiplex high-throughput SNP analysis at moderate multiplex ratios is accessible to matrix absorption laser desorption-ionization time-of-flight mass spectroscopy (MALDI-TOF MS; Tost and Gut, 2005).

Novel methods for exploitation of germplasm resources

Although substantial phenotypic variation for important target traits is present within the gene pool of perennial ryegrass and white clover, more extreme phenotypic variation may be observed in closely related species. For instance, higher levels of heat and drought tolerance are observed in tall fescue (*Festuca arundinacea* Schreb.) than in perennial ryegrass. The *Lolium* and *Festuca* genera are closely related within the Poeae tribe (of the Poodae super-tribe of the Pooideae subfamily of the Poaceae; Soreng and Davis, 1998). Fertile intergeneric hybrids and derived introgression lines between ryegrass species and tall fescue have been produced through the use of pentaploid hybrid bridges (Humphreys, 1989) and by generation of symmetrical and asymmetrical somatic hybrids (Spangenberg *et al*., 1994). The agronomic effectiveness of *Festulolium* hybrids has, however, been limited through the relatively uncontrolled nature of the hybridization process. Identification of SNP haplotypes associated with superior allele content can be extended from perennial ryegrass to other *Lolium* and *Festuca* species, both for within-taxon improvement and for optimum selection of parental genotypes for intergeneric crosses. Validated SNP locus assays from perennial ryegrass candidate genes have been tested for cross-taxon analysis in *Festuca* species, with reasonable levels of success. This result implies considerable levels of gene structure conservation between ryegrass and fescue species, and the potential for novel allele definition in exotic sources. This process may also be used to inform high-resolution searches for related novel alleles within perennial ryegrass, based on comparison of gene and protein structure. Similar approaches may be applied to the putative progenitors of white clover, and other related *Trifolium* taxa capable of sexual hybridization.

The effectiveness of such an 'allele mining' approach (Buckler *et al*., 2006) is dependent on the ability to perform high-throughput analysis of large germplasm collections, both in order to detect previously characterized alleles, and to identify novel alleles. Highly parallel DNA sequencing can address this requirement, especially through the use of the Life Sciences 454 platform (http://www.454.com). This system utilizes individual DNA molecule-dependent amplification at critical dilution in emulsion phase micelles, followed by sequestration of templates in picolitre-sized reactors, sequence data being generated by pyrosequencing analysis (Margulies *et al*., 2005). PCR amplicons have been demonstrated to be viable templates for this analysis. Over 25 million nucleotides may be read in a single run of the system, with average read lengths in the range from 100–150 bp. Although this scale of sequence analysis is potentially limiting for *de novo* assembly of complex genomes, and more suited to analysis of BAC clone pools, it is appropriate for resequencing of previously characterized regions for SNP haplotype analysis. The individual molecule-dependent analysis is also suited to the heterozygous nature of pasture plant species, for which haplotype structure in outbred genotypes must generally be determined by inferential methods. In addition, the highly parallel nature of the 454 sequencing platform will permit large-scale analysis of germplasm collections to a previously unrealized level, permitting assessment of genuine global diversity. The highly parallel nature of the analysis also presents a challenge for tracking of haplotype origin from individual genotypes. In order to address this problem, hierarchical pooling techniques based on multidimensional parameters, analogous to those used for PCR-based screening of BAC libraries to identify specific clones, may be designed and tested. Such a methodology will be of very high value for germplasm characterization in a range of agricultural species, and may legitimately be termed 'allele panning' rather than 'allele mining'.

Conclusions

The key components for implementation of functionally associated genetic markers in perennial ryegrass and white clover breeding have been established. They include: generation of substantial genomic sequence resources; methods for functional analysis and verification of candidate gene status, including transcriptomics and transgenesis; efficient methods for *in silico* and *in vitro* discovery and validation of gene-associated SNP loci; evaluation of intra- and inter-population diversity; estimation of SNP haplotype complexity and LD; development of robust protocols for phenotypic analysis of key agronomic traits; design and implementation of experiments for haplotype–phenotype correlation; and development of efficient strategies for gene-based marker deployment in existing germplasm improvement programmes. These processes are mature for perennial ryegrass, and despite the added complexities of genetic architecture in white clover, optimized methods for SNP discovery should be equally effective for the latter species. The strategies are

also suitable for other agronomically important temperate forage species, such as the grasses Italian ryegrass, tall fescue (*Festuca arundinacea* Schreb.), meadow fescue (*Festuca pratensis* Huds.), cocksfoot (*Dactylis glomerata* L.) and timothy (*Phleum pratense* L.), as well as the legumes red clover (*Trifolium pratense* L.) and lucerne (*Medicago sativa* L.).

Warm-season pasture grasses with C_4 photosynthetic capacity, such as kikuyu grass (*Pennisetum clandestinum*), weeping lovegrass (*Eragrostis curvula*), paspalum (*Paspalum dilatatum*), Bahia grass (*Pasplaum notatum*) and signal grass (*Brachiaria decumbens*) also provide potential targets. The typical mode of reproduction for such species is apomixis, which is a parthenogenetic mechanism in which seed development occurs in the absence of gamete fusion, such that the maternal genotype is asexually reproduced. Apomixis and sexuality may coexist in the same plant in some species (facultative apomicts), while other species only reproduce asexually (obligate apomicts). The genetic control of aposporous apomixis (in which the megametophyte develops from a somatic cell within the ovule) has been studied in a number of grass species (Pessino *et al*., 1999). Different chromosome races are frequently observed, corresponding to outbreeding sexual diploid genotypes and polyploid apomictic genotypes. As the apomictic genotypes produce viable male gametes, pair-crossing and genetic mapping studies may be performed when using pollen donors to sexual plants (Nogler, 1984), and functionally associated genetic markers may be effectively implemented during this phase of genetic diversity generation. A number of genetic mapping studies in genera such as *Panicum*, *Pennisetum*, *Brachiaria* and *Paspalum* (Savidan, 1981; do Valle *et al*., 1994; Pessino *et al*., 1997; Ozias-Akins *et al*., 1998; Martinez *et al*., 2001) have demonstrated the presence of putative single apomixis-specific genomic regions capable of transfer to divergent sexual germplasm. Elite genotypes generated through selection based on genic markers, such as for enhanced herbage quality, may consequently undergo large-scale clonal propagation. In the context of global climate change, genetic improvement of warm-season apomictic pasture grasses will be of key importance for the future of currently temperate pasture cultivation regions, such as those of southern Australia.

Acknowledgements

This original research described here was supported by funding from the Victorian Department of Primary Industries (DPI), Dairy Australia Ltd, the Geoffrey Gardiner Dairy Foundation, Meat and Livestock Australia Ltd and the Molecular Plant Breeding Cooperative Research Centre. The EST generation and characterization programme was jointly funded by DPI and AgResearch New Zealand. Frances Wilson and Courtney Inch (Agriseeds New Zealand) provided information and advice on specific components of commercial pasture plant breeding programmes. The authors wish to thank all colleagues past and present within DPI for their contributions, and Professor Michael Hayward for scientific support and advice.

References

Abberton, M.T. and Marshall, A.H. (2005) Progress in breeding perennial clovers for temperate agriculture. *Journal of Agricultural Science* 143, 117–135.

Andersen, J.R. and Lübberstedt, T. (2003) Functional markers in plants. *Trends in Plant Science* 8, 554–560.

Armstead, I.P., Turner, L.B., King, I.P., Cairns, A.J. and Humphreys, M.O. (2002) Comparison and integration of genetic maps generated from F_2 and BC_1-type mapping populations in perennial ryegrass. *Plant Breeding* 121, 501–507.

Armstead, I.P., Turner, L.B., Farrell, M., Skøt, L., Gomez, P., Montoya, T., Donnison, I.S., King, I.P. and Humphreys, M.O. (2004) Synteny between a major heading-date QTL in perennial ryegrass (*Lolium perenne* L.) and the *Hd3* heading-date locus in rice. *Theoretical and Applied Genetics* 108, 822–828.

Attwood, S.S. (1940) Genetics of cross-incompatibility among self-incompatible plants *of Trifolium repens*. *Journal of the American Society of Agronomy* 32, 955–968.

Attwood, S.S. (1941) Controlled self- and cross-pollination of *Trifolium repens*. *Journal of the American Society of Agronomy* 33, 538–545.

Attwood, S.S. (1942a) Oppositional alleles causing self-incompatibility in *Trifolium repens*. *Genetics* 27, 333–338.

Attwood, S.S. (1942b) Genetics of pseudo-self-incompatibility and its relation to cross-incompatibility in *Trifolium repens* L. *Journal of Agricultural Research* 64, 699–709.

Barrett, B., Griffiths, A., Schreiber, M., Ellison, N., Mercer, C., Bouton, J., Ong, B., Forster, J., Sawbridge, T., Spangenberg, G., Bryan, G. and Woodfield, D. (2004) A microsatellite map of white clover (*Trifolium repens* L.). *Theoretical and Applied Genetics* 109, 596–608.

Barrett, B.A., Baird, I.J. and Woodfield, D.R. (2005) A QTL analysis of white clover seed production. *Crop Science* 45, 1844–1850.

Battersby, B.J., Lawrie, G.A., Johnston, A.P.R. and Trau, M. (2002) Optical bar-coding of colloidal suspensions: applications in genomics, proteomics and drug discovery. *Chemical Communications* 14, 1435–1441.

Baumann, U., Juttner, J., Bian, X.-Y. and Langridge, P. (2000) Self-incompatibility in the grasses. *Annals of Botany* 85 (Supplement A), 203–209.

Bian, X.-Y., Friedrich, A., Bai, J.-R., Baumann, U., Hayman, D., Barker, S.J. and Langridge, P. (2004) High-resolution mapping of the *S* and *Z* loci of *Phalaris coerulescens*. *Genome* 47, 918–930.

Bray, R.A. and Irwin, J.A.G. (1999) *Medicago sativa* L. (lucerne) cv. Hallmark. *Australian Journal of Experimental Agriculture* 39, 643–644.

Bryan, G.J., Stephenson, P., Collins, A., Kirby, J., Smith, J.B. and Gale, M.D. (1999) Low levels of DNA sequence variation among adapted genotypes of hexaploid wheat. *Theoretical and Applied Genetics* 99, 192–198.

Buckler, E.S., Gaut, B.S. and McMullen, M.D. (2006) Molecular and functional diversity of maize. *Current Opinion in Plant Biology* 9, 172–176.

Buetow, K.H., Edmonson, M.N. and Cassidy, A.B. (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genetics* 21, 323–325.

Caldwell, K.S., Dvorak, J., Lagudah, E.S., Akhunov, E., Luo, M.-C., Wolters, P. and Powell, W. (2004) Sequence polymorphism in polyploid wheat and their D-genome diploid ancestor. *Genetics* 167, 941–947.

Caldwell, K.S., Russell, J., Langridge, P. and Powell, W. (2006) Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. *Genetics* 172, 557–567.

Carruthers, V.R. and Neil, P.G. (1997) Milk production and ruminal metabolites from cows offered two pasture diets supplemented with non-structural carbohydrate. *New Zealand Journal of Agricultural Research* 40, 513–521.

Chalmers, J., Johnson, X., Lidgett, A. and Spangenberg, G.C. (2003) Isolation and characterisation of a sucrose:sucrose 1-fructosyltransferase gene from perennial ryegrass (*Lolium perenne* L.). *Journal of Plant Physiology* 160, 1385–1391.

Chalmers, J., Lidgett, A., Johnson, X., Jennings, K., Cummings, N., Forster, J. and Spangenberg, G. (2005) Molecular genetics of fructan metabolism in temperate grasses. *Plant Biotechnology Journal* 3, 459–474.

Cheng, D.W., Armstrong, K.C., Tinker, N., Wight, C.P., He, S., Lybaert, A., Fedak, G. and Molnar, S.J. (2002) Genetic and physical mapping of *Lrk10*-like receptor kinase sequences in hexaploid oat (*Avena sativa* L.). *Genome* 45, 100–109.

Cogan, N.O.I., Vecchies, A.C., Yamada, T., Dobrowolski, M.D., Smith, K.F. and Forster, J.W. (2005a) QTL analysis of mineral content in perennial ryegrass (*Lolium perenne* L.). In: Humphreys, M.O. (ed.) *Molecular Breeding for the Genetic Improvement of Forage Crops and Turf*. Wageningen Academic Publishers, Wageningen, The Netherlands, p. 153.

Cogan, N.O.I., Smith, K.F., Yamada, T., Francki, M.G., Vecchies, A.C., Jones, E.S., Spangenberg, G.C. and Forster, J.W. (2005b) QTL analysis and comparative genomics of herbage quality traits in perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics* 110, 364–380.

Cogan, N.O.I., Abberton, M.T., Smith, K.F., Kearney, G., Marshall, A.H., Williams, A., Michaelson-Yeates, T.P.T., Bowen, C., Jones, E.S., Vecchies, A.C. and Forster, J.W. (2006a) Individual and multi-environment combined analyses identify QTLs for morphogenetic and reproductive development traits in white clover (*Trifolium repens* L.). *Theoretical and Applied Genetics* 112, 1401–1415.

Cogan, N.O.I., Ponting, R.C., Vecchies, A.C., Drayton, M.C., George, J., Dobrowolski, M.P., Sawbridge, T.I., Spangenberg, G.C., Smith, K.F. and Forster, J.W. (2006b) Gene-associated single nucleotide polymorphism (SNP) discovery in perennial ryegrass (*Lolium perenne* L.). *Molecular Genetics and Genomics* 276, 101–122.

Cogan, N.O.I., Drayton, M.C., Ponting, R.C., Vecchies, A.C., Bannan, N.R., Sawbridge, T.I., Smith, K.F., Spangenberg, G.C. and Forster, J.W. (2007) Validation of *in silico*-predicted genic single nucleotide polymorphism in white clover (*Trifolium repens* L.). *Molecular Genetics and Genomics* 277, 413–425.

Cook, J.P., Wichman, D.M., Martin, J.M., Bruckner, P.L. and Talbert, L.E. (2004) Identification of microsatellite markers associated with a stem solidness locus in wheat. *Crop Science* 44, 1397–1402.

Cornish, M.A., Hayward, M.D. and Lawrence, M.J. (1979) Self-incompatibility in ryegrass. I. Genetic control in diploid *Lolium perenne* L. *Heredity* 43, 95–106.

Cronn, R.C. and Wendel, J.F. (1998) Simple methods for isolating homoeologous loci from allopolyploid genomes. *Genome* 41, 756–762.

Do Valle, C.B., Glienke, C. and Leguizamon, G.O.C. (1994) Inheritance of apomixis in *Bracharia*, a tropical forage grass. *Apomixis Newsletter* 7, 42–43.

Dobrowolski, M.P. and Forster, J.W. (2007) Linkage disequilibrium-based association mapping in forage species. In: Oraguzie, N.C., Rikkerink, E., Gardiner, S.E. and De Silva, N.H. (eds) *Association Mapping in Plants*. Springer, New York, pp. 197–209.

Doyle, J.J. and Luckow, M.A. (2003) The rest of the iceberg. Legume diversity and evolution in a phylogenetic context. *Plant Physiology* 131, 900–910.

Doyle, P.T., Stockdale, C.R., Lawson, A.R. and Cohen, D.C. (2000) *Pastures for Dairy Production in Victoria*. The State of Victoria, Department of Natural Resources and Environment. Rodney Printers, Tatura, Victoria, Australia.

Draper, J., Mur, L.A.J., Jenkins, G., Ghosh-Biswas, G.C., Bablak, P., Hasterok, R., Routledge, A.P.M. (2001) *Brachypodium distachyon*: a new model system for functional genomics in grasses. *Plant Physiology* 127, 1539–1555.

Dumsday, J.L., Smith, K.F., Forster, J.W. and Jones, E.S. (2003) SSR-based genetic linkage analysis of resistance to crown rust (*Puccinia coronata* Corda f. sp. *lolii*) in perennial ryegrass (*Lolium perenne* L.). *Plant Pathology* 52, 628–637.

Dunbar, S.A. (2006) Applications of luminex xMAPTM technology for rapid, high-throughput multiplexed nucleic acid detection. *Clinica Chimica Acta* 363, 71–82.

Eckard, R. (1998) *A critical review of research on the nitrogen nutrition of dairy pastures in Victoria*. The Institute of Land and Food Resources, University of Melbourne and Agriculture Victoria, Department of Natural Resources and Environment, Ellinbank, Victoria, Australia.

Edwards, D., Forster, J.W., Cogan, N.O.I., Batley, J. and Chagné, D. (2007) Single nucleotide polymorphism discovery in plants. In: Oraguzie, N.C., Rikkerink, E., Gardiner, S.E. and De Silva, N.H. (eds) *Association Mapping in Plants*. Springer, New York, pp. 53–76.

Ellison, N.W., Liston, A., Steiner, J.J., Williams, W.M. and Taylor, N.L. (2006) Molecular phylogenetics of the clover genus (*Trifolium* – Leguminosae). *Molecular Phylogenetics and Evolution* 39, 688–705.

Endo, T.R. and Gill, B.S. (1996) The deletion stocks of common wheat. *Journal of Heredity* 87, 295–307.

Evans, L.T. (1999) Gibberellins and flowering in long day plants, with special reference to *Lolium temulentum*. *Australian Journal of Plant Physiology* 26, 1–8.

Faville, M., Vecchies, A.C., Schreiber, M., Drayton, M.C., Hughes, L.J., Jones, E.S., Guthridge, K.M., Smith, K.F., Sawbridge, T., Spangenberg, G.C., Bryan, G.T. and Forster, J.W. (2004) Functionally-associated molecular genetic marker map construction in perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics* 110, 12–32.

Feuillet, C., Reuzeau, C., Kjellbom, P. and Keller, B. (1998) Molecular characterisation of a new type of receptor-like kinase (wlrk) gene family in wheat. *Plant Molecular Biology* 37, 943–953.

Flint-Garcia, S.A., Thornsberry, J.M. and Buckler, E.S.I. (2003) Structure of linkage disequilibrium in plants. *Annual Review of Plant Biology* 54, 357–374.

Forster, J.W., Jones, E.S., Kölliker, R., Drayton, M.C., Dumsday, J., Dupal, M.P., Guthridge, K.M., Mahoney, N.L., van Zijll de Jong, E. and Smith, K.F. (2001a) Development and implementation of molecular markers for forage crop improvement. In: Spangenberg, G. (ed.) *Molecular Breeding of Forage Crops*. Kluwer Academic Press, Dordrecht, The Netherlands, pp. 101–133.

Forster, J.W., Jones, E.S., Kölliker, R., Drayton, M.C., Dupal, M.P., Guthridge, K.M. and Smith, K.F. (2001b) DNA profiling in outbreeding forage species. In: Henry, R. (ed.) *Plant Genotyping – The DNA Fingerprinting of Plants*. CAB International, Wallingford, UK, pp. 299–320.

Forster, J.W., Jones, E.S., Batley, J. and Smith, K.F. (2004) Molecular marker-based genetic analysis of pasture and turf grasses. In: Hopkins, A., Wang, Z.-Y., Sledge, M. and Barker, R.E. (eds) *Molecular Breeding of Forage and Turf*. Kluwer Academic Press, Dordrecht, The Netherlands, pp. 197–239.

Francki, M.G., Walker, E., Forster, J.W., Spangenberg, G. and Appels, R. (2006) Fructosyltransferase and invertase genes evolved by gene duplication and rearrangements: rice, perennial ryegrass, and wheat gene families. *Genome* 49, 1081–1091.

Frugier, F., Poirier, S., Satiat-Jeunemaître, B., Kondorosi, A. and Crespi, M. (2000) A Krüppel-like zinc finger protein is involved in nitrogen-fixing root nodule organogenesis. *Genes and Development* 14, 475–482.

Fuong, F.T., Voylokov, A.V. and Smirnov, V.G. (1993) Genetic studies of self-fertility in rye (*Secale cereale* L.). 2. The search for molecular marker genes linked to self-incompatibility loci. *Theoretical and Applied Genetics* 87, 619–623.

Gallagher, J.A. and Pollock, C.J. (1998) Isolation and characterisation of a cDNA clone from *Lolium temulentum* L. encoding for a sucrose hydrolytic enzyme which shows alkaline/neutral invertase activity. *Journal of Experimental Botany* 49, 789–795.

Gallagher, J.A., Cairns, A.J. and Pollock, C.J. (2004) Cloning and characterisation of a putative fructosyltransferase and two putative invertase genes from the temperate grass *Lolium temulentum* L. *Journal of Experimental Botany* 55, 557–569.

Gao, S.Q., Xu, H.J., Cheng, X.G., Chen, M., Xu, Z.S., Li, L.C., Ye, X.G., Du, L.P., Hao, X.Y. and Ma, Y.Z. (2005) Improvement of wheat drought and salt tolerance by expression of a stress-inducible transcription factor *Gm*DREB of soybean (*Glycine max*). *Chinese Science Bulletin* 50, 2714–2723.

George, J., Cogan, N.O.I., Smith, K.F., Spangenberg, G.C. and Forster, J.W. (2006a) Genetic map integration and comparative genome organisation of white clover (*Trifolium repens* L.) with model legume species. *Plant and Animal Genome XIV*, San Diego, California, P542.

George, J., Dobrowolski, M.P., van Zijll de Jong, E., Cogan, N.O.I., Smith, K.F. and Forster, J.W. (2006b) Assessment of genetic diversity in cultivars of white clover (*Trifolium repens* L.) detected by simple sequence repeat polymorphism. *Genome* 49, 919–930.

Gupta, P.K., Rustgi, S. and Kulwal, P.L. (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. *Plant Molecular Biology* 57, 461–485.

Guthridge, K.M., Dupal, M.D., Kölliker, R., Jones, E.S., Smith, K.F. and Forster, J.W. (2001) AFLP analysis of genetic diversity within and between populations of perennial ryegrass (*Lolium perenne* L.). *Euphytica* 122, 191–201.

Hammond, S.M., Caudy, A.A. and Hannon, G.J. (2001) Post-transcriptional gene silencing by double-stranded DNA. *Nature Reviews Genetics* 2, 1110–1119.

Heath, R., Huxley, H., Stone, B. and Spangenberg, G. (1998) cDNA cloning and differential expression of three caffeic acid *O*-methyltransferase homologues from perennial ryegrass (*Lolium perenne* L.). *Journal of Plant Physiology* 153, 649–657.

Heath, R., McInnes, R., Lidgett, A., Huxley, H., Lynch, D., Jones, E.S., Mahoney, N.L. and Spangenberg, G.C. (2002) Isolation and characterisation of three 4-coumarate:CoA-ligase homologue cDNAs from perennial ryegrass (*Lolium perenne* L.). *Journal of Plant Physiology* 159, 773–779.

Holzberg, S., Brosio, P., Gross, C. and Pogue, G.P. (2002) Barley stripe mosaic virus-induced silencing in a monocot plant. *Plant Journal* 30, 315–327.

Humphreys, M.W. (1989) The controlled introgression of *Festuca arundinacea* genes into *Lolium multiflorum*. *Euphytica* 42, 105–116.

Hutchinson, J., Rees, H. and Seal, A.G. (1979) An assay of the activity of supplementary DNA in *Lolium*. *Heredity* 43, 411–421.

Jansen, G., Hazendonk, E., Thijssen, K.L. and Plasterk, R.H.A. (1997) Reverse genetics by chemical mutagenesis in *Caenorhabditis elegans*. *Nature Genetics* 17, 119–121.

Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends in Genetics* 17, 388–391.

Jeanneau, M., Gerentes, D., Foueillassar, X., Zivy, M., Vidal, J., Toppan, A. and Perez, P. (2002) Improvement of drought tolerance in maize: towards the functional validation of the *Zm-Asr1* gene and increase of water use efficiency by over-expressing C4-PEPC. *Biochimie* 84, 1127–1135.

Jenkins, G., Head, J. and Forster, J.W. (2000) Probing meiosis in hybrids of *Lolium* (Poaceae) with a discriminatory repetitive genomic sequence. *Chromosoma* 109, 280–286.

Jensen, L.B., Muylle, H., Arens, P., Andersen, C.H., Holm, P.B., Ghesquiere, M., Julier, B., Lübberstedt, T., Nielsen, K.K., De Riek, J., Roldán-Ruiz, I., Roulund, N., Taylor, C., Vosman, B. and Barre, P. (2005a) Development and mapping of a public reference set of SSR markers in *Lolium perenne* L. *Molecular Ecology Notes* 5, 951–957.

Jensen, L.B., Andersen, J.R., Frei, U., Xing, Y., Taylor, C., Holm, P.B. and Lübberstedt, T. (2005b) QTL mapping of vernalisation response in perennial ryegrass (*Lolium perenne* L.) reveals co-location with an orthologue of wheat *VRN1*. *Theoretical and Applied Genetics* 110, 527–536.

Johnson, X., Lidgett, A., Chalmers, J., Guthridge, K., Jones, E. and Spangenberg, G.C. (2003) Isolation and characterisation of an invertase gene from perennial ryegrass (*Lolium perenne* L.). *Journal of Plant Physiology* 160, 903–911.

Jones, E.S., Dupal, M.P., Kölliker, R., Drayton, M.C. and Forster, J.W. (2001) Development and characterisation of simple sequence repeat (SSR) markers for perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics* 102, 405–415.

Jones, E.S., Mahoney, N.L., Hayward, M.D., Armstead, I.P., Jones, J.G., Humphreys, M.O., King, I.P., Kishida, T., Yamada, T., Balfourier, F., Charmet, C. and Forster, J.W. (2002a) An enhanced molecular marker-based map of perennial ryegrass (*Lolium perenne* L.) reveals comparative relationships with other Poaceae species. *Genome* 45, 282–295.

Jones, E.S., Dupal, M.D., Dumsday, J.L., Hughes, L.J. and Forster, J.W. (2002b) An SSR-based genetic linkage map for perennial ryegrass (*Lolium perenne* L.). *Theoretical and Applied Genetics* 105, 577–584.

Jones, E.S., Hughes, L.J., Drayton, M.C., Abberton, M.T., Michaelson-Yeates, T.P.T., Bowen, C. and Forster, J.W. (2003) An SSR and AFLP molecular marker-based genetic map of white clover (*Trifolium repens* L.). *Plant Science* 165, 447–479.

Kadarmideen, H.N., von Rohr, P. and Janss, L.L.G. (2006) From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian Genome* 17, 548–564.

Kathiresan, A., Lafitte, H.R., Chen, J.X., Mansueto, L., Bruskiewich, R. and Bennett, J. (2006) Gene expression microarrays and their application in drought stress research. *Field Crops Research* 97, 101–110.

Kölliker, R., Jones, E.S., Drayton, M.C., Dupal, M.P. and Forster, J.W. (2001) Development and characterisation of simple sequence repeat (SSR) markers for white clover (*Trifolium repens* L.). *Theoretical and Applied Genetics* 102, 416–424.

Korzun, V., Malyshev, S., Voylokov, A.V. and Börner, A. (2001) A genetic map of rye (*Secale cereale* L.) combining RFLP, isozyme, protein, microsatellite and gene loci. *Theoretical and Applied Genetics* 102, 709–717.

Kota, R., Rudd, S., Facius, A., Kolesov, G., Thiel, T., Zhang, H., Stein, N., Mayer, K. and Gräner, A. (2003) Snipping polymorphisms from large EST collections in barley (*Hordeum vulgare* L.). *Molecular Genetics and Genomics* 270, 24–33.

Kubik, C., Meyer, W.A. and Gaut, B.S. (1999) Assessing the abundance and polymorphism of simple sequence repeats in perennial ryegrass. *Crop Science* 39, 1136–1141.

Labate, J.A. and Baldo, A.M. (2005) Tomato SNP discovery by EST mining and resequencing. *Molecular Breeding* 16, 343–349.

Lane, L.A., Ayres, J.F. and Lovett, J.V. (2000) The pastoral significance, adaptive characteristics and grazing value of white clover (*Trifolium repens* L.) in dryland conditions in Australia: a review. *Australian Journal of Experimental Agriculture* 40, 1033–1046.

Lauvergeat, V., Barre, P., Bonnet, M. and Ghesquiére, M. (2005) Sixty simple sequence repeat markers for use in the *Festuca-Lolium* complex of grasses. *Molecular Ecology Notes* 5, 401–405.

Lawless, K., Cogan, N.O.I., Drayton, M.C., George, J., Bannan, N.R., Wilkinson, T.C., Smith, K.F., Spangenberg, G.C. and Forster, J.W. (2006) *In vitro* discovery and characterisation of gene-associated SNPs for genetic improvement of white clover (*Trifolium repens* L.). *Proceedings of the Third International Legume Genetics and Genomics Conference*, Brisbane, Australia, April 2006, p. 67.

Li, X., Song, Y., Century, K., Straight, S., Ronald, P., Dong, X., Lassner, M. and Zhang, Y. (2001) A fast neutron deletion mutagenesis-based reverse genetics systems for plants. *Plant Journal* 27, 235–242.

Lidgett, A., Jennings, K., Johnson, X., Guthridge, K., Jones, E. and Spangenberg, G. (2002) Isolation and characterisation of fructosyltransferase gene from perennial ryegrass (*Lolium perenne*). *Journal of Plant Physiology* 159, 415–422.

Lu, C.M., Yang, W.Y., Zhang, W.J. and Lu, B.-R. (2005) Identification of SNPs and development of allelic specific PCR markers for high molecular weight glutenin subunit *Dtx1.5* from *Aegilops tauschii* through sequence characterisation. *Journal of Cereal Science* 41, 13–18.

Lübberstedt, T., Zein, I., Andersen, J., Wenzel, G., Krutzfeldt, B., Eder, J., Ouzunova, M. and Chun, S. (2005) Development and application of functional markers in maize. *Euphytica* 146, 101–108.

Lynch, D., Lidgett, A., McInnes, R., Huxley, H., Jones, E., Mahoney, N. and Spangenberg, G. (2002) Isolation and characterisation of three cinnamyl alcohol dehydrogenase homologue cDNAs from perennial ryegrass (*Lolium perenne* L.). *Journal of Plant Physiology* 159, 653–660.

Mackay, T.F.C. (2001) The genetic architecture of quantitative traits. *Annual Review of Genetics* 35, 303–309.

Mager, J., Schultz, R.M., Brunk, B.P. and Bartolomei, M.S. (2006) Identification of candidate maternal-effect genes through comparison of multiple microarray data sets. *Mammalian Genome* 17, 941–949.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L.I., Jarvie, T.P., Jirage, K.B., Kim, J.-B., Knight, J.R., Lanza, J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Ronan, M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F. and Rothberg, J.M. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380.

Martinez, E.J., Urbani, M.H., Quarin, C.L. and Ortiz, J.P.A. (2001) Inheritance of apospory in bahiagrass. *Paspalum notatum*. *Hereditas* 135, 19–25.

McCallum, C.M., Comai, L., Greene, E.A. and Henikoff, S. (2000) Targeting induced local lesions in genomes (TILLING) for plant functional genomics. *Plant Physiology* 123, 439–442.

McInnes, R., Lidgett, A., Lynch, D., Huxley, H., Jones, E., Mahoney, N. and Spangenberg, G. (2002) Isolation and characterisation of a cinnamoyl-CoA reductase gene from perennial ryegrass (*Lolium perenne*). *Journal of Plant Physiology* 159, 415–422.

Merchan, F., Breda, C., Perez Hormaeche, J., Sousa, C., Kondorosi, A., Aguilar, O.M., Megias, M. and Crespi, M. (2003) A *Krüppel*-like transcription factor gene is involved in salt stress responses in *Medicago* spp. *Plant and Soil* 257, 1–9.

Michaelson-Yeates, T.P.T., Marshall, A., Abberton, M.T. and Rhodes, I. (1997) Self-incompatibility and heterosis in white clover (*Trifolium repens* L.). *Euphytica* 94, 341–348.

Munro, J.M.M., Davies, D.A., Evans, W.B. and Scurlock, R.V. (1992) Animal production evaluation of herbage grass varieties. 1. Comparison of Aurora with Frances, Talbot and Melle perennial ryegrass grown alone and with clover. *Grass and Forage Science* 47, 259–273.

Muylle, H., Baert, J., Van Bockstaele, E., Moerkerke, B., Goetghebeur, E. and Roldán-Ruiz, I. (2005a) Identification of molecular markers linked with crown rust (*Puccinia coronata* f.sp. *lolii*) resistance in perennial ryegrass (*Lolium perenne*) using AFLP markers and a bulked segregant approach. *Euphytica* 143, 135–144.

Muylle, H., Baert, J., Van Bockstaele, E., Petijs, J. and Roldán-Ruiz, I. (2005b) Four QTLs determine crown rust (*Puccinia coronata* f.sp. *lolii*) resistance in a perennial ryegrass (*Lolium perenne*) population. *Heredity* 95, 348–357.

Neale, D.B. and Savolainen, O. (2004) Association genetics of complex traits in conifers. *Trends in Plant Science* 9, 325–330.

Nogler, G.A. (1984) Gametophytic apomixes. In: Johri, B.M. (ed.) *Embryology of Angiosperms*. Springer, Berlin, Germany.

Ogihara, Y., Hasegawa, K. and Tsuijimoto, H. (1994) High-resolution cytological mapping of the long arm of chromosome 5A in common wheat using a series of deletion lines induced by

gametocidal (*Gc*) genes of *Aegilops speltoides*. *Molecular Genetics and Genomics* 244, 253–259.

Oleykowski, C.A., Bronson Mullins, C.R., Godwin, A.K. and Yeung, A.T. (1998) Mutation detection using a novel plant endonuclease. *Nucleic Acids Research* 26, 4597–4602.

Oxtoby, E., Dunn, M.A., Pancoro, A. and Hughes, M.A. (2004) Nucleotide and derived amino acid sequence of the cyanogenic β-glucosidase (linamarase) from white clover (*Trifolium repens* L.). *Plant Molecular Biology* 17, 209–219.

Ozias-Akins, P., Roche, D. and Hanna, W.W. (1998) Tight clustering and hemizygosity of apomixis-linked molecular markers in *Pennisetum squamulatum* implies genetic control of apospory by a divergent locus that may have no allelic form in sexual genotypes. *Proceedings of the National Academy of Sciences of the USA* 95, 5127–5132.

Paran, I. and Zamir, D. (2003) Quantitative traits in plants: beyond the QTL. *Trends in Genetics* 19, 303–306.

Pardo, J.M., Cubero, B., Leidi, E.O. and Quintero, F.J. (2006) Alkali cation exchangers: roles in cellular homoeostatis and stress tolerance. *Journal of Experimental Botany* 57, 1181–1199.

Pessino, S.C., Ortiz, J.P.A., Leblanc, O., do Valle, C.B., Evans, C. and Hayward, M.D. (1997) Identification of a maize linkage group related to apomixis in *Bracharia*. *Theoretical and Applied Genetics* 94, 439–444.

Pessino, S.C., Ortiz, J.P.A., Hayward, M.D. and Quarin, C.L. (1999) The molecular genetics of gametophytic apomixis. *Hereditas* 130, 1–11.

Pflieger, S., Lefebvre, V. and Causse, M. (2001) The candidate gene approach in plant genetics: a review. *Molecular Breeding* 7, 275–281.

Picoult-Newberg, L., Ideker, T.E., Pohl, M.G., Taylor, S.L., Donaldson, M.A., Nickerson, D.A. and Boyce-Jacino, M. (1999) Mining SNPs from EST databases. *Genome Research* 9, 167–176.

Ponting, R.C., Drayton, M.D., Cogan, N.O.I., Dobrowolski, M.D., Spangenberg, G.C., Smith, K.F. and Forster, J.W. (2005) SNP discovery and haplotypic variation in full-length herbage quality genes of perennial ryegrass (*Lolium perenne* L). In: Humphreys, M.O. (ed.) *Molecular Breeding for the Genetic Improvement of Forage Crops and Turf*. Wageningen Academic Publishers, Wageningen, The Netherlands, p. 196.

Posselt, U.K. (1993) Hybrid production in *Lolium perenne* based on self-incompatibility. *Euphytica* 71, 29–33.

Powell, W. and Langridge, P. (2004) Unfashionable crop species flourish in the 21st century. *Genome Biology* 5, 233.

Price, A.H. (2006) Believe it or not, QTLs are accurate! *Trends in Plant Science* 11, 213–216.

Prioul, J.L., Pelleschi, S., Séne, M., Thévenot, C., Causse, M., de Vienne, D. and Leonard, A. (1999) From QTLs for enzyme activity to candidate genes in maize. *Journal of Experimental Botany* 50, 1281–1288.

Qi, L., Echalier, B., Friebe, B. and Gill, B.S. (2003) Molecular characterisation of a set of wheat deletion stocks for use in chromosome bin mapping of ESTs. *Functional and Integrative Genomics* 3, 39–55.

Qi, Y.J. and Hannon, G.J. (2005) Uncovering RNAi mechanisms in plants: biochemistry enters the foray. *FEBS Letters* 579, 5899–5903.

Rafalski, A. (2002) Applications of single nucleotide polymorphisms in crop genetics. *Current Opinion in Plant Biology* 5, 94–100.

Rodriguez-Navarro, A. and Rubio, F. (2006) High-affinity potassium and sodium transport systems in plants. *Journal of Experimental Botany* 57, 1149–1160.

Rogers, G.L., Porter, R.H.D. and Robinson, I. (1982) Comparison of perennial ryegrass and white clover for milk production. In: MacMillan, K.K. and Tuafa, V.K. (eds) *Proceedings of the Conference on Dairy Production from Pasture*. Clark and Matheson, Hamilton, New Zealand, pp. 213–214.

Rogers, M.E., Noble, C.L., Nicolas, M.E. and Halloran, G.M. (1993) Variation in yield potential and salt tolerance of selected cultivars and natural populations of *Trifolium repens* L. *Australian Journal of Agricultural Research* 44, 785–798.

Rogers, M.E., Noble, C.L., Halloran, G.M. and Nicolas, M.E. (1997) Selecting for salt tolerance in white clover (*Trifolium repens*): chloride ion exclusion and its heritability. *New Phytologist* 135, 645–654.

Rostocks, N., Mudie, S., Cardle, L., Russell, J., Ramsay, L., Booth, A., Svensson, J.T., Wanamaker, S.I., Walia, H., Rodriguez, E.M., Hedley, P.E., Liu, H., Morris, J., Close, T.J., Marshall, D.F. and Waugh, R. (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Molecular Genetics and Genomics* 274, 515–527.

Rus, A., Lee, B.H., Munoz-Mayor, A., Sharkhuu, A., Miura, K., Zhu, J.K., Bressan, R.A. and Hasegawa, P.M. (2004) *At*HKT1 facilitates Na + homoeostasis and K + nutrition *in planta*. *Plant Physiology* 136, 2500–2511.

Sackville-Hamilton, N.R., Skøt, L., Chorlton, K.H., Thomas, I.D. and Mizen, S. (2002) Molecular genecology of temperature response in *Lolium perenne*: 1. Preliminary analysis to reduce false positives. *Molecular Ecology* 11, 1855–1863.

Savidan, Y. (1981) Genetics and utilisation of apomixis for the improvement of guineagrass (*Panicum maximum* Jacq.) *Proceedings of the XIVth International Grassland Congress*, Lexington, Kentucky, pp. 182–184.

Sawbridge, T., Ong, E.-K., Binnion, C., Emmerling, M., Meath, K., Nunan, K., O'Neill, M., O'Toole, F., Simmonds, J., Wearne, K., Winkworth, A. and Spangenberg, G. (2003a) Generation and analysis of expressed sequence tags in white clover (*Trifolium repens* L.). *Plant Science* 165, 1077–1087.

Sawbridge, T., Ong, E.-K., Binnion, C., Emmerling, M., McInnes, R., Meath, K., Nguyen, N., Nunan, K., O'Neill, M., O'Toole, F., Rhodes, C., Simmonds, J., Tian, P., Wearne, K., Webster, T., Winkworth, A. and Spangenberg, G. (2003b) Generation and analysis of expressed sequence tags in perennial ryegrass (*Lolium perenne* L.). *Plant Science* 165, 1089–1100.

Seal, A.G. and Rees, H. (1982) The distribution of quantitative DNA changes associated with the evolution of the diploid Festuceae. *Heredity* 94, 179–190.

Schmid, K.J., Sörensen, T.R., Stracke, R., Törjerk, O., Altmann, T., Mitchell-Olds, T. and Weisshaar, B. (2003) Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Research* 13, 1250–1257.

Shen, R., Fan, J.B., Campbell, D., Chang, W.H., Chen, J., Doucet, D., Yeakley, J., Bibikova, M., Garcia, E.W., McBride, C., Steemers, F., Garcia, F., Kermani, B.G., Gunderson, K. and Oliphant, A. (2005) High-throughput SNP genotyping on universal bead arrays. *Mutation Research – Fundamental and Molecular Mechanisms of Mutagenesis* 573, 70–82.

Shi, H., Ishitani, M., Kim, C. and Zhu, J.-K. (2000) The *Arabidopsis thaliana* salt tolerance gene *SOS1* encodes a putative Na^+/H^+ antiporter. *Proceedings of the National Academy of Sciences of the USA* 97, 6896–6901.

Shi, H.Z., Kim, Y., Guo, Y., Stevenson, B. and Zhu, J.K. (2003) The *Arabidopsis SOS5* locus encodes a putative cell surface adhesion protein and is required for normal cell expansion. *Plant Cell* 15, 19–32.

Shinozuka, H., Hisano, H., Ponting, R.C., Jones, E.S., Cogan, N.O.I., Forster, J.W. and Yamada, T. (2005) Molecular cloning and genetic mapping of perennial ryegrass protein kinase CK2α-subunit genes. *Theoretical and Applied Genetics* 112, 167–177.

Sim, S., Chang, T., Curley, J., Warnke, S.E., Barker, R. and Jung, G. (2005) Chromosomal rearrangements differentiating the ryegrass genome from the Triticeae, oat and rice genomes using common heterologous RFLP probes. *Theoretical and Applied Genetics* 110, 1011–1019.

Skøt, L., Sackville-Hamilton, N.R., Mizen, S., Chorlton, K.H. and Thomas, I.D. (2002) Molecular genecology of temperature response in *Lolium perenne*: 2. Association of AFLP markers with ecogeography. *Molecular Ecology* 11, 1865–1876.

Skøt, L., Humphreys, J., Armstead, I.P., Humphreys, M.O., Gallagher, J.A. and Thomas, I.D. (2005a) Approaches for associating molecular polymorphisms with phenotypic traits based on linkage disequilibrium in natural populations of *Lolium perenne*. In: Humphreys, M.O. (ed.) *Molecular Breeding for the Genetic Improvement of Forage Crops and Turf*. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 157.

Skøt, L., Humphreys, M.O., Armstead, I., Heywood, S., Skøt, K.P., Sanderson, R., Thomas, I.D., Chorlton, K.H. and Sackville-Hamilton, N.R. (2005b) An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L.). *Molecular Breeding* 15, 233–245.

Somers, D.J., Kirkpatrick, R., Moniwa, M. and Walsh, A. (2003) Mining single-nucleotide polymorphisms from hexaploid wheat ESTs. *Genome* 49, 431–437.

Soreng, R.J. and Davis, J.I. (1998) Phylogenetics and character evolution in the grass family (Poaceae): simultaneous analysis of morphological and chloroplast DNA restriction site character sets. *Botanical Reviews* 64, 1–85.

Sorrells, M.E. and Wilson, W.A. (1997) Direct classification and selection of superior alleles for crop improvement. *Crop Science* 37, 691–697.

Spangenberg, G., Vallés, M.P., Wang, Z.-Y., Montavon, P., Nagel, J. and Potrykus, I. (1994) Asymmetric somatic hybridisation between tall fescue (*Festuca arundinacea* Schreb.) and irradiated Italian ryegrass (*Lolium multiflorum* Lam.) protoplasts. *Theoretical and Applied Genetics* 88, 509–519.

Spangenberg, G., Forster, J.W., Edwards, D., John, U., Mouradov, A., Emmerling, M., Batley, J., Felitti, S., Cogan, N.O.I., Smith, K.F. and Dobrowolski, M.P. (2005) Future directions in the molecular breeding of forage and turf. In: Humphreys, M.O. (ed.) *Molecular Breeding for the Genetic Improvement of Forage Crops and Turf*. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 83–97.

Starker, C.G., Parra-Colmenares, A.L., Smith, L., Mitra, R.K. and Long, S.R. (2006) Nitrogen fixation mutants of *Medicago truncatula* fail to support plant and bacterial symbiotic gene expression. *Plant Physiology* 140, 671–680.

Teulat, B., Borries, C. and This, D. (2001) New QTLs identified for plant water status, water soluble carbohydrate and osmotic adjustment in a barley population grown in a growth-chamber under two water regimes. *Theoretical and Applied Genetics* 103, 161–170.

Thornsberry, J.M., Goodman, M.M., Doebley, J., Kresovich, S., Nielsen, D. and Buckler, E.S. (2001) *Dwarf8* polymorphisms associate with variation in flowering time. *Nature Genetics* 28, 286–289.

Thorogood, D., Kaiser, W.J., Jones, J.G. and Armstead, I. (2002) Self-incompatibility in ryegrass 12: genotyping and mapping the *S* and *Z* loci of *Lolium perenne* L. *Heredity* 88, 385–390.

Till, B.J., Reynolds, S.H., Greene, E.A., Codomo, C.A., Enns, L.G., Johnson, J.E., Burtner, C., Odden, A.R., Young, K., Taylor, N.E., Henikoff, J.G., Comai, L. and Henikoff, S. (2003) Large-scale discovery of induced point mutations with high-throughput TILLING. *Genome Research* 13, 524–530.

Tost, J. and Gut, I.G. (2005) Genotyping single nucleotide polymorphisms by MALDI mass spectrometry in clinical applications. *Clinical Biochemistry* 38, 335–350.

Turner, L.B., Cairns, A.J., Armstead, I.P., Ashton, J., Skøt, K., Whittaker, D. and Humphreys, M.O. (2006) Dissecting the regulation of fructan metabolism in perennial ryegrass (*Lolium perenne*) with quantitative trait locus mapping. *New Phytologist* 169, 45–58.

Vaidyanathan, R., Kuruvilla, S. and Thomas, G. (1999) Characterisation and expression pattern of an abscisic acid and osmotic stress responsive gene from rice. *Plant Science* 140, 25–36.

Vance, V. and Vaucheret, H. (2001) RNA silencing in plants – defence and counterdefence. *Science* 292, 2277–2280.

Vijn, I. and Smeekens, S. (1999) Fructan: more than a reserved carbohydrate? *Plant Physiology* 120, 351–359.

Vogel, J.P., Qiang Gu, Y., Twigg, P., Lazo, G.R., Laudencia-Chingcuanco, D., Hayden, D.M., Donse, T.J., Vivian, L.A., Stamova, B. and Coleman-Derr, D. (2006) EST sequencing and phylogenetic analysis of the model grass *Brachypodium distachyon*. *Theoretical and Applied Genetics* 113, 186–195.

Vogel, K.P. and Pedersen, J.F. (1993) Breeding systems for cross-pollinated forage grasses. *Plant Breeding Reviews* 11, 251–274.

Voylokov, A.V., Korzun, V. and Börner, A. (1998) Mapping of three self-fertility mutations in rye (*Secale cereale* L.) using RFLP, isozyme and morphological markers. *Theoretical and Applied Genetics* 97, 147–153.

Warnke, S.E., Barker, R.E., Jung, G., Rouf Mian, M.A., Saha, M.C., Brilman, L.A., Dupal, M.D. and Forster, J.W. (2004) Genetic linkage mapping of an annual x perennial ryegrass population. *Theoretical and Applied Genetics* 109, 294–304.

Waugh, R., Leader, D.J., McCallum, N. and Caldwell, D. (2006) Harvesting the potential of induced biological diversity. *Trends in Genetics* 11, 71–79.

Wilkins, P.W. and Thorogood, D. (1992) Breakdown of self-incompatibility in perennial ryegrass at high temperature and its uses in breeding. *Euphytica* 64, 65–69.

Wilkins, P.W. and Humphreys, M.O. (2003) Progress in breeding perennial forage grasses for temperate agriculture. *Journal of Agricultural Science* 140, 129–150.

Yamada, T. and Forster, J.W. (2005) QTL analysis and trait dissection in ryegrasses (*Lolium* spp.). In: Humphreys, M.O. (ed.) *Molecular Breeding for the Genetic Improvement of Forage Crops and Turf*. Wageningen Academic Publishers, Wageningen, The Netherlands, pp. 43–53.

Yamada, T., Higuchi, A. and Fukuoka, A. (1989) Recurrent selection of white clover (*Trifolium repens* L.) using self-compatible plants. I. Selection of self-compatible plants and inheritance of a self-compatibility factor. *Euphytica* 44, 167–172.

Yamada, T., Jones, E.S., Cogan, N.O.I., Vecchies, A.C., Nomura, T., Hisano, H., Shimamoto, Y., Smith, K.F. and Forster, J.W. (2004) QTL analysis of morphological, developmental and winter hardiness-associated traits in perennial ryegrass (*Lolium perenne* L.). *Crop Science* 44, 925–935.

Yamada, T., Forster, J.W., Humphreys, M.W. and Takamizo, T. (2005) Genetics and molecular breeding in the *Lolium*/*Festuca* pasture grass species complex. *Grassland Science* 51, 89–106.

Yang, W.C., Bai, X.-D., Kabelka, E., Eaton, C., Kamoun, S., van der Knapp, E. and Francis, D. (2004) Discovery of single nucleotide polymorphisms in *Lycopersicon esculentum* by computer aided analysis of expressed sequence tags. *Molecular Breeding* 14, 21–34.

Young, N.D., Cannon, S.B., Sato, S., Kim, D., Cook, D.R., Town, C.D., Roe, B.A. and Tabata, S. (2005) Sequencing the genespaces of *Medicago truncatula* and *Lotus japonicus*. *Plant Physiology* 137, 1174–1181.

Zhang, D.-X. and Hewitt, G.M. (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology* 12, 563–584.

Zhang, Y., Sledge, M.K. and Bouton, J.H. (2007) Genome mapping of white clover (*Trifolium repens* L.) and comparative analysis within the Trifolieae using cross-species SSR markers. *Theoretical and Applied Genetics* 144, 1367–1378.

Zhu, H., Choi, H.-K., Cook, D.R. and Shoemaker, R.C. (2005) Bridging model and crop legumes through comparative genomics. *Plant Physiology* 137, 1189–1196.

11 Genotyping for Rice Eating Qualities

L.M.T. Bradbury, D.L.E. Waters and R.J. Henry

Introduction

The quality of plant products is a key target for plant breeders supplying competitive markets. Rice is an important species because it is both a model crop (with a complete genome sequence) and a species of great social and economic importance. Consumers have a preference for rice with desirable eating qualities. Rice with the largest number of individual desirable eating quality traits is in highest demand, increasing the wholesale price. Farmers respond to this demand by growing rice varieties displaying the greatest number of attractive qualities which in turn enables them to receive a premium price. Rice quality can be more important to farmers than yield alone, as is exemplified by low-yielding basmati rice which returns a profit of roughly three times that of non-fragrant non-basmati rice types. The higher price more than offsets the low yield.

The preferred rice eating quality traits will vary depending upon the individual consumer and the grain end use. Rice is prepared and consumed in a number of ways ranging from simple boiling or frying to risotto, rice cakes, sake and so on. Breeding for certain quality traits can be difficult and time-consuming in part because phenotypic selection is possible only when the grain is fully mature. Phenotyping often requires dehulling, milling, grinding, boiling or digestion, resulting in destruction and loss of useful seed. Phenotyping often also requires the use of large tissue samples and the use of various specialized pieces of expensive equipment. In addition, environmental variation is an ever-present complicating factor in the measurement of rice grain quality.

If the genetic basis of a trait is sufficiently well understood, genotyping is a simple, inexpensive and rapid way for breeders to introgress desirable grain characteristics into new or existing varieties because the confusing effects of the environment are removed and breeding programmes can minimize the

 Plant Genotyping II: SNP Technology (ed. R.J. Henry)

degree of replication within the programme, increasing the effective size of available resources. Genotyping can significantly reduce analysis time, cost and required sample size by offering a single platform test for all of the traits being analysed in contrast to phenotyping which can require specialized equipment for each trait. At its most simple, genotyping requires only PCR and agarose gel technology.

Amylose Content

Starch is composed of a mixture of amylose and amylopectin. The amylose content of rice grain determines the hardness, gloss and level of water absorption of cooked rice (Singh *et al*., 2000; Olsen *et al*., 2006). Rice with no amylose (around 0–2%) (Zhou *et al*., 2002; Olsen *et al*., 2006), known as waxy or sticky rice, becomes glossy, tender and sticky when cooked (Ayres *et al*., 1997; Olsen *et al*., 2006). These qualities are typical of some temperate japonica rice varieties (Tan *et al*., 1999; Hohn and Puchta, 2003; Olsen *et al*., 2006) and are preferred for many types of rice dishes, especially desserts (Olsen *et al*., 2006). Rice with intermediate amylose (10–20%) also becomes tender and cohesive when cooked, though not to the same extent as waxy rice (Ayres *et al*., 1997; Olsen *et al*., 2006). These qualities are typical of non-waxy temperate japonica rice varieties (Tan *et al*., 1999; Hohn and Puchta, 2003; Olsen *et al*., 2006) and are favoured for many types of rice dishes. Rice with high amylose (20–30%) cooks drier and fluffier and remains separate (Ayres *et al*., 1997; Olsen *et al*., 2006). These qualities are typical of indica and tropical japonica rice varieties (Tan *et al*., 1999; Hohn and Puchta, 2003; Olsen *et al*., 2006).

The traditional method for determining amylose content in rice grain requires milling, grinding, sieving, gelatinizing in a heated NaOH solution overnight, boiling, cooling, extraction of lipids before titration with an iodine solution in a colorimetric assay (Tan *et al*., 1999; Zhou *et al*., 2002; Itoh *et al*., 2003). This lengthy process requires calibration against a pure amylose standard and is difficult for routine use in large breeding programmes. Iodine can also react with amylopectin, although to a lower extent than with amylose, and so the values obtained using this method are often referred to as 'apparent amylose content' (Zhou *et al*., 2002). Alternative methods have been developed that utilize tagged antibodies raised against wheat granule-bound starch synthase 1 (GBSS1), an enzyme involved in production of amylose grain endosperm (Gale *et al*., 2004), but these are utilized in scientific studies rather than in breeding programmes as costs are prohibitive.

Molecular analysis has revealed that the *waxy* (*Wx*) gene of plants encodes the enzyme GBSS1 and levels of this enzyme are correlated with differing levels of amylose in the grain (Okagaki and Wessler, 1988; Wang *et al*., 1995; Bligh *et al*., 1998; Cai *et al*., 1998). The level of GBSS1 in rice grain is dependent on the levels of mature *waxy* mRNA (Wang *et al*., 1995; Bligh *et al*., 1998; Cai *et al*., 1998). Although there are multiple factors affecting amylose content in rice grain, one SNP (G to T) situated at the 5' splice site of intron 1 of the *waxy* gene has been found to be the major locus responsible for high or low

amylose content (Wang *et al*., 1995; Bligh *et al*., 1998; Cai *et al*., 1998; Isshiki *et al*., 1998). This SNP alters the splicing of the *waxy* pre-mRNA which in turn influences the level of mature *waxy* mRNA and GBSS1 protein and therefore the level of amylose in the grain. This SNP has been assayed by RFLP and AFLP (Sano *et al*., 1986; Bao *et al*., 2006) but most genotyping efforts have focused on a closely linked microsatellite located 55 bp upstream of the SNP site (Ayres *et al*., 1997; Bligh *et al*., 1998).

Eight different alleles of the *waxy* gene were identified based on this microsatellite by Ayres *et al*. (1997) who argued that this could explain more of the variation in amylose content, such as differences between no and intermediate amylose levels, than the SNP alone. Assaying this dinucleotide microsatellite (CT) involves the use of flanking primers in a PCR and polyacrylamide gel electrophoresis to differentiate between the PCR product sizes. Microsatellites $(CT)_{17}$ and $(CT)_{18}$ associate with the SNP allele 'T' and low ~15% amylose; $(CT)_{19}$ corresponds to SNP allele 'G' and ~17% amylose; $(CT)_{20}$, $(CT)_{8}$ and $(CT)_{14}$ correspond to SNP allele 'G' and medium (~21%) amylose and $(CT)_{11}$ corresponds to SNP allele 'G' and high amylose. Although the CT microsatellite has been the marker of choice for measuring rice amylose content it only explains 82.9% of the variation in apparent amylose content in non-waxy varieties and still requires genotyping of the G to T SNP (Ayres *et al*., 1997). Further sequence analysis and investigation of the causative 'cryptic splicing' of intron 1 from the *waxy* pre-mRNA (Cai *et al*., 1998) identified other causative mutations that can be used as perfect markers for genotyping amylose content. More recent work has developed an effective SNP assay that detects the G–T polymorphism in the granule-bound starch synthase (Bormans *et al*., 2002).

Gelatinization Temperature

Gelatinization temperature (GT) is another starch quality trait that affects rice cooking and eating quality. As its name suggests, it is a measure of the temperature at which starch gelatinizes and, as such, directly influences cooking time and eating quality. This trait is influenced by the other major starch component, amylopectin. While amylose consists of mainly linear polymers of α(1–4)-linked glucose molecules, amylopectin is composed of α(1–4)-linked glucose molecules with many α(1–6)-linked branches of glucose chains of varying length (Zhou *et al*., 2002).

Methods for determining rice starch GT usually require preparative milling and grinding to flour, followed by mixing with water and heating, before determining the GT using specialized equipment, such as a differential scanning calorimeter which detects when a phase transition occurs in the rice flour paste (Sievert and Holm, 1993; Waters *et al*., 2006). Umemoto *et al*. (2002) mapped a gene which controls rice starch GT and amylopectin chain length distribution and found it co-segregated with the soluble starch synthase IIa (SSIIa) encoding gene. Two exon 8 polymorphisms in this gene, G to A and GC to TT mutations (Umemoto and Aoki, 2005; Waters *et al*., 2006), together

explain variation in GT and SSIIa activity level across a range of rice varieties. Two separate allele-specific PCRs which genotype both of these alleles can accurately predict the GT class of individuals or rice varieties. See Chapter 6 (this volume) for more details.

Fragrance

Consumer demand for fragrant rice has increased significantly over recent years. Traditionally grown mainly for the Indian and Thai markets, Western consumers are now willing to pay a premium for fragrant rice. The recessive nature of the trait combined with the subtlety of rice aroma makes it difficult to breed for. Because of this rice breeders have an interest in developing simple and inexpensive methods to distinguish between fragrant and non-fragrant rice plants.

Some of the simpler methods involve heating the leaf or grain in water and smelling the vapours released. These methods are labour-intensive and require subjective analysis by a panel of experts. Unfortunately, an individual analyst's senses can quickly become saturated, rapidly reducing their ability to distinguish between fragrant and non-fragrant samples. Buttery *et al*. (1982, 1983) determined that the main chemical responsible for the aroma of fragrant rice varieties is the chemical 2-acetyl-1-pyrroline (2AP). Since this discovery, a number of more sophisticated chemical methods have been developed for determining the fragrance phenotype. Analysis of 2AP levels using gas chromatography is accurate and objective but is complicated by environmental influences, time of harvest, storage conditions (Bhattacharjee *et al*., 2002; Itani *et al*., 2004; Yoshihashi *et al*., 2005), degree of milling (Buttery *et al*., 1983; Widjaja *et al*., 1996a; Philpot *et al*., 2005) and requires expensive equipment and often large tissue samples (Lorieux *et al*., 1996; Widjaja *et al*., 1996b; Cordeiro *et al*., 2002).

Although there are reports of multiple loci controlling fragrance in rice, most agree there is one major locus for fragrance in rice with the possibility of other loci with minor influence. The major gene for rice fragrance was linked to a region on chromosome 8. Markers within this region were developed for use in PCR-based assays but were not 100% accurate as they were only linked to the fragrance gene. Recently the genetic cause of fragrance was identified (Bradbury *et al*., 2005a), allowing the generation of a perfect marker for genotyping fragrance in rice at any stage of the plant's life cycle (Bradbury *et al*., 2005b).

The genetic cause of fragrance in basmati and jasmine-style rice was found to be due to an 8 bp deletion and three SNPs in a gene encoding a betaine aldehyde dehydrogenase homologue (BAD2) (Fig. 11.1). The deletion causes a frame shift which generates a premature stop codon leading to the production of a truncated, non-functional enzyme (Bradbury *et al*., 2005a). BAD enzymes have been shown to be involved in the production of gamma-aminobutyric acid (Trossat *et al*., 1997; Livingstone *et al*., 2003). Inactivation of this enzyme most likely perturbs the pathway away from the production of gamma-aminobutyric acid and towards the production of 2AP (Bradbury *et al*., 2008).

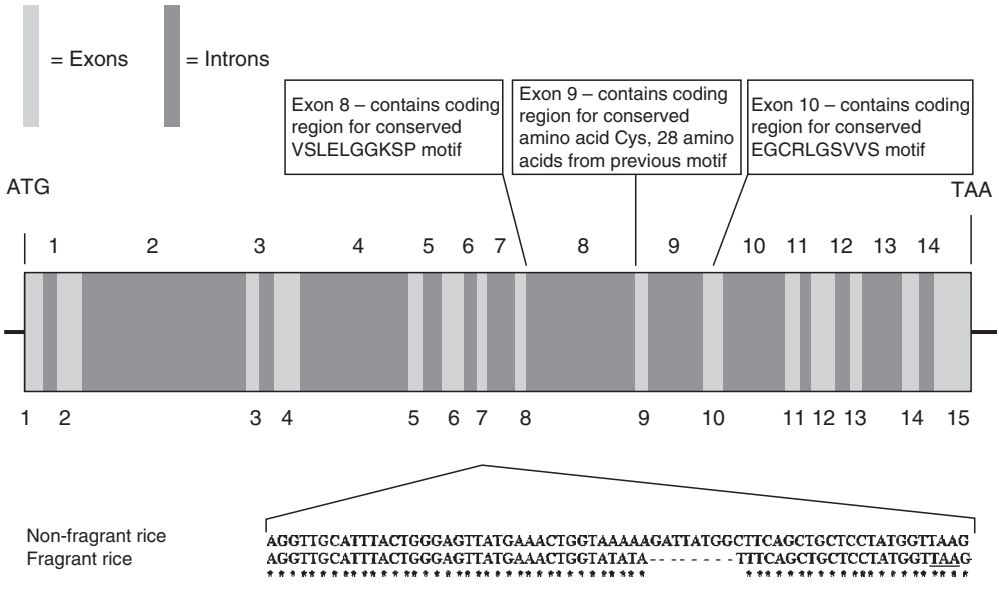


Fig. 11.1. Structure of the fragrance gene (*fgr*) (Knowledge-based Oryza Molecular Biological Encyclopedia (KOME) ID: J023088C02) showing initiation codon (ATG), 15 exons, 14 introns and the ATT termination site. The nucleotide sequence of exon 7 is shown for both non-fragrant and fragrant rice varieties. The fragrant variety shows a large deletion and three single nucleotide polymorphisms (SNPs), and then terminates prematurely (stop codon underlined), within this exon. The truncated protein encoded in fragrant rice varieties would therefore lack the highly conserved sequences encoded by exons 8, 9 and 10, which are believed to be critical for protein function. (From Bradbury *et al*., 2005a.)

The generation of an allele-specific PCR for this trait was relatively straightforward because an 8 bp deletion significantly interferes with allele-specific primer binding. Bradbury *et al*. (2005b) designed a four primer PCR that could distinguish between individuals and cultivars which are homozygous and heterozygous at the fragrance gene (Fig. 11.2). The assay involves the use of four primers, two flanking external primer pairs (external sense primer (ESP) and external anti-sense primer (EAP)) that anneal to

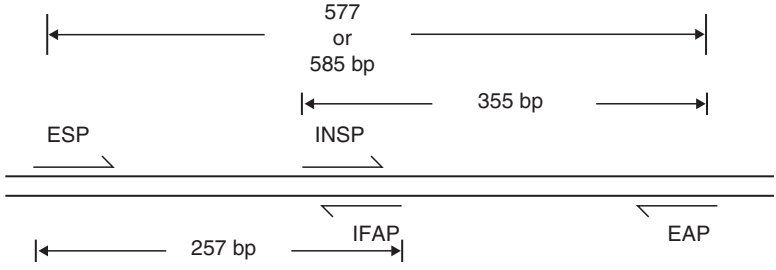


Fig. 11.2. Relative positions of PCR primers used in fragrance PCR. (From Bradbury *et al*., 2005b.)

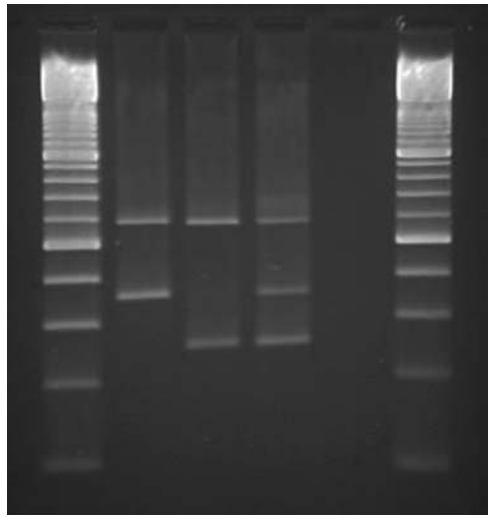


Fig. 11.3. Agarose gel showing (lanes 2–5) a non-fragrant variety (Nipponbare), a fragrant variety (Kyeema), a heterozygous individual (Kyeema/Gulfmont) and a negative control (water) flanked by Roche DNA Ladder XIV (100 bp). (From Bradbury *et al.*, 2005b.)

locations that are at opposite sides of, and are non-equidistant to, the mutated region and two competing internal primers (internal non-fragrant sense primer (INSP) and internal fragrant anti-sense primer (IFAP)) which anneal only to the non-fragrant and fragrant alleles respectively. The internal primers pair with the external primers to form products of varying size (Fig. 11.2). In all cases a product of approximately 580 bp is generated by the external flanking primers which acts as a positive control for the assay. The generation of the other two products is dependent on the genotype: in the case of a homozygous fragrant individual or variety, a band of 257 bp is produced by primers IFAP and ESP; in the case of a homozygous non-fragrant individual or variety, a band of 355 bp is produced by primers INSP and EAP. If a heterozygous individual is used, both INSP and IFAP will bind to the template DNA, generating both a 355 bp and a 257 bp product. The PCR products are easily separated on an agarose gel (Fig. 11.3), making this assay a simple and robust method for determining the fragrance status of rice varieties or individuals. The assay is fast to perform requiring denaturing, annealing and extension times of no more than 5 s each and can be performed on crude DNA extracts such as leaves boiled in PCR buffer for 10 min. It is low cost and can be utilized in laboratories which have access to PCR machines and agarose gel apparatus (Bradbury *et al.*, 2005b). Alternatively, more sophisticated high-throughput equipment could be used and this would allow more samples to be processed. This assay could also be adapted for use in a quantitative real-time PCR analysis in order to detect dilution of high-quality fragrant grain with inferior grain, a common practice that is a serious issue in the marketplace today.

References

Ayres, N.M., McClung, A.M., Larkin, P.D., Bligh, H.F.J., Jones, C.A. and Park, W.D. (1997) Microsatellites and a single-nucleotide polymorphism differentiate apparent amylose classes in an extended pedigree of US rice germ plasm. *Theoretical and Applied Genetics* 94, 773–781.

Bao, J.S., Corke, H. and Sun, M. (2006) Analysis of genetic diversity and relationships in waxy rice (*Oryza sativa* L.) using AFLP and ISSR markers. *Genetic Resources and Crop Evolution* 53, 323–330.

Bhattacharjee, P., Singhal, R.S. and Kulkarni, P.R. (2002) Basmati rice: a review. *International Journal of Food Science and Technology* 37, 1–12.

Bligh, H.F.J., Larkin, P.D., Roach, P.S., Jones, C.A., Fu, H.Y. and Park, W.D. (1998) Use of alternate splice sites in granule-bound starch synthase mRNA from low-amylose rice varieties. *Plant Molecular Biology* 38, 407–415.

Bormans, C.A., Rhodes, R.B., Kephant, D.D., McClung, A.M. and Park, W. (2002) Analysis of a single nucleotide polymorphism that controls the cooking quality of rice using a non-gel based assay. *Euphytica* 128, 261–267.

Bradbury, L.M.T., Fitzgerald, T.L., Henry, R.J., Jin, Q.S. and Waters, D.L.E. (2005a) The gene for fragrance in rice. *Plant Biotechnology Journal* 3, 363–370.

Bradbury, L.M.T., Henry, R.J., Jin, Q.S., Reinke, R.F. and Waters, D.L.E. (2005b) A perfect marker for fragrance genotyping in rice. *Molecular Breeding* 16, 279–283.

Bradbury, L.M.T., Henry, R.J. and Waters, D.L.E. (2008) Flavour development in rice. In: Frenkel, D.H. and Belanger, F. (eds) *Biotechnology in Flavour Production*. Blackwell Publishing, Oxford.

Buttery, R.G., Ling, L.C. and Juliano, B.O. (1982) 2-Acetyl-1-Pyrroline – an important aroma component of cooked rice. *Chemistry & Industry*, 958–959.

Buttery, R.G., Ling, L.C., Juliano, B.O. and Turnbaugh, J.G. (1983) Cooked rice aroma and 2-acetyl-1-pyrroline. *Journal of Agricultural and Food Chemistry* 31, 823–826.

Cai, X.L., Wang, Z.Y., Xing, Y.Y., Zhang, J.L. and Hong, M.M. (1998) Aberrant splicing of intron 1 leads to the heterogeneous 5' UTR and decreased expression of waxy gene in rice cultivars of intermediate amylose content. *Plant Journal* 14, 459–465.

Cordeiro, G.M., Christopher, M.J., Henry, R.J. and Reinke, R.F. (2002) Identification of microsatellite markers for fragrance in rice by analysis of the rice genome sequence. *Molecular Breeding* 9, 245–250.

Gale, K.R., Blundell, M.J. and Hill, A.S. (2004) Development of a simple, antibody-based test for granule-bound starch synthase Wx-B1b (Null-4A) wheat varieties. *Journal of Cereal Science* 40, 85–92.

Hohn, B. and Puchta, H. (2003) Some like it sticky: targeting of the rice gene waxy. *Trends in Plant Science* 8, 51–53.

Isshiki, M., Morino, K., Nakajima, M., Okagaki, R.J., Wessler, S.R., Izawa, T. and Shimamoto, K. (1998) A naturally occurring functional allele of the rice waxy locus has a GT to TT mutation at the 5' splice site of the first intron. *Plant Journal* 15, 133–138.

Itani, T., Tamaki, M., Hayata, Y., Fushimi, T. and Hashizume, K. (2004) Variation of 2-acetyl-1-pyrroline concentration in aromatic rice grains collected in the same region in Japan and factors affecting its concentration. *Plant Production Science* 7, 178–183.

Itoh, K., Ozaki, H., Okada, K., Hori, H., Takeda, Y. and Mitsui, T. (2003) Introduction of Wx transgene into rice wx mutants leads to both high- and low-amylose rice. *Plant and Cell Physiology* 44, 473–480.

Livingstone, J.R., Maruo, T., Yoshida, I., Tarui, Y., Hirooka, K., Yamamoto, Y., Tsutui, N. and Hirasawa, E. (2003) Purification and properties of betaine aldehyde dehydrogenase from *Avena sativa*. *Journal of Plant Research* 116, 133–140.

Lorieux, M., Petrov, M., Huang, N., Guiderdoni, E. and Ghesquiere, A. (1996) Aroma in rice: genetic analysis of a quantitative trait. *Theoretical and Applied Genetics* 93, 1145–1151.

Okagaki, R.J. and Wessler, S.R. (1988) Comparison of non-mutant and mutant *waxy* genes in rice and maize. *Genetics* 120, 1137–1143.

Olsen, K.M., Caicedo, A.L., Polato, N., McClung, A., McCouch, S. and Purugganan, M.D. (2006) Selection under domestication: evidence for a sweep in the rice waxy genomic region. *Genetics* 173, 975–983.

Philpot, K., Aryan, A. and Oliver, J. (2005) Grain quality of fragrant rice: distribution of 2-acetyl-1-pyrroline indifferent tissues of paddy rice. *The 12th Royal Australian Chemical Institute (RACI) Convention*. Sydney, Australia.

Sano, Y., Katsumata, M. and Okuno, K. (1986) Genetic studies of speciation in cultivated rice. 5. inter- and intra-specific differentiation in the waxy gene expression in rice. *Euphytica* 35, 1–9.

Sievert, D. and Holm, J. (1993) Determination of amylose by differential scanning calorimetry. *Starch-Starke* 45, 136–139.

Singh, R.K., Singh, U.S., Khush, G.S. and Rohilla, R. (2000) Genetics and biotechnology of quality traits in aromatic rices. In: Singh, R.K., Singh, U.S. and Khush, G.S. (eds) *Aromatic Rices*. Science Publishers, Enfield, New Hampshire and Oxford & IBH Publishing, New Delhi, India, pp. 47–70.

Tan, Y.F., Li, J.X., Yu, S.B., Xing, Y.Z., Xu, C.G. and Zhang, Q. (1999) The three important traits for cooking and eating quality of rice grains are controlled by a single locus in an elite rice hybrid, Shanyou 63. *Theoretical and Applied Genetics* 99, 642–648.

Trossat, C., Rathinasabapathi, B. and Hanson, A.D. (1997) Transgenically expressed betaine aldehyde dehydrogenase efficiently catalyzes oxidation of dimethylsulfoniopropionaldehyde and omega-aminoaldehydes. *Plant Physiology* 113, 1457–1461.

Umemoto, T. and Aoki, N. (2005) Single-nucleotide polymorphisms in rice starch synthase IIa that alter starch gelatinisation and starch association of the enzyme. *Functional Plant Biology* 32, 763–768.

Umemoto, T., Yano, M., Satoh, H., Shomura, A. and Nakamura, Y. (2002) Mapping of a gene responsible for the difference in amylopectin structure between japonica-type and indica-type rice varieties. *Theoretical and Applied Genetics* 104, 1–8.

Wang, Z.Y., Zheng, F.Q., Shen, G.Z., Gao, J.P., Snustad, D.P., Li, M.G., Zhang, J.L. and Hong, M.M. (1995) The amylose content in rice endosperm is related to the posttranscriptional regulation of the waxy gene. *Plant Journal* 7, 613–622.

Waters, D.L.E., Henry, R.J., Reinke, R.F. and Fitzgerald, M.A. (2006) Gelatinization temperature of rice explained by polymorphisms in starch synthase. *Plant Biotechnology Journal* 4, 115–122.

Widjaja, R., Craske, J.D. and Wootton, M. (1996a) Changes in volatile components of paddy, brown and white fragrant rice during storage. *Journal of the Science of Food and Agriculture* 71, 218–224.

Widjaja, R., Craske, J.D. and Wootton, M. (1996b) Comparative studies on volatile components of non-fragrant and fragrant rices. *Journal of the Science of Food and Agriculture* 70, 151–161.

Yoshihashi, T., Huong, N.T.T., Surojanametakul, V., Tungtrakul, P. and Varanyanond, W. (2005) Effect of storage conditions on 2-acetyl-1-pyrroline content in aromatic rice variety, Khao Dawk Mali 105. *Journal of Food Science* 70, S34–S37.

Zhou, Z.K., Robards, K., Helliwell, S. and Blanchard, C. (2002) Composition and functional properties of rice. *International Journal of Food Science and Technology* 37, 849–868.

12 Towards Universal Loci for Plant Genotyping

T. Pacey-Miller

Introduction

The accumulation of molecular data has made a huge impact on the fields of plant systematics, plant identification and breeding. It is, however, obvious that the use of different molecular tools is required depending on the question that is being asked. The level of phylogenetic resolution required will determine the loci used in the analysis. The rates of evolution vary considerably between different genomes, genes and gene regions. Thus, it is essential that the gene or gene region is chosen that contains the optimal level of genetic variation. With varying degrees of utility, a combination of these techniques may be needed to address taxonomic relationships, depending on the level of classification desired.

It is the basic assumption that variations within a defined genetic sequence result from mutations which represent an evolutionary event. Changes in DNA sequence comprise insertions, deletions, differences in copy number of repeated sequences and single base pair differences. The main benefit of sequence-based genotyping as opposed to marker-based genotyping is the enduring nature of the sequence. The sequence of the gene or the genome contains the entire collection of the variation within that region which can be compared as opposed to, for example, genotyping using a single nucleotide polymorphism (SNP).

The advances in DNA technology have enabled analysis of both nuclear and organelle DNA, each proving to be very useful in reconstructing phylogenies and as a tool for identification. Botanists produced the first classification based solely on gene sequences of angiosperms in 1998 (APG, 1998) which was revised in 2003 (APG II, 2003). The resulting patterns of congruent phylogenetic trees produced by many hundreds of taxa over many genes in different genomes were translated into a new and comprehensive classification. The aim was to produce a usable and predictive system by combining and maximizing the

 Plant Genotyping II: SNP Technology (ed. R.J. Henry)

information of phylogeny gathered by many authors, as opposed to being the subjective views of a single author (APG, 1998; APG II, 2003).

Focus has more recently shifted away from the phylogenetic uses of sequence analysis on to the process of genetic fingerprinting. New goals not only want to identify plant sources but locate genetic markers for desired traits. These applications have been the major focus for DNA technologies in assisting plant breeding programmes, maintaining crop consistency, plant forensics and in commercial activities. Also a goal in the forefront of the use of these techniques is the creation of a 'Barcode of Life'.

Sequence Variation

To obtain all of the genetic information within the genome requires an enormous effort. A more manageable process is to examine a much smaller handful of sites that contain enough variation to obtain enough information to discern a difference. Different regions of the genome evolve at different rates. It is known that mitochondrial evolves the slowest, chloroplast slightly faster and the nuclear genome the fastest (Wolfe *et al*., 1987). In addition to this, coding regions evolve more slowly than non-coding regions presumably due to selective constraints (Curtis and Clegg, 1984; Wolfe *et al*., 1987; Clegg and Zurawski, 1991; Clegg *et al*., 1991).

Chloroplast DNA

The chloroplast (cp) genome has been the most widely used genome for plant systematics. It has a relatively small size. It is a circular molecule which contains a large duplicated region of reverse orientation, separated by two regions of single copy DNA (Curtis and Clegg, 1984). It is also non-recombinant, generally uniparentally inherited, contains both coding (gene) and non-coding (intron and intergenic spacer) sequences and is singe copy. Its high structural stability has facilitated the design of universal primers (Taberlet *et al*., 1991), which have become the most widely used non-coding cpDNA sequences in plant systematics. As of February 2007, Web of Science lists 1088 citations of the Taberlet *et al*. (1991) paper.

Different regions of the chloroplast genome are used to infer relationships depending on the classification level required. Chloroplast genes that have been used more regularly at family level and above include *rbc*L, *atp*B, *mat*K and *ndh*F (Gielly and Taberlet, 1994; Johnson and Soltis, 1994; Steele and Vilgalys, 1994; Clark *et al*., 1995; Duvall and Morton, 1996; Nishikawa *et al*., 2002). Sequences of chloroplast non-coding regions such as introns (*rp*L16, *rpo*C1, *rp*S16,*trn*L,*trn*K) and intergenic spacers (*trn*T-*trn*L, *trn*L-*trn*F, *atp*B-*rbc*L, *psb*A-*trn*H) are more useful at lower taxonomic levels because they are less functionally constrained and therefore more free to vary, thereby potentially providing more phylogenetically informative characters per unit of sequence effort (Curtis and Clegg, 1984; Clegg and Zurawski, 1991; Downie

and Palmer, 1991; Soltis *et al*., 1991; Taberlet *et al*., 1991, 2006; Kelchner and Clark, 1997; Small *et al*., 1998, 2004; Shaw *et al*., 2005).

Mitochondrial DNA

The slow rate of sequence evolution and high rate of structural evolution (frequent gene rearrangements) means that mitochondrial genes are largely ignored by plant systematists as potential source of data.

Nuclear Ribosomal DNA

Used at higher taxonomic levels due to the slow rate of evolution, ribosomal genes exist in tandem arrays of genes composed of hundreds to thousands of copies per array (Small *et al*., 2004). This highly repetitive nature also gives it properties that affect its potential utility and reliability in phylogenetic studies (Small *et al*., 2004). Nuclear genes encoding ribosomal DNA (5S and 18S-5.8S-26S) and non-coding regions (internal transcribed spacer, ITS1 and ITS2) have been regions of choice showing broad utility (Wolfe *et al*., 1987; Appels and Clarke, 1992; Sastri *et al*., 1992; Baldwin *et al*., 1995; Hsiao *et al*., 1995; Buckler and Holtsford, 1996; Ko and Henry, 1996; Hughes *et al*., 2006).

Nuclear Genes

The need for phylogenetic markers from the nuclear genome to complement the growing cpDNA data sets has led to the examination of new genes for plant systematics. The mutation rate of nuclear genes is up to five times that of chloroplast genes and 20 times that of mitochondrial genes (Small *et al*., 2004), and therefore the sequencing effort required is more efficient with nuclear genes. There is also rate variation among regions within nuclear genes, for example 5' untranslated region (UTR), exons, introns and 3'UTR. Nuclear genes usually exist in families, but low copy nuclear genes in plants are a rich source of genetic variation. Low copy nuclear genes typically evolve independently and tend to be stable in position and copy number (Small *et al*., 2004). These genes have contributed to a better understanding of interspecific relationships of various plant groups and reconstructing allopolyploidization in plants (Sang, 2002). No 'universally' useful low copy nuclear gene sequence loci have been developed during the last 10 years (Sang, 2002). Each has had to be developed specifically for the taxonomic group of interest. Some examples of low copy nuclear genes used regularly include *Adh* (alcohol dehydrogenase), *G3PDH* (glyceraldehyde 3-phosphate dehydrogenase), *GBSS1* (granule-bound starch synthase), *MADS-box* genes, *PHY* (phytochrome) and *PGI* (phospho-glucose isomerase) (Freeling and Bennett, 1985; Wolfe *et al*., 1987; Mason-Gamer *et al*., 1998; Small *et al*., 1998, 2004; Mathews *et al*., 2000; Sang, 2002; McIntosh *et al*., 2005).

Comparative Sequence Analysis

Comparative sequence analysis can be used to efficiently transfer information from a model species to a crop species or species of interest. It can be used to identify genes or traits of interest or to assess within-species diversity so that the best alleles can be identified and assembled in superior varieties (Sorrells *et al*., 2007) for use with crop improvement strategies. It involves comparisons of structure and function to estimate similarity of biological organization. The main uses currently include cross-referencing genes between species maps, enhancing the resolution of comparative maps, studying patterns of gene evolution, identifying conserved regions of genomes and facilitating interspecies gene cloning (La Rota and Sorrells, 2004).

DNA Barcoding

DNA barcoding would involve characterizing species of organisms using a short DNA sequence from a standard and agreed-upon position in the genome. Barcoding incorporates tools for managing the earths changing biodiversity. An international collaboration exists for this purpose known as the Consortium for the Barcode of Life (CBOL) (http://barcoding.si.edu). This involves organizations such as museums, zoos, aquaria, herbaria and repositories of biological material for supply of material for analysis. The material is then sequenced and analysed by other members such as research institutes and molecular labs and the data compiled into databases such as Genbank and EMBL or the DNA data bank of Japan by computer experts. Specimens can then be identified by reference to databases. The mission is to compile a public library of sequences linked to named specimens and promote the development of portable devices for DNA barcoding (http://barcoding.si.edu/ and http://phe.rockefeller.edu/BarcodeConference/index.html).

To be the ideal system for barcoding there are several criteria that should be met. It needs to be sufficiently variable to discriminate among all species but conserved enough to be less variable within than between species (Taberlet *et al*., 2006). It should also be standardized with the same DNA region for different taxonomic groups (Taberlet *et al*., 2006). It should contain enough phylogenetic information to easily assign species to its taxonomic group (Taberlet *et al*., 2006). It should be extremely robust with highly conserved priming sites and highly reliable DNA amplifications and sequencing (Taberlet *et al*., 2006). In addition, the target DNA region should be short enough to allow amplification of degraded DNA (Taberlet *et al*., 2006). It is not easy to satisfy all five criteria at once; however, to different users each of these criteria will be of varying importance. For example, the value to a taxonomist requiring enough variation in a sequence to be phylogenetically informative is quite different to that of a forensic scientist who would place more value in standardized and robust methodologies.

Land plants have the reputation of being problematic for barcoding mainly due to low variability in the standard regions used for other organ-

isms such as algae, animals and fungi as well as in the typically used plant plastid phylogenetic markers (*rbc*L, *trn*L-F, etc.) (Chase *et al*., 2005). However, Chase *et al*. (2005) do suggest that this lack of utility for phylogenetic resolution does not necessarily exclude these loci from being useful as identity codes. Since, most potential users of this technology would be the researchers who needed a quick, easy and accurate system of identification as opposed to taxonomists or systematists, a relatively crude diagnostic method would be more than acceptable, at least until a more sophisticated barcoding technique was developed (Chase *et al*., 2005).

There have been some investigations into loci that are already regularly used for phylogenetic work into their potential for use as a 'barcode'. Taberlet *et al*. (2006) examined the power and limitations of chloroplast *trn*L intron for barcoding. This locus did show a relatively low resolution of 67.3% (Taberlet *et al*., 2006) but this is more than compensated by the prospect of highly conserved universal primers and robust amplification system. The P6 loop, which is a shorter fragment of this intron, was also examined in the same study for its potential. It was found to have 19.5% resolution which is quite low but still higher than existing alternative systems. It does have the additional benefit of being good for use with degraded DNA.

Kress *et al*. (2005) examined several loci that met the barcode criteria before proposing the use of the internal transcribed spacer (ITS) region as well as the plastid *trn*H-*psb*A intergenic spacer in combination as potential usable regions for plant DNA barcoding for flowering plants. The ITS region being the most commonly sequenced region for plant phylogenetic studies at the species level shows high levels of interspecific divergence (Kress *et al*., 2005). The *trn*H-*psb*A spacer is the most variable plastid region in angiosperms and is easily amplified across a broad range of land plants. Its short sequence length (average 465 bp) packaged with its high-level interspecific nucleotide differences makes it an ideal candidate. Despite these advantages, only limited potential at species-level identification was determined with each of these loci. This study provides evidence that there are potential gene sequences suitable for DNA barcoding of flowering plants; however, it may be necessary to use more than one locus for species-level discrimination.

Identification, Breeding and Forensics

In addition to the incredibly large-scale process of barcoding all living organisms, there are several smaller-scale, though no less significant, reasons for plant genotyping using DNA sequences. The goal of genetic fingerprinting is not only to identify plant sources but also to locate genetic markers for desired traits. These applications have been the major focus for DNA technologies in assisting plant breeding programmes, maintaining crop consistency, plant forensics and commercial activities.

It is often difficult to identify plant species in the field, particularly when collecting wild relatives of crop species for conservation of the material and eventual utilization in breeding programmes. Often visual inspection of the

material can be misleading, particularly at different stages of the growth cycle. It is often very small differences that can mean the classification of material to one species instead of another. It has become very clear that the identification of species in some germplasm collections is incorrect. The identification of germplasm can now be verified by sequencing.

Many different loci have been used for this purpose. Genus *Beta* was successfully identified using ITS1 sequence information, both species and subspecies can be determined. It was used to identify wild beet populations for use in a breeding programme and demonstrates the potential use of the technique in other crops (Shen *et al*., 1998). ITS was also used for the identification of below-ground parts of plants (roots, stems, buds, tubers). All were identified to genus level and in most cases to the level of species (Linder *et al*., 2000).

Vanilla species, which are important to the flavour and fragrance industries, were identified using several gene fragments *YCF*5 and *rpo*B and the intergenic spacers *psb*A-*trn*H and ITS. Leafless vanilla is very difficult to identify when not in flower and only some species are useful to the industry. The study uniquely identified 50 different accessions (Cameron, 2006).

In the herbal supplement market, identification between two very similar species becomes important when only one of them has the claimed medicinal benefit. For example, one of the ginseng species on the Korean market is thought to tonify the *Qi* (energy) more effectively against ageing, weakness and stress (Leem *et al*., 2005). Both samples are hard to identify particularly from the slice or powder or extract. The genetic variance between the two species was identified by examining the ITS with pyrosequencing. Medicinal quality can also vary in the same supplement species depending on where it is grown (Leem *et al*., 2005). It is possible to trace the origin of the particular herb or ingredient using gene sequencing.

Samples of commonly found plants in Taiwan were correctly identified, using pairwise comparisons, to species level, with the *trn*L intron and the *trn*L-*trn*F intergenic spacer. Of the 373 species examined all were identified correctly with the exception of three species pairs at the *trn*L locus and five species pairs at the *trn*L-*trn*F locus (Tsai *et al*., 2006b). This information was placed into a database to be used for identification of unknown samples.

The technology can be used for verification purposes. Seized goods were falsely claimed to be seeds of a medicinal plant that is closely related to the *Cannabis sativa* species. They were identified as illegal *Cannabis sativa* seeds using *trn*L-*trn*F intergenic spacer (Tsai *et al*., 2006a), and legal action was taken. By using gene identification, it was unnecessary to wait until the seeds had been germinated for visual and chemical identification to be undertaken, therefore producing a more expedient result.

The identification of trace evidence in forensic botany, most commonly grasses, to provide links between crime scenes and individuals (Ward *et al*., 2005) is becoming increasingly important. Gene identification can be used on these often very small samples. Identification to cultivar level is important especially when the majority of these specimens are in wide distribution in the urban environment as with turf grasses. Here arises the difficulty, and the development of a DNA-based identification using several PCR assays that

works like a biological key is in progress, in which moving through to the next step is dependent on the outcome of the previous step (Ward *et al*., 2005).

Increasing in necessity is a method for the recognition of novel conventional or transgenic organisms for protection of patented or 'Identity Preserved' lines, as well as detecting transgenics and tracing dispersal (Gressel and Ehrlich, 2002). For example, several genetically modified potatoes are registered in Canada though not currently in commercial production. Distinguishing genetically modified (GM) cultivars from non-GM cultivars and other GM cultivars is made possible by analysing the sequence at the insertion site of the transgene corresponding to a specific transformation event (Cote *et al*., 2005).

Identifying Mixtures

Identification of cereal species is essential for the handling, marketing and processing of grain and for the protection of plant breeders' rights. It is becoming increasingly important in today's society to be able to verify the exact ingredients of a product or even to trace the products back to the farm from whence they came (Ko *et al*., 1994; Terzi *et al*., 2005), in some cases not only to identify but also quantify cereal species in food products. Gene identification is useful in cases where contamination or mixing occurs and the species are difficult to identify by visual inspection, in particular in milled cereals or processed food products. Ko and Henry (1996) use the 5S ribosomal RNA spacer to detect individual species of seven different cereals within one mixture.

It would be essential to be able to fingerprint and identify grain at the genotype and variety levels. This would then allow detection of admixtures and prevent errors in handling grain in storage (e.g. malting and feed barleys can be distinguished) (Ko *et al*., 1994). This could also ensure seed purity for breeding and commercial production.

Genotyping Gene Loci Using Pyrosequencing

As a tool for the sequencing of gene loci, pyrosequencing is a high-throughput method that can increase the sequencing output. Pyrosequencing is a DNA synthesis technique that is based on the detection of released pyrophosphate (PP_i) during DNA synthesis (Ronaghi *et al*., 1996, 1998). Pyrosequencing is a sequence-by-synthesis method in which a series of enzymatic reactions yields detectable light, which is proportional to the incorporated nucleotides (Ahmadian *et al*., 2000).

Traditionally pyrosequencing has been performed using a single-stranded template which was achieved by the physical separation of the two strands of a PCR reaction, one of which was labelled with a biotinylated primer. Unfortunately this process is not only time-consuming, but the cost of biotinylating one of the primers in a pair is excessive. A double-stranded

protocol was developed (Nordstrom *et al*., 2000a,b) in which was proposed a new enzyme-related clean-up method; however, success was varied in trials (Pacey-Miller and Henry, 2003). As an alternative an eleven base tag sequence 5'-(Biotin)GCCCCCGCCCG-3' which had been biotin labelled at the 5' end and HPLC-purified was developed which could be annealed to one of the strands (Pacey-Miller and Henry, 2003). Until a more reliable double-stranded method is developed, the system of using a biotinylated tag provides the most reliable and least labour-intensive and cost-effective high-throughput approach for pyrosequencing in plant tissues.

Pyrosequencing is an ideal high-throughput method for the sequencing of short fragments. McIntosh *et al*. (2005) developed a universal protocol for the identification of cereals using pyrosequencing. *GBSS1* is a well-conserved gene within Poaceae and due to its commercial importance has undergone much investigation. Its utility in phylogenetics and sequence polymorphism surrounding desired traits has been investigated (Mason-Gamer *et al*., 1998; McLauchlan *et al*., 2001). Three regions of the *GBSS1* locus of four grasses (wheat, rice, barley and maize) showing varying degrees of polymorphism were pyrosequenced. The structure of this gene is such that it contains many small exons and introns. Both cyclic and non-cyclic dispensations (Pacey-Miller and Henry, 2003; McIntosh *et al*., 2005) were used to generate 'fingerprints' or pyrograms for the four grasses in question, rice, maize, wheat and barley. Individual peaks, but more often a combination of peaks, are used to differentiate between these plant species. Specific peak heights are not comparable between different runs; therefore, it is the presence or absence of nucleotide incorporation at the SNP positions that is assessed. Unique phylograms for each of these grasses were obtainable (Fig. 12.1). The polymorphism between aligned single region sequences of *GBSS1* coding regions could identify variation at the genus level but was limited at the species level depending on the genus examined. By combining coding sequence data sets from all three fragments targeted, an accurate fingerprint at the species level is obtainable. No application was observed for varietal identification and a more rapidly evolving gene would be required for this purpose. The detection of small insertions and deletions can also be determined using this process (Guo *et al*., 2003).

This study (McIntosh *et al*., 2005) was aimed at plant identification rather than phylogenetic relationships and the potential for a commercial application for pyrosequencing was investigated. Pyrosequencing has already proven its effectiveness as a tool for targeting small regions of DNA including SNPs. The usefulness in the analysis of manufactured plant food products was investigated by testing these methods on mixed, plant-derived materials (McIntosh *et al*., 2005). The results thus far have enabled the identification of the individual plant constituents of materials derived from multiple sources. This utility will be very useful in industry where the ever-increasing demand on quality control must be met. The future of this plant identification procedure is its potential for identification outside the grass family.

Pyrosequencing and technologies like it will lead the way in the search for universal loci for plant genotyping. Ultimately it would be desirable to be able to use a minimal number of loci for sample identification purposes.

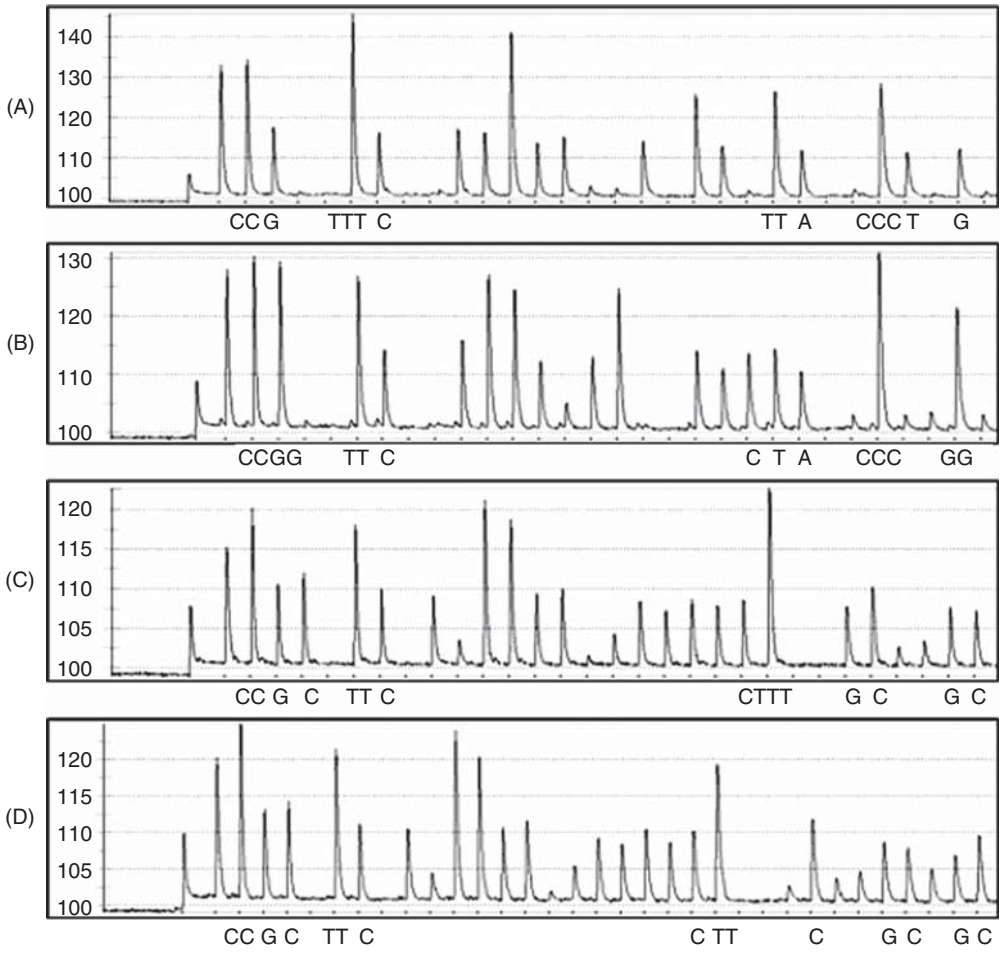


Fig. 12.1. Pyrosequencing analysis of grasses. Pyrograms of rice (A), maize (B), barley (C) and wheat (D) produced by extending with a sequencing primer using a cyclic dispensation. The peaks highlighted on the x-axis indicate some of the main polymorphic regions. Y-axis units are arbitrary, representing light intensity.

However, the more likely scenario would be that depending on the level of classification desired, a combination of loci using varying molecular analysis techniques may be needed to address taxonomic relationships.

References

Ahmadian, A., Gharizadeh, B., Gustafsson, A., Sterky, F., Nyren, P., Uhlen, M. and Lundeberg, J. (2000) Single-nucleotide polymorphism analysis by pyrosequencing. *Analytical Biochemistry* 280, 103–110.

APG (1998) An ordinal classification for the families of flowering plants. *Annals of the Missouri Botanical Garden* 85, 531–553.

APG II (2003) An update of the angiosperm phylogeny group classifications for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141, 399–436.

Appels, R. and Clarke, B.C. (1992) The 5S DNA units of bread wheat (*Triticum aestivum*). *Plant Systematics and Evolution* 183, 195–208.

Baldwin, B., Sanderson, M., Porter, J., Wojciechowski, M., Campbell, C. and Donoghue, M. (1995) The ITS region of nuclear ribosomal DNA; a valuable source of evidence on angiosperm phylogeny. *Annals of the Missouri Botanical Garden* 82, 247–277.

Buckler, E. and Holtsford, T. (1996) Zea systematics: ribosomal ITS evidence. *Molecular Biology and Evolution* 13, 612–622.

Cameron, K.M. (2006) DNA barcoding as a method for Vanilla (Orchidaceae) species identification. *Botany 2006*, California State University, California.

Chase, M., Salamin, N., Wilkinson, M., Dunwell, J., Kesanakurthi, R., Haidar, N. and Savolainen, V. (2005) Land plants and DNA barcodes: short-term and long-term goals. *Philosophical Transactions of the Royal Society of London, Series B* (doi:10.1098/rstb.2005.1720).

Clark, L.G., Weiping, Z. and Wendel, J.F. (1995) A phylogeny of the grass family (Poaceae) based on *ndhF* sequence data. *Systematic Biology* 20, 436–460.

Clegg, M. and Zurawski, G. (1991) Chloroplast DNA and the study of plant phylogeny. In: Soltis, P., Soltis, D. and Doyle, J. (eds) *Molecular Systematics of Plants*. Chapman & Hall, New York, pp. 1–13.

Clegg, M., Learn, G. and Golenberg, E. (1991) Molecular evolution of chloroplast DNA. In: Selander, R., Clark, A. and Whittam, T. (eds) *Evolution at the Molecular Level*. Sinauer, Sunderland, Massachusetts, pp. 135–149.

Cote, M.J., Meldrum, A., Raymond, P. and Dollard, C. (2005) Identification of genetically modified potato (*Solanum tuberosum*) cultivars using event specific polymerase chain reaction. *Journal of Agricultural Food Chemistry* 53, 6691–6696.

Curtis, S. and Clegg, M. (1984) Molecular evolution of chloroplast DNA sequences. *Molecular Biology and Evolution* 1, 291–301.

Downie, S. and Palmer, J. (1991) Use of chloroplast DNA rearrangements in reconstructing plant phylogeny. In: Soltis, P., Soltis, D. and Doyle, J. (eds) *Molecular Systematics of Plants*. Chapman & Hall, New York, pp. 14–35.

Duvall, M.R. and Morton, B.R. (1996) Molecular phylogenetics of poaceae: an expanded analysis of rbcL sequence data. *Molecular Phylogenetics and Evolution* 5, 352–358.

Freeling, M. and Bennett, C. (1985) Maize Adh1. *Annual Review of Genetics* 19, 297–323.

Gielly, L. and Taberlet, P. (1994) The use of chloroplast DNA to resolve plant phlogenies: noncoding versus rbcL sequences. *Molecular Biology & Evolution* 11, 769–777.

Gressel, J. and Ehrlich, G. (2002) Universal inheritable barcodes for identifying organisms. *Trends in Plant Science* 7, 542–544.

Guo, D., Qi, Y., He, R., Gupta, P. and Milewicz, D. (2003) High throughput detection of small genomic insertions or deletions by pyrosequencing. *Biotechnology Letters* 25, 1703–1707.

Hsiao, C., Chatterton, N., Asay, K. and Jensen, K. (1995) Molecular phylogeny of the pooideae (Poaceae) based on nuclear rDNA (ITS) sequence. *Theoretical and Applied Genetics* 90, 389–398.

Hughes, C., Eastwood, R. and Bailey, C.D. (2006) From famine to feast? Selecting nuclear DNA sequence loci for plant species-level phylogeny reconstruction. *Philosophical Transactions of the Royal Society of London, Series B* 361, 211–225.

Johnson, L. and Soltis, D. (1994) matK DNA sequence and phylogenetic reconstruction of Saxifragaceae s.s. *Systematic Botany* 19, 143–156.

Kelchner, S. and Clark, L. (1997) Molecular evolution and phylogenetic utility of the chloroplast rpl16 intron in Chusquea and the Bambusoideae (Poaceae). *Molecular Phylogenetics and Evolution* 8, 385–397.

Ko, H. and Henry, R. (1996) Specific 5S ribosomal RNA primers for plant species identification in admixtures. *Plant Molecular Biology Reporter* 14, 33–43.

Ko, H., Henry, R., Graham, G., Fox, G., Chadbone, D. and Haak, I.C. (1994) Identification of cereals using the polymerase chain reaction. *Journal of Cereal Science* 19, 101–106.

Kress, W.J., Wurdack, K., Zimmer, E., Weigt, L. and Janzen, D. (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the USA* 102, 8369–8374.

La Rota, M. and Sorrells, M. (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Functional and Integrative Genomics* 4, 34–46.

Leem, K., Kim, S., Yang, C. and Seo, J. (2005) Genetic identification of *Panax gingeng* and *Panax quinquefolius* by pyrosequencing methods. *Bioscience Biotechnology and Biochemistry* 69, 1771–1773.

Linder, C., Moore, L. and Jackson, R. (2000) A universal molecular method for identifying underground plant parts to species. *Molecular Ecology* 9, 1549–1559.

Mason-Gamer, R.J., Weil, C.F. and Kellogg, E.A. (1998) Granule-bound starch synthase: structure, function, and phylogenetic utility. *Molecular Biology and Evolution* 15, 1658–1673.

Mathews, S., Tsai, R. and Kellogg, E.A. (2000) Phylogenetic structure in the grass family (Poaceae): evidence from the nuclear gene phytochrome B. *American Journal of Botany* 87, 96–107.

McIntosh, S., Pacey-Miller, T. and Henry, R. (2005) A universal protocol for identification of cereals. *Journal of Cereal Science* 41, 37–46.

McLauchlan, A., Ogbonnaya, F., Hollingsworth, B., Carter, M., Gale, K., Henry, R., Holton, T., Morell, M., Rampling, L., Sharpe, J., Shariflou, M., Jones, M. and Appels, R. (2001) Development or robust PCR-based markers for each homo-allele of granule-bound starch synthase and their application in wheat breeding programs. *Australian Journal of Agricultural Research* 52, 1409–1416.

Nishikawa, T., Salomon, B., Komatsuda, T., von Bothmer, R. and Kadowaki, K. (2002) Molecular phylogeny of the genus *Hordeum* using three chloroplast DNA sequences. *Genome* 45, 1157–1166.

Nordstrom, T., Nourizad, K., Ronaghi, M. and Nyren, P. (2000a) Method enabling pyrosequencing on double-stranded DNA. *Analytical Biochemistry* 282, 186–193.

Nordstrom, T., Ronaghi, M., Forsberg, L., de Faire, U., Morgenstern, R. and Nyren, P. (2000b) Direct analysis of single-nucleotide polymorphism on double-stranded DNA by pyrosequencing. *Biotechnology & Applied Biochemistry* 31, 107–112.

Pacey-Miller, T. and Henry, R. (2003) Single-nucleotide polymorphism detection in plants using a single stranded pyrosequencing protocol with a universal biotynylated primer. *Analytical Biochemistry* 317, 165–170.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry* 242, 84–89.

Ronaghi, M., Uhlen, M. and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate. *Science* 281, 363–365.

Sang, T. (2002) Utility of low-copy nuclear gene sequences in plant phylogenetics. *Critical Reviews in Biochemistry & Molecular Biology* 37, 121–147.

Sastri, D.C., Hilu, K., Appels, R., Lagudah, E.S., Playford, J. and Baum, B.R. (1992) An overview of evolution in plant 5S DNA. *Plant Systematics and Evolution* 183, 169–181.

Shaw, J., Lickey, E., Beck, J., Farmer, S., Liu, W., Miller, J., Siripun, K., Winder, C., Schilling, E. and Small, R. (2005) The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany* 92, 142–166.

Shen, Y., Newbury, H.J. and Ford-Lloyd, B.V. (1998) Identification of taxa in the genus *Beta* using ITS1 sequence information. *Plant Molecular Biology Reporter* 16, 147–155.

Small, R., Ryburn, J., Cronn, R., Seelanan, T. and Wendel, J. (1998) The tortoise and the hare: choosing between noncoding plastome and nuclear ADH sequences for phylogeny reconstruction in a recently diverged plant group. *American Journal of Botany* 85, 1301–1315.

Small, R., Cronn, R. and Wendel, J. (2004) L.A.S. Johnson Review No. 2. Use of nuclear genes for phylogeny reconstruction in plants. *Australian Systematic Botany* 17, 145–170.

Soltis, D., Soltis, P. and Milligan, B. (1991) Intraspecific chloroplast DNA variation: systematic and phylogenetic implication. In: Soltis, P., Soltis, D. and Doyle, J. (eds) *Molecular Systematics of Plants*. Chapman & Hall, New York, pp. 117–150.

Sorrells, M., La Rota, M., Bermudez-Kandianis, C., Greene, R., Kantety, R., Munkvold, J., Miftahudin, Mahmoud, A., Ma, X., Gustafson, P., Qi, L., Echalier, B., Gill, B., Matthews, D., Lazo, G., Chao, S., Anderson, O., Edwards, H., Linkiewicz, A., Dubcovsky, J., Akhunov, E., Dvorak, J., Zhang, D., Nguyen, H., Peng, J., Lapitan, N., Gonzalez-Hernandez, J., Anderson, J., Hossain, K., Kalavacharla, V., Kianian, S., Choi, D.W., Close, T., Dilbirligi, M., Gill, K., Steber, C., Walker-Simmons, M., McGuire, P. and Qualset, C. (2007) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Research* 13, 1818–1827.

Steele, K. and Vilgalys, R. (1994) Phylogenetic analyses of polemoniaceae using nucleotide sequences of the plastid gene *mat*K. *Systematic Botany* 19, 126–142.

Taberlet, P., Geilly, L., Pautou, G. and Bouvet, J. (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* 17, 1105–1109.

Taberlet, P., Coissac, E., Pomanon, F., Geilly, L., Miquel, C., Valentini, A., Vermat, T., Corthier, G., Brochmann, C. and Willerslev, E. (2006) Power and limitations of the chloroplast *trn*L (UAA) intron for plant DNA barcoding. *Nucleic Acids Research* e1–e8.

Terzi, V., Morcia, C., Gorrini, A., Stanca, A.M., Shrewry, P. and Faccioli, P. (2005) DNA-based methods for identification and quantification of small grain cereal mixtures and fingerprinting of varieties. *Journal of Cereal Science* 41(3), 213–220.

Tsai, L.C., Hsieh, H.M., Wang, J.C., Huang, L.H., Linacre, A. and Lee, J.C.I. (2006a) Cannabis seed identification by chloroplast and nuclear DNA. *Forensic Science International* 158, 250–251.

Tsai, L.C., Yu, Y.C., Hsieh, H.M., Wang, J.C., Linacre, A. and Lee, J.C.I. (2006b) Species identification using sequences of the trnL intron and the trnL-trnF IGS of chloroplast genomeamong popular plants in Taiwan. *Forensic Science International* 164, 193–200.

Ward, J., Peakall, R., Gilmore, S.R. and Robertson, J. (2005) A molecular identification system for grasses: a novel technology for forensic botany. *Forensic Science International* 152(2–3), 121–131.

Wolfe, K., Li, W.-H. and Sharp, P. (1987) Rates of nucleotide substitution very greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Sciences of the USA* 84, 9054–9058.

13 DNA Banks as a Resource for SNP Genotyping

N. Rice, S. Kasem and R.J. Henry

Introduction

Plant genomics has developed rapidly and as a result many high-throughput analytical methods are available. The increase in the adoption of these genomic platforms has seen the storage of DNA and associated by-products of the procedures increase. However, there are only a relatively small number of groups that make their collections available to third parties (Andersson *et al*., 2006). DNA Banks is the name given to such collections which aim to store DNA for long periods and allow withdrawals which facilitate the research efforts of others. It is in this way that DNA Banks can provide a key resource to exploit the high-throughput capabilities of plant genomics and in particular genotyping through the use of single nucleotide polymorphisms (SNPs) marker systems. This chapter will provide a brief overview of the current status of DNA Banks, operations at the Australian Plant DNA Bank (www.dnabank.com.au) and sample handling issues relevant to plant genomics research.

DNA Banks

Biological collections are valuable resources for access to material and information that underpins biological research programmes. DNA collections (or DNA Banks) are a new form of biological collection providing a link between traditional biological collections and the application of the array of molecular tools emerging from areas such as genomics and transcriptomics. DNA Banks have previously been defined as a store of extracted genomic DNA (Rice *et al*., 2006b). DNA Banks have grown directly from the use of molecular-based techniques and these centralized collections assist by reducing the need for duplication of resources and costs associated with the need for recollection and/or preparation of samples.

 Plant Genotyping II: SNP Technology (ed. R.J. Henry)

Only a relatively small number of DNA collections are formalized and represent a readily accessible resource for researchers usually via a World Wide Web portal (Andersson *et al*., 2006; Rice *et al*., 2006b; Hodkinson *et al*., 2007). This is evident from the results of a recent Bioversity (formerly International Plant Genetic Resource Institute) survey which reports that, out of 243 respondents, only 51 claimed to store DNA, and of the three groups within Australia only the Australian Plant DNA Bank has currently made provisions for formal access via a web portal (Andersson *et al*., 2006). The need for DNA Banks has probably been greater in species such as plants for which DNA extraction is more difficult and some organisms, such as micro-organisms, may often be analysed directly without DNA extraction, stored quite easily for long periods as live cultures or as a specimen collection (Rice *et al*., 2006a).

Chapter 14 (this volume) outlines the numerous modifications of DNA extraction methods which may be necessary to extract DNA of reasonable quality and quantity from various plant species and tissue types. Although the extraction of DNA from plants is perceived as an easy task (Hodkinson *et al*., 2007), the volume of published techniques indicates that the polysaccharides, proteins and secondary metabolites in plant tissues can be species- and tissue-specific and means that the extraction of DNA suitable for downstream techniques and long-term storage is somewhat problematic.

DNA Extraction

The Australian Plant DNA Bank uses DNA extraction kit-based technologies for ease of use and recovery of high-quality DNA. These kits are often readily suitable for automation and the techniques are easily modified to suit a range of different species. The QIAGEN MagAttract 96 DNA Plant kit is the method currently used for automated extraction on the AVISO TheOnyx platform with an additional RPW wash and three ethanol washes. Plant Tissue is ground in deep 96-well plates using a Retsch Mixer Mill and a tungsten bead. Tissue can be ground frozen or dried using this technique.

Although many rapid extraction protocols exist which are suitable for SNP analysis, these methods are not favoured by the Australian Plant DNA Bank as they do not yield very pure DNA. As the DNA in the collection is ultimately held for long periods and for multiple end uses it is desirable to ensure that the DNA is as pure as possible. End uses of DNA from the Australian Plant DNA Bank for SNPs have been previously highlighted (Rice *et al*., 2006a).

DNA is quantified by using either UV absorbance or fluorescence. The Quan-iT Picogreen reagent is the preferred stain for quantification using fluorescence. It is increasingly popular to use fluorescent techniques as stains such as Quan-iT Picogreen as they are able to quantify only double-stranded DNA and are capable of accurate quantification of lower concentrations, i.e. 25 pg/ml (MolecularProbes, 2005).

Sample Handling and Storage Issues

One assumption that we make in relation to handling extracted DNA is that it should be stored frozen and we should minimize the number of freeze–thaw cycles. Genomic DNA was extracted from rice (*Oryza sativa* Poaceae) and then effects of freeze–thaw cycles and various temperatures on both the utility and stability of the DNA was studied (Kasem, 2004). The results of this study are presented in Tables 13.1 and 13.2, and it is interesting to note that DNA was able to go through around 70 cycles of freezing and thawing and still maintain reasonable integrity in terms of molecular weight (Table 13.2). The DNA was suitable for PCR although it is worth noting that only relatively small target regions were amplified (250–400 bp).

When stored at 50°C the genomic DNA sample degraded relatively quickly and no low or high molecular DNA fragments were visible on an agarose gel after 72 h (Table 13.1). Despite the rapid loss of integrity which was visible through an agarose gel, the DNA was still useful for PCR-based analysis up to day 7 for the microsatellite marker and up to day 17 for the GBSSI fragment (Table 13.1). These results indicate that despite exposure to non-ideal storage conditions the downstream utility of the DNA was not entirely lost.

It was expected that storage at low temperatures, −20°C and −80°C would be ideal to ensure minimal to no degradation and the results confirmed this for the period of the study (Table 13.1). Similarly DNA stored at 4°C also showed no loss of quality during the 16-week study period. The extracted DNA was also stored in solution at room temperature and was stable for 24 days before some degradation was noticeable (Table 13.1).

These results have important implications for how we may choose to handle DNA samples in relation to what we perceive as good laboratory practice and also place some importance on the establishment of a regular monitoring programme. This is particularly important for DNA Banks and in relation to our understanding of how quality may influence the end use of the DNA sample held. At the Australian Plant DNA Bank, we are willing to accept extracted DNA samples from donors. In doing this, we have little control over the quality of the sample entering the collection. This is not ideal but in the case of rare and extinct taxa we may have very few options in relation to access to genomic material. DNA samples which we know are of a lower quality than others can be flagged as suitable for only a smaller number of techniques, e.g. they may be suitable for certain PCR-based assays targeting small fragments but not suitable for restriction enzyme-based techniques and AFLPs.

Logistics of Sample Withdrawals

Currently DNA samples from the Australian Plant DNA Bank can be sent either in solution or lyophilized. Within Australia we have sent the samples

Table 13.1. Effects of temperatures and storage durations on DNA quality. (From Kasem, 2004, p. 241.)

Temperature	Duration	DNA quality in 1% agarose gel	Microsatellite PCR	GBSSI PCR
High temperature 50°C	24 h	HMW DNA was visible	Satisfactory result up to 7 days of storage	Obtained amplification products till day 17 of storage
	48 h	Smeared		
	72 h	Totally degraded		
	96 h to 9 days	Severely degraded. Losing of signals in gel images		
Room temperature 20–22°C	1–20 days	Intact HMW DNA	Produced expected bands for more than 2 months	Produced expected bands for more than 2 months
	24–33 days	Minor smearing was visible		
	41–65 days	Progression of smearing was observed		
Cold temperature +4°C	1–16 weeks	No significant alteration was observed in high molecular mass with course of time	A high yield of a PCR product of the correct size during 4 months of study period	A high yield of a PCR product of the correct size during 4 months of study period
Low temperature –20°C	1–16 weeks	No effect on quality	High yield of expected bands from all samples stored for different lengths of time	High yield of expected bands from all samples stored for different lengths of time
Very low temperature –80°C	1–16 weeks	No effect on quality		

via an overnight courier service or by express post. This has alleviated the need to send samples frozen as they are guaranteed to arrive within a 24-h period. Orders for DNA samples in the collection can be made online via the webpage (www.dnabank.com.au).

Sample withdrawals and then distribution is monitored by a Material Transfer Agreement (MTA). This is necessary to ensure that the recipient of

Table 13.2. Effects of the freeze–thaw process on DNA quality. (From Kasem, 2004.)

Thawing temperature	Number of cycles	DNA quality in 1% agarose gel	Microsatellite PCR	GBSSI PCR
Thawing at room temperature (20–22°C)	Cycle 1–Cycle 20	Intact up to 20 cycles	Freeze–thaw treatment for 200 cycles at room temperature; high-yield good-quality PCR products from the samples of each cycle	Freeze–thaw treatment for 200 cycles at room temperature; high-yield good-quality PCR products from the samples of each cycle
	Cycle 25	Started to lose high molecular length		
	Cycle 30–Cycle 65	Progression of loss of high molecular length and increasing of smearing		
	Cycle 70	Totally smeared and degraded DNA was visualized		
	Cycle 75–Cycle 90	Progression of degradation and smearing		
	Cycle 95–Cycle 120	Total degradation of the integrity		
	Cycle 125–Cycle 200	Extensive damage of integrity and became less visible under UV light		
Thawing at moderately high temperature (37°C)	Cycle 1–Cycle 25	Intact DNA up to 25 cycles	Freeze–thaw treatment for 100 cycles at 37°C; satisfactory PCR products from each template of each cycle	Freeze–thaw treatment for 100 cycles at 37°C; satisfactory PCR products from each template of each cycle

Continued

Table 13.2. *Continued*

Thawing temperature	Number of cycles	DNA quality in 1% agarose gel	Microsatellite PCR	GBSSI PCR
	Cycle 30	Showed a little tendency towards smear to low molecular weight		
	Cycle 35–Cycle 45	Low degree of smearing		
	Cycle 50–Cycle 70	Loss of high molecular weight		
	Cycle 70–Cycle 80	Moderate smearing		
	Cycle 85–Cycle100	Degraded and smeared DNA		

the DNA sample adheres to the international and local legislation and agreements which guide our operations. Within Australia, each state has developed its own policies concerning access and use of biological resources in line with the UN Convention on Biological Diversity (UNCED, 1992). Two Australian states, Queensland (Biodiscovery Act 2004 No. 19, 2004) and Northern Territory (Biological Resources Bill 2006, 2006), have developed legislation which often place conditions on how we transfer any DNA samples derived from samples collected within the State boundaries. It is necessary for access to biological resources to be governed by such agreements, policies and legalization so as to ensure that over collection of a species in the wild does not cause irreversible damage to the species or the ecosystem. It is to this point that a DNA collection can assist in the protection of fragile ecosystems and the species within.

There is a need for the development and adoption of standards for DNA Banking (Rice *et al*., 2006b; Hodkinson *et al*., 2007). These standards would need to include many aspects but one immediate problem is the possible exhaustion of genomic DNA samples. This is particularly a problem when it is no longer possible to repeat the extraction from fresh or stored tissue. One possible solution, although it will need thorough investigation, is the use of whole genome amplification methods to provide a copy of the original DNA sample to the requestor (Rice *et al*., 2006b).

Bioinformatics and Sample Tracking

The Australian Plant DNA Bank has custom-built relational tables which are interfaced to the World Wide Web (www.dnabank.com.au). These relational tables are now embedded in a wider in-house Laboratory Information System

(LIMS). The advantage this gives is that the DNA Bank can now track the movement of samples from the collection through to request from an in-house user. The sample can then be placed back in the collection and its history of use be flagged prior to assignment to another researcher. This allows more efficient use of individual aliquots of DNA by the associated research group. It also allows better management of sample use in today's climate of legal logistics, i.e. where each sample is monitored by an MTA which prohibits transfer.

External requestors of DNA samples can still access the collection via the World Wide Web catalogue. It is highly appreciated if the users of the collection can filter results in the form of publications and web links back to the curator so that the information about each sample can be added to making the collection more valuable.

Wherever possible, links to the donor's or donor institution's webpage are held with each accession. This has the advantage of allowing valuable links through to herbaria which may hold the voucher specimen, seed banks which may hold the seed from which the sample was derived and databases like Genbank which hold the results of molecular projects.

It is a long-term goal of the Australian Plant DNA Bank to ensure that a network of collections exist throughout Australia. This model is similar to the original global network proposed by Adams (1997). This concept is well supported by all collaborative groups and is currently progressing. This would allow facilitated access to not only a wide range of DNA samples but would also mean that the taxonomic experts would be directly contributing to nodes within the regions that held the diverse taxa at the centre of their interests.

DNA Banks and Their Role in the Conservation of Plant Genes

A DNA Bank conserves genomic DNA or stores tissue for the purpose of extracting genomic DNA. The collection cannot be readily used for re-creation of an extinct species using conventional plant breeding methods. Despite the current limitation of stored DNA not being easily used to regenerate plants (Graner *et al*., 2006; Rice *et al*., 2006b), a future change in the techniques available may enable the use of DNA for regeneration of species. It is possible that DNA Banks will be a source of DNA for use in transformation and cloning (Godwin, 1997; Rabiya, 2000). The DNA Bank approach to the storage of plant tissues and genomic extracts does enable relatively secure, efficient and cheap storage of large numbers of samples (Brown *et al*., 1997). These large collections of plant materials and genomic extracts can then be used as a basis for plant molecular research and allows a central access point. The advantages of storing plant material as part of a plant DNA Bank collection are threefold (Adams, 1997):

1. The collection provides a constant supply of material for future genomic extractions.
2. Extraction of genomic DNA can be prioritized and the collection can still grow through the incorporation of new plant samples.

3. If the plant material is to be stored with the aim of being used for extraction of DNA, then it will be preserved to facilitate downstream applications and may allow for improvements in techniques.

Preserving only the isolated genomic DNA, rather than combinations of plant tissue and genomic DNA, may be more space efficient and less subject to long-term degradation but it is not feasible to consider that a collection of plant DNA extracts will exist without the original plant material from which it was derived. The best practice is for the DNA samples to be linked and supported by the relevant herbarium specimens. It can be argued that all molecular studies when they publish sequence data should show a link to a herbarium voucher for the original plant sample and a link to a DNA Bank where the sample lodged is now a 'DNA' voucher.

It has been highlighted that preserving the DNA of species is in essence about the storage of information (Mattick *et al*., 1992). Mattick *et al*. (1992) estimated that the genetic variation in different species is in the range of 1 million–10 million bp and that a DNA Bank holding the 10 million species on Earth would have an associated 1015 pieces of information contained in the database. It is now demonstrated that phylogenotypic diversity within species is a more sensitive and significant marker of changing biodiversity than the measures of species richness alone (Forest *et al*., 2007). Therefore, plant DNA Banks have an enormous potential to collect and conserve a great deal of information about global plant genetic diversity and in the future it is possible that these collections will be 'molecular snapshots' which can be dated to the time that the samples were collected. The ability to interrogate temporal DNA samples is a powerful and novel tool for ongoing monitoring of the impact of environmental (and particularly climate) change on biodiversity.

DNA Collections as a Tool for Plant Breeding

One of the main objectives of any *ex situ* collection is species conservation and therefore it is also important not to overlook the need to conserve species to support plant breeding programmes. Molecular analyses play an important role in the management of germplasm collections and in the selection of material for breeding programmes, which also includes crop wild relatives (Richards, 2004). Recent examples of DNA analysis of diversity of plant germplasm associated with the Australian Plant DNA Bank include studies of sugarcane (Cordeiro *et al*., 2003), barley (Bundock and Henry, 2004) and sorghum (Dillon *et al*., 2005).

Protection of crop wild relatives has received little attention but they represent extremely valuable resources which are critical to world food security (Henry, 2005). Domestication of crop species has led to a loss of genetic diversity in the cultivated varieties leaving a gene pool with little potential to provide resistance to biotic and abiotic stresses (Nevo, 2005). This is particularly relevant to current changes in environmental pressures and predicted climatic shifts. Molecular studies on wild barley genetic resources, *Hordeum*

spontaneum, from three countries have been studied and the results indicate that this material is genetically more diverse than landraces (Nevo, 2005). By using DNA to determine phylogenetic relationships it is also possible to identify the closest potential sources for new genes in plant breeding programmes. Molecular studies in the Australian Vitaceae identified that *Cissus hypoclauca* and *Cissus sterculifolia* are closely related to, and as a result potentially suitable for improvement of, cultivated grapes (Rossetto *et al*., 2001). Similarly, Australian native *Sorghum* spp. have been identified as a potentially good source of new genes for improvement of *Sorghum bicolor* (Dillon *et al*., 2001). Phylogenetic studies present novel options for extending the crop gene pool of this important species via wide-crossing (Rice *et al*., 2006a). DNA Banks when linked with traditional germplasm resources provide a valuable tool for the evaluation of suitable material for breeding programmes. They provide a mechanism for molecular biology to mine the DNA sequence thus assisting the selection of suitable material from core collections and wild relatives and can be useful in pre-emptive breeding programmes.

Molecular-based Plant Identification with an Application to Protection of Intellectual Property

The application of molecular-based techniques to the plant genome has resulted in a significant increase in knowledge of within and between species variation. It is the understanding of this variation which leads to its broader application in relation to the protection of Intellectual Property (IP) and even forensic investigation. The protection of IP is particularly important for varieties and cultivars which have some economic importance. The registration of new varieties is often based upon agronomic and morphological characteristics. Molecular-based DNA profiling techniques do have an application in the protection of IP in relation to unauthorized commercial use. The success of these DNA profiles will largely depend on the ability of the testing laboratory to access verified reference samples and this could be an issue if at the time of registration a reference sample was not lodged with an appropriate genetic resource collection. Random amplified polymorphic DNAs (RAPDs) have been used to support a legal case in relation to the unauthorized commercialization of a patented strawberry variety and the results were accepted as evidence in court (Congiu *et al*., 2000). There is sufficient support for the submission of a sample for DNA storage in a collection such as a DNA Bank which could be called on as a reference sample for the identification of a variety and thereby protection of IP.

The application of molecular-based marker techniques to forensic science is an area which is receiving more attention and will continue to grow. The ability for DNA-based evidence to provide information about a crime scene is reality. Microsatellite markers have been used to compare plant samples from a suspect's car to live trees at a crime scene (Craft *et al*., 2007). In this particular case, the leaves found in the car gave a different profile from the live trees at the crime scene and hence the evidence could not be used against the suspect

(Craft *et al*., 2007). It is also possible to use genes which are more routinely used for phylogenetics and systematics such as *rbc*L, *trn*L intron, *adh*1, *mat*K, *chl*L and *trn*L-*trn*F (Clegg and Zurawski, 1991; Clegg *et al*., 1991; Taberlet *et al*., 1991; Buckler and Holtsford, 1996; Kajita *et al*., 1998; Kusumi *et al*., 2000; Pacey-Miller and Henry, 2003) to identify the plant species. Ideally these regions could be used to generate a unique DNA barcode (Kress *et al*., 2005). A case example of a well-conserved single copy nuclear gene is the grass family granule-bound starch synthase 1 (GBSS1). Comparison of the entire GBSS1 coding region between closely related species has revealed 3–5% polymorphism, which is sufficient for species discrimination (McIntosh *et al*., 2005). The use of DNA profiles to identify the species in the protection of IP is an area which will benefit from DNA Banks as a means of gaining access to reference samples which will aid the construction of molecular data sets as evidence.

Conclusion

Despite their obvious research and conservation value, DNA Banks continue to be funded in an ad hoc manner and cannot survive without the support of a larger research group or institution. The current DNA Banks have been criticized as operating independently from one another (Hodkinson *et al*., 2007). While this is the case, it is not necessarily intentional. Rather this probably reflects the relatively low profile that DNA Banks have as an *ex situ* genetic resource.

Ideally, it would be advantageous for DNA Banks to operate as a global network, similar to what was originally proposed by Adams *et al*. (1997). Within Australia we are attempting to establish more formalized networks of DNA Banks and this will capture existing efforts in herbaria, botanic gardens and research institutions. Molecular data is increasingly being used to support reclassification of taxonomic groups. With the recent update of the angiosperm, phylogeny based on DNA evidence (APGII2003, 2003) leads to suggestions that DNA samples within DNA Banks may act in a similar way to a voucher specimen in herbaria, and strengthens the need to ensure that strong collaborative links exist.

To ensure the continued growth of DNA Banks research into the long-term effects of storage on DNA functionality need to be continued. At the moment we can only speculate as to the length of time that DNA can be stored at −80°C and the increasing number of room temperature storage techniques need to be independently evaluated for all biological systems.

References

Adams, R.P. (1997) Conservation of DNA: DNA Banking. In: Callow, J.A., Ford-Lloyd, B.V. and Newbury, H.J. (eds) *Biotechnology and Plant Genetic Resources Conservation and Use*. CAB International, Wallingford, UK, pp. 163–174.

Andersson, M.S., Fuquen, E.M. and de Vicente, C.M. (2006) *DNA Banks – Providing Novel Options for Genebanks?* Topical Reviews in Agricultural Biodiversity, International Plant Genetic Resources Institute, Rome, Italy.

APGII2003 (2003) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. *Botanical Journal of the Linnean Society* 141, 399–436.

Biodiscovery Act 2004 No. 19 (2004) Office of the Queensland Parliamentary Counsel, Queensland, Australia.

Biological Resources Bill 2006 (2006) Northern Territory of Australia, Northern Territory, Australia.

Brown, A.H.D., Brubaker, C.L. and Grace, J.P. (1997) Regeneration of germplasm samples: wild versus cultivated plant species (Germplasm regeneration: developments in population genetics and their implications). *Crop Science* 37, 7–13.

Buckler, E. and Holtsford, T. (1996) Zea systematics: ribosomal ITS evidence. *Molecular Biology and Evolution* 13, 612–622.

Bundock, P.C. and Henry, R.J. (2004) Single nucleotide polymorphism haplotype diversity and recombination in the *isa* gene of barley. *Theoretical and Applied Genetics* 109, 543–551.

Clegg, M. and Zurawski, G. (1991) Chloroplast DNA and the study of plant phylogeny. In: Soltis, P.S., Soltis, D.E. and Doyle, J.J. (eds.) *Molecular Systematics of Plants*. Chapman & Hall, New York, pp. 1–13.

Clegg, M.T., Learn, G.H. and Golenberg, E.M. (1991) Molecular evolution of chloroplast DNA. In: Selander, R.K., Clark, A.G. and Whittam, T.C. (eds) *Evolution at the Molecular Level*. Sinauer, Sunderland, Massachusetts, pp. 135–149.

Congiu, L., Chicca, M., Cella, R., Rossi, R. and Bernacchia, G. (2000) The use of random amplified polymorphic DNA (RAPD) markers to identify strawberry varieties: a forensic application. *Molecular Ecology* 9, 229–232.

Cordeiro, G.M., Pan, Y.-B. and Henry, R.J. (2003) Sugarcane microsatellites for the assessment of genetic diversity in sugarcane germplasm. *Plant Science* 165, 181–189.

Craft, K.J., Owens, J.D. and Ashley, M.V. (2007) Application of plant DNA markers in forensic botany: genetic comparison of Quercus evidence leaves to crime scene trees using microsatellites. *Forensic Science Journal* 165, 64–70.

Dillon, S.L., Lawrence, P.K. and Henry, R.J. (2001) The use of ribosomal ITS to determine phylogenetic relationships within *Sorghum*. *Plant Systematics and Evolution* 230, 97–110.

Dillon, S.L., Lawrence, P.K. and Henry, R.J. (2005) The new use of *Sorghum bicolor*-derived SSR markers to evaluate genetic diversity in 17 Australian *Sorghum* species. *Plant Genetic Resources* 3, 19–28.

Forest, F., Grenyer, R., Rouget, M., Davies, T.J., Cowling, R.M., Faith, D.P., Balmford, A., Manning, J.C., Proches, S., van der Bank, M., Reeves, G., Hedderson, T.A.J. and Savolainen, V. (2007) Preserving the evolutionary potential of floras in biodiversity hotspots. *Nature* 445, 757–760.

Godwin, I.D. (1997) Gene identification, isolation and transfer. In: Callow, J.A., Ford-Lloyd, B.V. and Newbury, H.J. (eds) *Biotechnology and Plant Genetic Resources Conservation and Use*. CAB International, Wallingford, UK, pp. 203–234.

Graner, A., Andersson, M.S. and de Vicente, M.C. (2006) A model for DNA banking to enhance the management, distribution and use of ex situ stored PGR. In: de Vincente, M.C. (ed.) *DNA Banks – Providing Novel Options for Genebanks?* Topical Reviews in Agricultural Biodiversity, International Plant Genetic Resources Institute, Rome, Italy.

Henry, R.J. (2005) Conserving genetic diversity in plants of environmental, social or economic importance. In: Henry, R.J. (ed.) *Plant Diversity and Evolution*. CAB International, Wallingford, UK, pp. 317–325.

Hodkinson, T.R., Stephen Waldren, S., Parnell, J.A.N., Kelleher, C.T., Salamin, K. and Salamin, N. (2007) DNA banking for plant breeding, biotechnology and biodiversity evaluation. *Journal of Plant Research* 120, 17–29.

Kajita, T., Kamiya, K., Nakamura, K., Tachida, H., Wickneswari, R., Tsumura, Y., Yoshimaru, H. and Yamazaki, T. (1998) Molecular phylogeny of dipetrocarpaceae in Southeast Asia based on nucleotide sequences of matK, trnL intron, and trnL-trnF intergenic spacer region in chloroplast DNA. *Molecular Phylogenetics and Evolution* 10, 202–209.

Kasem, S. (2004) *The Stability of Stored DNA*. Southern Cross University, Lismore, New South Wales, Australia, Master of Science Thesis, 75 pp.

Kress, W.J., Wurdack, K.J., Zimmer, E.A., Weight, L.A. and Janzen, D.H. (2005) Use of barcodes to identify flowering plants. *PNAS* 102, 8369–8374.

Kusumi, J., Tsumara, Y., Yoshimaru, H. and Tachida, H. (2000) Phylogenetic relationships in Taxodiaceae and Cupressaceae sensu stricto based on matK gene, chlL gene, trnL-trnF IGS region, and trnL intron sequences. *American Journal of Botany* 87, 1480–1488.

Mattick, J.S., Ablett, E.M. and Edmonson, D.L. (1992) The gene library: preservation and analysis of genetic diversity in Australasia. In: Adam, R.P. and Adms, J.E. (eds) *Conservation of Plant Genes: DNA Banking and* In Vitro *Biotechnology*. Academic Press Inc., Royal Botanic Gardens, Kew, UK.

McIntosh, S., Pacey-Miller, T. and Henry, R. (2005) A universal protocol for identification of cereals. *Journal of Cereal Science* 41, 37–46.

MolecularProbes (2005) Quant-iT PicoGreen dsDNA Reagent and Kits. Invitrogen.

Nevo, E. (2005) Genomic diversity in nature and domestication. In: Henry, R.J. (ed.) *Plant Diversity and Evolution*. CAB International, Wallingford, UK, pp. 287–316.

Pacey-Miller, T. and Henry, R. (2003) Single-nucleotide polymorphism detection in plants using a single stranded pyrosequencing protocol with a universal biotynylated primer. *Analytical Biochemistry* 317, 165–170.

Rabiya, S. (2000) DNA Banks Noah's Ark at –200°C. *HMS Beagle* 84.

Rice, N., Cordeiro, G.M., Shepherd, M., Bundock, P.C., Bradbury, L.M.T., Pacey-Miller, T., Furtado, A. and Henry, R.J. (2006a) DNA banks and their role in facilitating the application of genomics to plant germplasm. *Plant Genetic Resources* 4, 64–70.

Rice, N., Henry, R.J. and Rossetto, M. (2006b) DNA banks: a primary resource for conservation research. In: de Vincente, M.C. (ed.) *DNA Banks – Providing Novel Options for Genebanks?* Topical Reviews in Agricultural Biodiversity, International Plant Genetic Resources Institute, Rome, Italy.

Richard, C. (2004) Molecular technologies for managing and using genebank collections. In: deVicente, M.C. (ed.) The evolving role of gene banks in the East developing field of molecular genetics. *Issues in Genetic Resources No. 11 International Plant Genetic Resources Institute (IPGRI)*, Rome, Italy, pp. 13–18.

Rossetto, M., Jackes, B.R., Scott, K.D. and Henry, R.J. (2001) Intergeneric relationships in the Australian Vitaceae: new evidence from cpDNA analysis. *Genetic Resources and Crop Evolution* 48, 307–314.

Taberlet, P., Gielly, L., Patou, G. and Bouvet, J. (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. *Plant Molecular Biology* 17, 1105–1109.

UNCED (1992) Convention on biological diversity. *United Nations Conference on Environment and Development*. United Nations, Geneva.

14 DNA Extraction from Plant Tissue

S. Kasem, N. Rice and R.J. Henry

Introduction

Access to genomic DNA, of high quality, for downstream analyses requires isolation from plant tissue. Currently used protocols vary from very simple extractions (Thomson and Henry, 1995) to slightly more complex procedures (Graham *et al*., 1994) delivering a greater quantity or purity of DNA. Numerous protocols have been published and the subtle and large changes in the protocols can be partly attributed to the species-level differences between the type and concentration of secondary metabolites (Aljanabi and Martinez, 1997; Stein *et al*., 2001). The post-extraction analytical techniques are often what determine the choice of method for DNA isolation (Chaves *et al*., 1995; Strozycki and Legocki, 1995; Henry, 1997). Each analysis requires a different level of quality and quantity of extracted DNA. The polymerase chain reaction (PCR) applied to allele-specific PCR or simple sequence repeat (SSR) analysis requires small quantities of DNA and can be performed with crude or degraded DNA (Reiss *et al*., 1995; Weising *et al*., 1995). In contrast, PCR-based markers such as randomly amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP) and inter-simple sequence repeats (ISSR) require small amounts of high-quality DNA (Henry, 2001a). Highly purified high molecular weight DNA and larger quantities are needed for end labelling and southern blotting (Stein, 1993), and for restriction fragment length polymorphism (RFLP), genomic library construction and DNA cloning protocols (Rogstad *et al*., 2001). This chapter will discuss the approaches which have been used in the isolation of DNA from plant tissues and a summary of approaches is presented in Appendix 14 (Tables 14.1A–14.12A).

Plant Tissue

DNA of good quality can be isolated from various parts of a fresh plant, herbarium specimens or even from ancient plant sources (Deguilloux *et al*., 2002)

 Plant Genotyping II: SNP Technology (ed. R.J. Henry)

(Table 14.1A). Fresh leaf is the preferred tissue for DNA extraction as they contain low concentration of metabolites and polysaccharides (Jobes *et al*., 1995). It is generally accepted that seeds contain moderate to less nucleic acids, in comparison to embryos, leaves and cotyledons (Rogers and Bendich, 1985; Rollo *et al*., 1987). The influence of the leaf maturity on the yield of DNA is dependent on both species and tissue with young leaf producing a higher yield of DNA when compared to extraction from other plant tissues (Asemota, 1995). In contrast, more mature leaf has been reported as yielding higher concentrations of DNA in sweet potato (Varadarajan and Prakash, 1991) and *Vitis* (Lodhi *et al*., 1994). The starch-rich endosperm of seeds usually produces a low yield of DNA due to the endosperm composition creating difficulties in extracting DNA, but the starch storage tissue polyploid cotyledons of pea and broad bean provide good amounts of DNA (Rogers and Bendich, 1985). Dry plant tissue can also be a good source of DNA (Thomson and Henry, 1993).

DNA Extraction Techniques

Although there are many published DNA isolation and purification techniques these methods have common elements (Fig. 14.1) including disruption of the tissues (Table 14.2A), DNA release into extraction buffer and purification of the DNA molecule (Fig. 14.1). Even species from the same genera may require a different isolation protocol (Weising *et al*., 1995). These protocols differ in time, quality of the final product (Milligan, 1998) and the reagents in the isolation buffer (Tables 14.3A–14.12A).

Tissue Disruption

The disruption of the tissue is the first step in the process for extraction of the DNA molecule. Several approaches have been published for the effective tissue disruption and the choice is dependent upon the resultant quality of DNA, cost, labour, sample type, sample size and reliability of the grinding process. The quality and the quantity of the extracted DNA are influenced by the proportion of the disrupted cell walls and the efficiency of the method chosen to break the cell walls (Jhingan, 1992). Plant cells contain a rigid cell wall which needs shear forces to break it down and to release DNA in the extraction solution (Harborne, 1989; Denijs *et al*., 1996). The grinding of the tissue is a delicate procedure as more grinding will increase the overall yield of DNA but it increases the chance of breaking the high molecular weight DNA into shorter fragments (Boffey, 1986; Richards, 1988; Rogstad *et al*., 2001). Ultimately, the method must use a gentle force to quickly and efficiently disrupt the cell membrane (Boffey, 1986; Richards, 1988). In order to minimize the degradation from photolytic activity, tissue disruption is usually done at low temperatures (Murray and Thompson, 1980; Lefort and Douglas, 1999).

Tissue disruption in a ceramic mortar using a heavy round-ended pestle by applying physical forces (Harborne, 1989) is the most common method to grind

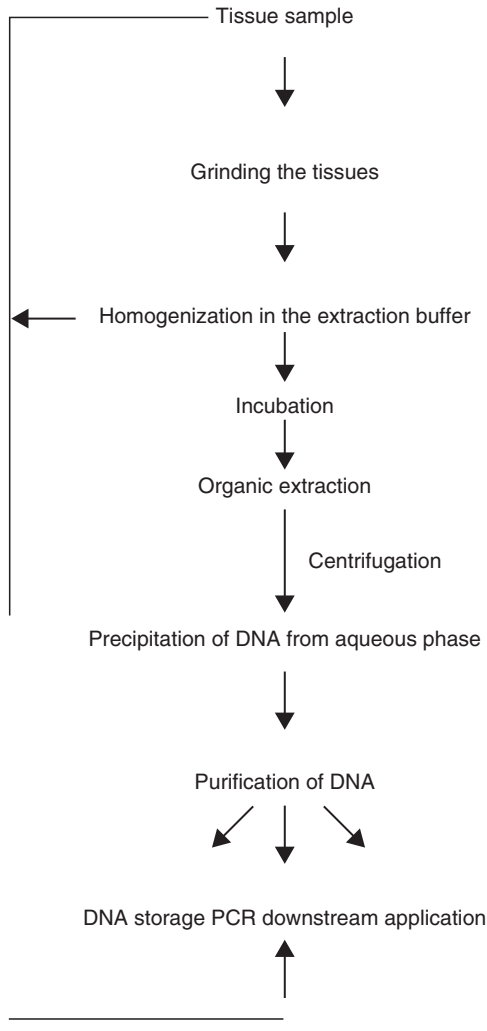


Fig. 14.1. Flow diagram of a typical plant DNA extraction procedure.

fresh, frozen, dried or lyophilized tissues. Lyophilized tissues are easy to grind and therefore reduce grinding time (Guidet, 1994). Usually tissue pulverization by mortar and pestle is done in the presence of liquid nitrogen which minimizes oxidation of polyphenols, inhibits nuclease activity and aids the grinding process (Murray and Thompson, 1980; Weising *et al*., 1995; Lefort and Douglas, 1999). Mondragon *et al*. (2000) suggested that during grinding of succulent tissue, continuous addition of liquid nitrogen decreases mucilage secretion resulting in easier manipulation of the samples at a later stage. A mortar and pestle has always been suggested for grinding starch-rich tubers (Asemota, 1995). Dry ice can be used as an alternative to liquid nitrogen, but crushing the tissue is more difficult (Couch and Fritz, 1990; Stein, 1993). It has been reported that a

combination of liquid nitrogen and glass powder or sterile sand is useful (Murray and Thompson, 1980; Katterman and Shattuck, 1983; Guillermaut and Marchal-Drouard, 1992; Ouenzar *et al*., 1998; Woodhead *et al*., 1998). Collision between cell wall and the solid abrasive (sand/glass) enhances cell disruption, facilitates smooth grinding and thus ensures fine powder from tough and fibrous tissues (Murray and Thompson, 1980; Towner, 1991; Guillermaut and Marchal-Drouard, 1992; Aitchitt *et al*., 1993; Weising *et al*., 1995; Ouenzar *et al*., 1998; Hu and Zhou, 2001). When liquid nitrogen is unavailable pre-cooled plant tissue can be ground in pre-cooled extraction buffer using a mortar and pestle (Ouenzar *et al*., 1998). For fibrous material pre-cutting of the leaves with scissors is recommended (Weising *et al*., 1995). For hard tissues overnight soaking in water facilitates easier grinding using a mortar and pestle (Rogers and Bendich, 1988). When mortar and pestle is chosen for tissue disruption the DNA yields become dependent on the force applied by the person carrying out the experiment (Denijs *et al*., 1996), with excessive mechanical force causing shearing of the DNA molecule (Lodhi *et al*., 1994). The use of a mortar and pestle makes this disruption technique a limiting step when a large number of samples are to be processed in combination with the difficult nature of some tissues which cannot be ground adequately using a mortar and pestle. To avoid the shortcomings of a mortar and pestle, various commercial mechanical disruption techniques are now being used (Table 14.2A).

Jhingan (1992) and Williams and Ronald (1994) reported a non-grinding DNA isolation protocol which relies on the compound potassium ethyl xanthogenate to release DNA from leaves without tissue homogenization. This fast and simple procedure does not involve any deproteination and is able to extract clean DNA suitable for PCR amplification and Southern blotting (Jhingan, 1992) from various types of genera. The main feature of this technique is the incubation of the tissues with the extraction buffer, centrifugation to remove the cell debris, precipitation with ethanol, resuspension in TE and re-precipitation with ethanol (Jhingan, 1992). If necessary, chloroform extraction step may also be included (Milligan, 1998).

Extraction Buffer Content

The first obvious step in a plant DNA extraction protocol is to disrupt cell membrane in order to allow the release of the cellular constituents including DNA in the extraction medium (Rogers and Bendich, 1988; Towner, 1991). Extraction buffer/lysis buffer generally contains a mixture of the following:

1. Detergents which allow the release of membrane-bound DNA;
2. Buffering agent, e.g. Tris (pH 7.0–8.0), to control pH of the extraction solution;
3. Chelating agent – EDTA to inactivate nuclease activity;
4. Salt to solubilize DNA and other molecules in the lysis buffer;
5. Reducing agent, e.g. 2-mercaptoethanol (2-ME), DTT to prevent peroxidase or polyphenoloxidase activities, added to extraction buffer just before use;

6. Other compounds to remove contaminants depending upon the species under study (Henry, 2001b; Sosa and Oliveira, 1992).

Detergents

The use of detergents in a DNA extraction protocol results in a higher yield of DNA, superior quality and improved longevity of preserved DNA compared to isolation procedures which rely on only tissue grinding in a buffer or ion chelating agent (Rollo *et al*., 1994). Detergents assist with the disruption of tissues by removing the lipid molecule from the cell membrane and allowing DNA and other cellular material to be released into the solution (Brown, 2001; Henry, 2001a). Different protocols utilize different detergents to extract plant DNA (Henry, 2001a). The most commonly used detergents are:

- Cetyltrimethyl ammonium bromide (CTAB);
- Sodium dodecyl sulphate (SDS);
- Sarkosyl/*N*-lauroyl sarcosine;
- Triton X-100; and
- commercial biological laundry detergent (Persil Megaperls and Forsch).

Cationic CTAB is widely used in plant DNA isolation protocols and has the ability to work on various types of fresh tissues (e.g. seedlings, leaves, cotyledons, seeds, grains, endosperm, embryos, tissue culture callus and pollen tissues) and preserved plant tissues (i.e. frozen, lyophilized and dehydrated) (Richards, 1988). CTAB produces higher yields than other detergents from any type of starting material (Huang *et al*., 2000). It not only releases DNA from cellular membranes and proteins (Rogers and Bendich, 1985) but also separates polysaccharides from the extracted DNA molecule (Pirttila *et al*., 2001). At low salt concentration the positively charged CTAB solubilizes the cell membrane and forms a complex with negatively charged DNA, later the DNA can be precipitated from this complex mixture by increasing the salt concentration (Towner, 1991). During homogenization of *Cuphea* tissue, Webb and Knapp (1990) used a very low concentration of CTAB (0.5% w/v) and found that at this low concentration, CTAB does not lose its efficiency or precipitation ability. In addition, low concentrations of CTAB assist in reduction of foaming in the homogenizer, minimize viscosity of the solution and decrease centrifugation time and speed. Higher concentrations of CTAB are recommended for polysaccharide-rich plant tissues (Doyle and Doyle, 1990; Stewart and Via, 1993; De la Cruz *et al*., 1997). However, Chaudhry *et al*. (1999) stated that higher concentrations of CTAB may decrease the DNA yield in cotton leaves. Another study indicated that 2% CTAB (w/v) is able to produce a good yield but 3% CTAB (w/v) gave DNA of good quality that was more readily digested by restriction enzymes (Steenkamp *et al*., 1994). Weir *et al*. (1996) reported that among several different detergents, CTAB was the most efficient in producing high molecular weight DNA from various fruit trees.

Anionic detergents like SDS and sarkosyl have the ability to break cell membranes and to denature proteins (Hong *et al*., 1995; Dey *et al*., 1997). SDS

is commonly used in combination with potassium acetate (KOAc), phenol or proteinase K (Towner, 1991). Kang *et al*. (1998) found that SDS was the most efficient detergent to extract DNA from dry seed. Direct addition of SDS and sarkosyl detergents with an extraction media causes premature cell lysis (Richards, 1988; Jobes *et al*., 1995) and shearing of the DNA (Manning, 1991) molecule; therefore, it is suggested to add them after resuspension of the tissue in an extraction media (Manning, 1991; Jobes *et al*., 1995; Bekesiova *et al*., 1999). Later addition of SDS provides cleaner DNA than direct addition with extraction buffer (Jobes *et al*., 1995) because prior to the DNA being released from the tissue into the solution, the detergent molecule will bind directly to organelle polyphenols during incubation. Collins and Symons (1992) found that among the detergents like SDS or CTAB, only sarkosyl was able to lyse nuclear membrane of grapevine leaves, and they also suggested that during precipitation by ethanol a low pH avoids the co-precipitation of sarkosyl.

A combination of detergents can lead to the efficient formation of insoluble complexes allowing elimination of protein and polysaccharides (Lefort and Douglas, 1999). For this reason, several studies used a mixture of SDS and CTAB (De la Cruz *et al*., 1995, 1997; Jobes *et al*., 1995; Lim *et al*., 1997; Lefort and Douglas, 1999). Mondragon *et al*. (2000) reported that the combination of CTAB and SDS was effective in the elimination of polysaccharide from mucilage-containing tissues. Steenkamp *et al*. (1994) stated that the combination of CTAB and SDS reduced the concentration of polysaccharide but SDS can cause interference in the endonuclease digestion and inhibit the *Taq* polymerase enzyme in PCR amplification (Blanchard and Nowotny, 1994). Combinations of CTAB and sarkosyl have also been used to lyse nuclei of different plant species (Blanchard and Nowotny, 1994; Fulton *et al*., 1995; Scott and Playford, 1996; Bekesiova *et al*., 1999; Chaudhry *et al*., 1999; Li *et al*., 2001).

The non-ionic detergent Triton X-100 is usually used to lyse organelle membranes in the nuclear DNA isolation procedure (Katterman and Shattuck, 1983; Collins and Symons, 1992; Peterson *et al*., 1997). Triton X-100 does not destroy enzymes or proteins (Dey *et al*., 1997) but can protect nuclear integrity while lysing both the chloroplast and mitochondrial membranes (Jofuku and Goldberg, 1988; Peterson *et al*., 1997). Guidet *et al*. (1990) reported that they were able to achieve high molecular weight nuclear DNA when they added Triton X-100 in the buffer. Commercial washing powder (i.e. Sunil, Omo, Persil Megaperls, Dash) contains a mixture of detergents, enzymes and chelating agents and thus can be an alternative of any other standard extraction buffer and can effectively replace any conventional detergent like SDS and CTAB (Bahl and Pfenninger, 1996; Pusch, 1997; Nordell *et al*., 1999). These studies stated that when compared to other conventional buffer systems no detectable difference was found in the quality and quantity of extracted DNA and they were able to achieve high molecular weight DNA suitable for PCR amplification and restriction enzyme digestion.

Buffering agent

Usually 100 mM Tris HCl at pH 7.0–8.0 is used as a buffering agent in most studies. A study of Steenkamp *et al*. (1994) indicated that optimum DNA from *Vitis* species was achieved when the molarity of Tris HCl in the extraction buffer was raised from 100 mM to 1 M for 3% of CTAB at pH 8.0. This study also stated that buffering capacity of Tris HCl increased when the molarity increased but at the same time a molarity higher than 1.5 produced yellow pigments and the DNA was more problematic to dissolve in TE buffer. An extraction buffer containing Tris and Boric acid (pH 7.6) is effective in the elimination of polyphenols and other secondary metabolites as a borate complex in strawberry leaves (Manning, 1991).

Katterman and Shattuck (1983) first adopted a low pH citrate buffer to isolate nuclei. A low pH citrate buffer enhances the binding capability of the nuclear components to one another and maintains nuclear integrity in the whole procedure (Dounce, 1955). A low pH citrate buffer can also prevent precipitation of carbohydrate and eliminate polyphenols (Mercado *et al*., 1999). Many studies (Couch and Fritz, 1990; Guillermaut and Marchal-Drouard, 1992; Rether *et al*., 1993; Ziegenhagen *et al*., 1993; Jobes *et al*., 1995) have reported that a low pH (pH 5.5) extraction buffer containing sodium citrate/sodium acetate prevents ionization and polyphenol oxidation and is useful to precipitate cell debris and other contaminants from various polysaccharide and polyphenol-rich species (Guillermaut and Marchal-Drouard, 1992).

Inhibition of nuclease activities

Nuclease activity can cause shearing of DNA molecules (Weising *et al*., 1995). Keeping tissue frozen during or prior to homogenization reduces the risk of nuclease-activated degradation of DNA molecules (Jofuku and Goldberg, 1988). Activities of endogenous nuclease are dependent on the pH of the extraction buffer (Jofuku and Goldberg, 1988). The pH of the lysis buffer needs to be monitored carefully, especially after the addition of the plant material (Steenkamp *et al*., 1994). Some of the nuclease enzymes (i.e. soybean) require neutral pH, therefore raising the pH of extraction buffer sometimes aids inhibition of nucleases (Jofuku and Goldberg, 1988). Magnesium (Mg^{2+}) ions are a cofactor for most of the nuclease activity (Weising *et al*., 1995; Milligan, 1998), therefore addition of bivalent ion chelator EDTA in extraction buffer prevents DNA degradation (Lahiri and Schnabel, 1993). Some studies have reported that without EDTA, no DNA is yielded (Hong *et al*., 1995; Rogstad *et al*., 2001), whereas higher concentrations of EDTA can cause removal of Mg^{2+} which is essential for restriction digestion enzyme activity (Hengen, 1994). Adams *et al*. (1999) and Flournoy *et al*. (1996) reported that ethanol can denature DNases irreversibly. Different plant species contain different kinds of DNases, for which it is sometimes difficult to inactivate all of them by using EDTA only; addition of a small amount of ethanol during grinding can solve this problem (Adams *et al*., 1999).

Stabilizer

In nuclei isolation protocols, the isolation buffer contains stabilizing agents to maintain nuclei integrity while lysing other cytoplasmic contaminants (Katterman and Shattuck, 1983). Several reducing sugars such as glucose (Katterman and Shattuck, 1983; Couch and Fritz, 1990; Collins and Symons, 1992; Paterson *et al*., 1993), sucrose (Sangwan *et al*., 1998), mannitol and sorbitol (Scott and Playford, 1996; Mercado *et al*., 1999; Li *et al*., 2001) have been used for this purpose. A large amount of reducing sugars in the extraction medium also ensures isolation of purified DNA (Katterman and Shattuck, 1983), by inhibiting polyphenol oxidase and preventing browning (Stein, 1993). Chaudhry *et al*. (1999) used both glucose and sorbitol in the extraction media and in the lysis buffer as a stabilizer, reducing agent and osmotica respectively. Li *et al*. (2001) found that sorbitol containing extraction buffer along with 40–60 s vortex mixing yielded more DNA from cotton leaves than any other reducing sugars containing extraction buffer. Polyamine spermine and spermidine (Rhonda *et al*., 1992) also serve as nuclear membrane stabilizer during nuclei isolation procedure (Jofuku and Goldberg, 1988). Hexylene glycol (2-methyl-2,4-pentanediol) was used as a stabilizer in nuclei extraction medium by Peterson *et al*. (1997).

Other chemicals

For some tissues, it was found that CTAB was unable to break cell membrane (Fu *et al*., 1998) and a satisfactory yield of DNA could not be obtained by conventional SDS or CTAB-based methods (Howland *et al*., 1991). The combination of urea with a low concentration of SDS has proven useful to disrupt the cell walls of dried roots or rhizomes (Fu *et al*., 1998), old roots of lupin (Strozycki and Legocki, 1995), older leaves of *Capsicum* spp. (Prince *et al*., 1997) and old, damaged birch tissues (Howland *et al*., 1991).

Enzymes

Potassium or sodium ethyl xanthogenate has also been used by Williams and Ronald (1994) to break down the rigid cell walls of plants. Plant cell walls consist of 90% polysaccharide and 10% protein (Heldt, 1997). Xanthates: dissolve cell walls (Williams and Ronald, 1994) by reacting with the hydroxyl group of polysaccharides, changing them to water-soluble polysaccharides (Jhingan, 1992); degrade cell wall proteins (Williams and Ronald, 1994) by reacting with their amino groups (Jhingan, 1992); and inhibit DNase activities by binding metal ions (Williams and Ronald, 1994) essential for enzyme activity.

Purification

Solvent purification

Solvent purification is usually conducted as a two-step process; the first step involves treating the cell extracts with an organic agent like phenol (Phe) or chloroform (Cho) or mixtures of phenol, chloroform and isoamylalcohol (IAA). The second step involves centrifugation of the lysis mixture (just after the addition of organic solvents in the cell extract) with the purpose of separating the DNA-containing aqueous phase from the organic phase, containing denatured protein and lipid contaminants. Phenol denatures protein and eliminates cellular debris; chloroform also denatures protein; and isoamylalcohol facilitates the separation from aqueous phase to organic phase and reduces foaming during centrifugation (Sosa and Oliveira, 1992). When phenol is used, the DNA needs to be further extracted with chloroform (Michaels, 1994) or Cho/IAA (24:1) to remove the phenol from the aqueous phase and to precipitate the remaining proteins (Towner, 1991). It is very likely that the phenol layer contains some DNA and in that case one or two extractions of phenol by TE buffer enhance DNA recovery (Towner, 1991). The study of Webb and Knapp (1990) proved that back- or re-extraction from phenol phase doubles the yield of DNA in *Cuphea* species. Sometimes mixtures of Phe/Cho (1:1) are preferred over phenol (Goodwin and Lee, 1993; Ouenzar *et al*., 1998) because DNA is much less soluble in Phe/Cho than in the phenol alone (Towner, 1991). Strozycki and Legocki (1995) indicated that separate addition of chloroform and phenol helped them to obtain better results in isolating genomic DNA from lupin roots. Phe/Cho/IAA (25:24:1) has been reported (Collins and Symons, 1992; Pich and Schubert, 1993; Aljanabi *et al*., 1999) as being more effective in denaturing proteins than phenol only (Keller and Manak, 1989). Many studies (Kanazawa and Tsutsumi, 1992; Fulton *et al*., 1995; Kim *et al*., 1997; Kang *et al*., 1998; Woodhead *et al*., 1998; Lefort and Douglas, 1999; Sangwan *et al*., 2000) preferred Cho/IAA (24:1) extraction instead of phenol or Phe/Cho extraction. The reason is that the latter solvent causes reduction of quality and quantity of isolated DNA (Schierenbeck, 1994). Many researchers (Katterman and Shattuck, 1983; Webb and Knapp, 1990; Cheng *et al*., 1997) are now choosing to use Cho/octanol as an alternative of Cho/IAA because octanol is more effective in isolating nuclei than isoamylalcohol (Richards, 1988). Sometimes repetition of the Cho/octanol extraction is needed to remove cloudiness due to addition of polyphenol-absorbent PVP in the extraction procedures (Lodhi *et al*., 1994; Porebski *et al*., 1997). Paterson *et al*. (1993) suggested that two extractions with Phe/Cho followed by a third extraction with Phe/Cho/IAA facilitate endonuclease cleavage or DNA amplification by removing more contaminants. Congiu *et al*. (2000) found that two consecutive extractions by Phe/Cho/IAA and Cho/IAA produced better amplification products from DNA isolated from strawberry leaves rather than three extractions with Phe, Phe/Cho, Phe/Cho/IAA. Wettasinghe and Peffley (1998) stated that additional Phe/Cho/IAA and Cho/IAA extractions can cause shearing of DNA and loss of yield and they suggested a protocol which does not require

solvent extraction. When extracting DNA from sugarcane (Rhonda *et al*., 1992), it has been observed that two extractions with Cho/IAA alcohol improved the purity of DNA without affecting the yield.

Precipitation

Following the organic extraction steps, DNA is precipitated by ethanol or isopropanol (Keller and Manak, 1989) in the presence of sodium chloride, sodium acetate, or ammonium acetate (Kirby, 1990; Sosa and Oliveira, 1992) at −20°C for a period of 1 h to overnight (Keller and Manak, 1989). Ethanol precipitation has the advantage of leaving short chain and monomeric nucleic acids in the solution, thus fragmented RNA produced by treatment with RNase is lost at this stage (Brown, 2001). Alternatively, isopropanol can be used and it is more efficient than ethanol in isolating high molecular weight DNA from polysaccharides (Rueda *et al*., 1998).

When the yield is high the precipitated DNA can be recovered by spooling the DNA out of the solution with a glass rod. Centrifugation is the preferred method when small amounts of DNA are precipitated (Towner, 1991; Brown, 2001). Campilho *et al*. (1997) reported that during isopropanol precipitation, when DNA strands were visible, wash medium was directly added quickly after discarding the precipitation buffer before centrifugation, and subsequent washing was carried out if the DNA was flocculent in the solution. The DNA pellet is usually rinsed with 70% ethanol to remove any traces of salt and then dried under vacuum (Keller and Manak, 1989; Towner, 1991). Khanuja *et al*. (1999) stated that initial precipitation with isopropanol followed by ethanol washing yields more DNA. Krishna and Jawali (1997) indicated that DNA suitable for PCR amplification was obtained from cotton and legume seed when they precipitated DNA using isopropanol in the presence of ammonium acetate rather than only using isopropanol. Pirttila *et al*. (2001) reported potassium acetate with isopropanol can efficiently precipitate DNA free of polysaccharides and other secondary metabolites. Hong *et al*. (1995) stated that ethanol precipitation in the presence of ammonium acetate, instead of sodium acetate, causes more carbohydrate co-precipitation with DNA. Dellaporta *et al*. (1983) described that isopropanol precipitation in the presence of sodium acetate yields high molecular weight DNA which is a tight fibrous molecule which is easy to dry and wash. Many studies (Lodhi *et al*., 1994; Schierenbeck, 1994; Jobes *et al*., 1995; Porebski *et al*., 1997; Aljanabi *et al*., 1999; Khanuja *et al*., 1999) use ethanol/isopropanol precipitation in the presence of high concentration of sodium chloride (NaCl) to get carbohydrate-free DNA from polysaccharide-rich species. Huang *et al*. (2000) stated that for woody species, a silica matrix precipitation in the presence of guanidine thiocyanate, at pH 6.5, is more efficient for precipitation of DNA free of polysaccharides and other secondary metabolites than using ethanol or isopropanol. In this process silica is placed in the column, DNA binds with the silica and remains in the column and all other contaminants pass through. The column is then washed with guanidine thiocyanate and DNA is recovered by elution with water (Brown,

2001). Many studies (Murray and Thompson, 1980; Richards, 1988; Webb and Knapp, 1990; Barnwell *et al*., 1998; Sangwan *et al*., 2000) used CTAB precipitation buffer instead of isopropanol for polysaccharide-rich species. When the detergent CTAB is added in the extraction buffer it forms an insoluble complex with the DNA. The DNA-CTAB complex is then precipitated at low salt concentrations, and the contaminants remain in the solution (Richards, 1988). After centrifugation pellets are resuspended in the high-salt (1 M) TE buffer to break down the complex and DNA is further purified by re-precipitation using ethanol (Webb and Knapp, 1990; Barnwell *et al*., 1998; Sangwan *et al*., 1998; 2000) or isopropanol (Richards, 1988) or by using cesium chloride (CsCl; Murray and Thompson, 1980).

Purification of extracted DNA

Depending on the required purity of DNA for further analysis, purification can be achieved using a CsCl gradient (Couch and Fritz, 1990; Howland *et al*., 1991; Maliyakal, 1992). Purification through CsCl allows elimination of residual carbohydrates, denatured proteins and RNA (Sosa and Oliveira, 1992); however, it is expensive, time-consuming and reduces ultracentrifugation (Couch and Fritz, 1990; De la Cruz *et al*., 1995; Weising *et al*., 1995; Huang *et al*., 2000). Sangwan *et al*. (1998) found that DE-52 column chromatography produces fully restriction-digestible DNA, though yield is reduced. A comparative study by Csaikl *et al*. (1998) found that purification by anion exchange chromatography using a Qiagen kit produced DNA of higher yield and quality from seven different species compared to other conventional methods. Guillermaut and Marchal-Drouard (1992) stated that, regardless of the DNA extraction method, RPC-5 column chromatography can improve the purity and yield of DNA suitable for restriction enzyme digestion, PCR amplification and RFLP analysis (Guillermaut and Marchal-Drouard, 1992). Sharma *et al*. (2000) used DEAE cellulose suspension in order to achieve pure DNA from mature tree leaves or needles. Some studies (Rowland and Nguyen, 1993; Schierenbeck, 1994) found that after initial precipitation using ethanol–ammonium acetate, a final precipitation by 13% polyethylene glycol (PEG) (w/v) improved the DNA purity, making it more suitable for downstream application, and the yield of DNA in woody species was not reduced significantly. Another study (Li *et al*., 1994) employed Sephacryl S-100 chromatography followed by precipitation with 20% (w/v) PEG in the presence of 1.2 M NaCl, and yielded DNA of high purity from polysaccharide-rich leaf tissues. Cheng *et al*. (1997) stated that the presence of a low concentration of spermine at the final purification and precipitation step yields DNA of higher quality from bark of various woody species. Kejnovsky and Kypr (1997, 1998) reported that a buffer with alkaline pH (pH > 8.0), sub-millimolar concentration of phosphate and millimolar concentration of alkaline cations (i.e. Zn, Ca, Co, Mn) can increase DNA sedimentation (when DNA concentration is less than 40 µg/ml). Their study also suggested that, among the cations, zinc (Zn) is the most effective for the deposition of DNA and 1 mM $ZnCl_2$ in the presence of 0.1 nM phosphate (Na-Phosphate) can increase DNA sedimentation.

Elimination of Polysaccharides

Polysaccharides in plants are diversified compounds which vary with species and with different developmental stages (Michaels, 1994) and cause the majority of problems in regard to purity of extracted DNA. Carbohydrates in plants are clear and sticky in nature and can easily be identified by their viscous consistency in the solution (Jobes *et al*., 1995). Younger tissues are preferred for extraction of DNA from polysaccharide-rich species (Mondragon *et al*., 2000) because they contain fewer polysaccharides than older or mature tissues. During cell disruption starch is released (Jobes *et al*., 1995) and co-precipitates with DNA; it is extremely difficult to separate from DNA (De la Cruz *et al*., 1997) or to carry out the extraction procedure (Richards, 1988). These compounds also inhibit various enzyme's activities (Rether *et al*., 1993), such as restriction enzymes (Richards, 1988), ligases and polymerase (Fang *et al*., 1992), and may also interfere during concentration of the DNA solution (Fang *et al*., 1992). Pectin can cause problems in chromatography by blocking the flow of the column and also cause problems in electrophoresis (Rether *et al*., 1993). Mucilage-pectin-like complex polysaccharides bind with water and prevent DNA extraction in mini-preparation methods (Mondragon *et al*., 2000). Table 14.1A shows the reported approaches to avoid the problems associated with polysaccharides during the extraction procedure.

Ion exchanger

Purification using DEAE cellulose has been used by few studies, to obtain polysaccharide-free DNA from mature tree leaves, needles, roots, embryos, seeds, pollen, callus (Drouard and Guillemaut, 1995) and from groundnut leaves (Sharma *et al*., 2000). In this technique, DNA binds with DEAE cellulose and any contaminants remain in the solution. This technique is suitable for mid- to mini-scale isolation procedures and yields DNA of sufficient quality for PCR amplification, Southern hybridization and RFLP analysis (Sharma *et al*., 2000). Along with CTAB extraction and CTAB precipitation buffer, DE-52 anion exchanger was also used by Sangwan *et al*. (1998) for further purification of DNA from antimalarial plant *Artemisia annua* leaves.

Chromatography

Sephacryl (S-1000) chromatography was employed to remove the problems associated with polysaccharides in carbohydrate-rich species like *Ficus*, *Citrus*, *Stenomesson* and *Caliphruica* (Li *et al*., 1994). Polysaccharides were partially separated from DNA after performing the chromatography and the remaining carbohydrates were removed by 20% PEG (w/v)/1.7 M NaCl precipitation. RPC-5 column was used to obtain highly purified DNA from leaves, seeds, embryos and callus of various woody species (Guillermaut and Marchal-Drouard, 1992).

High molar salt

Exclusion of polysaccharides by using high concentrations of NaCl is an easy, quick and inexpensive method for plant DNA isolation (Milligan, 1998). NaCl can be added directly after phenol extraction or after dissolving DNA in TE buffer (Fang *et al*., 1992). A high salt concentration allows precipitation of DNA while the polysaccharide remains in the solution (Fang *et al*., 1992). One molar NaCl is sufficient to aid the removal of polysaccharides by increasing their solubility in ethanol so that they do not co-precipitate with DNA, but 2 M NaCl is suggested in polysaccharide-rich species by Fang *et al*. (1992). Lodhi *et al*. (1994) found that higher concentrations of NaCl may be required by *Amelopsis* and *Vitis* species for effective removal of polysaccharides. Wang *et al*. (1996) used 2.5 M NaCl and reported that it effectively removed polysaccharides from silica gel-dried grapevine leaves. In *Fragaria* species, 5 M NaCl was used to remove polysaccharides and, depending on the species, a second salt precipitation may be beneficial (Porebski *et al*., 1997). Polysaccharide-free DNA from fungi, insects, shrimps and leaves of wheat, barley, potato, beans, pear and almond was obtained by using 6 M NaCl in the extraction (Aljanabi and Martinez, 1997). Jobes *et al*. (1995) reported that potassium acetate precipitation together with high molar NaCl/ethanol precipitation yields polysaccharide-free DNA. High molar NaCl suppresses precipitation of polysaccharides and keeps them soluble in the ethanol; on the other hand, potassium acetate removes protein and polysaccharides as potassium dodecyl sulphate precipitate.

Detergent

Higher concentrations of the detergent CTAB in the extraction buffer are recommended for polysaccharide-rich plant tissues (Doyle and Doyle, 1990; Stewart and Via, 1993; De la Cruz *et al*., 1997). The detergent CTAB not only disrupts the cell walls to release DNA but is also able to separate polysaccharides from extracted DNA (Pirttila *et al*., 2001). At low salt concentration, CTAB binds with DNA (Murray and Thompson, 1980; Rogers and Bendich, 1985) and, at a later stage, DNA is precipitated in increased salt concentration leaving other compounds and contaminants in the solution. A 5% CTAB (w/v) solution containing 0.7 M NaCl has been used in some studies to remove polysaccharides from wheat, barley, canola, oat leaves (Procunier *et al*., 1991) and various woody plant tissues (Katterman and Shattuck, 1983; Schierenbeck, 1994). Use of a CTAB extraction buffer, initially proteinase K-SDS or SDS-potassium acetate (KOAc) protocol, was reported by De la Cruz *et al*. (1995, 1997) to reduce polysaccharide contamination of DNA extracted from cacti and tropical tree species. Use of buffers with three different CTAB concentrations has been applied in some studies (Rogers and Bendich, 1985; Richards, 1988; Barnwell *et al*., 1998; Sangwan *et al*., 1998) to avoid polysaccharide contaminants from the cacti family *S. telephium* genus, antimalarial plant *Artemisia annua* and other plants. These

studies used a 2% CTAB (w/v) concentration in the extraction buffer followed by addition of 10% CTAB (w/v) to the supernatant collected after centrifugation and a 1% CTAB (w/v) as precipitation buffer. After incubation and centrifugation the pellet was dissolved in high salt TE buffer to precipitate DNA and to keep polysaccharide and other contaminants in the solution. Mondragon *et al*. (2000) reported that a 25 min incubation at 65°C with 2% CTAB (w/v) extraction buffer and addition of 2% CTAB (w/v) separation buffer reduced viscosity of the solution and also provided polysaccharide-free DNA from the cacti family. Lim *et al*. (1997) used both CTAB and SDS detergents in the extraction buffer and CTAB precipitation buffer to recover polysaccharide-free DNA from CAM ferns. Rogstad *et al*. (2001) reported that DNA yield was significantly improved and pectinase contaminants were dramatically reduced in woody species when they repeated the CTAB extraction three times, collecting the supernatant each time and later joining them together before organic extraction. Some studies included combination of CTAB and SDS in the extraction buffer to reduce polysaccharide contamination in the extracted DNA from polysaccharide-rich plant tissues (Steenkamp *et al*., 1994; Lefort and Douglas, 1999; Mondragon *et al*., 2000).

Combination of CTAB and high molar salt

Aljanabi *et al*. (1999) used 20% CTAB (w/v) and 6 M NaCl to remove polysaccharides and prevent the interaction between DNA and polysaccharide in sugarcane species. Tel-Zur *et al*. (1999) used CTAB extraction buffer with a high salt concentration (4 M NaCl) in combination with three wash steps of the ground tissues with extraction buffer to eliminate polysaccharide from cacti family. Du Plessis *et al*. (1999) reported that the quality of extracted DNA from *Acacia karroo* was increased when they used several CTAB extractions, increased the buffer to tissue ratio (4 ml:100 mg) and used a mixture of 5 M NaCl/ethanol for the precipitation of the DNA.

Enzymes

Removal of polysaccharides by hydrolytic enzymes and subsequent extraction by Phe/Cho yields high-quality DNA (Rogers *et al*., 1996). Some authors advocate the use of hydrolytic enzymes to remove polysaccharides from nucleic acid solution, such as the use of pectinase to breakdown pectin like substances into smaller molecules which then can easily be separated by gel purification (Rogstad *et al*., 2001). Rether *et al*. (1993) and Woodhead *et al*. (1998) incubated the samples with glycosidic enzymes – Caylase M3 at 37°C overnight to degrade polysaccharides from berry species and other various polysaccharide-rich species.

Alcohol

Michaels (1994) reported that precipitation of polysaccharides by 0.35 volume of 100% ethanol (v/v) was successful in eliminating polysaccharide contamination from *Arabidopsis*, tobacco and cucumber. In the same study, it was observed that a high salt concentration (1.5 M NaCl) can remove 80% of polysaccharides, whereas ethanol precipitation showed no contamination at all and the yield was 95% higher. Another study (Manning, 1991) observed that precipitation of carbohydrates using a low concentration of 2-butoxyethanol (2-BE) was able to achieve 97% starch-free DNA from strawberry, cherry, pear, potato, carrot and tomato species compared to ethanol precipitation. Manning (1991) also reported that boric acid in the extraction buffer aided the removal of polysaccharides through the production of boric acid–polysaccharide complexes. A final precipitation of DNA using a mixture of 13% PEG (w/v) and high molarity NaCl (4 M) yields DNA which is free from polysaccharide contaminants (Rowland and Nguyen, 1993; Schierenbeck, 1994). Many studies precipitate DNA using isopropanol or ethanol in the presence of high molarity NaCl (Lodhi *et al*., 1994; Wang *et al*., 1996; Aljanabi *et al*., 1999; Khanuja *et al*., 1999; Mondragon *et al*., 2000), as the presence of the NaCl enhances the solubility of polysaccharides in the ethanol resulting in a reduction of co-precipitation of carbohydrates with the DNA (Du Plessis *et al*., 1999).

Nuclei isolation technique

Many studies suggested that prior to lysis, isolation of nuclei can prevent binding between DNA and cytoplasmic polysaccharides (Collins and Symons, 1992; Paterson *et al*., 1993; Scott and Playford, 1996; Mercado *et al*., 1999).

Elimination of Polyphenol and Other Secondary Metabolites

Distribution and accumulation of phenolic compounds may vary in different tissues of a plant (Haslam, 1989) as well as in different plant species of the plant kingdom (Harborne, 1989). Older tissues contain more polyphenolic and terpenoid compounds than younger tissues (Porebski *et al*., 1997; Sharma *et al*., 2000). These compounds are brown in colour, are released from vacuoles during tissue grinding (Kanazawa and Tsutsumi, 1992) and can irreversibly bind with DNA (Couch and Fritz, 1990; Mercado *et al*., 1999). The presence of these phenolic and terpenoid compounds makes DNA unsuitable for restriction digestion or PCR amplification (Jobes *et al*., 1995; Aljanabi *et al*., 1999) and leads to the generation of poor fingerprinting profiles (Dobelling, 2000). Due to UV radiation from sunlight (Katterman and Shattuck, 1983), field-grown plants contain larger amounts of tannins and polyphenolic compounds than plants grown in a green house (Chaves *et al*., 1995). Different studies have reported the use of different concentrations and combinations of polyphenol absorbents,

antioxidants and reducing agents to circumvent polyphenol contaminations from the species under study; a summary is shown in Table 14.3A.

Approaches

Polyphenols form complexes with DNA very quickly (Couch and Fritz, 1990) and maintaining the freezing temperature during and/or prior to tissue homogenization minimizes polyphenol oxidation (Lodhi *et al*., 1994; Aljanabi *et al*., 1999). Some studies reported the use of polyvinyl polypyrropolydone (PVPP; Howland *et al*., 1991), polyvinyl pyrrolidone (PVP; Wang *et al*., 1996) and 2-ME (Kim *et al*., 1997) during grinding of the tissues in order to reduce polyphenol contamination at an early stage of the extraction protocol. Stange *et al*. (1998) reported that they were not able to obtain DNA suitable for RAPD amplification from various woody species unless they included various polyphenol absorbents, reductants and antioxidants like diethyldithiocarbamic acid (DIECA), 2-ME and ascorbic acid in the CTAB extraction buffer. DIECA acts as an inhibitor of nuclease activities by competing with oxygen for the copper of the enzyme (Howland *et al*., 1991; Milligan, 1998). Therefore, many studies use DIECA in the extraction buffer to reduce polyphenol oxidase activity (Howland *et al*., 1991; Paterson *et al*., 1993; Permingeat *et al*., 1998). PVP or PVPP serves as polyphenol absorbent; PVP gives higher yields than PVPP (Kim *et al*., 1997; Porebski *et al*., 1997). Hong *et al*. (1995) reported that 0.2% concentration of PVPP (w/v) is sufficient to absorb polyphenolic compounds, but in some cases 0.05% PVPP (w/v) can yield higher quality DNA. In a study on defatted oilseed, it was found that 1% PVPP (w/v) satisfactorily removed polyphenol contaminants (Sangwan *et al*., 2000). PVP binds the polyphenols through hydrogen bonding making it easier to eliminate them from the solution (Steenkamp *et al*., 1994; Porebski *et al*., 1997). Direct addition of PVP in the extraction media (Jobes *et al*., 1995) and low pH (< 5.0) initiates more binding between the PVP and plant polyphenols (Mercado *et al*., 1999). Concentration of PVP must initially be optimized and this is species-dependent (Steenkamp *et al*., 1994). Cheng *et al*. (1997) suggested that effective removal of polyphenols from bark of various species required a minimum of 2% PVP (w/v), whereas Bandana and Ahuja (1999) found that 1% PVP (w/v) was enough for effective removal of polyphenol from dried tea leaves. A mixture of 4% PVP (w/v) along with ascorbic acid, DIECA and 2-ME was used to eliminate polyphenols from the cacti family (De la Cruz *et al*., 1997) and 10% PVP (w/v) was used in the removal of phenolic compounds in the extraction of DNA from sugarcane meristem (Aljanabi *et al*., 1999). Some studies (Alshayji *et al*., 1994; Jobes *et al*., 1995) included dithoitheritol (DTT) along with PVP to inhibit phenol oxidation (Jobes *et al*., 1995) and to protect DNA from quinones, disulphides, peroxidases and phenol oxidases (Milligan, 1998). Alshayji *et al*. (1994) found that, along with DTT and 6% PVP (w/v), potassium-meta-bisulphite was the effective reducing agent for eliminating polyphenols from DNA extracted from mature palm leaves. Jobes *et al*. (1995) used DTT and PVP to inhibit oxidation and to absorb polyphenol, respectively.

Bovine serum albumin (BSA) denatures degrading enzymes (Milligan, 1998) and was used as a polyphenol absorbent in species of date palm (Ouenzar *et al*., 1998) and in rainforest plant species (Scott and Playford, 1996). 2-ME also protects DNA from quinones, disulphides, peroxidases and polyphenol oxidases (Milligan, 1998) and therefore retards polyphenol oxidation. Many of the DNA extraction studies found that PVP and 2-ME are able to produce DNA free from polyphenol contamination (Maliyakal, 1992; Pich and Schubert, 1993; Schierenbeck, 1994). In comparison, the sole use of 2-ME proved to be sufficient in removing polyphenol contamination from tuber species and from lotus leaves (Varadarajan and Prakash, 1991; Asemota, 1995). Lefort and Douglas (1999) found that 1% 2-ME (v/v) is needed to denature endogenous nuclease and to reduce oxidation of phenolics. Cheng *et al*. (1997) suggested that effective elimination of polyphenol from bark of various species needs more than 1% of 2-ME (v/v), and 2% of 2-ME (v/v) is sufficient to produce DNA free of polyphenol contaminations. It was found that exclusion of 2-ME does not affect the quantity of DNA extracted but the quality of DNA is reduced without it (Hong *et al*., 1995). It is suggested that using of 2% 2-ME (v/v) along with 2% PVP (w/v) can eliminate polyphenol from bark tissues of various woody species, silica gel-dried leaves of woody plants and grapevine leaves (Cheng *et al*., 1997). Some studies included reducing agent cysteine instead of 2-ME (Rether *et al*., 1993; Drouard and Guillemaut, 1995), because cysteine is more efficient and is less toxic than 2-ME in the protection of DNA against oxidation (Scotti *et al*., 2001). Some studies employed sodium sulfite as a reducing agent to inhibit polyphenol oxidation in sugarcane meristem and polyphenol, tannins-rich woody plants (Aljanabi *et al*., 1999; Du Plessis *et al*., 1999). Sodium meta bisulfite (Peterson *et al*., 1997) and sodium bisulfite (Fulton *et al*., 1995; Bekesiova *et al*., 1999; Sperisen *et al*., 2000; Li *et al*., 2001) have been used in recent studies to prevent polyphenol oxidation in tomato, cotton, carnivorous plants and other herbaceous plants. Manning (1991) used 2-butoxyethanol (2-BE) at higher concentration of sodium ions to precipitate polyphenols from nucleic acids. In the same study the author also used 2-ME and boric acid in the extraction buffer. Boric acid forms a complex with polyphenols facilitating removal from the solution (Manning, 1991). Permingeat *et al*. (1998) found that inclusion of glucose (0.5 M) in the CTAB extraction buffer improved the DNA quality and reduced polyphenol oxidation in cotton leaves. Activated charcoal has been reported as an effective endogenous polyphenol absorbent (Oakley *et al*., 1994; Vroh Bi *et al*., 1996), and is more effective when it is added prior to incubation, at a concentration of 10–40 mg/g of fresh tissue (Vroh Bi *et al*., 1996). Several extraction protocols have used ascorbic acid as an antioxidant (Paterson *et al*., 1993; Milligan, 1998; Permingeat *et al*., 1998; Chaudhry *et al*., 1999). Many studies have adopted nuclei isolation techniques to eliminate most of the cytoplasmic polyphenol compounds at an earlier stage of the procedure (Katterman and Shattuck, 1983; Peterson *et al*., 1997; Chaudhry *et al*., 1999). A low pH isolation buffer, containing various polyphenol absorbents and oxidase inhibitors, has been used by some studies to facilitate the elimination of polyphenols (Couch and Fritz, 1990; Mercado *et al*., 1999).

Elimination of Protein

Solvent phenol and chloroform help to denature proteins while isoamylalcohol aids the separation of precipitated proteins in the organic phase and the nucleic acids remain in the aqueous phase (Sosa and Oliveira, 1992; Brown, 2001). Tel-Zur *et al*. (1999) used phenol chloroform extraction to remove proteins from DNA extracted from cacti family and the phenol was removed by subsequent chloroform extraction. Due to presence of large protein molecules, in some cases one phenol extraction is not enough and therefore several extractions and centrifugation steps are required (Brown, 2001). Centrifugation may cause shearing of DNA and loss of yield (Wettasinghe and Peffley, 1998; Brown, 2001); alternatively, a proteinase K digestion (Schierenbeck, 1994; Jobes *et al*., 1995) can be performed. Proteinase K digests polypeptides into small molecules which are then easily removed by a phenol extraction (Brown, 2001). Several studies (Langridge *et al*., 1991; Varadarajan and Prakash, 1991; Pich and Schubert, 1993; Ziegenhagen *et al*., 1993; Steenkamp *et al*., 1994; Asemota, 1995; Drouard and Guillemaut, 1995; De la Cruz *et al*., 1997) have precipitated proteins using high concentrations of potassium acetate with SDS followed by centrifugation to separate and remove the protein as a massive pellet along with other cellular debris (Jobes *et al*., 1995). Soni and Murray (1994) stated that efficient activity of proteinase K requires 2% SDS (w/v) minimum. Kang *et al*. (1998) found that prior to grinding and lysis of dry seed, incubation with extraction buffer containing proteinase K gave DNA of good quality and quantity for RFLP analysis from soybean and sesame seed. Porebski *et al*. (1997) used proteinase K and an additional Phe/Cho extraction to eliminate protein. Jobes *et al*. (1995) used proteinase K, potassium acetate and two sequential Phe/Cho extractions to eliminate contaminating protein from polysaccharide and polyphenol-rich species. Proteinase K treatment along with three organic extractions by Phe/Cho/IAA and Cho/IAA was required to remove any excess proteins from lotus species (Kanazawa and Tsutsumi, 1992). Li and Arumuganathan (2000) used a high concentration of NaCl (5 M) after proteinase K digestion rather than phenol to remove proteins from sorted maize chromosomes. Chaves *et al*. (1995) reported that dichloromethane can be a cheap and effective substitute to chloroform for the elimination of proteins. This study also reported that quantitatively no difference was found when proteins were eliminated either by dichloromethane or by chloroform; and neither was it to affect the performance of digestion by restriction enzymes.

Elimination of RNA

During the DNA isolation procedure a considerable quantity of RNA is often extracted. RNA contamination can cause overestimation of the amount of DNA extracted when quantified using UV absorbance or it can appear as low molecular weight DNA in agarose gel electrophoresis (Keller and Manak, 1989; Kirby, 1990). Contaminating RNA may cause suppression of PCR

amplification, and lead to mis-priming of DNA templates during thermal cycle sequencing (Jobes *et al*., 1995). Removal of RNA therefore should be an important consideration in any DNA isolation procedure (Jobes *et al*., 1995). Generally to overcome RNA contamination, DNA samples are first dissolved in low concentration TE buffer and then treated with RNAses at 37°C from 30 min to 1 h. RNase treated DNA is then extracted with Cho/IAA (24:1) and the DNA is then re-precipitated with alcohol in the presence of sodium or ammonium acetate. RNase A is a frequently used enzyme to digest RNA into small fragments (Rether *et al*., 1993; Lefort and Douglas, 1999) of ribonucleoside which do not contaminate DNA if digested for about 40 min (Tel-Zur *et al*., 1999). Porebski *et al*. (1997) found that a 1 h treatment with RNase A is sufficient enough to degrade RNA into small ribonucleosides which are not detectable in agarose gels. Removal of RNA by RNase A increases the efficiency of the amplification process (Guidet, 1994) and reduces DNA degradation (Lin and Ritland, 1995). Wettasinghe and Peffley (1998) found that RNase T1 was more effective in eliminating RNA from DNA extracted from onion. Mixtures of RNase A and RNase T1 were also found to be effective for the removal of RNA (Murray and Thompson, 1980; Li *et al*., 1994; Drouard and Guillemaut, 1995). Currently, many studies are using lithium chloride LiCl instead of RNAase to precipitate large molecules of RNA rather than digest it (Jobes *et al*., 1995; Strozycki and Legocki, 1995; Lefort and Douglas, 1999; Pirttila *et al*., 2001).

References

Adams, R.P., Zhong, M. and Fei, Y. (1999) Preservation of DNA in plant specimens: inactivation and reactivation of DNases in field specimens. *Molecular Ecology* 8, 681–684.

Aitchitt, M., Ainsworth, C.C. and Thangavelu, M. (1993) A rapid and efficient method for the extraction of total DNA from mature leaves of date palm (*Phoenix dactylifera* L.). *Plant Molecular Biology Reporter* 11, 317–319.

Aljanabi, S.M. and Martinez, I. (1997) Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Research* 25, 4692–4693.

Aljanabi, S.M., Forget, L. and Dookun, A. (1999) An improved and rapid protocol for the isolation of polysaccharide and polyphenol free sugarcane. *Plant Molecular Biology Reporter* 17, 281–281.

Alshayji, Y., Saleem, M., Alamad, S., Alawadhi, S. and Alsalameen, F. (1994) Isolation and analysis of total genemic DNA from the date palm (P-dactyliferal) and related species. *Acta Biotechnologica* 14, 163–168.

Asemota, H.N. (1995) A fast simple and efficient miniscale method for the preparation of DNA from tissues of YAM (*Dioscorea* spp.). *Plant Molecular Biology Reporter* 13, 214–218.

Bahl, A. and Pfenninger, M. (1996) A rapid method of DNA isolation using laundry detergent. *Nucleic Acids Research* 24, 1587–1588.

Bahnweg, G., Schulze, S., Moller, E.M., Rosenbrock, H., Langebartels, C. and Sandermann, H.J. (1998) DNA Isolation from recalcitrant materials such as tree roots, bark, and forest soil for the detection of fungal pathogen by polymerase chain reaction. *Analytical Biochemistry* 262, 79–82.

Bandana, S.M. and Ahuja, P.S. (1999) Isolation and PCR amplification of genomic DNA from market samples of dry tea. *Plant Molecular Biology Reporter* 17, 171–178.

Barnwell, P., Blanchard, A.N., Bryant, J.A., Smirnoff, N. and Weir, A.F. (1998) Isolation of DNA from the highly mucilaginous succulent plant shape *Sedum telephium*. *Plant Molecular Biology Reporter* 16, 133–133.

Bekesiova, I., Nap, J.-P. and Mlynarova, L. (1999) Isolation of high quality DNA and RNA from leaves of the carnivorous plant *Drosera rotundifolia*. *Plant Molecular Biology Reporter* 17, 269–277.

Bennett, P.V., Hada, M., Hidema, J., Lepre, A.M., Pope, L.C., Quaite, F.E., Sullivan, J.H., Takayanagi, S. and Sutherland, B.M. (2001) Isolation of high molecular length DNA: alfalfa, pea, rice, sorghum, soybean, and spinach. *Crop Science* 41, 167–172.

Berthold, D.A., Best, B.A. and Malkin, R. (1993) A rapid DNA preparation for PCR from *Chlamydomonas reinhardtii* and *Arabidopsis thaliana*. *Plant Molecular Biology Reporter* 11, 338–344.

Blanchard, M.M. and Nowotny, V. (1994) High throughput rapid yeast DNA extraction – application to yeast artificial chromosomes as polymerase chain reaction templates. *Genetic Analysis* 11, 7–11.

Boffey, S. (1986) Molecular biology technique. In: Wilson, K. and Goulding, K.H. (eds) *A Biologist's Guide to Principles and Techniques of Practical Biochemistry*. Edward Arnold, London, pp. 153–196.

Brown, T.A. (2001) *Gene Cloning and DNA Analysis. An Introduction*. Blackwell Science, Manchester, UK.

Brune, L.D. (1992) An alternative, rapid method of plant DNA extraction for PCR analyses. *Nucleic Acids Research* 20, 4676.

Burr, K., Harper, R. and Linacre, A. (2001) One-step isolation of plant DNA suitable for PCR amplification. *Plant Molecular Biology Reporter* 19, 367–371.

Byrne, M., Macdonald, B. and Francki, M. (2001) Incorporation of sodium sulfite into extraction protocol minimizes degradation of Acacia DNA. *Biotechniques* 30, 742–744, 748.

Campilho, A., Almeida, J.M., Santos, I. and Salema, R. (1997) A method for extraction of DNA from *Castanea* spp. and PCR protocol for RAPD reproducibility. *Agronomia Lusitana* 46, 97–99.

Chaudhry, B., Yasmin, A., Husnain, T. and Riazuddin, S. (1999) Miniscale genomic DNA extraction from cotton. *Plant Molecular Biology Reporter* 17, 280–281.

Chaves, A.L., Vergara, C.E. and Mayer, J.E. (1995) Dichloromethane as an economic alternative to chloroform in the extraction of DNA from plant tissue. *Plant Molecular Biology Reporter* 13, 18–25.

Chen, D.-H. and Ronald, P.C. (1999) A rapid DNA minipreparation method suitable for AFLP and other PCR application. *Plant Molecular Biology Reporter* 17, 53–57.

Cheng, F.S., Brown, S.K. and Weeden, N.F. (1997) A DNA extraction protocol from various tissues in woody species. *Hortscience* 32, 921–922.

Cheung, W.Y., Hubert, N. and Landry, B.S. (1993) A simple and rapid DNA microextraction method for plant, animal and insect suitable for RAPD and other PCR analyses. *PCR Methods and Applications* 3, 69–70.

Collins, G.G. and Symons, R.H. (1992) Extraction of nuclear DNA from grape vine leaves by a modified procedure. *Plant Molecular Biology Reporter* 10, 233–235.

Colosi, J.C. and Schaal, B.A. (1993) Tissue grinding with ball bearing and vortex mixer for DNA extraction. *Nucleic Acids Research* 21, 1051–1052.

Congiu, L., Chicca, M., Cella, R., Rossi, R. and Bernacchia, G. (2000) The use of random amplified polymorphic DNA (RAPD) markers to identify strawberry varieties: a forensic application. *Molecular Ecology* 9, 229–232.

Couch, J.A. and Fritz, P.J. (1990) Isolation of DNA from plants high in polyphenolics. *Plant Molecular Biology Reporter* 8, 11–15.

Csaikl, U.M., Bastian, H., Brettschneider, R., Gauch, S., Meir, A., Schauerte, M., Scholz, F., Sperisen, C., Vornam, B. and Ziegenhagen, B. (1998) Comparative analysis of different

DNA extraction protocols: a fast, universal maxi-preparation of high quality plant DNA for genetic evaluation and phylogenetic studies. *Plant Molecular Biology Reporter* 16, 69–86.

De la Cruz, M., Whitkus, R.M. and Motabravo, L. (1995) Tropical tree DNA isolation and amplification. *Molecular Ecology* 4, 787–789.

De la Cruz, M., Ramirez, F. and Hernandez, H. (1997) DNA Isolation and amplification from Cacti. *Plant Molecular Biology Reporter* 15, 319–325.

Deguilloux, M.F., Pemonge, M.H. and Petit, R.J. (2002) Novel perspectives in wood certification and forensics: dry wood as a source of DNA. *Proceedings of the Royal Society of London – Series B: Biological Sciences* 269, 1039–1046.

Dellaporta, S.L., Wood, J. and Hicks, J.B. (1983) A plant DNA minipreparation: Version II. *Plant Molecular Biology Reporter* 1, 19–21.

Dempster, E.L., Pryor, K.V., Francis, D., Young, J.E. and Rogers, H.J. (1999) Rapid DNA extraction from ferns for PCR-based analyses. *Biotechniques* 27, 66.

Denijs, M., Nabben, L. and Wernars, K. (1996) Isolation of fusarium DNA for molecular analysis with and without mechanical cell disruption. *Journal of Microbiological Methods* 27, 13–17.

Dey, P.M., Brownleader, M.D. and Harborne, J.B. (1997) The plant the cell and its molecular components. In: Dey, P.M. and Harborne, J.B. (eds) *Plant Biochemistry*. Academic Press, San Diego, California, pp. 1–47.

Dilworth, E. and Frey, J.E. (2000) A rapid method for high throughput DNA extraction from plant material for PCR amplification. *Plant Molecular Biology Reporter* 18, 61–64.

Dobelling, U. (2000) Simultaneous RNA and DNA extraction from biopsy material, culture cells, plants and bacteria. In: Rapley, R. (ed.) *The Nucleic Acid Protocols Handbook*. Humana Press, Totowa, New Jersey, pp. 53–56.

Dounce, A.L. (1955) The nucleic acids. In: Chargaff, E. and Davidson, J.N. (eds) *The Nucleic Acids*. Academic Press, New York, pp. 53–193.

Doyle, J.J. and Doyle, J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemical Bulletin* 19, 11–15.

Doyle, J.J. and Doyle, J.L. (1990) Isolation of plant DNA from fresh tissue. *Focus* 12, 13–15.

Drouard, L.M. and Guillemaut, P. (1995) A powerful but simple technique to prepare polysaccharide-free DNA quickly and without phenol extraction. *Plant Molecular Biology Reporter* 13, 26–30.

Du Plessis, S., Buys, M.H. and Nel, M. (1999) Optimised DNA isolation from *Acacia karroo* (Fabaceae). *South African Journal of Botany* 65, 437.

Edwards, K., Johnstone, C. and Thompson, C. (1991) A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Research* 19(6), 1349.

Fang, G., Hammer, S. and Rebecca, R. (1992) A quick and inexpensive method for removing polysaccharides from plant genomic DNA. *Biotechniques* 13, 52–56.

Flournoy, L.E., Adams, R.P. and Pandy, R.N. (1996) Interim and archival preservation of plant specimens in alcohols for DNA studies. *Biotechniques* 20, 657–660.

Fu, R.Z., Wang, J., Sun, Y.R. and Shaw, P.C. (1998) Extraction of genomic DNA suitable for PCR analysis from dried plant rhizomes/roots. *Biotechniques* 25, 796–798, 800–801.

Fulton, T.M., Chunwongse, J. and Tanksely, S.D. (1995) Microprep. protocol for extraction of DNA from tomato and other harbaceous plants. *Plant Molecular Biology Reporter* 13, 207–209.

Gauch, S., Hermann, R., Feuser, P., Oelmuller, U. and Bastin, H. (1998) Isolation of nucleic acids using silica gel based membrane: methods based on the use of QIAamp spin columns. In: Karp, A., Isaac, P.G. and Ingram, D.S. (eds) *Molecular Tools for Screening Biodiversity*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 59–63.

Goodwin, D.C. and Lee, S.B. (1993) Microwave minipreparation of total genomic DNA from fungi, plants, protists and animals for PCR. *Biotechniques* 15, 438ff.

Graham, G.C., Mayer, P. and Henry, R.J. (1994) A simplified method for the preparation of fungal genomic DNA for PCR and RAPD analysis. *Biotechniques* 16, 48–50.

Guidet, F. (1994) A powerful new technique to quickly prepare hundreds of plant extracts for PCR and RAPD analyses. *Nuclic Acids Research* 22, 1772–1773.

Guidet, F., Rogowsky, P. and Langridge, P. (1990) A rapid method of preparing megabase plant DNA. *Nucleic Acids Research* 18, 49–55.

Guillermaut, P. and Marchal-Drouard, L. (1992) Isolation of plant DNA: a fast, inexpensive and reliable method. *Plant Molecular Biology Reporter* 10, 60–65.

Harborne, J.B. (1989) General procedures and measurement of total phenolics. In: Harborne, J.B. and Dey, P.M. (eds) *Methods in Plant Biochemistry*. Academic Press, London, pp. 1–28.247

Haslam, E. (1989) *Plant Polyphenols Vegetable Tannis Revisited*. Cambridge University Press, New York.

Heldt, H.W. (1997) *Plant Biochemistry and Molecular Biology*. Oxford University Press, New York, pp. 3–501.

Hengen, P.N. (1994) Methods and reagents – on the magic of mini preps. *Trends in Biochemical Sciences* 19, 182–183.

Henry, R.J. (1997) *Practical Applications of Plant Molecular Biology*. Chapman & Hall, London, pp. 248.

Henry, R.J. (2001a) Plant DNA extraction. In: Henry, R.J. (ed.) *Plant Genotyping: The DNA Fingerprinting of Plants*. CAB International, Wallingford, UK, pp. 239–249.

Henry, R.J. (ed.) (2001b) *Plant Genotyping: The DNA Fingerprinting of Plants*. CAB International, Wallingford, UK, pp. 323.

Hong, Y.K., Kim, S.-D., Polne-Fuller, M. and Gibor, A. (1995) DNA extraction conditions from prophyra perforata using LiCl. *Journal of Applied Phycology* 7, 101–107.

Howland, D.E., Oliver, R.P. and Davy, A.J. (1991) A method of extraction of DNA from birch. *Plant Molecular Biology Reporter* 9, 340–344.

Hu, Y.J. and Zhou, Z.G. (2001) Extraction of RAPD-friendly DNA from *Laminaria japonica* (Phaeophyta) after enzymatic dissociation of the frozen sporophyte tissues. *Journal of Applied Phycology* 13, 415–422.

Huang, J.C., Ge, X.-J. and Sun, M. (2000) Modified CTAB protocol using a silica matrix for isolation of plant genomic DNA. *BioTechniques* 28, 432–434.

Ikeda, N., Bautista, N.S., Yamada, T., Kamijima, O. and Ishii, T. (2001) Ultra-simple DNA extraction method for marker-assisted selection using microsatellite markers in rice. *Plant Molecular Biology Reporter* 19, 27–32.

Jackson, D.P., Hayden, J.D. and Quirke, P. (1991) Extraction of nucleic acid from fresh and archival material. In: Mcpherson, M.J., Quirke, P. and Taylor, G.R. (eds) *PCR a Practical Approach*. IRL Press, Oxford, pp. 29–49.

Jhingan, A.K. (1992) A novel technology for DNA isolation. *Methods Molecular Cell Biology* 3, 15–22.

Jobes, D.V., Hurley, D.L. and Thien, L.B. (1995) Plant DNA isolation – a method to efficiently remove polyphenolics, polysaccharides and RNA. *Taxon* 44, 379–386.

Jofuku, K.D. and Goldberg, R.B. (1988) Analysis of plant gene structure. In: Shaw, C.H. (ed.) *Plant Molecular Biology – A Practical Approach*. IRL Press, Oxford, pp. 37–65.

Kanazawa, A. and Tsutsumi, N. (1992) Extraction of restrictable DNA from plants of the genus *Nelumbo*. *Plant Molecular Biology Reporter* 10, 316–323.

Kang, H.W., Cho, Y.G., Yoon, U.H. and Eun, M.Y. (1998) A rapid DNA extraction method for RFLP and PCR analysis from a single dry seed. *Plant Molecular Biology Reporter* 16, 90–90.

Katterman, F.R.H. and Shattuck, V.I. (1983) An effective method of DNA isolation from the mature leaves of *Gossypium* species that contain large amounts of phenolic terpenoids and tannis. *Preparative Biochemistry* 13, 347–359.

Kejnovsky, E. and Kypr, J. (1997) DNA extraction by zinc. *Nucleic Acids Research* 25, 1870–1871.

Kejnovsky, E. and Kypr, J. (1998) A novel method of DNA extraction from solution based on cosedimentation of DNA with insoluble metal phosphates. *Chemical Papers – Chemicke Zvesti* 52, 303.

Keller, G.H. and Manak, M.M. (1989) Sample preparation. In: Keller, G.H. and Manak, M.M. (eds) *DNA Probes*. Stockton Press, New York, pp. 29–70.

Khanuja, S.P.S., Shasany, A.K., Darokar, M.P. and Kumar, S. (1999) Rapid isolation of DNA from dry and fresh samples of plants producing large amounts of secondary metabolities and essential oils. *Plant Molecular Biology Reporter* 17, 1–7.

Kilpatrick, C.W. (2002) Non-cryogenic preservation of mammalian tissue for DNA extraction: as assessment of storage methods. *Biochemical Genetics* 40, 53–62.

Kim, C.S., Lee, C.H., Shin, J.S., Chung, Y.S. and Hyung, N.I. (1997) A simple and rapid method for isolation of high quality genomic DNA from fruit trees and conifers using PVP. *Nucleic Acids Research* 25, 1085–1086.

Kirby, L.T. (1990) *DNA Fingerprinting: An Introduction*. Stockton Press, New York.

Krishna, T.G. and Jawali, N. (1997) DNA isolation from single or half seeds suitable for random amplified polymorphic DNA analyses. *Analytical Biochemistry* 250, 125–127.

Lahiri, D.K. and Schnabel, B. (1993) DNA isolation by rapid method from human blood samples: effect of Mg Cl_2, EDTA, storage time and temperature on DNA yield and quality. *Biochemical Genetics* 31, 322–388.

Lange, D.A., Penuela, S., Denny, R.L., Mudge, J., Concibido, V.C., Orf, J.H. and Young, N.D. (1998) A plant DNA isolation protocol suitable for polymerase chain reaction based marker-assisted breeding. *Crop Science* 38, 217–224.

Langridge, U., Schwall, M. and Langridge, P. (1991) Squashes of plant tissue as a substrate for PCR. *Nucleic Acids Research* 19(24), 6954.

Lefort, F. and Douglas, G.C. (1999) An efficient micro-method of DNA isolation from mature leaves of four hardwood tree species *Acer, Fraxinus, Prunus* and *Quercus*. *Annals of Forest Science* 56, 259–263.

Li, H., Luo, J., Hemphill, J.K., Wang, J.T. and Gould, J.H. (2001) A rapid and high yielding DNA miniprep for cotton (*Gossypium* spp.). *Plant Molecular Biology Reporter* 19, 183a–183e.

Li, L. and Arumuganathan, K. (2000) High recovery of large molecular weight DNA from stored maize chromosomes. *Plant Molecular Biology Reporter* 18, 41–45.

Li, Q.B., Chai, Q. and Guy, C.L. (1994) A DNA extraction method for RAPD analysis from plants rich in soluble polysaccharides. *Plant Molecular Biology Reporter* 12, 215–220.

Li, X.-Y., Su, X.-Z. and Chen, F. (2002) Rapid extraction of genomic DNA from leaves and bracts of dove tree (*Davidsonia involcrata*). *Plant Molecular Biology Reporter* 20, 185a–185e.

Lim, S.H., Looi, L.K.C., Ong, B.L. and Wee, Y.C. (1997) A method of DNA isolation from epiphytic cam ferns for use in random amplified polymorphic DNA analysis. *Biologia Plantarum* 39, 637–639.

Lin, J.-Z. and Ritland, K. (1995) Flower petal allow simpler and better isolation of DNA for plant RAPD analyses. *Plant Molecular Biology Reporter* 13, 210–213.

Lin, R.C., Ding, Z.S., Li, L.B. and Yun, I.K. (2001) A rapid and efficient DNA minipreparation suitable for screening transgenic plants. *Plant Molecular Biology Reporter* 19, 379a–379e.

Lodhi, M.A., Ye, G.N., Weeden, N.F., Reisch, B.I., Ye, G.N., Weeden, N.F. and Reisch, B.I. (1994) A simple and efficient method for DNA extraction from grapevine cultivars, *Vitis* species and *Ampelopsis*. *Plant Molecular Biology Reporter* 12, 6–13.

Maliyakal, E.J. (1992) An efficient method for isolation of RNA and DNA from plants containing polyphenolics. *Nucleic Acids Research* 20, 2381.

Manning, K. (1991) Isolation of nucleic acids from plants by differential solvent precipitation. *Analytical Biochemistry* 195, 45–50.

Manubens, A., Lobos, S., Jadue, Y., Toro, M., Messina, R., Lladser, M. and Seelenfreund, D. (1999) DNA isolation and AFLP fingerprinting of nectarine and peach varieties. *Plant Molecular Biology Reporter* 17, 255–267.

Mercado, J.A., Mansouri, E.I., Bermudez, J.S., Alfaro, F.P. and Quesada, M.A. (1999) A convenient protocol for extraction and purification of DNA from *Fragaria*. *In Vitro Cellular & Developmental Biology – Plant* 35, 152–153.

Michaels, S.D. (1994) Removal of polysaccharides from plant DNA by ethanol precipitation. *Biotechniques* 17, 274–276.

Michaels, S.D. and Amasino, R.M. (2001) High throughput isolation of DNA and RNA in 96-well format using a paint shaker. *Plant Molecular Biology Reporter* 19, 227–233.

Milligan, B.G. (1998) Total DNA isolation. In: Hoelzel, A.R. (ed.) *Molecular Genetic Analysis of Populations – A Practical Approach*. IRL Press, Oxford.

Mondragon, J.C., Doudareva, N. and Bordelon, B.P. (2000) DNA extraction from several cacti. *Hortscience* 35, 1124–1126.

Murray, M.G. and Thompson, W.F. (1980) Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* 8, 4321–4325.

Nkongolo, K.K., Klimaszewska, K. and Graton, W.S. (1998) DNA yields and optimization of RAPD patterns using spruce embryogenic lines, seedlings, and needles. *Plant Molecular Biology Reporter* 16, 1–9.

Nordell, K.J., Jackelen, A.M.L., Condren, S.M., Lisensky, G.C. and Ellis, A.B. (1999) Liver and onions: DNA extraction from animal and plant tissues. *Journal of Chemical Education* 76, 400B.

Oakley, E.J., Lazarus, C.M. and Macdonald, H. (1994) The use of activated charcoal to remove endogenous fluorescence from tobacco callus extracts. *Plant Molecular Biology Reporter* 12, 14–19.

Oard, J.H. and Dronavalli, S. (1992) Rapid isolation of rice and maize DNA for analysis by random primer PCR. *Plant Molecular Biology Reporter* 10, 236–241.

Ouenzar, B., Hartmann, C., Rode, A. and Benslimane, A. (1998) Date palm DNA minipreparation without liquid nitrogen. *Plant Molecular Biology Reporter* 16, 263–269.

Paris, M. and Carter, M. (2000) Cereal DNA – a rapid high throughput extraction method for marker assisted selection. *Plant Molecular Biology Reporter* 18, 357–360.

Pasakinskiene, I. and Paplauskiene, V. (1999) Floral meristems as a source of enhanced yield and quality of DNA in grasses. *Plant Cell Report* 18, 490–492.

Paterson, A.H., Brubaker, C.L. and Wendel, J.F. (1993) A rapid method for extraction of cotton (*Gossypium* spp.) Genomic DNA suitable for RFLP or PCR analysis. *Plant Molecular Biology Reporter* 11, 122–127.

Permingeat, H.R., Romagnoli, M.V., Sesma, J.I. and Vellejos, R.H. (1998) A simple method for isolating DNA of high yield and quality from cotton (shape *Gossypium hirsutum* L.) leaves. *Plant Molecular Biology Reporter* 16, 89–89.

Perry, M.D., Davey, M.R., Power, J.B., Lowe, K.C., Bligh, H.F.J., Roach, P.S. and Jones, C. (1998) DNA isolation and AFLP™ genetic fingerprinting of *Theobroma cacao* (L.). *Plant Molecular Biology Reporter* 16, 49–59.

Peterson, D.G., Boehm, K.S. and Stack, S.M. (1997) Isolation of milligram quantities of nuclear DNA from tomato (*Lycoperscion esculentum*), a plant containing high levels of polyphenolic compounds. *Plant Molecular Biology Reporter* 15, 148–153.

Pich, U. and Schubert, I. (1993) Midiprep method for isolation of DNA from plants with a high content of polyphenolics. *Nucleic Acids Research* 21, 3328.

Pirttila, A.M., Hirsikorpi, M., Kamarainen, T., Jaakola, L. and Hohtola, A. (2001) DNA isolation methods for medicinal and aromatic plants. *Plant Molecular Bilology Reporter* 19, 273a–273f.

Porebski, S.L., Bailey, G. and Baum, B.R. (1997) Modification of CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Molecular Biology Reporter* 15, 8–15.

Prince, J.P., Zhang, Y., Radwanski, E.R. and Kyle, M.M. (1997) A versatile and high yielding protocol for the preparation of genomic DNA from capsicum spp. (pepper). *Hortscience* 32, 937–939.

Procunier, J.D., Xu, J. and Kasha, K.J. (1991) A rapid and reliable DNA extraction method for higher plants. *Barley Genetics Newsletter* 20, 74–75.

Pusch, C. (1997) A simple and fast procedure for high quality DNA isolation from gels using laundry detergent and inverted columns. *Electrophoresis* 18, 1103–1104.

Qiagen (2000) *Plant Nucleic Acids Purification Technical Hints and Applications*. Qiagen, Valencia, California.

Reiss, R.A., Schwert, D.P. and Ashworth, A.C. (1995) Field preservation of Coleoptera for molecular genetic analyses. *Environmental Entomology* 24, 716–719.

Rether, B., Delmas, G. and Laouedj, A. (1993) Isolation of polysaccharide-free DNA from plants. *Plant Molecular Biology Reporter* 11, 333–337.

Reyes, A., Linacero, R. and Ochando, M.D. (1997) Molecular genetic and integrated control: a universal genomic DNA micro extraction for pCR, RAPD, restriction and southern analysis. *IOBC/WPRS Bulletin* 20, 274–284.

Rhonda, J., Sorbel, H.B.W.S., Keim, P. and Irvine, J.E. (1992) A rapid DNA extraction method for sugarcane and its relatives. *Plant Molecular Biology Reporter* 10, 66–72.

Richards, E. (1988) Preparation of genomic DNA from plant tissue. In: Ausubel, M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A. and Struhl, K. (eds) *Current Protocols in Molecular Biology*. Greene Publishing Associates and Wiley, New York.

Rogers, H.J., Burns, N.A. and Parkes, H.C. (1996) Comparison of small scale methods for the rapid extraction of plant DNA suitable for PCR analysis. *Plant Molecular Biology Reporter* 14, 170–183.

Rogers, S.O. and Bendich, A.J. (1985) Extraction of DNA from milligram amounts of fresh, herbarium, and mummified plant tissues. *Plant Molecular Biology Manual A6* 5, 69–76.

Rogers, S.O. and Bendich, A.J. (1988) Extraction of DNA from plant tissue. *Plant Molecular Biology Manual* A6, 1–10.

Rogstad, S.H. (1992) Saturated NaCl-CTAB solution as a means of field preservation of leaves for DNA analyses. *Taxon* 41, 701–708.

Rogstad, S.H., Keane, B., Keiffer, C.H., Hebard, F. and Sisco, P. (2001) DNA extraction from plants: the use of pectinase. *Plant Molecular Biology Reporter* 19, 353–359.

Rollo, F., Venanzi, F.M. and Amici, A. (1994) *Ancient DNA*. Springer, New York.

Rollo, F.A., Marca, L.A. and Amici, A. (1987) Nucleic acids in mummified plant seeds screening of twelve specimens by gel electrophoresis, molecular hybridization and DNA cloning. *Theoretical and Applied Genetics* 73, 501–505.

Rowland, L.J. and Nguyen, B. (1993) Use of polyethylene glycol for purification of DNA from tissue of woody plants. *BioTechniques* 14, 735–736.

Rueda, J., Linacero, R. and Vazquez, A.M. (1998) Plant total DNA extraction. In: Karp, A., Isaac, P.G. and Ingram, D.S. (eds) *Molecular Tools for Screening Biodiversity*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 10–14.

Saghai-Maroof, M.A., Soliman, R.A., Jorgensen, R.A. and Allard, R.W. (1984) Ribosomal DNA spacer length polymorphisms in barley: Mendelian inheritance, chromosomal location and population dynamics. *Proceedings of the National Academy of Sciences USA* 81, 8014–8018.

Saini, H.S., Shepherd, M. and Henry, R.J. (1999) Microwave extraction of total genomic DNA from barley grains far use in PCR. *Journal of the Institute of Brewing* 105, 185–190.

Sangwan, N.S., Sangwan, R.S. and Kumar, S. (1998) Isolation of genomic DNA from the antimalarial plant *Artemisia annua*. *Plant Molecular Biology Reporter* 16, 365.

Sangwan, R.S., Yadav, U. and Sangwan, N.S. (2000) Isolation of genomic DNA from defatted oil seed residue of opium poppy (*Papaver sominiferum*). *Plant Molecular Biology Reporter* 18, 265–270.

Schierenbeck, K.A. (1994) Modified polyethylene glycol DNA extraction procedure for silica gel dried tropical woody plants. *Biotechniques* 16, 392–394.

Scott, K.D. and Playford, J. (1996) DNA extraction technique for PCR in rainforest plant species. *Biotechniques* 20, 974ff.

Scotti, N., Cardi, T. and Drouard, L.M. (2001) Mitochondrial DNA and RNA isolation from small amounts of potato tissue. *Plant Molecular Biology Reporter* 19, 67a–67h.

Sharma, K.K., Lavanya, M. and Anjaiah, V. (2000) A method for isolation and purification of peanut genomic DNA suitable for analytical applications. *Plant Molecular Biology Reporter* 18, 393a–393h.

Soni, R. and Murray, J.A.H. (1994) Isolation of intact DNA and RNA from plant tissues. *Analytical Biochemistry* 218, 474–476.

Sosa, P.A. and Oliveira, M.C. (1992) DNA extraction from micro algae. *Applied Phycology Forum* 9, 7–9.

Sperisen, C., Gugerli, F., Buchler, V. and Matays, G. (2000) Comparison of two rapid DNA extraction protocols for gymnosperms for applications in population genetic and phylogenetic studies. *Forest Genetics* 7, 133–136.

Stange, C., Prehn, D. and Arce-Johnson, P. (1998) Isolation of *Pinus radiata* genomic DNA suitable for RAPD analysis. *Plant Molecular Biology Reporter* 16, 366–366.

Steenkamp, J., Wiid, I., Lournes, A. and Van Helden, P. (1994) Improved method for DNA extraction from *Vitis vinifera*. *American Journal of Enology and Vita culture* 45, 102–106.

Stein, D.B. (1993) Isolation and comparison of nucleic acids from land plants: nuclear and organellar genes. In: Zimmer, E.A., Thomas, J.W., Cann, R.L. and Wilson, A.C. (eds) *Methods in Enzymology*. Academic Press, San Diego, California. pp. 153–167.

Stein, N., Herren, G. and Keller, B. (2001) A new DNA extraction method for high-throughput marker analysis in a large-genome species such as *Triticum aestivum*. *Plant Breeding* 120, 354–356.

Steiner, J.J., Poklemba, C.J., Fjellstrom, R.G. and Elliot, L.F. (1995) A rapid one tube genomic DNA extraction process for PCR and RAPD analysis. *Nucleic Acids Research* 23, 2569–2570.

Stewart, C.N. and Via, L.E. (1993) A rapid CTAB DNA isolation technique useful for RAPD fingerprinting and other PCR applications. *Plant Molecular Biology Reporter* 14, 748–749.

Strozycki, P.M. and Legocki, A.B. (1995) An efficient method of genomic DNA isolation from plant tissues. *Acta Biochimica Polonica* 42, 329–331.

Sweeney, P., Golembiewski, R. and Danneberger, K. (1996) Random amplified polymorphic DNA analysis of dry turf grass seed. *Hortscience* 31, 400–401.

Tel-Zur, N., Abbo, S., Myslabodski, D. and Mizrahi, Y. (1999) Modified CTAB procedure for DNA isolation from epiphytic cacti of the genera *Hylocereus* and *Selenicereus* (Cactaceae). *Plant Molecular Biology Reporter* 17, 249–254.

Thomson, D. and Henry, R.J. (1993) Use of DNA from dry leaves for PCR and RAPD analysis. *Plant Molecular Biology Reporter* 11, 202–206.

Thomson, D. and Henry, R.J. (1995) Single-step protocol for the preparation of plant tissue for analysis by PCR. *Biotechniques* 19, 394–400.

Towner, P. (1991) Purification of DNA. In: Brown, T.A. (ed.) *Essential Molecular Biology*. IRL Press, Oxford, pp. 47–68.

Varadarajan, G.S. and Prakash, C.S. (1991) A rapid and efficient method for the extraction of total DNA from the sweet potato and its related species. *Plant Molecular Biology Reporter* 9, 6–12.

Vroh Bi, I., Harvengt, L., Chandelier, A., Mergeai, G. and du Jardin, P. (1996) Improved RAPD amplification of recalcitrant plant DNA by the use of activated charcoal during DNA extraction. *Plant Breeding* 115, 205–206.

Wang, J.Y., Wu, Z.-L., Xing, Y.-Y., Zheng, F.-G., Guo, X.-L., Zhang, W.-G. and Hong, M.-M. (1990) Nucleotide sequence of rice *waxy* gene. *Nucleic Acids Research* 18, 5898.

Wang, X.D., Wang, Z.i-P. and Zou,Y.P. (1996) An improved procedure for the isolation of nuclear DNA from leaves of wild grapevine dried with silica gel. *Plant Molecular Biology Reporter* 14, 369–373.

Webb, D.M. and Knapp, S.J. (1990) DNA extraction from previously recalcitrant plant genus. *Plant Molecular Biology Reporter* 8, 180–185.

Weir, B.J., Pierre, R.G.S. and Chibbar, R.N. (1996) Isolation of DNA for RAPD analysis from leaves of the Saskatoon (*Amelanchier alnifolia* Nutt.) and other horticultural crops. *Canadian Journal of Plant Science* 76, 819–824.
Weising, K., Nybom, H., Wolf, K. and Meyer, W. (1995) *DNA Fingerprinting in Plants and Fungi*. CRC Press, London.
Wettasinghe, R. and Peffley, E.B. (1998) A rapid and efficient extraction method for onion DNA. *Plant Breeding* 117, 299–303.
Williams, C.E. and Ronald, P.C. (1994) PCR template DNA isolated quickly from monocot and dicot leaves with tissue homogenization. *Nucleic Acids Research* 22, 1917–1918.
Woodhead, M., Davies, H.V., Brennan, R.M. and Taylor, M.A. (1998) The isolation of genomic DNA from blackcurrant (*Ribes nigrum* L.). *Molecular Biotechnology* 9, 243–246.
Wulff, E.G., Torres, S. and Vigil, E.G. (2002) Protocol for DNA extraction from potato tubers. *Plant Molecular Biology Reporter* 20, 187.
Yoshihashi, T. (2000) Simple and rapid extraction from milled rice and its application to thai aromatic rice (*Oryza sativa* L.) variety, Khao dwak mali. *JIRCAS Journal* 8, 41–47.
Ziegenhagen, B., Gulliemaut, P. and Scholz, F. (1993) A procedure for minipreparation of genomic DNA from needles of silver fir. *Plant Molecular Biology Reporter* 11, 117–121.

Appendix 14

Table 14.1A. Examples of reported plant tissues which have been used for isolation of genomic DNA.

Tissue	References
Dry seed/grain	Kang *et al*., 1998; Sangwan *et al*., 2000; Sweeney *et al*., 1996; Yoshihashi, 2000
Cotyledons	Rogers and Bendich, 1985
Herbarium specimens including seeds and spores	
Dried leaf tissue	Steiner *et al*., 1995
Sporophyte and seedling	Bennett *et al*., 2001; Hu and Zhou, 2001
Shoots, roots, bark and bracts	Bahnweg *et al*., 1998; Li *et al*., 2002; Manubens *et al*., 1999
Floral meristem	Pasakinskiene and Paplauskiene, 1999
Flower and flower buds	Scotti *et al*., 2001
Pollen	Rogers and Bendich, 1985
Young branches	Mondragon *et al*., 2000
Rhizomes	Fu *et al*., 1998
Tubers	Varadarajan and Prakash, 1991; Wulff *et al*., 2002
Fully expanded leaf, fresh leaves	Doyle and Doyle, 1987; Lange *et al*., 1998

Table 14.2A. Components for tissue grinding.

Apparatus	Comments	References
Mortar and pestle	Common place in conjunction with grinding medium and co-disruptor	Bekesiova *et al.*, 1999; Collins and Symons, 1992; Kang *et al.*, 1998; Krishna and Jawali, 1997; Li *et al.*, 2001; Paterson *et al.*, 1993; Rogstad *et al.*, 2001; Ziegenhagen *et al.*, 1993
Microcentrifuge tube with glass rod, plastic pestle, wooden stick	Often with ext. buff. or liq. N Yield can be very variable	Bekesiova *et al.*, 1999; Brune, 1992; Collins and Symons, 1992; Gauch *et al.*, 1998; Kang *et al.*, 1998; Krishna and Jawali, 1997; Li *et al.*, 2001; Oard and Dronavalli, 1992; Paterson *et al.*, 1993; Rogstad *et al.*, 2001; Ziegenhagen *et al.*, 1993
Vortex	Multiple samples in parallel (more than one 96-well plate) Easy, quick and more reliable Reduce time by 90%	Colosi and Schaal, 1993; Drouard and Guillemaut, 1995
High sample number capacity – mixer mill and paint shaker style with the presence of glass beads, tungsten carbide or steel beads	Significant increase in yield	Csaikl *et al.*, 1998; Michaels and Amasino, 2001; Paris and Carter, 2000; Qiagen, 2000; Ziegenhagen *et al.*, 1993
Polytron homogenizer		Dellaporta *et al.*, 1983; Ouenzar *et al.*, 1998; Webb and Knapp., 1990
Grinding medium, lysis buffer	Can be used with sand and mortar and pestle Quality may not be as good as using liq. N	Bekesiova *et al.*, 1999; Doyle and Doyle, 1990; Huang *et al.*, 2000; Scott and Playford, 1996; Weising *et al.*, 1995

Liq. N, liquid nitrogen; ext. buff., extraction buffer

Table 14.3A. Elimination of polyphenol and other secondary metabolites.

		Agents used													
Species	Protocol	PVPP	PVP	BSA	Char-coal	2-ME	Cys-tiene	Na-sulfite	Na-bisulfite	DTT	DIECA	Low pH ext. buff.	A. acid	Remarks	References
Gossypium hirsutum	SDS-KOAc		✓			✓									Fu *et al.*, 1998; Maliyakal, 1992
Dried rhizomes, roots	Urea buffer		6%✓			✓								PVP was added prior to EtOH precipitation	Prince *et al.*, 1997
Capsicum spp.	Urea buffer								✓						Howland *et al.*, 1991
Betula spp.	Urea P-buffer	✓									✓			PVPP was added during tissue grinding	Barnwell *et al.*, 1998; Mercado *et al.*, 1999
Various recalcitrant spp.	CTAB		2%✓		✓	5%✓								Activated charcoal used to absorb polyphe and other secondary metabolites	Kanazawa and Tsutsumi, 1992
Nelumbo genus	CTAB					5%✓								5% 2-ME was sufficient to remove polyphe contamination from lotus	Sangwan *et al.*, 2000
Papaver sominiferum	CTAB	1% ✓												Successfully removes polyphe from oilseed	Woodhead *et al.*, 1998
Ribes nigrum L.	CTAB			✓						✓					Bandana and Ahuja, 1999; Barnwell *et al.*, 1998

Continued

Table 14.3A. *Continued*

Species	Protocol	Agents used												Remarks	References
		PVPP	PVP	BSA	Char-coal	2-ME	Cys-tiene	Na-sulfite	Na-bisulfite	DTT	DIECA	Low pH ext. buff.	A. acid		
Sedum telephium and dried tea leaves	CTAB		1% ✓											1% PVP was found optimal for quality DNA	Campilho *et al.*, 1997; Dempster *et al.*, 1999; Khanuja *et al.*, 1999; Lefort and Douglas, 1999; Lodhi *et al.*, 1994; Mondragon *et al.*, 2000; Porebski *et al.*, 1997
Wide range of spp.	CTAB		✓			✓									Schierenbeck, 1994
Silica gel-dried woody leaves	CTAB		4%✓			2%✓									Byrne *et al.*, 2001
Acacia spp.	CTAB							✓						Strongly suggested for sodium sulphite to prevent DNA degradation	Aljanabi *et al.*, 1999
Sugarcane meristem	CTAB		10%✓					✓						High molar PVP and sodium sulphite to prevent oxidation of phenolic compounds	Tel-Zur *et al.*, 1999

Forest tree spp. and cacti	CTAB			✓				Cheng *et al.*, 1997; Pirttila *et al.*, 2001
Various woody spp.	CTAB	2%✓		2%✓			2%PVP and 2% 2-ME are suggested for removal of polyphenolics from barks	Bekesiova *et al.*, 1999
Drosera rotundifolia	CTAB				✓ 1 M			Permingeat *et al.*, 1998
Gossypium hirsutum L.	CTAB						0.5 M glucose was used in the ext. buff. as a reducing agent	Guillermaut and Marchal-Drouard, 1992
Polysac-rich spp.	SDS-KOAc	2%✓		✓		✓		Manning, 1991
Date palm leaf tissue	SDS-Phe		✓	✓			Mannitol used as a reducing agent in the ext. buff.	Asemota, 1995; Michaels, 1994; Varadarajan and Prakash, 1991
Wide range of spp.	SDS-Phe			✓				Alshayji *et al.*, 1994
Date palm spp.	SDS-ProtK	6%✓			✓		Found that k-meta bisulfite was the most effective in removing polyphe contaminants	Alshayji *et al.*, 1994

Continued

Table 14.3A. *Continued*

Species	Protocol	Agents used												Remarks	References
		PVPP	PVP	BSA	Char-coal	2-ME	Cys-tiene	Na-sulfite	Na-bisulfite	DTT	DIECA	Low pH ext. buff.	A. acid		
Fragaria spp.	SDS-ProtK					✓								Boric acid in ext. buff. and high concentration of 2-BE to eliminate phenolic problem	Manning, 1991
Wide range of spp.	SDS-KOAc		2%✓							1 M✓		✓			Schierenbeck, 1994
Wide range of spp.	SDS-KOAc		6%✓			1%✓								Designed for phenolic-rich species	Pich and Schubert, 1993
Ficus, Citrus spp.	SDS-KOAc					✓									Li *et al*., 1994
Fruit trees and conifers	CTAB		6%✓			1%✓								2-ME was added during grinding and PVP prior to EtOH precipitation	Kim *et al*., 1997
Tropical tree spp.	CTAB		4%✓			✓					✓		✓	To reduce large contaminants of polyphe	De la Cruz *et al*., 1995
Abies alba Mill.	SDS-Phe		2%✓									✓		Low pH ext. buff. to avoid ionization and polyphe oxidation	Ziegenhagen *et al*., 1993

Cacti family	CTAB	4%✓		✓			✓		✓	Higher concentration of PVP,DIECA, A. acid and 2-ME to reduce polyphe oxidation	De la Cruz *et al.*, 1997
Hardwood tree spp.	CTAB	1%✓			1%✓					Found that 1% 2-ME is optimum to keep DNA in non-oxidative environment	Lefort and Douglas, 1999
Cocoa leaves	N.lysis	2%✓	1%✓				✓	✓			Couch and Fritz, 1990
Cotton leaves						✓					Li *et al.*, 2001
Cotton leaves		✓		✓					✓		Chaudhry *et al.*, 1999
Cotton leaves		✓		✓			✓		✓		Paterson *et al.*, 1993
Rainforest plant spp.	N.lysis		✓								Scott and Playford, 1996
Dried vine leaves	N.lysis	6%✓			2%✓					6% PVP added during grinding	Wang *et al.*, 1996

ProtK, proteinase K; spp., species; ext. buff., extraction buffer; K, potassium; N, nuclear; Phe, phenol; A. acid, ascorbic acid; polyphe, polyphenol; polysac, polysaccharide

Table 14.4A. Elimination of polysaccharides.

Species	Protocol	Approaches	References
Forest tree species	CTAB	Incubation of the DNA samples with pectinase solution (Gel purification may be needed)	Rogstad, 1992
Ficus spp., *Citrus* spp.	SDS-KOAc	Seprachyl S-100 column and 20% PEG/NaCl precipitation (Suitable for mature plant tissues)	Li *et al*., 1994
Arabidopsis, cucumber, tobacco	SDS-Phe	0.35 M 100% EtOH precipitation (20% higher yield than precipitation by high salt)	Michaels and Amasino, 2001
Wide range of spp.	SDS-KOAc	High molar salt/EtOH precipitation (90% carbohydrates were precipitated)	Fang *et al*., 1992
Wide range of spp.	SDS-KOAc	Incubation DNA samples with hydrolytic enzyme at 37°C for o.n (Column chromatography may be necessary)	Kilpatrick, 2002; Woodhead *et al*., 1998
Wide range of spp.	SDS-KOAc	Purification by DEAE cellulose (Suitable for fresh, frozen or lyophilized tissues)	Drouard and Guillemaut, 1995; Sharma *et al*., 2000
Sugarcane	CTAB	High con.CTAB (20%) used. High molar (6 M) NaCl/EtOH precipitation and addition of sarkosyl	Aljanabi *et al*., 1999
Wide range of spp.	SDS-ProtK	High molar NACl/EtOH precipitation (High molar salt suppress the precipitation of polysac)	Aljanabi and Martinez, 1997; Jobes *et al*., 1995
Mature strawberry leaves and *Vitis* leaves	CTAB	High molar NaCl/EtOH precipitation	Lodhi *et al*., 1994; Porebski *et al*., 1997)
Fragaria spp.	SDS-Phe	Low con. of 2-BE to precipitate polysac (Boric acid in the ext. buff. also aided in removal of CHO)	Manning, 1991

Continued

Table 14.4A. *Continued*

Species	Protocol	Approaches	References
Various woody plants	CTAB	High con. CTAB solution and 13% PEG/4 M NaCl precipitation (no further purification is necessary)	Rowland and Nguyen, 1993; Schierenbeck, 1994
Various tissues in woody spp.	CTAB	Increased NaCl in ext. buff.	Cheng *et al.*, 1997
Wide range of spp.	SDS-KOAc	RPC-5 column chromatography	Guillermaut and Marchal-Drouard, 1992
Tropical trees and shrubs	CTAB	SDS included in the extraction procedure	De la Cruz *et al.*, 1995; Lefort and Douglas, 1999
Woody plant spp.	CTAB	Purification by silica matrix	Huang *et al.*, 2000
Wide range of spp.	CTAB	High con. of CTAB in the ext. buff.	Aitchitt *et al.*, 1993; Campilho *et al.*, 1997; Congiu *et al.*, 2000; Khanuja *et al.*, 1999; Steenkamp *et al.*, 1994
Acacia karoo	CTAB	High con. CTAB ext. buff. and high molar NaCl/ EtOH precipitation	Du Plessis *et al.*, 1999
Polysac-rich plants	CTAB	CTAB ext. buff., CTAB solution and CTAB precipitation buffer	Barnwell *et al.*, 1998; Kanazawa and Tsutsumi, 1992; Sangwan *et al.*, 1998
Cacti	CTAB	NaCl/EtOH precipitation	Mondragon *et al.*, 2000
Cacti family	SDS-KOAc	Initial extraction with CTAB buffer	De la Cruz *et al.*, 1997
Date palm	SDS-Phe	PEG in ext. buff.	Ouenzar *et al.*, 1998
Various recalcitrant spp.	CTAB	High molar NaCl in ext. buff.	Tel-Zur *et al.*, 1999, Vroh Bi *et al.*, 1996
Sugarcane		High salt, high con. CTAB and high con. SDS used	Aljanabi *et al.*, 1999
Dioscorea spp.	SDS-KOAc	High molar salt in the ext. buff.	Asemota, 1995
Fragaria spp.	N.lysis	High molar salt and high con.CTAB washing buffer	Mercado *et al.*, 1999
Rainforest spp. and dried *Vitis* leaves	N.lysis	High molar salt in ext. buff.	Scott and Playford, 1996; Wang *et al.*, 1990

PEG, polyethylene glycol; ext. buff, extraction buffer; ProtK, proteinase K; Phe, phenol; con., concentration; N, nuclear; EtOH, ethanol; polysac, polysaccharide

Table 14.5A. Plant DNA isolation via nuclei isolation.

			Nuclei extraction				Nuclei lysis				
Species	Tissue type	Tissue disruption	Buffer	Detergent	Stabilizer	Other agents	Detergent	Other agents	Precipitation and purification	Modification	References
Gossypium	Leaf powder	M&P and liq. N and ext. buff.	Low pH-5.0 citrate buffer	Triton X-100	High molar glucose		SDS		Organic extraction, EtOH precipitation	Additional purification may be necessary for low pH ext. buff. to avoid polyphe contamination	Katterman and Shattuck, 1983
Theobroma cacao L.	Fresh, frozen lyophilized tissue	M&P and dry ice	Low pH-6.0 citrate buffer		Glucose	PVP, BSA, DIECA	SDS		CsCl centrifugation	Specially designed for polyphe-rich plants	Csaikl *et al.*, 1998
Vitis vinifera	Fresh young leaves	Motor-driven conical homo-genizer and ext. buff.	Low pH-6.0 citrate buffer	Triton X-100	Glucose	2-ME	Sarkosyl		Organic extraction, EtOH washing	RNA digestion was included; tailored for polyphe-rich grape leaves	Collins and Symons, 1992
Gossypium	Fresh and frozen	In microcentri-fuge tube with ext. buff.	pH-8.0 TE buffer		Glucose	PVP, A. acid, 2-ME, DIECA	CTAB	PVP, A. acid, 2-ME, DIECA	Organic extraction and alcohol precipitation	Does not need anymore purification; several agents employed to avoid endogenous polyphe problem	Paterson *et al.*, 1993

Wide range of species	Seedlings	In eppendorf tube with household drill	pH-7.5,TE buffer	Sorbitol		CTAB sarkosyl	Sodium bi-sulfite	Organic extraction and alcohol precipitation and purification	High molar salt used in the ext. buff. designed for tomato and other herba-ceous plants	Fulton *et al.*, 1995
Vitis vinifera	Silica gel-dried leaves	With PVP in centrifuge tube	TE buffer, pH-8.0		NaCl	CTAB	2-ME	Organic extraction and NaCl/ EtOH precipitation	High PVP and high 2-ME used to avoid polyphe contamination and high salt to eliminate polysac	Steenkamp *et al.*, 1994
Wide range of genera	Fresh, lyophi-lized, herbarium leaf tissue	M&P and sand and ext. buff.	pH-8.0 TE buffer	Sorbitol	BSA, PEG	CTAB sarkosyl		Organic extraction and alcohol precipitation	High molar salt used in ext. buff. specially designed for rainforest plant species	Scott and Playford, 1996
Lycopersicon esculentum	Seedling	Homogenization in ext. buff. with blender	Low pH-6.0 potas-sium per chlo-rate, $MgCl_2$	Hexylene glycol	PVP, 2-ME DIECA, Na-bi sulphite	Triton X-100	Na-metabi-sulfite, 2-ME	Organic extraction and alcohol purification	Tissue treated with ethyl ether and nuclei concentrated with percoll gradient layer protein and RNA digestion included	Peterson *et al.*, 1997

Continued

Table 14.5A. *Continued*

Species	Tissue type	Tissue disruption	Nuclei extraction				Nuclei lysis		Precipitation and purification	Modification	References
			Buffer	Detergent	Stabilizer	Other agents	Detergent	Other agents			
Fragaria spp.	Mature leaves	Liq. N, M&P	Low pH-5.0, acetate buff.		Sorbitol		CTAB sarkosyl		Organic extraction, alcohol precipitation	High salt ext. buff. used. Designed to reduce polyphe and CHO from mature leaves of strawberry spp.	Mercado *et al.*, 1999
Gossypium spp	Fresh/ frozen	M&P and liq. N	pH-8.0 TE buff.		Glucose	PVP, 2-ME A.acid	CTAB sarkosyl	2-ME, PVP A.acid,	Organic extraction and alcohol precipitation	Lysis buffer contains ProtK. designed to dilute effects of phenolic levels	Chaudhry *et al.*, 1999
Gossypium spp.	Young leaves	Liq. N in microcentri-fuge tube	pH-8.0 TE buff.		Sorbitol	Na-bi sulphite	CTAB sarkosyl		Organic extraction and alcohol precipitation	High molar salt used in ext. buff. Sorbitol and vortex mixing provided higher yield. Designed for cotton spp.	Li *et al.*, 2001
Theobroma cacao	Young frozen leaves	Liq. N in microcentri-fuge tube	pH-8.0 TE buff.	Triton X-100	Spermine spermi-dine sucrose	2-ME PMSF	CTAB	Na-DIECA	Organic extraction and CTAB precipitation	Designed to obtain AFLP genetic fingerprinting suitable DNA from cocoa leaves	Perry *et al.*, 1998

Ext. buff., extraction buffer; M&P, mortar and pestle; liq. N, liquid nitrogen; CHO, carbohydrate; A.acid, ascorbic acid; EtOH, ethanol; polyphe, polyphenol; polysac, polysaccharide

Table 14.6A. Isolation of DNA by protein precipitation technique.

Species	Tissue type	Grinding	Primary precipitation	Primary purification	RNA digestion	Secondary precipitation	Secondary purification	Remarks	References
Arachis hypogaea L.	Mature leaf tissue	In liq. N	Isopropanol	DEAE cellulose	RNase	Isopropanol	EtOH	Designed to overcome polysac and polyphe problem	Sharma *et al.*, 2000
Wide range of spp.	Fresh, frozen, dried and lyophilized tissues	M&P with sand and liq. N	Isopropanol	RPC-5 column chro.		EtOH	EtOH	Designed to avoid polysac and polyphe from spruce needles. 2% PVP in the ext. buff.	Guillermaut and Marchal-Drouard, 1992
Dioscorea spp.	Fresh or lyophilized	In M. tube in the presence of sand and ext. buff.	Isopropanol	EtOH	RNase	NaOAc/ isopropanol	EtOH	Ext. buff. contains 1 M NaCl to increase polysac precipitation	Asemota, 1995
Ficus, Citrus, *Stenomesson*, *Caliphuria*	Mature leaf tissue	In liq. N	Isopropanol	Sephacryl S-100	RNase	PEG/NaCl	EtOH	Tailored for polysac-rich plants. CHO and remaining contaminants were eliminated by PEG/NaCl precipitation	Li *et al.*, 1994
Wide range of polysac-rich spp.	Variety of tissues	M&P with liq. N	EtOH	Incubation with Caylase M3	RNase	Organic extraction	EtOH	Designed for polysac-free DNA. PVP in the ext. buff.	Rether *et al.*, 1993

Continued

Table 14.6A. *Continued*

Species	Tissue type	Grinding	Primary precipitation	Primary purification	RNA digestion	Secondary precipitation	Secondary purification	Remarks	References
Vicia faba, Solanum tuberosum, Lycopersicon esculentum	Leaf and stem	M&P with liq. N	Isopropanol	Organic extraction		Isopropanol	EtOH	High con. of PVP and 2-ME to eliminate polyphe problem	Pich and Schubert, 1993
Wide range of spp.	Variety of tissues	In M. tube in the presence of glass-vortexed		DEAE cellulose	RNase	NaOAc/ isopropanol	Isopropanol	Ext. buff. contains 2% of PVP and cysteine designed to obtain polysac-free DNA	Drouard and Guillemaut, 1995
Cucumis melo, C. sativus	Young leaves	M&P with liq. N	Isopropanol	organic extraction		NaCl/EtOH	EtOH	Tailored to get polysac-free genomic DNA	Fang *et al*., 1992
Ipomoea batatas (L.)	Leaves	M&P with liq. N	Isopropanol			NaOAc/ isopropanol	EtOH	Few times washing with EtOH and additional centrifugation included to enhance recovery of DNA	Varadarajan and Prakash, 1991
Cotton seeds	Seed	In PCR tube with metal rod	NH_4OAc/ isopro panol	EtOH				Precipitation by NH_4OAc/ isopropanol instead of isopropanol supported amplification	Kang *et al*., 1998

M&P, mortar and pextle; liq. N, liquid nitrogen; EtOH, ethyl alcohol; ext. buff., extraction buffer; polysac, polysaccharide polyphe, polyphenol; con., concentration; M. tube, microcentrifuge tube.

Table 14.7A. Isolation of DNA by protein digestion technique.

Species	Tissue type	Tissue disruption	RNA digestion	Solvent extraction	Precipitation	Purification	Modifications	Remarks	References
Arabidopsis	Variety of tissues	M&P with liq. N	Rnase	Phe, Phe/Cho, Phe/Cho/IAA	NH_4OAc/EtOH	EtOH	LiCl included in the isolation buffer	Suitable for dicotyledon species	Soni and Murray, 1994
Phoenix, Cocos	Mature leaves	M&P with liq. N	Rnase		Isopropanol	EtOH	K-meta bisulfite, PVP, DTT included in the ext. buff. KOAc precipitation included	Designed to eliminate polyphe and other metabolites from mature leaves of coconut and date palm	Alshayji *et al.*, 1994
Aceraceae, *Magnolia,* Hydrocharitaceae, *Pinus,* Taxodiaceae	Leaf tissue	M&P with liq. N	LiCl		Primary: isopropanol Secondary: NaCl/EtOH	Chloroform purification	DTT, PVP included in the buffer high molar (5 M) salt used in precipitation step	Specially designed for elimination of polysac, polyphe and RNA from genomic DNA	Jobes *et al.*, 1995
Wide range of polysac species	Fresh leaves	Polytron homogenizer			NaCl/ isopropanol	EtOH	High molar salt (6 M) used in precipitation step	Designed to avoid co-precipitation of polysac	Aljanabi and Martinez, 1997

M&P, mortar and pestle; liq. N, liquid nitrogen; Phe, phenol; Cho, chloroform; IAA, Isoamylalcohol; polysac, polysaccharide; polyphe, polyphenol; ext. buff., extraction buffer

Table 14.8A. DNA isolation with CTAB and ethanol/ammonium acetate wash.

Species	Tissue type	Tissue disruption	CTAB con.	Organic extraction	Primary precipitation	Wash	RNA	Secondary precipitation	Modification	Remarks	References
Vitis vinifera	Fresh leaves	In M&P with liq. N	3%	Cho: IAA	Isopropanol	EtOH/ NH_4OAc	RNase A	EtOH	3% CTAB and 1 M Tris HCl used in the ext. buff. Modified from Doyle and Doyle (1987)	High molar CTAB to get more digestible DNA and 1 M TE to increase buffering capacity	Steenkamp *et al*., 1994
Coffee, *Gossypium, Hevea, Musa, Manihot* spp.	Young leaves	In M&P with liq. N	2%	Cho: IAA	Isopropanol	EtOH/ NH_4OAc	RNase A	EtOH	2% PVP, 1% 2-ME and 2 M NaCl used in the ext. buff.	Activated charcoal included after incubation to remove resinous and coloured material	Vroh Bi *et al*., 1996
Castanea spp.	Fresh mature young leaves,	In M&P with liq. N	3%	Cho: IAA	Isopropanol	EtOH/ NH_4OAc		EtOH	1% PVP used in the ext. buff.	Designed for chestnut tree; modified from Doyle and Doyle (1990)	Campilho *et al*., 1997
Lolium, Festuca	Floral meristem	In M&P with liq. N	3%	Cho: IAA	Isopropanol	EtOH/ NH_4OAc	RNase	NaOAc/ EtOH	1% PVP added in the ext. buff.	Floral meristem provided higher yield than any other tissue; modified from Doyle and Doyle (1990)	Pasakinskiene and Paplauskiene, 1999

Con., concentration; M&P, mortar and pestle; liq. N, liquid nitrogen; Phe, phenol; Cho, chloroform; IAA, isoamylalcohol; EtOH, ethanol

Table 14.9A. Plant DNA isolation with CTAB precipitation.

Species	Tissue type	Tissue disruption	Organic extraction	Secondary precipitation	RNA	Purification	Modification	Remarks	References
Nelumbo	Young leaves	In M&P with liq. N	Cho: IAA	Isopropanol		EtOH	5% 2-ME included in ext. buff. ProtK digestion employed	Designed to overcome protein and polyphe contamination from *Lotus* spp.	Jackson *et al.*, 1991
Sedium telephium	Frozen leaves	In M&P with liq. N		EtOH	12 M LiCl	EtOH	1% PVP included in ext. buff.	Yield was low but quality was high	Barnwell *et al.*, 1998
Artemisia annua	Fresh young leaves	In M&P with liq. N	Cho: IAA	EtOH	RNase	De-52 anion exchanger after organic purification	1% PVPP used in the ext. buff.	Designed for medicinal and aromatic plants which have high levels of secondary metabolites	Sangwan *et al.*, 1998
Papaver somniferum	Defatted seed	In M&P with liq. N	Cho: IAA	EtOH/ NaOAc	RNase A	Organic purification	1% PVPP added in the ext. buff.	Before extraction procedure oilseed was defatted first. Suitable for other oilseed and lipid-rich plants	Sangwan *et al.*, 2000

M&P, mortar and pestle; liq. N, liquid nitrogen; ext. buff., extraction buffer; ProtK, proteinase K; Phe, phenol; Cho, chloroform; IAA, isoamylalcohol; polyphe, polyphenol; EtOH, ethanol

Table 14.10A. Modifications of CTAB protocol.

Species	Tissue type	Tissue disruption	CTAB con.	Organic extraction	Precipitation	Wash	RNA	Second precipitation and purification	Modification	Remarks	References
Saccharum spp.	Leaf tissue	In homogenization	10%	Cho: IAA-twice	Isopropanol	EtOH		NaCl/ EtOH. EtOH	10% SDS and 5 M NaCl added in the procedure; modified from Doyle and Doyle (1987)	High CTAB to remove RNA, polysac, and proteins. Designed for sugarcane and its related spp.	Rhonda *et al.*, 1992
Phoenix dactylifera L.	Mature leaves	In M&P with liq. N	3%	Cho: IAA-	Isopropanol	EtOH	RNase	NaOAc/ EtOH	High con. 2-ME (1%) in the ext. buff. Modified from Murray and Thompson (1980), Saghai-Maroof *et al.* (1984)	Specially tailored to extract DNA from mature leaves of date palm and its related spp.	Aitchitt *et al.*, 1993
Several blueberry spp.	Young leaves								Modified from Doyle and Doyle (1987). Addition of more CTAB (5%) after organic ext. and final precipitation with 13% PEG/4 M NaCl	Specially designed for blueberry spp. to overcome polysac problem	Rowland and Nguyen, 1993

Woody angiosperm	Silica gel-dried leaves	In M&P with liq. N	2%	Cho: IAA	Isopropanol				Modified from Doyle and Doyle (1987). 4% PVP and 2% 2-ME in ext. buff. addition of more CTAB (5%) after organic ext. and final precipitation with 13% PEG/4 M NaCl	Designed for silica gel-dried woody leaves. Before extraction tissues were pretreated with buffer containing sarkosyl and sorbitol	Schierenbeck, 1994
Vitis spp. and *Amelopsis*	Young leaves	In M&P with liq. N	2%	Cho: octanol twice	5 M NaCl/ EtOH	RNase	EtOH		High molar NaCl to remove polysac contaminants and PVP to remove polyphe	Suitable for other fruit species	Lodhi *et al*., 1994
Fragaria spp.	Leaf tissues	In M&P with liq. N	2%	Cho: octanol twice	5 M NaCl/ EtOH	RNase A	organic purification		Modified from Doyle and Doyle (1990). Increased con. of PVP added in the ext. buff. ProtK digestion included	Designed for Rosaceae family which contain high level of polysac and polyphe	Porebski *et al*., 1997
Ribes nigrum L.	Young leaves	In M&P with liq. N and sand	2%	Cho: IAA	Isopropanol		EtOH		After first precipitation; polysac was digested with Calyse M3 followed by organic purification	Designed for blackcurrant leaves	Woodhead *et al*., 1998

Continued

Table 14.10A. *Continued*

Species	Tissue type	Tissue disruption	CTAB con.	Organic extraction	Precipitation	Wash	RNA	Second precipitation and purification	Modification	Remarks	References
Gossypium hirsutum L.	Young leaves	In M&P with liq. N	2%	Cho:IAA	Isopropanol		RNase A	NaOAc/ EtOH	0.35 M glucose used in the ext. buff. to avoid browning reaction	Designed for cotton leaves	Permingeat *et al.*, 1998
Wide range of plant spp.	Leaves, stems, whole flower and their part	In M&P with liq. N	2.50%	Cho:IAA	5 M NaCl/ isopropanol	EtOH	RNase A	Organic purification	High con. Of CTAB and NaCl used in the ext. buff. with 1% of PVP	Designed for aromatic plants which produce large number of polysac and secondary metabolites	Khanuja *et al.*, 1999
Camellia sinensis L.	Dried tea leaves	In M&P with liq. N	2%	Cho:IAA	Isopropanol		RNase	Organic purification	1% PVP added in the ext. buff.	Soaking and washing with water helped to reduce brownish colour	Bandana and Ahuja, 1999
Drosera rotundifolia	Fresh leaf tissue	In M&P with liq. N	2%	Cho:IAA	NaOAc/ EtOH	EtOH	RNase		High molar salt used in ext. buff. 5% sarkosyl, sorbitol and Na-bisulfite included in the procedure	Designed for carnivorous plants	Bekesiova *et al.*, 1999

Hylocerus, Selenicereus	Fresh roots	In M&P with liq. N	1.80%	Cho:IAA	NaOAc/ EtOH	EtOH	RNase A	Organic purification	High molar salt (4 M) used in ext. buff. sarkosyl, sorbitol and 1% 2-ME included in ext. buff.	Three times rinsing with ext. buff. removed most of the polysac contaminants	Tel-Zur *et al.*, 1999
Various hardwood tree species	Fresh mature leaves	In M&P with liq. N	1%	Cho:IAA	Isopropanol	EtOH	RNase		1% PVP, 2%SDS, 1% 2-ME and LiCl included in ext. buff.	Found that combination of chemicals provided better yield than one detergent, one reductant and one salt	Lefort and Douglas, 1999
Fragaria spp.	Leaf tissue	In ext. buff.	4%	Phe; Phe/ Cho IAA; Cho/IAA	NaOAc/ EtOH	EtOH			Modified from Doyle and Doyle (1990)		Congiu *et al.*, 2000
Saccharum spp.	Fresh meri-stem	Ultra turax hom ogenizer with buffer	2%	Phe/ Cho/IAA	6 M NaCl/ isopropanol	EtOH			Modified from Doyle and Doyle (1990); high molar salt added in ext. buff. 20% CTAB, 5% sarkosyl, 10% PVP included in the procedure	Designed to overcome polysac and polyphe contamination from sugarcane leaves	Aljanabi *et al.*, 1999

Continued

Table 14.10A. *Continued*

Species	Tissue type	Tissue disruption	CTAB con.	Organic extraction	Precipitation	Wash	RNA	Second precipitation and purification	Modification	Remarks	References
Triticum aestivum, Brassica napus, Avena sativa, Hordeum	Young leaves	In M&P with liq. N and sand	2%	Cho:IAA	EtOH				Modified from Murray and Thompson (1980); after organic extraction 5% CTAB solution was added and then re-extracted with organic solvent ProtK digestion was included		Procunier *et al.*, 1991
Pinus spp.	Spruce seed-lings	In liq. N in M.C. tube	2%	Cho:Oct	Isopropanol			NH_4OAc/ EtOH	After first precipitation, pellet was resuspended in TE and second organic extraction was done and then re-precipitated with NH_4OAc/EtOH	Specially designed for *Pinus* seedlings, needles and embryogenic lines	Nkongolo *et al.*, 1998

Polypodiaceae	Fresh frond tissue	In liq. N. with M&P	1.88%	Cho	EtOH		RNase	Organic purification and EtOH precipitation	Ext. buff. Contained PVP and SDS. After first organic extraction 10% CTAB solution was used and Cho extraction was repeated	Designed for extraction for RAPD suitable DNA from epiphytic ferns	Lim *et al.*, 1997
Adiantum capillus-veneris	Fresh, frozen or dried frond	In liq. N. with M&P	2%	Cho:IAA	NaCl/ isopropanol	EtOH	RNase	EtOH/ NaOAc	PVP, 2-ME included in ext. buff. and high salt precipitation	Specially designed for fern spp. for RAPD analyses	Dempster *et al.*, 1999
Drosera rotundifolia, Artemisia dracunculua	Fresh tissue	In liq. N with M&P	2%	Cho:IAA twice	Isopropanol		LiCl	ETOH	Ext. buff. contains high con. of PVP and 2-ME. LiCl also included in ext. buff. Before precipitation with isopropa-nol, KOAc was included	Specially designed for medicinal and aromatic plants	Pirttila *et al.*, 2001

M&P, mortar and pestle; liq. N, liquid nitrogen; ext. buff., extraction buffer; con., concentration; polysac, polysaccharide; polyphe, polyphenol; Phe, phenol; ProtK, Protinase K; M.C., microcentrifuge tube; IAA, isoamylalcohol; Cho, chloroform; EtOH, ethanol

Table 14.11A. DNA extraction protocols for specific species or tissues.

Species	Tissue type	Tissue disruption	Contents of ext. buff.	Organic purification	Precipitation	Wash	Remarks	References
Brassica spp.	Leaf tissue	In M.C. with motor-driven knots pestle in the presence of ext. buff.	Tris EDTA, sarkosyl, 2 M NaCl, Na-bisulphite		NH_4OAc/ Isopro-panol	EtOH	DNA yield was uniform in the same tissues and comparable with conventional method	Cheung *et al.*, 1993
Castanea dentata, Vaccinium mallow, Pelargonium hortorum, Arachis hypogaea	Fresh leaf tissue	In eppendorf tube with pipette tip mounted on cordless tube	Tris EDTA, CTAB, PVP A. acid, DTT, 2-ME	Cho:IAA	Isopropanol		Ext. buff. is designed to avoid polysac and polyphe contamination	Stewart and Via, 1993
Various monocot and dicot species		Homogenization in ext. buff. in M.C.tube with electric motor-driven pestle	Tris EDTA, sucrose SDS		Isoprpanol	EtOH	KOAc included in the procedure. RNase A can be added, specially for higher yield from small amount of tissue	Reyes *et al.*, 1997
Oryza spp.	Leaf tissues	In eppendorf tube with pellet pestle mounted on drill, liq. N	Tris EDTA, NaCl, PVP A. acid, DTT,	Cho: IAA	Isopropanol	EtOH	Designed for high quality of DNA from mature rice, maize, tomato, pepper and mint leaves	Chen and Ronald, 1999

Pinus radiata	Shoots	In M&P with liq. N	Tris EDTA, CTAB, PVP DIECA, A.acid, NaCl	Cho: IAA	Isopropanol	EtOH	Designed for *Pinus radiata*, grinding in mortar and pestle provided higher yield than home made pestle	Stange *et al.*, 1998
Nicotiana tabacum	Leaves	In eppendorf tube with Liq. N	Tris EDTA, NaCl, SDS, 2-ME, PVP	Phe-Cho-IAA twice	Isopropanol	EtOH	Specially designed for screening transgenic plants	Lin *et al.*, 2001
Phoenix dactylifera L.	Young leaves	In M&P with sand/ glass powder	Tris EDTA, mannitol PEG, BSA, 2-ME, SDS,	Phe-Cho	Isopropanol	EtOH	NaOAc was included together with SDS	Ouenzar *et al.*, 1998
Various plant species	Dry seed	In M.C. tube with glass rod in ext. buff.	Tris EDTA, NaCl, SDS ProtK. After in cubation CTAB buffer added containing PVP	Cho: IAA and Phe	Isopropanol	EtOH	This method is designed for extraction from single dry seed of rice for PCR analysis. Also suitable for RFLP analysis	Kang *et al.*, 1998
Gossypium spp.	Seed	In PCR tube with a metal rod in ext. buff.	Tris EDTA, NaCl, SDS				After incubation period KOAc included centrifuged and supernatant diluted in TE and used for direct PCR amplification	Krishna and Jawali, 1997
Mangrove spp.	Fresh leaves	In M.C. tube with ext. buff. and sand	Tris EDTA, NaCl, CTAB	Cho:IAA	GuSCN, silica suspension	EtOH	Suitable for wide range of recalcitrant spp.; high purity suitable for cloning and sequencing	Huang *et al.*, 2000

M.C., microcentrifuge tube; Phe, phenol; Cho, choloroform; IAA, isoamylalcohol; polyphe, polyphenol; polysac, polysaccharide; ext. buff., extraction buffer; A. acid, ascorbic acid; liq. N, liquid Nitrogen; ProtK, protinase K

Table 14.12A. Crude isolation of plant DNA for PCR amplification.

Species	Contents of ext. buff.	Extraction procedure	References
Various plant species turf grass seed	Tris EDTA, NaCl and SDS	Leaf discs were macerated in eppendorf tube, incubation for 1 h with ext. buff. containing SDS; centrifugation precipitation with isopropanol, redissolved in TE and samples were ready for PCR	Edwards *et al.*, 1991
Brassica spp.	Tris EDTA, ProtK, sarkosyl	Lyophilized leaf tissue grounded in M.C. tube with glass rod in the presence of ext. buff., extract is then digested by Rnase A and boiled and diluted solution is then used for PCR reaction	Guidet, 1994
Various plant species	NaCl and Tris EDTA	Samples were squashed on nylon membrane by stainless steel rod then washed couple of times with Tris EDTA and water and immediately used for PCR	Rether *et al.*, 1993
Brassica napus	Tris EDTA, SDS and ProtK	Tissues are crushed in the tube with mini drill in the presence of lysis buffer, after incubation for 1 h at 55°C, samples are used for PCR reaction	Brune, 1992
Oryza sativa	PEX, Tris EDTA, NaCl	Samples are homogenized in the PEX ext. buff. followed by NaOAc/EtOH precipitation, resuspension in TE and samples are then ready for PCR	Williams and Ronald, 1994
Various species	Tris EDTA, Sarkosyl, PVPP	Tissues are lysed in eppendorf tube with shaker in the presence of glass bead and lysed with lysis buffer followed by incubation and straight for PCR reaction	Steiner *et al.*, 1995
Grass species	Micro LYSIS	In PCR tubes samples are taken with Micro LYSIS followed by heating and cooling in order to lyse the cells open and then contents of tube are used for PCR	Burr *et al.*, 2001
Barley and wheat spp.		Leaf samples are lysed in matrix mill, treated with NaOH neutralized with Tris EDTA and samples are ready for PCR	Paris and Carter, 2000

Continued

Table 14.12A. *Continued*

Species	Contents of ext. buff.	Extraction procedure	References
Arabidopsis thaliana	5% Chelex	Samples are incubated with 100% EtOH for 1 min then addition of chelex-100, vortexed followed by cooling and centrifugation and then supernatant can be used for direct PCR	Berthold *et al.*, 1993
Malus domestica	Tris EDTA, Tween 20, Qiagen ProtK	Leaf samples in microtitre plate with glass bead and ext. buff. several incubation and centrifugation and then samples are ready for PCR reaction.	Dilworth and Frey, 2000
Oryza sativa	Tris EDTA	Leaf tissues are taken in the centrifuge tube with TE buffer placed on the boiling water, centrifuge and supernatant can be used for PCR amplification	Ikeda *et al.*, 2001
Barley grain	Tris EDTA and potassium chloride	Whole or half seed is placed in the microcentrifuge tube; tube is then placed in the microwave oven for 60 s; tubes are cooled vortexed to be used for PCR reaction	Saini *et al.*, 1999

PEX, potassium ethyl xanthogenate; ProtK, Proteinase K; ext. buff., extraction buffer

15 Future Prospects for Plant Genotyping

R.J. Henry

Introduction

The technology for plant genotyping and the range of resulting applications have progressed rapidly as outlined in this book. As the technology provides tests that are easier and faster to perform and lower in cost, the range of applications expands. This chapter will explore these issues and attempt to identify the prospects for, and the impact of, future developments.

Requirements of Different Applications

Plant genotyping has diverse applications (Henry, 2005) and as a result requires different approaches and technologies for these different applications. Some applications such as plant breeding aim to identify differences between individual plants, while applications such as policing of intellectual property rights (such as plant breeder's rights) require distinction of varieties or cultivars that may include a wide range of genotypes in sexually reproducing species.

Different applications have differing requirements for quantitative analysis, time of analysis and cost (Table 15.1). Applications in food processing may require almost real-time analysis to allow management of processes in response to genotyping data. In contrast, plant breeding applications may need results only in time to make decisions to enable planting of the next generation. Many applications involving identification of genotype are qualitative while testing of seed purity may require highly quantitative analysis of the composition of mixed genotypes in seed lots.

DNA Banks as Reference Tools

Plant identification at the species variety or individual level requires the availability of authentic reference samples of genotypes to be identified to

  Plant Genotyping II: SNP Technology (ed. R.J. Henry)

Table 15.1. Requirements for plant genotyping in different contexts.

Application	Quantitative analysis	Real time	Acceptable cost
Plant breeding	Some applications	No	Very low
Biological research	No	No	Medium
Seed lot testing	Yes	No	Medium
Silo testing	Some applications	Yes	Medium
Food processing	Yes	Some	Medium
Product analysis	Yes	No	High
IP protection	Yes	No	Very high

allow comparison with unknowns. The cost and time required to assemble reference samples makes many plant genotyping applications unattractive. DNA Banks (Rice *et al*., 2006) provide an important resource for such reference samples and in many cases may be the key to cost-effective identification. DNA Banks need to hold information that allows the source and identity of the DNA to be verified. This data should include links to herbarium voucher numbers of seed lot numbers in seed banks. Plant varieties grown from seed in agricultural production may change over time relative to the foundation seed lot released by the breeder. This genetic drift may be monitored using genotyping tests using seed or DNA from reference collections and may need to be considered when applying genotyping to variety identification. The barcoding of samples in such collections has been based upon analysis of the sequence of a well-conserved gene in most biological collections. The search for the most suitable loci for routine analysis of a wide range of plant species continues (McIntosh *et al*., 2005; Tsai *et al*., 2006).

The Australian Plant DNA Bank (biobank.com) stores DNA from Australian wild plants and species of economic importance. The collection includes a wide range of varieties of crop species such as wheat, barley, rice and sorghum. DNA Banks also hold examples representing the diversity of plants allowing easy access to samples for phylogenetic (Saarela *et al*., 2007) or forensic analysis (Jobling and Gill, 2004).

Sampling/Storage/DNA Extraction

Sampling needs to ensure reliable sample tracking from source to DNA sample and genotyping data. Storage will be more secure if more highly purified DNA is isolated and the sample is stored at low temperatures. Alternatively storage of the plant tissue prior to the isolation of DNA may be a better option if the tissues can be easily dried. The sampling of material needs to follow procedures that allow evidence to be presented in a court if legal disputes over the sample identity are involved.

Samples are stored at −80°C in the Australian Plant DNA Bank. This is a conservative approach, since well-prepared samples of plant DNA are generally

stable at much higher temperatures. Details of extraction and DNA storage options are provided in Chapters 13 and 14 (this volume).

Applications in Plant Improvement

DNA markers are widely used in plant breeding (Henry, 2004) and it is likely that the application of markers to plant improvement will continue to grow as analysis costs are reduced and linkages to phenotypes of value become better established. SNP analysis allows markers (often directly associated with useful traits) to be assayed in or near any gene of interest (candidate gene). Plant breeding requires the analysis of large populations and large numbers of genetic loci for some applications. This provides a requirement for low-cost genotyping. Data is usually only required in time for decisions on which genotypes to include in the next phase of the breeding (next generation) allowing for analysis that is not immediate. Technologies that provide large quantities of low-cost data are favoured, so the technology needs to be high-throughput. However, compared to more operational applications of genotyping, plant breeding can tolerate methods that have very long analysis times.

Examples of the application of markers in plant improvement have been provided in Chapters 10 and 11 (this volume).

Gene Discovery/Bioprospecting

Analysis of SNP is widely applied to discovery of relationship between genetic variation and the phenotype of an organism. In plants this approach can be used to identify the genetic basis of commercially important traits.

Discovery of the genetic basis of fragrance (Bradbury *et al*., 2005) and gelatinization temperature (Waters *et al*., 2006) in rice are good examples of the outcomes of this type of investigation. These studies defined specific sequence differences that were perfect markers for traits that were difficult to measure. The rate of these discoveries is increasing with the availability of more DNA sequence data and improved sequencing technologies.

Environmental Monitoring

Genotyping may also allow analysis of the variation in wild plant populations and be useful in studies of the evolution and environmental adaptation of plants (Cronin *et al*., 2007) and plant genetic resources. Concerns about climate change may result in greater emphasis on these studies in the future. SNP analysis of large numbers of loci provides a tool for monitoring changes in population structure using wild plant populations and determining the best strategies to manage the conservation of biodiversity. Selection of plants with the genes necessary for performance in new environments may also become a major focus of plant improvement programmes. Production of

crops in more marginal environments is likely to become necessary to meet growing future demand for crops for food, feed and industrial applications.

Applications of Plant Genotyping in IP Protection

Plant genotyping may be used to protect plant varieties that have been produced by plant breeding. Genotyping is often employed when a breach of Intellectual Property (IP) rights is suspected. This policing application is made easier if genotyping methods for identification of the variety have been developed during the breeding of the variety and are provided at the time of commercial release. Consideration of the ease of identification for commercial protection of a plant prior to commercial release can help to ensure that the variety is selected to be uniform at the genetic loci that may be used in the distinction of the variety from others.

Disputes over the identity of plants in commercial production may involve costly legal action, making even the most difficult, expensive or laborious genotyping methods options. In these cases, the choice of genotyping method depends much more on the clarity of the outcome than other considerations. Sample handling and documentation is extremely important if evidence on plant genotyping is to be presented in court. Convincing evidence may require the genotyping of large numbers of varieties from which the disputed sample needs to be distinguished.

Applications in Plant Processing

The processing of plants in the production of food and other products can be aided by knowledge of genotype. Genotyping for this type of application needs to be capable of providing a genetic identity within operational time constraints.

Barley is a good example of a plant species for which the identity of varieties is important. The malting and brewing industry uses varieties that have been established to have characteristics making them suitable for their use. Malting involves germination of the seed and because different varieties may germinate at different rates a pure sample of a single variety is required to ensure a uniform processing of barley to produce malt of a good quality for brewing. Malting barley is traded internationally on the basis of variety.

A range of genotyping techniques have been developed and applied to barley because of the need to ensure separate malting of different varieties (Henry, 1997).

Microsatellite markers have been widely applied in barley genetics and breeding and this has provided a basis for many genotyping tests to determine variety (Ablett *et al*., 2003). SNP-based genotyping has recently become more attractive in barley with the analysis of large numbers of SNP in barley varieties (Rostocks *et al*., 2005).

Applications in Product Testing (End-product Analysis)

Genotyping can be used to test the identity of ingredients used to produce food products. This may be useful in analysis of competing products in the market.

Fruits and vegetables are often marketed with quality claims based upon the performance or characteristics of specific genotypes. The monitoring of variety identity through the whole production chain from farmer to end consumer may require genotyping. For example, pineapples guaranteed to be sweet on the basis of their genotype may allow consumers to purchase the product with greater confidence.

Analysis of the species and variety composition of processed products such as breakfast cereals may allow identification of the source of ingredients used by a competing food manufacturer. This may be important in product formulation to match or beat the cost inputs or quality of the product.

New Technologies

The techniques for genotyping have continued to develop and diversify and this is likely to continue into the future. Many older techniques (Henry, 1997) remain useful. The book *Plant Genotyping* (Henry, 2001) focused mainly on the use of SSR analysis, a technique that at that time had become well established as a standard technique for plant genotyping. This book is largely about SNP analysis reflecting the increased emphasis on SNP genotyping in plants.

The Focus on SNP Analysis

Genotyping methods increasingly focus on SNP analysis (Sobrino *et al*., 2005). Most other methods (e.g. microsatellite or SSR analysis) rely upon separation of DNA fragments based upon size. SNP analysis can be performed using a much wider range of detection strategies (Table 15.2). These range from simple tests that can be performed with minimal equipment (Bradbury *et al*., 2005) to very high-throughput applications requiring significant laboratory instrumentation (Ragoussis *et al*., 2006).

Table 15.2. Platforms for analysis of DNA sequence differences (at the single nucleotide polymorphism (SNP) level). (From Henry, 2007.)

Platform	Approach	Reference
Gel electrophoresis	Allele-specific PCR	Bundock *et al*., 2006
Real-time PCR	Specific PCR or detection	Kennedy *et al*., 2006a,b
Capillary electrophoresis	Primer extension	Batley *et al*., 2003
Mass spectrometry	Primer extension	Ragoussis *et al*., 2006
Microarray	Circularizable ligation probes	Banner *et al*., 2003
Flow cytometry	Bead/microparticle	Lee *et al*., 2004

SNP analysis techniques have continued to improve to the point that SNP discovery has been limiting the application of SNP in most plant species. Recent developments such as TILLING (Comai and Henikoff, 2006; Cordeiro *et al*., 2006b) and advances in DNA sequencing technology (Emrich *et al*., 2007) have substantially removed this constraint.

Innovations in the TILLING approach are covered in Chapter 3 (this volume).

RT PCR

Real-time (RT) PCR allows automated and relatively quantitative analysis of genotype and a wide range of applications have been developed with several new innovations being applied to plants (Kennedy *et al*., 2006a,b). The technology is robust, being applicable to samples with limited preparation, in high throughput and at relatively low cost. The main limitation of this approach has been the development cost for assays for specific loci.

PCR Versus Non-PCR

PCR allows genotyping based upon very small amounts of the sample and most current genotyping methods use PCR to amplify DNA sequences for analysis. However, PCR does not always amplify different sequences in a mixture equally and methods that do not require PCR are likely to be preferred for quantitative analysis.

Microarrays

Microarrays have been developed primarily for analysis of gene expression and have been proposed as genotyping tools because they allow analysis of large numbers of loci at the same time. The main limitations of this technology are cost lack, limited ability to provide quantitative analysis and a requirement for a high-quality DNA sample and/or extensive sample preparation for analysis.

Circularizable Ligation Probes

Circularizable ligation probes (Padlock probes) allow very highly multiplexed analysis (Nilsson *et al*., 1994) in combination with microarray detection (Banner *et al*., 2003). This will greatly reduce the cost of analysis of large numbers of loci in samples and provide a platform for novel applications requiring such analysis.

Mass Spectroscopy

Mass spectroscopy provides a highly discriminating method for analysis and identification of DNA fragments for genotyping applications (Ragoussis *et al*.,

2006). This provides a high-throughput, low-cost analysis platform that is probably only able to be beaten by nanotechnology-based approaches.

The sugarcane genome is highly polyploidy and provides an example of a genome requiring quantitative analysis of allele frequency. Pyrosequencing was found to be a useful system for sugarcane analysis (Cordeiro *et al.*, 2006a) but was not always reliable for rare alleles. More recently the mass array approach has been successfully applied in this complex system.

Nanotechnology

Nanotechnology offers opportunities to develop highly multiplexed genotyping strategies. The potential for this has been outlined in Chapter 9 (this volume). These technologies are likely to be the key to those applications of genotyping that require low cost analysis. This will provide another path to very low cost analysis of large numbers of loci for each sample.

Sequencing by Synthesis

SNP discovery has been rate-limiting in the application of genotyping in many systems. Developments in DNA sequencing technology are dramatically improving the efficiency of SNP discovery. DNA sequencing has been widely performed by the dideoxy Sanger sequencing method. In the last few years new technologies have started to appear. These have been based upon sequencing by synthesis and have dramatically increased the throughput and reduced the cost of DNA sequencing. These methods may have the potential to be further refined to make even greater gains in sequencing efficiency.

Conclusions

The ultimate method of plant genotyping would provide the entire sequence of the genome and interpretation of the sequence in real time at a very low cost. While progress towards the perfect technology continues, it is worth considering the more likely outcomes of technologies that will emerge and be perfected in the next few years. Determination of the complete DNA sequence of plant genotypes is likely to be routinely possible for research applications and will inform the design of tests for more real-time and low costs test in the production and processing of plants.

References

Ablett, G.A., Karakousis, A., Banbury, L., Cakir, M., Holton, T.A., Langridge, P. and Henry, R.J. (2003) Application of SSR markers in the construction of Australian barley genetic maps. *Australian Journal of Agricultural Research* 54, 1187–1195.

Banner, J., Isaksson, A., Waldenstrom, E., Jarvius, J., Landegren, U. and Nilsson, M. (2003) Parallel gene analysis with allele-specific padlock probes and tag microarrays. *Nucleic Acids Research* 31, e103.

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. *Molecular Breeding* 11, 111–120.

Bradbury, L.M.T., Henry, R.J., Jin, Q. and Waters, D.L.E. (2005) A perfect marker for fragrance genotyping in rice. *Molecular Breeding* 16, 279–283.

Bundock, P.C., Cross, M.J., Shapter, F.M. and Henry, R.J. (2006) Robust allele-specific PCR markers developed for SNP's in expressed barley sequences. *Theoretical and Applied Genetics* 112, 358–365.

Comai, L. and Henikoff, S. (2006) TILLING: practical single-nucleotide mutation detection. *Plant Journal* 45, 684–694.

Cordeiro, G.M., Eliott, F., McIntyre, C.L., Casu, R.E. and Henry, R.J. (2006a) Characterization of single nucleotide polymorphisms in sugarcane ESTs. *Theoretical and Applied Genetics* 113, 331–343.

Cordeiro, G.M., Elliott, F. and Henry, R.J. (2006b) An optimised ecotilling protocol for polyploids or pooled samples using a capillary electrophoresis system. *Analytical Biochemistry* 355, 145–147.

Cronin, J., Bundock, P., Henry, R.J. and Nevo, E. (2007) Adaptive climatic molecular evolution in wild barley at the Isa defense locus. *Proceedings of the National Academy of Sciences of the USA* 104, 2773–2778.

Emrich, S.J., Barbazul, W.B., Li, L. and Schnable, P.S. (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Research* 17, 69–73.

Henry, R.J. (1997) *Practical Applications of Plant Molecular Biology*. Chapman & Hall, London, p. 248.

Henry, R.J. (ed.) (2001) *Plant Genotyping: The DNA Fingerprinting of Plants*. CAB International, Wallingford, UK, pp. 323.

Henry, R.J. (2004) Genetic improvement of cereals. *Cereal Foods World* 49, 122–129.

Henry, R.J. (2005) Importance of plant diversity. In: Henry, R.J. (ed.) *Plant Diversity & Evolution: Genotypic & Phenotypic Variation in Higher Plants*. CAB International, Wallingford, UK, pp. 1–6.

Henry, R.J. (2007) Genomics as a tool for cereal chemistry. *Cereal Chemistry* 84, 365–369.

Jobling, M.A. and Gill, P. (2004) Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics* 5, 739–751.

Kennedy, B., Arar, K., Reja, V. and Henry, R.J. (2006a) LNA for optimising strand displacement probes for quantitative real-time PCR. *Analytical Biochemistry* 348, 294–299.

Kennedy, B., Waters, D.L.E. and Henry, R.J. (2006b) Screening for the rice blast resistance gene Pi-ta using LNA displacement probes and real-time PCR. *Molecular Breeding* 18, 185–193.

Lee, S.-H., Walker, D.R., Cregan, P.B. and Boerma, H.R. (2004) Comparison of four flow cytometric SNP detection assays and their use in plant improvement. *Theoretical and Applied Genetics* 110, 167–174.

McIntosh, S.R., Pacey-Miller, T. and Henry, R.J. (2005) A universal protocol for identification of cereals. *Journal of Cereal Science* 41, 37–46.

Nilsson, M., Malmgren, H., Samiotaki, M., Kwiatkowski, M., Chowdhary, B.P. and Landegren, U. (1994) Padlock probes – Circularizing oligonucleotides for localized DNA detection. *Science* 265, 2085–2088.

Ragoussis, J., Elvidge, G.P., Kuar, K. and Colella, S. (2006) Matrix-assisted laser desorption/ionisation time of flight mass spectrometry in genomics research. *PLOS Genetics* 2, 920–929.

Rice, N., Cordeiro, G.M., Shepherd, M., Bundock, P.C., Bradbury, L.M.E., Crawford, A.C., Pacey-Miller, T., Furtado, A. and Henry, R.J. (2006) DNA banks & their role in facilitating the application of genomics to plant germplasm. *Plant Genetic Resources* 4, 64–70.

Rostocks, N., Mudie, S., Cardle, L., Russell, R., Booth, A., Svensson, J.T., Wanamaker, S.I., Rodriguez, E.M., Liu, H., Morris, J., Close, T.J., Marshall, D.-F.F. and Waugh, R. (2005) Genome-wide SNP discovery and linkage analysis in barley based on genes responsive to abiotic stress. *Molecular Genetics and Genomics* 274, 515–527.

Saarela, J.M., Rai, H.S., Doyle, J.A., Endress, P.K., Mathews, S., Marchant, A.D., Briggs, B.G. and Graham, S.W. (2007) Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. *Nature* 446, 312–315.

Sobrino, B., Brion, M. and Carracedo, A. (2005) SNPs in forensic genetics:a review on SNP typing methodologies. *Forensic Science International* 154, 181–194.

Tsai, L.-C., Yu, Y.-C., Hsieh, H.-M., Wang, J.-C., Linacre, A. and Lee, J.C.-I. (2006) Species identification using sequences of the trnL intron and the trnL-trnF IGS of chloroplast genome among popular plants in Taiwan. *Forensic Science International* 164, 193–200.

Waters, D.L.E., Henry, R.J., Reinke, R.F. and Fitzgerald, M.A. (2006) Gelatinization temperature of rice explained by polymorphisms in starch synthase. *Plant Biotechnology Journal* 4, 115–122.

Index