



BMT Guidelines (proj.3)

ORIGINAL: English

DATE: June 9, 2005

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS
GENEVA

**GUIDELINES FOR
MOLECULAR MARKER SELECTION
AND
DATABASE CONSTRUCTION
("BMT GUIDELINES")**

DRAFT

Document prepared by the Office of the Union

to be considered by the

*Technical Working Party on Automation and Computer Programs (TWC),
at its twenty-third session to be held in Ottawa, from June 13 to 16, 2005*

and the

*Working Group on Biochemical and Molecular Techniques, and DNA-Profiling in Particular,
at its ninth session to be held in Washington D.C., United States of America,
from June 21 to 23, 2005*

A	INTRODUCTION.....	3
B	GENERAL PRINCIPLES.....	3
1.	Selection of a Molecular Marker Methodology.....	3
2.	Selection of Molecular Markers.....	4
2.1	General Criteria.....	4
2.2	Specific criteria.....	4
2.3	Access to the technology.....	5
3.	Material to be Analysed.....	6
3.1	Source of plant material.....	6
3.2	Type of plant material.....	6
3.3	Sample size.....	6
3.4	DNA reference sample.....	7
4.	Standardisation of Analytical Protocols.....	7
4.1	Introduction.....	7
4.2	Quality criteria.....	7
4.3	Evaluation Phase.....	8
4.4	Scoring of molecular data.....	8
5	Databases.....	9
5.1	Type of database.....	9
5.2	Transfer of data to the database.....	9
5.3	Data access / ownership.....	9
5.4	Data analysis.....	9
5.5	Validating the database.....	9
6.	Summary.....	9
C.	CROP SPECIFIC BMT GUIDELINES.....	10

A INTRODUCTION

The purpose of this document (BMT Guidelines) is to provide guidance for molecular marker selection and database construction. The BMT Guidelines are intended to address the need to build databases containing molecular profiles of varieties, possibly produced in different laboratories using different technologies. This sets high demands on the quality of the markers and on the necessity of reproducing data using these markers in situations where equipment and/or reaction chemicals might change. In addition, specific precautions need to be taken regarding the quality of data entered into a database.

With regard to the possible use of molecular markers in the examination of Distinctness, Uniformity and Stability (“DUS”) the current situation within UPOV is explained in the Annex to this document.

B GENERAL PRINCIPLES

1. Selection of a Molecular Marker Methodology

1.1 The most important criteria for choosing a methodology are:

- (a) reproducibility of data production between laboratories and detection platforms (types of equipment);
- (b) repeatability over time;
- (c) discrimination power;
- (d) possibilities for databasing;
- (e) accessibility of methodology.

1.2 As improvements in technology and new equipment become available, it is important for the continued sustainability of databases that the data produced are independent of the equipment used to produce them. This is, for example, the case with DNA sequencing data. Initially, radioactively labeled primers and sequencing gels were used to produce such data, whereas this can now be done using fluorescent dyes followed by separation on high throughput, largely automated, capillary gel electrophoresis systems. Despite these differences, the data produced with the various techniques are consistent with each other and independent of the techniques used to produce them. This can also apply to data produced using, e.g. DNA microsatellites (simple sequence repeats, SSR) or Single Nucleotide Polymorphisms (SNPs). This repeatability and reproducibility is important in the construction, operation and longevity of databases. Only in this way can a centrally maintained database, populated with verified data from a range of sources, be constructed in a cost-effective way such that the significant investment required in its establishment is only made once.

1.3 On the basis of these constraints, whilst the use of currently available multi-locus techniques where co-dominance cannot easily be recorded and reproducibility between laboratories using different equipment can be problematic is not ruled out, it is viewed as presenting difficulties in the DUS testing context. Consequently, the majority of this document is confined to considerations and recommendations with regard to the well-defined and researched uses of SSRs (microsatellites) and, for the future, to sequencing information (i.e. single nucleotide polymorphisms, SNPs).

2. Selection of Molecular Markers

2.1 *General Criteria*

Some criteria for choosing a specific marker or set of markers apply to all marker methods. These include, in particular:

- (a) useful level of polymorphism (indicated, for example, by PIC (polymorphism information content) value obtained on a set of representative varieties);
- (b) repeatability within, and reproducibility between, laboratories in terms of scoring data;
- (c) known distribution of the markers throughout the genome (i.e. map position), which whilst not being essential, is useful information;
- (a) the avoidance, as far as possible, of markers with “null” alleles (i.e. an allele whose effect is an absence of normal gene product at the molecular level), which again is not essential, but advisable.

2.2 *Specific criteria*

2.2.1 Microsatellite Markers (SSRs)

2.2.1.1 The analysis of simple sequence repeats (SSRs or microsatellites) using the polymerase chain reaction (PCR) is now widely used and has several advantages. Microsatellites are tandemly repeated DNA sequences, usually with a repeat unit of 2-4 base pairs (e.g. GA, CTT and GATA). In many species, multiple alleles have been shown to exist for some microsatellites, due to variations in the copy number of this repeat unit. Microsatellites can be analysed by PCR using specific primers, a procedure known as the sequence-tagged-site microsatellite (STMS) approach. The alleles (PCR products) can be separated by agarose or polyacrylamide gel electrophoresis. In order to develop sequence-tagged site microsatellites, information about the sequence of the DNA flanking the microsatellite is needed. This information can sometimes be acquired from existing DNA sequence databases, but otherwise has to be obtained empirically.

2.2.1.2 SSR markers are expressed co-dominantly, are generally easy to score (record) and can readily be mapped. They have been shown to be capable of analysis in different laboratories, and (given sufficient attention to experimental conditions) are robust and repeatable. In addition, they can be analyzed using automated, high throughput, non-radioactive DNA sequencers, based either on gel electrophoresis or capillary electrophoresis, and several can be analyzed simultaneously (multiplexing).

2.2.1.3 For effective microsatellite analysis, selecting high quality markers is essential. This includes a consideration of, *inter alia*:

- (a) the degree of “stuttering” (production of a series of one or more bands, differing by 1 repeat unit in size);
- (b) (n+1) peaks; Taq-polymerase often adds 1 bp to the end of a fragment. This can be prevented by using “pigtailed” primers;

- (c) the size of the amplification product;
- (d) effective separation between the various alleles in different detection systems;
- (e) reliable and reproducible scoring of the alleles in different detection systems;
- (f) the level of polymorphism (number of alleles detected) between varieties (note that this requires analysis of a significant number of varieties).

2.2.1.4 For scoring SSRs in different laboratories and using different detection equipment, it is crucial that reference alleles (i.e. sets of varieties) are defined and included in all analyses. These reference alleles are necessary because molecular weight standards behave differently in the various detection systems currently available and are therefore not appropriate for allele identification.

2.2.1.5 Primers used in a particular laboratory should be synthesized by an assured supplier, to reduce the possibility of different DNA profiles being produced from primers synthesized through different sources.

2.2.2 Single nucleotide polymorphism (SNP)

Single nucleotide polymorphisms (SNPs) can be detected via DNA sequencing, a routine technique which generally shows very high levels of repeatability over time and reproducibility between laboratories. However, detection of specific SNPs is currently carried out with a range of techniques, many of which are not yet routine. By their nature, SNPs have two allelic states, which makes scoring relatively straightforward and reliable and should reduce or remove many of the problems noted above. It also means that a large number of markers may need to be analyzed, either singly or in multiplexes, to allow the efficient profiling of a particular genotype.

2.2.3 Amplified fragment length polymorphism (AFLP)

The analysis of amplified fragment length polymorphisms, AFLP, involves the selective amplification of DNA restriction fragments. DNA is extracted, restricted using two enzymes in combination and then enzyme-specific adapters are added (ligated) to the fragments. PCR amplification is made selective by the use of primer pairs that are specific for these adapters and in addition contain 1-3 arbitrary bases that extend into the DNA fragments. These primer pairs will only amplify those fragments in which the primer extensions match the sequence flanking the restriction site. The PCR products are then separated by gel electrophoresis. AFLP profiles require no prior DNA sequence information and the number and nature of fragments amplified are altered by the choice of primer pair. Many loci can be sampled simultaneously and the method can be made robust by using stringent conditions in the PCR. However, AFLPs are dominant markers, the gel patterns produced can be very complex, radioactivity is widely used as a label (although there are alternatives) and there are also issues surrounding the availability of the technique (which is patented) for non-research purposes. For all of these reasons, no specific guidance on the use of AFLPs has been developed.

2.3 *Access to the technology*

Some molecular markers and materials are publicly available. However, a large investment is necessary to obtain, for example, high quality SSR markers and consequently markers and other methods and materials may be covered by intellectual property rights. UPOV has developed guidance for the use of characteristics which are examined by patented methods

(see document TGP/7/1 Annex 3, GN 14) and this guidance should be followed for the purposes of these guidelines. It is recommended that matters concerning intellectual property rights should be addressed at the start of any developmental work.

3. Material to be Analysed

The source and type of the material and how many samples need to be analyzed are the main issues with regard to the material to be analyzed.

3.1 Source of plant material

It is recommended that, wherever possible, the plant material to be analyzed is obtained from the sample of the variety submitted for examination for the purposes of Plant Breeders' Rights or for official registration. Use of such material will require the permission of the relevant authority and/or the breeder. Where appropriate, the individual plants from which the samples are taken should be traceable in case some of the plants subsequently prove not to be representative of the variety.

3.2 Type of plant material

The type of plant material to be sampled and the procedure for sampling the material for DNA extraction will, to a large extent, depend on the crop or plant species concerned. For example, in seed-propagated varieties, seed may be used as the source of DNA, whereas, in vegetatively propagated varieties, the DNA may be extracted from leaf material. Whatever the source of material, the method for sampling and DNA extraction should be standardized and documented. Furthermore, it should be verified that the sampling and extraction methods produce consistent results by DNA analysis.

3.3 Sample size

It is essential that the samples taken for analysis are representative of the variety.

3.3.1 Vegetatively propagated varieties

In principle, a single sample could be analyzed for vegetatively propagated varieties, as all individuals should be identical, except for mutations which are not usually detected using the current molecular tools. However, it is advisable in all cases to analyze at least duplicate samples.

3.3.2 Self-pollinated and mainly self-pollinated varieties

It is not always appropriate to assume that self-pollinated and mainly self-pollinated varieties are homozygous at all loci, including SSR or SNP loci. Heterogeneity can arise from, for example, residual heterozygosity, cross-pollination or physical admixture. For this reason, it is generally recommended that a number of single seeds be analyzed, as this will reveal any such heterozygosity. However, there may be reasons, including cost, to analyze a bulk sample of an agreed number of individuals to represent the DNA profile of a variety. Such analysis may alleviate potential problems due to the effects of maintenance and/or shifts in allele frequency. These issues are considered further below (Constructing the database).

3.3.3 Cross-pollinated varieties

Similar considerations apply in principle to varieties of cross-pollinated species. It is generally recommended that individual seeds/plants are analyzed because differences between varieties may be the result of differences in allele (or band) frequencies, as well as the presence or absence of particular alleles/bands. However, there are statistical approaches that can be applied and in the future, methods based on e.g. SNP analysis might allow the estimation of allele frequencies in bulked seed samples. As above, the analysis of a bulked sample of an agreed number of seeds will provide an overall DNA profile of a variety, which would be expected to be buffered against shifts in allele frequency.

3.4 DNA reference sample

It is recommended that a DNA reference sample collection should be created from the plant material sampled according to sections 3.1 to 3.3. This has the benefit that the DNA reference sample, rather than live plant material, can be stored and supplied to others. Furthermore, it would eliminate variations due to different extraction conditions.

4. Standardization of Analytical Protocols

4.1 Introduction

This document is not intended to provide detailed technical protocols for the production of DNA profiles of varieties. In principle, any suitable analytical methodology can be used, but it is important that the methodology is validated in an appropriate way. This may be via a full 'ISTA-type' ring test, or by developing a performance-based approach. In either case, there are some useful general considerations.

Any method used for genotyping and the construction of databases should be technically simple to perform, reliable and robust, allowing easy and indisputable scoring of marker profiles in different laboratories. This requires a level of standardization, for instance in the selection of markers, reference alleles and allele calling/scoring.

4.2 Quality criteria

4.2.1 It is not essential to standardize analytical protocols, but rather to agree on certain quality criteria concerning, for example:

- (a) the quality of DNA (this can be assessed for instance by measuring the absorbance ratio of the extract at 260/280 nm; the 260/230 nm ratio is also recommended by some);
- (b) the primer sequences used (with or without pigtailed, position of the primer, type of label used etc.)
- (c) the polymerase to be used in PCR-based methodologies (it may be advantageous to develop a list of assured products);
- (d) for PCR-based methodologies, the amount/concentration of each PCR component and other components (e.g. PCR buffer, MgCl₂, dNTP, primer, Taq polymerase, DNA template);
- (e) PCR cycling conditions (including length and temperature of initial denaturation; number of cycles; length and temperature of denaturation, annealing - for each primer pair - and extension; and length and temperature of final extension).

4.2.2 These agreed quality criteria should be set out in a protocol.

4.3 *Evaluation Phase*

4.3.1 Introduction

In order to select suitable markers and produce acceptable laboratory protocols for a given species, a preliminary evaluation phase involving more than one laboratory (i.e. a ring test) is recommended. This phase should be mainly concerned with selecting an agreed set of markers, and will usually involve the evaluation of existing markers, either published or available via other means. The number of markers to be evaluated will vary and depends on the possibilities presented by different species. The markers should derive from reliable sources (e.g. peer-reviewed publications) and be sourced from assured suppliers. The final choice of a number to be evaluated will be a balance between costs and the requirement to produce a satisfactory set of agreed markers at the end of the process. The objective is to produce an agreed set of markers that can be reliably and reproducibly analyzed, scored and recorded in different laboratories, potentially using different types of equipment and different sources of chemical reagents, etc.

4.3.2 Variety choice

An appropriate number of varieties, based on the genetic variability within the varieties concerned, should be selected as the basis for the evaluation phase. The choice of varieties should reflect an appropriate range of diversity and where possible should include some closely related and some morphologically similar varieties, to enable the level of discrimination in such cases to be assessed.

4.3.3 Interpretation of results

The next evaluation stage should also include a blind ring test if possible, to assess the whole methodology in an objective way. Any marker which causes difficulties in any of the laboratories involved in this evaluation phase should be rejected for subsequent use. Ideally, as from empirical experience most errors in the analysis of large variety collections seem to arise from scoring errors, construction of databases should be based on duplicate samples (e.g. different sub-samples of seed from the same variety), analyzed in different laboratories. Since the sub-samples (or DNA extracts from them) can be exchanged in the event of any discrepancy, this approach is very effective in highlighting sampling errors, or those due to heterogeneity within the samples, and eliminates possible laboratory artifacts (systematic errors).

4.4 *Scoring of molecular data*

4.4.1 A protocol for allele/band scoring should be developed in conjunction with the evaluation phase. The protocol should address how to score the following:

- (a) rare alleles (i.e. those at a specific locus which appear with a frequency below an agreed threshold (commonly 5-10%) in a population);
- (b) null alleles (an allele whose effect is an absence of normal gene product at the molecular level);

- (c) “faint” bands (i.e. bands where the intensity falls below an agreed threshold of detection, set either empirically or automatically, and the scoring of which may be open to question);
- (d) missing data (i.e. any locus for which there are no data recorded for whatever reason in a variety or varieties);
- (e) monomorphic bands (those alleles/bands which appear in every variety analyzed, i.e. are not polymorphic in a particular variety collection).

4.4.2 In addition. for markers such as SSRs, it is useful to establish a minimum and maximum size range for scoring markers for each primer pair (locus). Also, in cases where a gel-based system is used for visualizing marker bands, a suitable size (molecular weight) ‘ladder’ is recommended, to simplify the interpretation of results between laboratories.

5 Databases

5.1 Type of database

There are many ways in which molecular data can be stored, therefore, it is important that the database structure is developed in a way which is compatible with all intended uses of the data.

5.2 Transfer of data to the database

To reduce the number of errors in data transfer and transcription, it is advisable to automate transfer of data to databases as much as possible.

5.3 Data access / ownership

It is recommended that all matters concerning ownership of data and access to data in the database should be addressed at the outset of any work.

5.4 Data analysis

The purpose for which the data will be analyzed will determine the method of analysis, therefore, no specific recommendations are made within these guidelines.

5.5 Validating the database

When the first phase of the database is complete it is recommended to conduct a ‘blind test’, i.e. distribute a number of samples to different laboratories and ask them to use the agreed protocol in conjunction with the database to identify them.

6. Summary

The following is a summary of the approach recommended for the selection and use of molecular markers to construct central and sustainable databases of DNA profiles of varieties (i.e. databases that can be populated in the future with data from a range of sources, independent of the technology used).

- (a) consider the approach on a crop-by-crop basis;
- (b) agree on an acceptable marker source;

- (c) agree on acceptable detection platforms/equipment;
- (d) agree on laboratories to be included in the test;
- (e) agree on quality issues (see section 4.2);
- (f) verify the source of the plant material used (see section 3);
- (g) agree which markers are to be used in a preliminary collaborative evaluation phase, involving more than one laboratory and different detection equipment (see section 2);
- (h) conduct an evaluation (see section 4.3);
- (i) develop a protocol for scoring the molecular data (see section 4.4);
- (j) agree on the plant material/reference set to be analyzed, and the source(s);
- (k) analyze the agreed variety collection, in different laboratories/different detection equipment, using duplicate samples, and exchanging samples/DNA extracts if problems occur.
- (l) use reference varieties/DNA sample/alleles in all analyses;
- (m) verify all stages (including data entry) – automate as much as possible;
- (n) conduct a ‘blind test’ in different laboratories using the database;
- (o) adopt the same procedures for adding new data.

C. CROP SPECIFIC BMT GUIDELINES

It is recommended that, as appropriate, the general principles set out in section B of this document be used to develop crop specific BMT guidelines for approval within UPOV in the following way:

- (a) The Technical Committee (TC), on the basis of a recommendation by the relevant *Ad Hoc* Crop Subgroup on Molecular Techniques (Crop Subgroup), to commission the development of crop specific BMT guidelines and appoint a leading expert to co-ordinate, with all interested experts, the development of crop specific BMT guidelines in accordance with the principles set out in section B of this document.
- (b) Draft crop specific BMT guidelines to be reviewed by the relevant Crop Subgroup and the Working Group on Biochemical and Molecular Techniques and DNA Profiling in Particular (BMT) and, subject to agreement, to be submitted for approval by the TC.
- (c) Crop specific BMT guidelines approved by the TC to be published.

[End of document]