E

BMT Guidelines (proj.6)
ORIGINAL: English only
DATE: May 30, 2006

UPOV

INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS

GENEVA

DRAFT

GUIDELINES FOR DNA-PROFILING: MOLECULAR MARKER SELECTION AND DATABASE CONSTRUCTION ("BMT GUIDELINES")

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-fourth session, to be held in Nairobi, Kenya, from June 19 to 22, 2006

Working Group on Biochemical and Molecular Techniques, and DNA-Profiling in Particular (BMT) at its tenth session to be held in Seoul, Republic of Korea, from November 21 to 23, 2006

TABLE OF CONTENTS

A INTRODUCTION

The purpose of this document (BMT Guidelines) is to provide guidance for developing harmonized methodologies with the aim of generating high quality molecular data for a range of applications. The BMT Guidelines are also intended to address the construction of 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 The kinds of molecular techniques readily applicable for variety profiling are constrained by the requirement for the data to be repeatable, reproducible and consistent. Thus, while various multi-locus DNA profiling techniques have been successfully used for research, co-dominance cannot easily be recorded in many of them, and the reproducibility of complex banding patterns between laboratories using different equipment can be problematic. These factors are viewed as presenting difficulties in the context of variety profiling. Consequently, this document focuses on 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). Other techniques which rely on DNA sequence information, such as CAPS (cleaved amplified polymorphic sequences) and SCARS (sequence-characterized amplified regions) may also fulfill the above criteria but their use in DNA profiling of plant varieties has not yet been explored.

2. Selection of Molecular Markers

2.1 General Criteria

The following general criteria for choosing a specific marker or set of markers are intended to be appropriate for molecular markers irrespective of the use of the markers, although it is recognized that specific uses may impose certain additional criteria:

(a) useful level of polymorphism (indicated, for example, by PIC (polymorphism information content: see Glossary) 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 and helps to avoid the selection of markers that may be linked;
(d) the avoidance, as far as possible, of markers with "null" alleles (i.e. an allele whose effect is an absence of a PCR product at the molecular level), which again is not essential, but advisable.

2.2 Criteria for specific types of molecular markers

2.2.1 Microsatellite Markers (SSRs)

2.2.1.1 The analysis of simple sequence repeats (SSRs or microsatellites: see Glossary) using the polymerase chain reaction (PCR) is now widely used and has several advantages.

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 (see Glossary);
(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);
(g) avoidance of linkage.

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: see Glossary) 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 only two allelic states in diploid plants, although this may vary in polyploids where there will be dosage effects. That makes the scoring of SNPs 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 and effective profiling of a particular genotype.

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 products or methodologies which are the subject of intellectual property rights (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.

4. 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.

4.1 Source of plant material

The plant material to be analyzed should be an authentic, representative sample of the variety and, where possible, should be obtained from the sample of the variety used for examination for the purposes of Plant Breeders' Rights or for official registration. Use of samples of material submitted for examination for the purposes of Plant Breeders' Rights or for official registration will require the permission of the relevant authority, breeder and/or maintainer, as appropriate. 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.

4.2 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.

4.3 Sample size

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

4.3.1 Vegetatively propagated varieties

In principle, a single sample could be analyzed for vegetatively propagated varieties, as all individuals should be identical. However, it is advisable in all cases to analyze at least duplicate samples. If any differences are found, additional samples should be analyzed.

4.3.2 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.

4.3.3 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.

4.3.4 Hybrids

The appropriate method for ensuring that samples taken for analysis of hybrids are representative of the variety will depend on the type of hybrid. It should be recognized that hybrid varieties will be heterozygous at the loci coding for DNA markers, but could still be phenotypically uniform. The number of samples chosen for analysis will depend on the precise issue being addressed and the type of hybrid variety being assessed. The information concerning different types of hybrid varieties, provided in document TG/1/3 “General Introduction to the Examination of Distinctness, Uniformity and Stability and the Development of Harmonized Descriptions of New Varieties of Plants” (see Chapter 6.4.3), should be considered in that respect.

4.4 DNA reference sample

It is recommended that a DNA reference sample collection should be created from the plant material sampled according to sections 4.1 to 4.3. This has the benefit that the DNA reference sample can be stored and supplied to others.

5. Standardization of Analytical Protocols

5.1 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 an internationally recognized method of validation, 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.

5.2 Quality criteria

5.2.1 It is important to agree on certain quality criteria concerning, for example:

(a) the quality of DNA (this can be assessed for instance by measuring the absorbence 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 pigtails, 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, $MgCl_2$, 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):

5.2.2 The detailed methodology should be set out in a protocol.

5.3 Evaluation Phase

5.3.1 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. an internationally recognized method of validation, e.g. a ring test according to internationally agreed standards) is recommended. This phase should be mainly concerned with selecting a 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.

5.3.2 Variety choice

An appropriate number of varieties, based on the genetic variability within the species and type of variety 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.

5.3.3 Interpretation of results

The next evaluation stage should, if possible, include an internationally recognized method of validation 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).

5.4 Scoring of molecular data

5.4.1 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 PCR 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).

5.4.2 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 within and between laboratories. However, that should not be considered as a replacement for an allele reference collection.

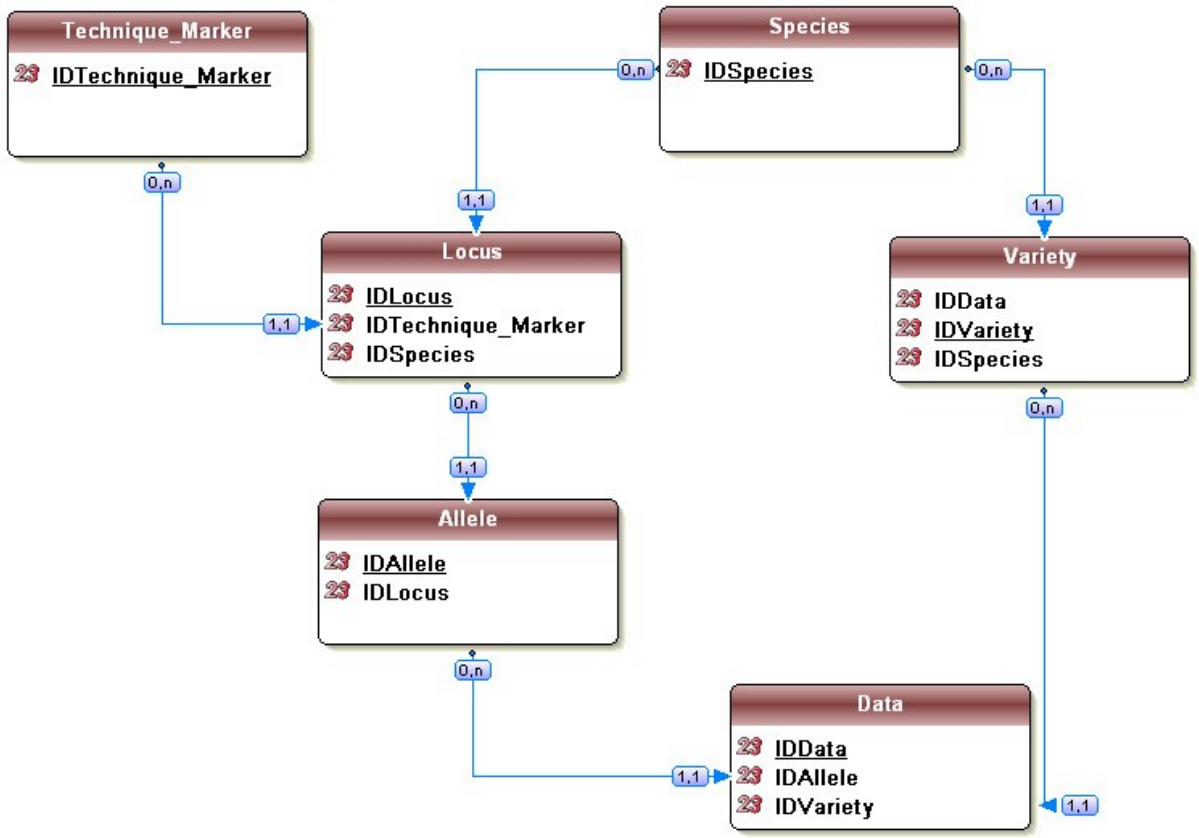
6. Databases

6.1 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.

6.2 Database model

The database model should be defined by IT database experts in conjunction with the users of the database. As a minimum the database model should contain six core objects: Species; Variety; Technique; Marker; Locus; and Allele.



6.3 Data Dictionary

6.3.1 In a database, each of the objects becomes a table in which fields are defined. For example:

(a) Technique/Marker code:

indicates the code or name of the technique or type of marker used, *e.g. SSR, SNP, etc.*

(b) Locus code:

indicates name or code of the locus for the species concerned, *e.g. gwm 149, A2, etc.*

(c) Allele code:

indicates name or code of the locus for the species concerned, *e.g. 1, 123, etc.*

(d) Data value:

indicates a data value for a given sample on a given locus-allele, *e.g. 0 (absence), 1(presence), 0.25 (frequency) etc.*

(e) Variety

the variety is the object for which the data have been obtained.

(f) Species

the species is indicated by the botanical name or the national common name, which sometimes also refers to the type of variety (e.g. use, winter/spring type etc.). The use of the UPOV code would avoid problems of synonyms and would, therefore, be beneficial for coordination.

6.3.2 In each table, the number of fields, their name and definition, the possible values and the rules to be followed, need to be defined in the "data dictionary".

6.4 Table Relationship

6.4.1 The links between the tables are an important aspect of the database design. The links between tables can be illustrated as follows:

Table	Link	Table	Description
Woman	*0* or *1 to n* *(0, n)*	Child	*0*: A woman may have no child *1 to n*: a woman may have 1 to n children (she is then a mother)
Child	*1 to 1* *(1,1)*	Woman	A given child has only one biological mother

6.4.2 The following table indicates the relationship between the six minimum core objects, as proposed in the database model in Section 6.2:

Table	*Link*	Table	description
Technique/marker	*0* or *1 to n*	Locus	*0*: A technique/marker can be present in Technique/marker, even if no locus/allele is yet used in the database *1 to n*: a given type of marker can provide 1 to n useful loci
Locus	*1 to 1*	Technique/marker	A given locus is defined within the scope of a given technique/marker
Locus	*1 to n*	Allele	For each Locus 1, or more than 1, allele can be described
Allele	*1 to 1*	Locus	A given Allele is defined within the scope of a given Locus
Allele	*0* or *1 to n*	Data	*0*: a given Allele can be defined, but without data *1 to n*: a given allele can be found in 1 to n data
Data	*1 to 1*	Allele	data corresponds to a given allele
Variety	*0* or *1 to n*	Data	*0*: the variety has no data *1 to n*: the variety has data
Data	*1 to 1*	Variety	data corresponds to a given variety
Data	*1 to 1*	Species	data is obtained for a given variety, then for the species of the variety.
Species	*0* or *1 to n*	Data	*0*: a species can have no data. *1 to n*: a species can have 1 to n data.

6.5 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.

6.6 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 beginning of any work.

6.7 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.

6.8 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.

7. 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 type and 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 5.2);
(f) verify the source of the plant material used (see section 4);
(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 5.3);
(i) develop a protocol for scoring the molecular data (see section 5.4);
(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 procedures for adding new data.

GLOSSARY

Microsatellites, or simple sequence repeats (SSRs)

Microsatellites, or simple sequence repeats (SSRs) 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 analyzed 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.

Single nucleotide polymorphisms (SNPs)

Single nucleotide polymorphisms (SNPs) (pronounced "snips") are DNA sequence variations that occur when a single nucleotide (A,T,C, or G) in the genome sequence is altered. For example a SNP might change the DNA sequence AAGGCTAA to ATGGCTAA. Generally, for a variation to be considered a SNP, it must occur in at least 1% of the population. The potential number of SNP markers is very high, meaning it should be possible to find them in all parts of the genome. SNPs can occur in both coding (gene) and non-coding regions of the genome. The discovery of SNPs involves comparative sequencing of numbers of individuals from a population. More commonly, potential SNPs are identified by comparing aligned sequences from the available sequence databases. Although they can be detected by relatively straightforward PCR + gel electrophoresis, high throughput and micro-array procedures are being developed for automatically scoring hundreds of SNP loci simultaneously.

Cleaved amplified polymorphic sequences (CAPS)

Cleaved amplified polymorphic sequences (CAPS) are DNA fragments amplified by PCR using specific 20-25 bp primers, followed by digestion with a restriction endonuclease. Subsequently, length polymorphisms resulting from variation in the occurrence of restriction sites are identified by gel-electrophoresis of the digested products. In comparison with markers such as RFLPs, polymorphisms are more difficult to identify because of the limited size of the amplified fragments (300-1800 bp). CAPS analysis, however, does not require Southern blot hybridization and radioactive detection. CAPS have generally been applied predominantly in gene mapping studies to date.

Sequence-characterized amplified regions (SCARS)

Sequence-characterized amplified regions (SCARS) are DNA fragments amplified by PCR using specific 15-30 bp primers, designed from previously identified polymorphic sequences. By using longer PCR primers, SCARs avoid the problem of low reproducibility. They are also usually co-dominant markers. SCARs are locus specific and have been applied in gene mapping studies and marker assisted selection.

Polymorphic information content (PIC) values

Polymorphic information content (PIC) values are a measure of the allelic diversity at a locus, used to estimate and compare the discrimination power of molecular markers. The PIC value of a PCR-based marker can be calculated as:

$$1 - \sum_{j=1}^{n} P_{ij}^2$$

where P_{ij} is the frequency of the *j*th PCR pattern for genotype *i*.

Pig-tailing

In SSR analysis, "pig-tailing" is the addition of a short oligonucleotide sequence to the primers used in the PCR, as a way of improving the clarity of the amplification products and reducing artifacts.

Null allele

In SSR analysis, a "null allele" is an allele at a particular locus whose effect is seen as an absence of a PCR product.

Stutter bands

In SSR analysis, "stutter bands" is the occurrence of a series of one or more bands, differing by 1 repeat unit in size, following PCR.

[End of document]