UPOV

International Union for the Protection of New Varieties of Plants

E

UPOV/INF/17/2

Original: English
Date: September 21, 2021

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

Document adopted by the Council
on September 21, 2021
by correspondence

TABLE OF CONTENTS

A. INTRODUCTION

The purpose of this document (BMT Guidelines) is to provide guidance on harmonized principles for the use of molecular markers with the aim of generating high quality molecular data for a range of applications. Only DNA molecular markers are considered in this document.

The BMT Guidelines are also intended to address the construction of databases containing molecular profiles of plant varieties, possibly produced in different laboratories using different technologies. In addition, the aim is to set high demands on the quality of markers and on the desire for generating reproducible data using these markers in situations where equipment and/or reaction chemicals might change. Specific precautions need to be taken to ensure quality entry into a database.

B. GENERAL PRINCIPLES

For DNA profiling of a plant variety, a set of molecular markers and a method to detect them are required. Two different sets of molecular markers detected with the same method will result in two different DNA profiles for a particular variety. In contrast, two different methods to detect the specific alleles of a given molecular marker set are expected to result in identical DNA profiles. Standardization of the detection method and technology is not required as long as the performance meets the quality criteria and the resulting DNA profiles are consistent. Irrespective of the technology used to detect defined marker sets, the genotype of a particular variety should not be affected.

Molecular marker sets, marker detection methods and subsequently the database developmental process can be subdivided into 5 different phases:

1. Selection of molecular markers
2. Selection of detection method
3. Validation and harmonization of the detection method
4. Construction of the database
5. Data exchange

This document describes these different phases in more detail. It is considered that these phases are independent from the stage of development of genotyping technologies and future improvements in high-throughput sequencing.

1. Selection of Molecular Markers

1.1 Sets of varieties for the selection process

For DNA profiling of plant varieties and database construction, molecular markers should be selected according to the objective. To start the marker selection process an appropriate number of varieties (development set) is needed to reflect at the most the diversity observed within the group/crop/species/type for which the markers are intended to be discriminative. Further selection is performed by profiling additional varieties (validation set) to measure the performance of the markers. Criteria for the choice of the validation set could be:

(a) genetically very similar varieties or lines, NILs, RILs
(b) parental lines and offspring
(c) genetically close but morphologically distinct varieties (e.g. mutants)
(d) some morphologically close varieties with different pedigree
(e) different lots of the same variety
(f) different origins of the same variety

1.2 Molecular markers – performance criteria

The following general criteria for selecting a specific marker or set of markers are intended to be appropriate irrespective of the use of the markers:

(a) Repeatability, reproducibility and robustness within and between laboratories in terms of scoring data;

(b) Possible sources of molecular markers

- Molecular markers derived from public resources
- Molecular markers derived from non-public resources, screening and selection of commercially available species-specific chips and arrays.
- Molecular markers selected from newly generated sequence data;

(c) 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;

(d) Allowance of easy, objective and indisputable scoring of marker profiles. These good performing markers are preferred over complex marker profiles that are sensitive to interpretation. Clear black and white answers also allows for easier harmonization;

(e) Co-dominant markers are generally preferred over dominant markers as they have a higher discriminative power;

(f) Markers located in coding and/or in non-coding regions; and

(g) The use of molecular markers is species-specific and should take into account the features of propagation of the species.

It is recognized that specific uses may impose certain additional considerations that may include but are not limited to:

i. The number of markers should be balanced with the accuracy of the genotype required for the objective. The number of markers to reach the necessary resolution or discriminative power depends on marker-type (dominant/co-dominant; bi-/multi-allelic), species and the quality of the marker performance;

ii. Coverage of the genome and the linkage disequilibrium should reflect the objectives. Knowing the physical and/or genetic position of the selected markers on the genome is not essential but enables a good selection of markers.

2. Selection of the Detection Method

2.1 DNA profiling methods - general considerations

2.1.1 Important considerations for choosing DNA profiling methods that generate high quality molecular data are:

(a) reproducibility of data production within and between laboratories and detection platforms (different types of equipment);
(b) repeatability over time;
(c) discrimination power;
(d) time and labor intensity;
(e) robustness of performance in time and conditions (sensitiveness to subtle changes in the protocol or condition);
(f) flexibility of the method, possibility to vary in the number of samples and/or number of markers;
(g) interpretation of the data produced is independent of the equipment;
(h) sustainability of databases;
(i) accessibility of methodology;
(j) independence of a specific machine, specific chemistry, specific supplier, particular partners or products;
(k) suitable for automation;
(l) suitable for multiplexing; and
(m) cost effective (costs, number of samples and number of markers are in balance).

2.2 Access to the Technology

Some molecular markers and materials are publicly available. However, a large investment is likely to be necessary to obtain high quality markers, consequently markers and other methods and/or 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 and these should be followed. It is recommended that matters concerning intellectual property rights should be addressed at the start of any developmental work.

3. Validation and harmonization of a marker set and detection method

3.1 Validation and harmonization – general considerations

Molecular markers and detection methods should be robust and give rise to consistent DNA profiles. Performance of molecular markers and genotyping methods are evaluated in a validation process. In case of shared database, consistency of the DNA profiles in different laboratories is evaluated in the harmonization process using different equipment and chemistries. The usage of validated markers and methods will lead to harmonized results.

3.2 Performance considerations - validation of markers and methods

The selected marker set should be fit-for-purpose. The accuracy should be measured. To determine the suitability of a method and DNA marker set several points should be considered:

(a) Discriminative capacity/informativeness;

(b) Repeatability; where identical test results are obtained with the same method, on identical test items, in the same laboratory, by the same operator, using the same equipment within short intervals of time.

(c) Reproducibility; where test results are obtained with the same method, on identical test items, within the same laboratory or between different laboratories, with different operators, using different equipment.

(d) Robustness; a measure of its capacity to remain unaffected by small, but deliberate deviations from the experimental conditions described in the procedure parameters and provides an indication of its reliability during normal usage; and

(e) Error-rate.

Definitions of the performance characteristics are based on: ISO 16 577:2016

3.3 Consistency considerations

To achieve consistency of results, the process of harmonization of markers and methods between different laboratories in the case of a shared database (ring test) should consider:

(a) Use of a defined collection of varieties representing a wide range of alleles as a reference in all labs to test consistency between labs;

(b) Inclusion of duplicates, sub-samples, individual plants of a variety to check the consistency of the DNA profiles and estimate the error-rate between labs;

(c) Agreements on the scoring of molecular data. The necessity to develop a protocol for allele/band scoring between labs depends on the used marker type (e.g. essential for SSR). The protocol could address how to score the following:

i. rare alleles (i.e. those at a specific locus which appear with a frequency below an agreed threshold (commonly 5-10%) in a population);

ii. null alleles (an allele whose effect is an absence of PCR product at the molecular level);

iii. "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);

iv. missing data (i.e. any locus for which there are no data recorded for whatever reason in a variety or varieties); and

v. monomorphic bands or non-informative allele scores (those alleles/bands which appear in every variety analyzed, i.e. are not polymorphic in a particular variety collection).

4. Construction of a Species-Specific Database

The data that is stored in a database and how it is stored should reflect the process of producing the data. Therefore, database construction should consider different levels of data processing (*i.e.* raw data, sequence data…). The database should store the end results, e.g. the DNA profile as well as how it was derived both in terms of laboratory method description and the computational steps.

4.1 Recommendations for database design

Design of databases could consider the following aspects:

(a) The database architecture should be flexible, e.g. allow for storing both flat files as well as compressed archives.

(b) Separate tables and entries are required for laboratory experimental work, data processing and the allele scores.

(c) Storage of information at different levels for example allele scores and any rules for interpretation behind the decision and links to the raw data (tiff files, bam files) that were produced.

(d) For sequencing data, variant call files in VCF or BCF format corresponding to the standard version 4.2 or higher. Header entries should contain the name and version of the different scripts used for both sequence read mapping, read filtering, variant calling and variant filtering in such a way that a bioinformatician can repeat the analysis.

(e) In case of replicate samples where the DNA profile does not match, the record needs to be flagged or filtered out where appropriate. The rules applied for these cases need to be documented in a publicly accessible code repository that is referenced from the variant call file. Frequencies could also be used for heterogeneous varieties.

(f) Validation of the VCF and or BCF data against relevant specifications.

(g) Easy to share data, (e.g. API).

4.2 Requirements of the plant material

The source, type of the material and how many samples to be stored and shared in the database should be considered.

4.2.1 Source of plant material

The plant material to be analyzed should be an authentic, representative sample of the variety and, when 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 these samples will require the permission of the relevant authority, breeder and/or maintainer, as appropriate. The plant material from which the samples are taken should be traceable in case some of the samples subsequently prove not to be representative of the variety.

4.2.2 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 documented. Furthermore, it should be verified that the sampling and extraction methods produce consistent results by DNA analysis.

4.2.3 Sample size and type (bulk or individual samples)

It is essential that the samples taken for analysis are representative of the variety. Consideration should be given to the features of propagation (see the General Introduction).

4.2.4 DNA reference sample

A DNA reference collection may be created from the plant material sampled. The method for sampling should follow recommended procedures and quality criteria for DNA extraction should be set. Both need to be documented.

The DNA samples should be stored in such a way as to prevent degradation (e.g. storing it at -80°C). The transfer of DNA reference samples is described in document TGP/5: Section 1.

4.3 Processing of sequence data

A detailed log of the data processing pipeline may include:

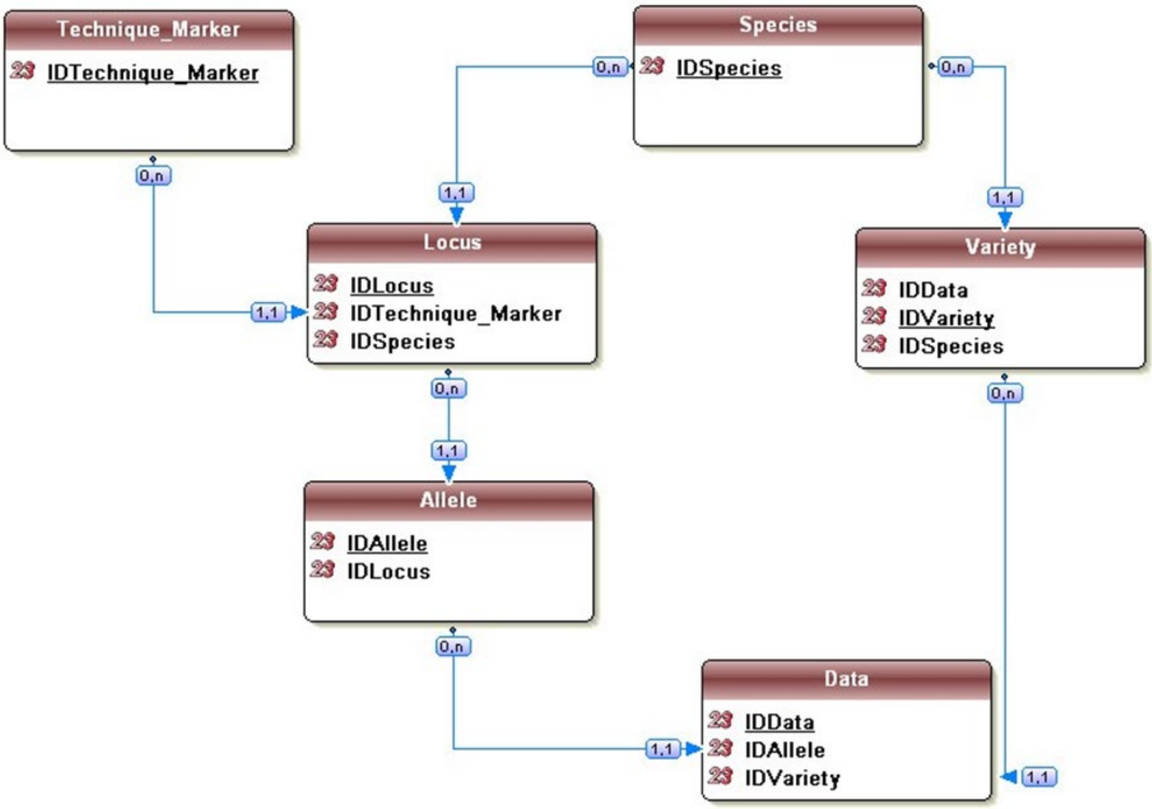
(a) type and versions of tools;
(b) command line used for the tool including thresholds;
(c) reproducibility counts;
(d) possibility for sharing the data and process;
(e) raw alignment data (BAM or CRAM files) should be stored where possible;
(f) multi-sample VCF files are not suitable, one VCF file per variety must be present;
(g) if VCF files are stored, all positions (both variants & non-variants) and their depth should be stored;
(h) both heuristic and probabilistic approaches should be considered and compared for detection methods;
(i) databases should facilitate input and output of variant call data in standardized format (VCF or BCF);
(j) the data processing pipeline should result in a detailed log file which should be stored in conjunction to the variant call data;
(k) if possible, raw data should be stored so that data processing can be repeated with new or updated tools; and
(l) a p-value or uncertainty for a given allele should be stored.

4.4 Type of database

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

4.5 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; Marker detection method; Marker; Locus; and Allele. For variants obtained from sequencing data, VCF files can be stored in a relational or no SQL database. In this case, each database record for a variant has a defined genome version, chromosome, position, reference allele.



4.6 Data Dictionary

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

(a) Marker type: indicates the code or name of the technique or type of marker used, e.g. SSR, SNP, etc.

(b) Reference genome position or Locus code: Preferably, a genome assembly version, chromosome and position should be provided if a reference genome is available for the species concerned, e.g. SL2.50ch05:63309763 for tomato *Solanum lycopersicum* assembly version 2.50 on chromosome 5 position 63309763. If no reference genome is available or the location is unknown, a name or code of the locus for the species concerned can be used, e.g. gwm 149, A2, etc.

(c) Genotype: For SNP profiles, the allele composition of the SNP or MNP should be given, e.g. A/T or A/A. For other techniques, genotype indicates the name or code of the allele of a given locus for the species concerned, e.g. 1, 123, etc.

(d) Allele depths or Data value: For SNPs obtained from next generation sequencing data this should indicate the depth of coverage for alleles e.g. 10/20 for an A/T allele in which the A is covered by 10 reads and the T by 20. Otherwise, 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: Variety denomination or breeder’s reference: the variety is the object for which the data have been obtained.

(f) Type of variety: e.g. Inbred Line or Hybrid

(g) 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 is recommended to avoid problems of synonyms.

4.6.2 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".

4.7 Data access - ownership

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

5. Data Exchange

5.1 Data exchange scenarios

For cooperation purposes, the data model should allow different types of scenarios including the exchange of data produced from a standardized set of markers for a specific crop (Scenario 1), and search and view data of selected varieties generated from the same standardized set of markers (Scenario 2). Technical details on both scenarios are described in the Annex: Data exchange scenarios and data transfer methods.

5.2 Data exchange methods

5.2.1 Fingerprint data transmission may contain a range of information, such as loci, samples, DNA, fingerprint data and fingerprint profiles. Method of data transmission needs to be determined by the content to be transferred and should consider the following:

(a) amount of data
(b) complexity of data
(c) requirements for query or search functions

Technical details on data transfer methods are described in the Annex: Data exchange scenarios and data transfer methods.

5.2.2 Commonly used data formats include: zip, csv, json and xml. Their respective characteristics are as follows:

(1) The zip format allows a variety of data information files in the original format and due to its large data compression ratio and ease of transmission is suitable for large and complex data.

(2) The csv format is more suitable for data information in simple data format, which has the advantage of having less invalid data and faster processing speeds.

(3) The json and xml formats can contain more complex character data information and more redundant information, but both offer good readability.

6. Summary

The following is a summary of the approach recommended for high quality DNA profiling of varieties including the selection and use of molecular markers as well as the construction of shared and sustainable molecular databases (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;
(f) verify the source of the plant material used;
(g) agree which markers are to be used in a preliminary collaborative evaluation phase, involving more than one laboratory and different detection equipment;
(h) conduct an evaluation;
(i) develop and agree a protocol for scoring the molecular data;
(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 references (varieties, DNA samples and alleles, as appropriate) 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 procedures for adding new data.

C. LIST OF ACRONYMS

API	Application Programming Interface
BAM	Binary Alignment Map
BCF	Binary Call Format
CRAM	Compressed Reference-oriented Alignment Map
MNP	Multiple Nucleotide Polymorphism
NGS	Next Generation Sequencing
NIL	Near Isogenic Line
RIL	Recombinant Inbred Line
SAM	Sequence Alignment Map
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
SSR	Simple Sequence Repeats
TIFF	Tagged Image File Format
VCF	Variant Call Format

[Annex follows]

ANNEX

DATA EXCHANGE SCENARIOS AND TRANSFER METHODS

A: Data exchange scenarios

Scenario 1: exchange of data produced from a standardized set of markers for a specific crop

In order to exchange data about the marker set used for a specific crop, the following web service can be used: https://office.org/locus?upov_code={upovcode}&type={marker type}&method={observation method}

For example, to obtain marker set information for maize using SSR and CE method, the following URL should be accessed:
https://office.org/locus?upov_code=ZEAAA_MAY&type=SSR&method=CE

The result would be:

```
{"techniqueid":
"CN_SSR_ZEAA_MAY_CE_V_1",
"description": "Laboratory method description"
["locusid": "M01",
"alleles":
["alleleid": "238/256",
"examplevariety":
],
["alleleid": "238/271",
"examplevariety":
],
["alleleid": "246/246",
"examplevariety":
],
["alleleid": "246/248",
"examplevariety":
],
["alleleid": "246/250",
"examplevariety":
],
["alleleid": "246/254",
"examplevariety":
],
["alleleid": "246/256",
"examplevariety":
],
["alleleid": "246/260",
"examplevariety":
],
["alleleid": "246/277",
"examplevariety":
],
["alleleid": "246/284",
"examplevariety":
],
["alleleid": "246/288",
"examplevariety":
],
["alleleid": "248/250",
"examplevariety":
],
["alleleid": "248/256",
"examplevariety":
],
["alleleid": "248/271",
"examplevariety":
],
["alleleid": "248/290",
"examplevariety":
],
["alleleid": "250/250",
"examplevariety":
],
["alleleid": "250/252",
"examplevariety":
],
["alleleid": "250/256",
"examplevariety":
],
["alleleid": "250/275",
"examplevariety":
],
["alleleid": "252/256",
"examplevariety":
],
["alleleid": "252/260",
"examplevariety":
],
["alleleid": "252/271",
"examplevariety":
],
["alleleid": "252/273",
"examplevariety":
],
["alleleid": "252/282",
"examplevariety":
],
["alleleid": "254/254",
"examplevariety":
],
["alleleid": "254/271",
"examplevariety":
],
["alleleid": "254/284",
"examplevariety":
],
["alleleid": "254/286",
"examplevariety":
],
["alleleid": "256/256",
"examplevariety":
],
["alleleid": "256/264",
"examplevariety":
],
["alleleid": "256/266",
"examplevariety":
],
["alleleid": "256/271",
"examplevariety":
],
["alleleid": "256/284",
"examplevariety":
],
["alleleid": "256/286",
"examplevariety":
],
["alleleid": "258/258",
"examplevariety":
],
["alleleid": "264/284",
"examplevariety":
],
["alleleid": "271/292",
"examplevariety":
]
],

["locusid"="M02”.
"alleles": […]
]} vi
```

Scenario 2: search and view data of selected varieties generated from the same standardized set of markers

In order to search and view molecular data of a variety, the following web service can be used:
https://office.org/variety?id={irn}&techniqueid={technique_code} vi

For example,
https://office.org/variety?id=XU_30201800000140 &techniqueid= CN_SSR_ZEAA_MAY_CE_V_1 vi

The result would be:

```
{"techniqueid": "CN_SSR_ZEAA_MAY_PAGE ",
"varietyid": " XU_30201800000140 ",
"computationalsteps": "xxxxxxxxxxxx"
"data":
[
"id": "M01",
"value" : "254/254"
],
[
"id": "M02",
"value" : "347/347"
],
[
"id": "M03",
"value" : "292/292"
],
[
"id": "M04",
"value" : "361/361"
],
…
} vi
```

B: Data transfer methods

The following provides an example of constructing a fingerprint packet in a zip format for data transmission. This method first needs to use independent IDs to identify samples, DNA, fingerprint data and fingerprint atlas. After that, the json format data file contains all the loci, samples and DNA information. Each fingerprint data is stored independently in its own json format file. The fingerprint ID will be bound to the corresponding locus of the fingerprint data, and all fingerprint data files and fingerprint spectrum files will be stored independently in the corresponding directory. So the format structure of the fingerprint data packet is as follows:

```
zip/markers.json
zip/samples.json
zip/dnas.json
zip/genes/gene_id_1.json
zip/genes/gene_id_2.json
......
zip/genes/gene_id_n.json
zip/maps/map_id_1.png
zip/maps/map_id_2.png
......
zip/maps/map_id_m.png
```

The zip format fingerprint packet can be extended to include more information. The core of the packet is the fingerprint data file, which is the core of the correlation, so that the correlation between the parts can be correctly parsed, allowing data transmission across different systems.

[End of Annex and of document]