

Survey of marker applications

C.T. Hash and P.J. Bramel-Cox

International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Genetic Resources and Enhancement Program, Patancheru 502 324, Andhra Pradesh, India

Introduction

Molecular markers are rapidly being adopted by crop improvement researchers globally as an effective and appropriate tool for basic and applied studies addressing biological components in agricultural production systems (Jones et al., 1997; Mohan et al., 1997; Prioul et al., 1997). Molecular markers offer specific advantages in assessment of genetic diversity and in trait-specific crop improvement. Use of markers in applied breeding programs can range from facilitating appropriate choice of parents for crosses, to mapping/tagging of gene blocks associated with economically important traits (often termed “quantitative trait loci” (QTLs)). Gene tagging and QTL mapping in turn permit marker-assisted selection (MAS) in backcross, pedigree, and population improvement programs (Mohan et al., 1997). This is especially useful for crop traits that are otherwise difficult or impossible to deal with by conventional means. The near-isogenic products of marker-assisted backcrossing programs provide genetic tools for crop physiologists and crop protection scientists to use in improving our understanding of the mechanisms of various abiotic stress tolerances (Jones et al., 1997; Prioul et al., 1997) and resistances to biotic production constraints such as diseases, insect pests, nematodes, and parasitic weeds like striga. QTL mapping of yield and quality components, as well as components of other physiologically or biochemically complex pathways, can provide crop breeders with a better understanding of the basis for genetic correlations between economically important traits (linkage and/or pleiotropic relationships between gene blocks controlling associated traits; e.g., flowering time and biomass; inflorescence size and inflorescence number). This can facilitate more efficient incremental improvement of specific individual target traits. Further, specific genomic regions associated with QTLs of large effect for one target trait can be identified having minimal effects on otherwise normally correlated traits, permitting an improvement in the first trait that need not be accompanied by counterbalancing reductions in others. Finally, these molecular marker tools can also be used in ways that allow us to more effectively discover and efficiently exploit the evolutionary relationships between organisms, through comparative genomics.

In: B.I.G. Haussmann, H.H. Geiger, D.E. Hess, C.T. Hash, and P. Bramel-Cox (eds.). 2000. Application of molecular markers in plant breeding. Training manual for a seminar held at IITA, Ibadan, Nigeria, from 16-17 August 1999. International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), Patancheru 502 324, Andhra Pradesh, India.

Details of specific applications of molecular markers to genetic diversity assessment, QTL mapping, and marker-assisted genetic enhancement and breeding will be covered in later presentations in this training seminar. This presentation will describe a range of the molecular markers systems available, and the plant genome databases that are under development to facilitate wide exploitation of the mass of genetic information being generated from laboratory and field experiments around the world. This will be followed by a discussion of the relative utility (including costs and effectiveness for specific uses) of the various types of molecular markers for diversity assessment studies, choice of parents for crosses, QTL mapping, and MAS. Molecular markers encompass a wide range of tools, and those most appropriate for a particular application will depend on the crop, its breeding behaviour, and the specific purpose of the study. We hope to provide you with enough information to be able to make rational, cost-effective decisions on the application of molecular marker technology to areas of your research interests.

Descriptions of types of molecular markers

Protein markers

Protein markers, including seed storage proteins, structural proteins, and isozymes were among the first group of molecular markers exploited for genetic diversity assessment and genetic linkage map development. They are the basis for a newly emerging research area called proteomics. They also provide some of the most cost-effective tools for data point generation, especially when iso-electric focusing equipment is used to precisely distinguish between very similar versions of proteins. The major limitations of these markers are

- that much of the genome (including much of the most polymorphic portions of it that are less subject to evolutionary restrictions) does not code for genes,
- different biochemical procedures are required to visualise allelic differences for enzymes having different functions, and
- many proteins are several post-transcriptional steps removed from underlying DNA sequence polymorphism and thus can mask variation present at that level (e.g., differences in tri-nucleotide sequences coding for the same amino acid, interon sequences that are post-transcriptionally removed from the mRNA, and post-translational modification can all contribute to reduced polymorphism expression at the protein level compared to that at the DNA level).

DNA markers

Most points on molecular marker-based genetic linkage maps are anonymous DNA polymorphisms (e.g., restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), and micro-satellite markers) and do not correspond to any gene of known function. However, some molecular markers (including coding DNA (cDNA) and expressed sequence tag (EST) markers, as well as the protein markers described above) do pinpoint individual genes. Anonymous DNA markers are generated by a wide variety of techniques, differing greatly in their reliability (repeatability and robustness), difficulty, expense, and the nature of the polymorphism that they detect. Because of these differences, they also vary greatly in their suitability for various uses. They may be hybridisation based (e.g., RFLP), or polymerase chain reaction (PCR) based (e.g., RAPD and AFLP); they may detect single locus, oligo-locus, or multiple locus differences; and the markers detected may be inherited in a presence/absence, dominant, or co-dominant manner. Brief descriptions of each of a number of the more widely used DNA marker groups are given below based on information contained in the Plant Genome website (<http://www.nal.usda.gov/pgdic/tutorial/lesson4.htm>) and recent reviews of molecular markers useful in mapping plant genomes (Karp et al., 1997; Malyshev and Kartel, 1997; and Mohan et al., 1997).

AFLP (amplified fragment length polymorphism):

This PCR-based technique requires no sequencing or cloning. It is similar to RAPD (see below), but the primer consists of a longer fixed portion (circa 15 base pairs) and a short (2-4 base pairs) random portion. The fixed portion gives the primer stability (and hence repeatability) and the random portion allows it to detect many loci. Polymorphism is detected as band presence/absence (so it is usually interpreted as dominantly inherited, although claims for co-dominant inheritance are also made based on band intensity). AFLP markers are often inherited as tightly linked clusters in centromeric and telomeric regions of chromosomes, but randomly distributed AFLP markers also occur outside these clusters. The technique is difficult to master and is less appropriate than others for comparative mapping studies.

CAPS (cleaved amplified polymorphic sequences):

These secondary markers are identified with two oligonucleotide primers synthesised on the basis of known DNA sequences. Like SCAR primers (see below), they specifically amplify single fragments. However, polymorphism of CAPS is revealed by pre-amplification digestion of template DNA with several restriction endonucleases.

DAF (DNA amplification fingerprint):

In this modification of the RAPD technique, one or more 7- to 8-nucleotide primers are used to produce a relatively complex pattern. Amplification products are separated electrophoretically and visualised by silver staining. Digestion of template DNA with 1 to 3 restriction endonucleases enhances amplification of polymorphic DNA, allowing even near-isogenic lines to be distinguished.

EST (expressed sequence tag):

This PCR-based approach requires both cloning and sequence information. As part of gene sequencing projects, partial sequences of cDNA clones are generated. These are then used to design 18-20 base pair primers that provide a unique sequence “tagging” the gene. It detects a unique, expressed region of the genome. The EST marker is usually detected as a size difference in the amplified product, so is inherited in a co-dominant manner. Design and creation of useful primers can be expensive. They are good for mapping, discovery of genes associated with specific QTLs, and should contribute considerably to our understanding of trait mechanisms.

Microsatellite:

These PCR-based markers can require considerable investment to generate, but are then highly polymorphic and inexpensive to use in mapping and MAS. They result from a short (2-5 base) motif that is repeated multiple times and flanked by a unique DNA sequence. The repeated motif is used as a probe against genomic or cDNA libraries to identify clones in which it is present. These clones are then end-sequenced and primers are designed to amplify the unique DNA flanking the repeated sequence. The method is highly repeatable, identifies a single locus, and targets hypervariable regions of the genome. Polymorphism is usually due to differences in length of the amplified product. Many alleles are available for many of these marker loci. Further, when care is taken in designing the primers, it is possible to simultaneously genotype several (3-15) markers associated with amplification products of substantially different sizes. This is especially cost-effective when combined with fluorescent labelling methods. The start-up costs for this technique are large, but should be justifiable for crops where large-scale mapping and MAS are a practical necessity.

RAPD (random amplified polymorphic DNA):

This PCR-based technique requires neither cloning nor sequencing of DNA. It can detect several loci simultaneously. Short (8-12 base pairs) random primer sequences are used to amplify DNA, usually resulting in presence/absence polymorphism. Although suggested to be easy, inexpensive and fast, its reproducibility is sufficiently problematic (due to short primers being easily affected by annealing conditions) to make it inappropriate for any but phylogenetic studies unless great care is used to ensure stringent annealing conditions.

RFLP (restriction fragment length polymorphism):

This hybridisation-based technique requires use of a library of DNA fragments cloned into some vector. These fragments may be from the species under study or from related (even distantly related) species. The library may be based on genomic or cDNA. RFLP does not require sequencing. The DNA of the organisms under study are digested with one or more restriction endonucleases, the resulting fragments separated electrophoretically according to size, and probed with DNA clones from the library. Fragments matching the probe DNA are visualised by autoradiography or the use of fluorescent labelling techniques. The radioactive label-based visualisation methods are robust and allow multiple uses of the DNA separations resulting from a single restriction digest and electrophoresis run. Because of this, these are generally less expensive than the biotin- or deoxygenin-based fluorescent label methods. Like many protein markers, most RFLP markers are inherited in a co-dominant manner. Further, RFLP markers are especially appropriate in comparative mapping studies.

SCAR (sequence-characterised amplified region):

These PCR-based secondary markers are detected with two 24-nucleotide primers homologous to sequenced ends of a RAPD marker. They amplify a single fragment with high reproducibility. Many are co-dominant and their polymorphism can often be increased by digesting the PCR product with restriction enzymes having 4-nucleotide binding sites.

SSCP (single-strand conformation polymorphism):

This demanding technique is powerful and rapid but can only be used with relatively short DNA fragments. However, it can identify heterozygosity of DNA fragments of the same molecular weight and can even detect changes of a few nucleotide bases. It is currently used in diagnostics of inherited diseases in humans, but is not well developed for crop applications.

SSR (simple sequence repeat): See “microsatellite” above.

STR (short tandem repeat): See “microsatellite” above.

STS (sequence-tagged site):

These PCR-based markers detect a single, unique, sequence-defined point in the genome. They are obtained by sequencing terminal regions of genomic fragments and cDNAs expressing RFLP. Primers of 18-20 base pairs are designed to amplify this short, unique fragment. Polymorphism is often reduced compared to the original RFLP marker, but can be increased at some additional cost by restricting the PCR products to increase the number of bands detected. Since they are longer than RAPD primers and based on a specific sequence,

STS markers more reliably detect the same locus. They are good for both mapping studies and MAS, provided that polymorphism detected is adequate.

VNTR (variable number tandem repeat): See “microsatellite” above.

Introduction to plant genome databases

One of the key sources of information and links to studies being carried out generally in a specific crop species, or for a specific trait in more than one species, is the plant genome databases that are available for use on the internet. Access to the information retained in these databases as well as the contribution of experimental results to the various databases will result in

- a broader application of marker analysis from specific studies,
- an efficient method to identify possible markers for new users, and
- a link to understanding the exact nature of the trait, marker, QTL, evolutionary relationship, or other biological issues of interest.

These databases are available for users on the internet through www.nal.usda.gov and at individual sites for each crop or data collection.

Sorghum database

The SorghumDB Data Collection Site is at algodon.tamu.edu and located at the USDA ARS Southern Crops Research Laboratory, College Station, Texas, USA, and has been updated as of March 1999. You can use the databases for finding information in five general classes: colleague, research, metabolism, literature, and comments. You can also contribute information on these areas as well. The SorghumDB contains information on a number of different areas (Table 1). Distinctions are made between the same type of data received from different labs and experiments.

To use the SorghumDB you need to go to AGIS website, www.nal.usda.gov/pgdic/, then to specific databases found at the Cornell site or genome.cornell.edu/cgi-bin/WebAce/. For pearl millet the crop specific collection site is MilletGenes and UK CropNet, synteny.nott.ac.uk, and is a tool for both accessing and sharing databases from marker research in pearl millet. This database is more focused on data sharing and serving as a repository for data sets than meeting the broader objectives of the SorghumDB. This is indicative of the crop-specific nature of the plant genome databases and the needs of their respective user communities.

Table 1. Areas of information in the SorghumDB.

Class	Entries	Description of contents
Allele	51	Gene and sequence
Author	16778	Authors associated with paper entries
Clone	176	Genes and markers
Colleague	548	Names and addresses of sorghum-related people
Image	70	Pictures and scans
Journal	2409	Journal headings
Locus	586	Observed and molecular genetic markers
Map	118	Maps for observed and molecular markers
Map_data	11	Data for maps
Multimap	12	Comparative maps
Paper	15887	Sorghum related Agricola and direct entry literature
Pathology	70	Biotic and abiotic factors associated with sorghum
Pathogens	90	Sorghum disease pathogens
Probe	266	Information about probes used to screen germplasm
Sequence	94	Sorghum associated sequences from Genbank and others
Taxonomy	6	Sections of sorghum
Trait	20	Plant traits
Other_locus	486	

Both databases are accessed through the WorldWide Web interface provided by AGIS. These are the key features and definitions for various terms used in the databases:

- Map is the class used to describe a single linkage group or chromosome. Each linkage group from a mapping experiment is stored in its own map record. In the SorghumDB, map is used for the linkage groups while map_data is used for the background information, raw data for the maps and to describe the mapping populations. The best way to access this information is using the Browse mode.
- The models for each database describe the underlying structure of the database. The models let you see what the fields are called, what type of data they contain, and which data items are hypertext links to other data items.
- Markers represent single points on a map. Doing a Query By Example or QBE can access information on specific molecular markers. A large number of marker types are used and careful assessment needs to be made of the appropriate use of the information. In the SorghumDB, this is represented by the locus class.
- To find information on genes instead of just the markers use Query Builder or QB. Using this query method you can access information on mapped and unmapped genes and

graphic displays of metabolic pathways. In the SorghumDB the two classes are allele for gene and sequence and clone for genes and markers. Both QBE and QB are single-class search methods.

- To search all the classes in one or more databases use WAIS. This is appropriate when you are uncertain of the exact class where information needed is found or the format it is held in.
- Fuzzy searching is also done to search all classes of data but this method allows for some uncertainty in the exact search information.
- To search for very specific interval data or to build very specific information bases across classes, it is best to use the ACEDB query language. The use of this query method requires an exact knowledge of the syntax, class, and field names and allows you to search for one class and return information linked in another class. This query method can be used to look for syntenic relationships for homoeologous segments or create multimap displays to match aligned regions from different maps
- Physical mapping information on base pair distances or exact sequences is only available at this time on a very limited number of the databases. In the SorghumDB, this information is stored in the allele and sequence classes. The information stored in sequence is linked to external sequence-related databases.
- Use of genome databases are limited in sorghum and pearl millet by the availability of information and the limited diversity of the types of information known in these two crops. These limitations are probably less severe in databases for more widely studied crops like maize and rice, and, unfortunately, are non-existent in crops for which databases have not yet been developed (like yams and most food legumes). The future value of these databases will depend upon the willingness of scientists to contribute information and the long-term commitment to update and maintain the database collection sites and the database sites. Hopefully the information generated on striga will be added as it is found complete, especially from the QTL mapping.

Appropriate applications of molecular markers

Markers for diversity assessment

The application of the various types of markers in the assessment of diversity among genus, species, accessions within a species, or parental varieties from breeding programs will vary according to a number of criteria. Methods for selection of the most appropriate type(s) of marker to use are presented as a decision-making model in Karp et al. (1997). In this technical bulletin, the authors attempt to provide a logical framework to assess the application of different methodologies to different biological questions. The framework uses five decision levels to arrive at the best possible options. These decision levels are:

1. The type of diversity information needed
2. The level of variation expected or indicated
3. The accessibility of probes and primer sets
4. The time constraints of the specific project
5. The level of operational and financial investment available.

This decision-making chart is presented and discussed in detail by Karp et al. (1997). The specific nature of the salient features of the different molecular screening techniques are discussed in association with the decision making framework.

Markers for mapping

In a pinch, for map development you can use any type of molecular marker available. However, co-dominant markers (e.g., RFLP and microsatellites associated with long-sequence unique flanking regions) will give you more information from F₂ and backcross generations than will markers giving predominantly presence/absence or dominantly inherited polymorphism. However, when mapping homozygous populations of random inbred lines, co-dominant markers offer little advantage over presence/absence and other dominantly inherited markers. For comparative mapping within and across species, use of RFLP markers as map anchor loci appears to be the best choice as the polymorphism they detect appears to be evolutionarily conserved in a more predictable manner than that of loci detected by hypervariable SSR and AFLP markers. AFLP, and other highly polymorphic markers, can then be used to fill gaps in RFLP-based genetic linkage maps, following bulk segregant analysis (Michelmore et al., 1991) approaches.

Markers for marker-assisted selection

PCR-based markers are preferred for MAS because they permit the breeder to get by with smaller amounts of more crudely prepared DNA from each plant being genotyped, thus reducing the time, labour, and operational expense for DNA extraction. However, in order to ensure effective differentiation of individuals heterozygous (following selfing or backcrossing) or homozygous (following selfing) for marker alleles flanking the donor segment of interest, co-dominantly inherited markers such as RFLP and SSR are preferred. AFLP markers, which can be treated as clusters of genes that are individually dominantly inherited, are good for quickly obtaining information on many loci in linkage groups where recovery of recurrent parent marker genotype is required. However, a pair of co-dominantly inherited markers flanking each donor segment targeted for transfer (with or without a third co-dominant marker located at the QTL peak), will best provide the information required to

identify individual segregants heterozygous (or homozygous) for the region being introgressed into the recurrent parent background.

Concluding remarks

High through-put, low cost, highly reproducible marker technology provides a powerful tool that can be applied to biological research, including that related to biodiversity and crop improvement in complex crop-livestock production systems. By far the most important issues in such studies are the biological questions to be addressed using whatever tools are most appropriate to provide timely and cost-effective answers. However, we hope that this brief introduction will help you in choosing marker systems appropriate for addressing specific biological questions in which you have interest, when molecular markers offer comparative advantages relative to other possible approaches.

References

- Jones, N., H. Ougham, and H. Thomas. 1997. Markers and mapping: we are all geneticists now. *New Phytologist* 137: 165-177.
- Karp, A., S. Kresovich, K.V. Bhat, W.G. Ayad, and T. Hodgkin. 1997. Molecular tools in plant genetic resources conservation: a guide to the technologies. IPGRI Technical Bulletin No. 2. International Plant Genetic Resources Institute: Rome, Italy.
- Malyshev, S.V. and N.A. Kartel. 1997. Molecular markers in mapping of plant genomes. *Molecular Biology* 31:163-171.
- Michelmore, R.W., I. Paran, and R.V. Kessli. 1991. Identification of markers linked to disease resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions using segregating populations. *Proceedings of the National Academy of Science USA* 88: 9828-9832.
- Mohan, M., S. Nair, A. Bhagwat, T.G. Krishna, and M. Yano. 1997. Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding* 3: 87-103.
- Prioul, J.-L., S. Quarrie, M. Causse, and D. de Vienne. 1997. Dissecting complex physiological functions through the use of molecular quantitative genetics. *Journal of Experimental Botany* 48:1151-1163.