

I.1 The Principle: Identification and Application of Molecular Markers

P. LANGRIDGE and K. CHALMERS¹

1 Introduction

Plant breeding is based around the identification and utilisation of genetic variation. The breeder makes decisions at several key points in the process. First in deciding on the most appropriate parents to use for the initial cross or crosses and then in the selection strategy used in identifying the most desirable individuals amongst the progeny of the cross. The efficiency of the breeding and selection process can be assessed in many different ways including the ultimate success of the varieties released and the frequency with which new varieties are produced. A major cost and logistical issue in plant breeding are the actual number of lines that need to be carried through the evaluation and selection phases of a program. Large breeding programs for annual crops may carry hundreds of thousands of lines to produce a new variety only once every few years. Field trials can be expensive and evaluation of some traits, such as quality and yield stability can be expensive to assess. Molecular markers have proved to be a powerful tool in replacing bioassays and there are now many examples available to show the efficacy of such markers.

The use of molecular markers to track loci and genome regions in crop plants is now routinely applied in many breeding programs. The location of major loci is now known for many disease resistance genes, tolerances to abiotic stresses and quality traits. Improvements in marker screening techniques have also been important in facilitating the tracking of genes. For markers to be effective, they must be closely linked to the target locus and be able to detect polymorphisms in material likely to be used in a breeding program. The prime applications of markers in most breeding programs have been in backcross breeding where loci are tracked to eliminate specific genetic defects in elite germplasm, for the introgression of recessive traits and in the selection of lines with a genome make-up close to the recurrent parent. In progeny breeding, markers have proved valuable in building crucial parents and in enriching F_1 s from complex crosses. Markers have also improved the strategies for gene deployment and enhanced the understanding of the genetic control of complex traits such as components of quality and broad adaptation.

¹ Australian Centre for Plant Functional Genomics, University of Adelaide. Waite Campus, Glen Osmond, SA 5064, Australia

2 Status

Recent developments that have occurred in molecular markers for many crop species have major implications for the future of the technology. There are three key components that are particularly significant. First, for many species, we now have markers closely linked to many traits of importance in the breeding programs. Indeed, for major crop species, we have markers for more loci than can be screened in a conventional breeding program. Second, we have tools that allow marker scanning of the whole genome. Of particular importance has been the development of microsatellite or SSR markers that now form the basis for analysis and allow highly multiplexed SSR screens. This trend will continue as newer, cheaper marker screening based on SNPs become available. The technological advances have improved our capacity for whole genome screens. Third, through association mapping projects we have, or are in the process of developing, whole genome fingerprints for many key lines and varieties of importance in breeding programs. These studies are developing large databases of historic germplasm that should, over the next few years, start to reveal the ways in which breeding programs have selected for and against specific regions of the genome. We can see these developments, particularly in crops such as maize and barley, where markers for most of the major disease resistance clusters, for key components of feed or processing quality and for many loci conditioning tolerance to abiotic stresses are available.

The new marker systems have several important implications for the future of marker-assisted selection (MAS) and breeding strategies in general. Existing strategies for MAS were initiated with a view of markers as providing a rapid and cheap alternative to bioassays and they have largely been used in this role. While highly successful, this strategy does not fully exploit the technology. The key limitation to an expansion of the scale and complexity of marker use is the size of the populations that would be required if one were to try and select for alleles at a large number of loci simultaneously. A further important feature of recent advances has been related to how we best take advantage of the genome information that has been generated for major crop species. We know, for example that chromosome 2H in barley and group 7 chromosomes of wheat, carry clusters of genes, often in repulsion that we would like to break up. Again conventional use of markers has not been very effective in utilising such genome regions. Conversely, we know that there are some chromosomes where there is little allelic variation between lines and it is a waste of effort to try and break these up in a breeding program.

The key challenge of new work is to investigate strategies for whole genome breeding: how we can use genome-wide information in the form of graphical genotypes and known locations of key loci and marker tags for both desirable and undesirable alleles, to design optimal breeding strategies that integrate as much of the available information as possible.

3 Molecular Markers

Molecular markers have been taken, in recent years, to refer to assays that allow the detection of specific sequence differences between two or more individuals. However, it should be recognized that isoenzyme and other protein-based marker systems also represent molecular markers and were in wide use long before DNA markers became popular. One of the earliest type of DNA-based molecular markers, restriction fragment length polymorphisms (RFLPs), were based around the detection of variation in restriction fragment length detected by Southern hybridisation. The types of sequence variation detected by this procedure could be caused by single base changes that led to the creation or removal of a restriction endonuclease recognition site or through insertions or deletions of sufficient size to lead to a detectable shift in fragment size. This technique has been largely superseded by micro-satellite or simple-sequence repeat (SSR) markers and is now rarely used in screening material for breeding programs, but it remains an important research tool. SSR markers detect variation in the number of short repeat sequences, usually two or three base repeats. The number of such repeat units has been found to change at a high frequency and allows the detection of multiple alleles. The large expansion of DNA, particularly EST, sequence databases has now opened the opportunity for the identification of single nucleotide polymorphisms, SNPs. These occur at varying frequencies depending on the species and genome region being considered. In *Arabidopsis* SNP frequencies of 0.007–0.0104 have been measured (Kawabe et al. 1997; Purugganan and Suddith 1998) while in maize a range of 0.00047–0.0037 has been measured (Hilton and Gaut 1998; Wang et al. 1999). SNPs are widely seen as providing the key advantage of multiple detection systems many of which, such as mass spectroscopy, offer high throughput at low detection cost. Importantly, new array based screening methods, such as DArT (Jacoud et al. 2001) appear to offer still cheaper assays due to their very high multiplexing capability. Interestingly, molecular markers may be coming full circle with protein markers again being proposed as viable genetic markers for MAS. Mass spectrometric methods for mass fingerprinting of proteins and for the analysis of low molecular weight proteins, again opens the option for high throughput protein screening. In these cases, single amino acid changes in protein sequence can often be detected and this provides a means for revealing variation in the corresponding DNA coding sequence.

In each method, DNA sequence variation is being detected. However, each method analyses different aspects of DNA sequence variation and different regions of the genome. For example, RFLPs were detected using cDNA clones, namely coding sequence, but frequently detected variation that lay in regions flanking the genes. SSR markers have generally been from non-coding regions although the recent move to three-base repeats and the use of ESTs as the source of SSR markers is changing this. Other markers such as RAPD and

AFLP markers appear to frequently target repetitive regions of the genome. The stability of the sequence difference may also be an issue in some cases. SSRs are seen as being too unstable for some applications since the mutation rate may in some cases be high.

The decision about the most appropriate marker system to use will vary greatly depending on the species, the objective of the marker work and resources available.

4 Identifying Marker/Trait Associations

The most widely used methods for identifying marker/trait associations are based around the construction, phenotyping and genotyping, with molecular markers, of special populations. The steps in identifying marker/trait associations and developing the markers through to application are summarised in Table 1. The populations are generally constructed from two varieties that show a major difference in the traits targeted for mapping. The genetic structure of the segregating populations can be immortalised by producing double haploids or recombinant inbred lines. The populations produced then become a major resource for a wide range of studies. Many such populations have become international resources used by researchers around the world. The ITMI population used for wheat research is an example of this. The population made from a cross between the wheat variety Opata 85 $\times$ W7984 a synthetic wheat, has become the international reference for wheat genetic research (Langridge et al. 2001). New markers, such as SSR and SNP, are being placed on the population continually and the population has been screened for a wide range of disease, abiotic stress tolerance, physiological and quality traits. The beauty of these populations is that they continue to grow in value as they are more and more widely used. Such reference populations are now available for several crop and model plant species.

However, there are also problems with the use of such structured populations. Many of the reference populations were constructed to facilitate marker screening and were based on highly diverse parents, this was the case for the ITMI reference population

There are three important issues that will frequently impact on the most appropriate procedure to be used in finding marker/trait associations:

- There is a major cost in phenotyping. This clearly varies depending on the trait being analysed, but usually the more complex and expensive the phenotypic screening is, the more valuable will be markers for the trait. Costs of phenotyping can be particularly important for traits that require extensive field trials, such as yield or tolerance to some stresses, or require large amounts of material for analysis, such as malting quality in barley or animal feeding trials. Due to costs, the number of replicates and sites is often limited, reducing the sensitivity of some of the analyses.

Table 1. Steps in identifying marker/trait association**1. Defining the target**

Decision about marker development

- Is the trait of importance to breeding program or to biological research?

- Is a molecular marker needed?

- What is the cost of the bioassay relative to marker assay?
- Is the trait dominant versus recessive? – recessive traits may be hard to identify in a bio-assay and will be a prime target for marker development
- Perhaps there is no alternative to marker use:
 - Quarantine trait – e.g., resistance to a disease not present in the country
 - Pyramiding resistances – accumulating multiple genes for resistance to protect against resistance breakdown
 - Map-based cloning of genes – high resolution map is needed to minimize region that needs to be assessed
 - Gene deployment – where desirable alleles are available to several loci, but only one is really needed. How does one decide on the best one to use?

2. Identify germplasm for marker development

Available germplasm, with and without the trait

3. Population structures

Deciding on the best material to use for identifying the marker trait associations

- Knowledge of genetics

- Is the trait simply inherited or multigenic?
- What is the heritability?
- If this information is not available, a trial experiment may be needed

A simple cross can be constructed to measure segregation ratio and heritability

Complex traits and traits of low heritability are often prime targets for marker development as they are hard to assay otherwise

- Decide on best population structure

The structure of the population will be related to the trait and purpose

Populations structure will differ between:

- in-bred versus out-breeding species
- long generation versus short generation plants
- perennial versus annual plants
- Doubled haploids – one meiotic event per line
- F_2 s – two-meiotic events per plant
- Recombinant inbreds or single seed descent
- Complex crosses between highly heterozygous parents

Population size

- For single gene 50 F_2 s may be adequate
- Map-based cloning over 1000 required

Is an existing population already available?

- Screen parents of existing crosses and mapped populations

4. Phenotypic evaluation

- Is phenotypic evaluation possible for single plants?
- For some traits a large number of seeds or plants may be required or field trials at multiple sites, e.g., quality and yield and traits of low heritability
- For association mapping phenotypic information can be collected from existing programs or lines pooled that have a common phenotypes, e.g., lines adapted to a common environment or of common quality ranking