UPOV

International Union for the Protection of New Varieties of Plants

E

UPOV/INF/17/2 Draft 3

Original: English
Date: August 25, 2020

DRAFT
(REVISION)

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 Working Group on Biochemical and Molecular Techniques, and DNA-Profiling in Particular (BMT)
at its nineteenth session, to be held in Alexandria, United States of America,
from September 23 to 25, 2020

Disclaimer: this document does not represent UPOV policies or guidance

Note for Draft version

Endnotes are background information when considering this draft and will not appear in the final, published document.

~~**Strikethrough**~~ **(highlighted in grey)** indicates deletion from the text of document UPOV/INF/17/1.

Underlining **(highlighted in grey)** indicates insertion to the text of document UPOV/INF/17/1.

~~**Double strikethrough**~~ and **double underlining** indicate changes to document UPOV/INF/17/2 Draft 1.

~~**Double strikethrough**~~ and **double underlining** **(highlighted in yellow)** indicate changes to document UPOV/INF/17/2 Draft 2.

TABLE OF CONTENTS

A. INTRODUCTION

The purpose of this document (BMT Guidelines) is to provide guidance ~~for developing~~ on harmonized ~~methodologies~~ principles for the use of[i] molecular markers[ii] 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 ~~the~~[ii] 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. ~~Generated DNA profiles are usually stored in appropriate databases. DNA profiling methods develop very fast and new technologies will keep being discovered. As a consequence, the methods for molecular marker detection will change in the future and may shift from single sample endpoint methods towards whole genome sequences approaches~~.[ii] 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[ii] different phases:

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

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

~~1. Selection of a Molecular Marker Methodology~~

~~1.1 Important criteria for choosing a methodology are:~~

~~(a) reproducibility of data production between laboratories and detection platforms (different types of equipment);~~
~~(b) repeatability over time;~~
~~(c) discrimination power;~~
~~(d) possibilities for databasing; and~~
~~(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 interpretation of 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.~~

~~1.3 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 and is~~

~~very important in generating a centrally maintained database, populated with verified data from a range of sources.~~

~~1.4 The 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.~~

~~1.5 These factors present 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,[i] to sequencing information (i.e. single nucleotide polymorphisms, SNPs). Other techniques which rely on DNA sequence information, such as cleaved amplified polymorphic sequences (CAPS) and sequence-characterized amplified regions (SCARs) may also fulfill the above criteria but their use in DNA profiling of plant varieties has not yet been explored.~~ **Error! Bookmark not defined.**

~~2.~~ 1. Selection of Molecular Markers

1.1 Sets of varieties for the selection process [ii]

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

~~2.~~1.2 ~~General Criteria~~ Molecular markers – performance criteria [ii]

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

(a) ~~useful level of polymorphism;~~ 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, iv]

(b) repeatability, reproducibility and robustness [ii] within and 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~~ Coverage of the genome and the linkage disequilibrium should reflect the objectives. Knowing the physical and/or genetic [ii] position of the selected markers on the genome is not essential but enables a good [ii] selection of markers [v]; ~~and~~

(d) 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 [ii, vi];

(e) 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~~.~~;

(f) 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; [vii]

(g) Co-dominant markers are generally [ii] preferred over dominant markers as they have a higher discriminative power; [vii]

(h) Durability of the marker. When a marker is located in a genomic area that is not subject to selection by breeders, there is a better chance that the marker will be informative in a durable way;

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

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

~~*2.2 Criteria for specific types of molecular markers*~~

~~2.2.1 Microsatellite Markers~~

~~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 used and analyzed in different laboratories, and under specific experimental conditions are generally 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 suitable detection systems;~~
~~(e) reliable and reproducible scoring of the alleles in different detection systems;~~
~~(f) the level of polymorphism 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 as a result of using 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. By their nature, SNPs have only two allelic states in diploid plants, although this may vary in polyploids where there will be dosage effects. The simple makeup of SNPs makes the scoring of SNPs~~

~~relatively straightforward and reliable. 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.~~ [viii]

2. Selection of the Detection Method [ix]

2.1 ~~Genotyping~~ DNA profiling [ii] 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 of the method;
(d) time and labor intensity of the method;
(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; [ix]
(j) independence of a specific machine, specific chemistry, specific supplier, particular partners or products; [ii]
(k) suitable for automation;
(l) suitable for multiplexing; and
(m) cost effective~~;~~ (costs, number of samples and number of markers are in balance). [ix]

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~~, 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 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. Validation and harmonization of a marker set and detection method [x]

3.1 Validation and harmonization – general considerations [xi]

Molecular marker selection and detection method descriptions are based on performance: markers and methods should be robust and give rise to consistent DNA profiles. Performance of molecular markers and genotyping methods is evaluated in a validation process. In case of shared database, consistence 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. [ii, xi]

3.2 Performance considerations - validation of markers and methods

It is needed to determine how suitable the selected marker set is (fit-for-purpose). The accuracy should be measured. To determine the adequacy of a method and DNA marker set several points should be considered:

(a) Discriminative capacity/informativeness;
(b) Repeatability[1];

[1] Repeatability: *Precision* (the relative standard deviation of test results) obtained under *repeatability conditions*.

Repeatability conditions are conditions where 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[2];
(d) Robustness[3]; and
(e) Error-rate. [xii]

Definitions of the performance characteristics are based on: DOI: 10.13140/RG.2.1.2060.5608

3.3 *Consistence considerations - harmonization of markers and methods between different laboratories in case of shared database – ring test* [ii]

(a) Use defined collection of varieties representing a wide range of alleles as a reference in all labs to test consistency between labs [ii, xiii]

(b) Duplicates, sub-samples, individual plants of a variety to check the consistency of the DNA profiles and estimate the error-rate between labs [ii, xiii]

(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 ~~but less urgent for SNP markers~~). [ii, xiii] 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).

~~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 consider quality criteria concerning, for example:~~

~~(a) the quality of DNA;~~
~~(b) methods of DNA extraction~~
~~(c) the primer sequences;~~

[2] Reproducibility: *Precision* (the relative standard deviation of test results) obtained under *reproducibility conditions. Reproducibility conditions are conditions 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.*

[3] Robustness: The robustness of a method is 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.

~~(d) the polymerase to be used in PCR-based methodologies;~~
~~(e) for PCR-based methodologies, the amount/concentration of each PCR component and other components; and~~
~~(f) PCR cycling conditions.~~

~~5.2.2 The detailed methodology should be set out in a protocol.~~ [xiv]

4. Construction of a Species-specific Database [xv]

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 1) the end results, e.g. the DNA profile as well as how it was derived both in terms of 2) laboratory method description and 3) the computational steps for deriving a DNA profile. [ii]

4.1 Recommendations for database design [xv]

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. [xv]

(b) ~~Contains different tables, s~~Separate tables and entries are required for laboratory experimental work [ii] , data processing and the allele [ii] scores. [xv]

(c) Store information at different levels (allele scores / how the allele score was called (the rules or the interpretation rules behind a decision) / (links) to the raw data (tiff files, bam files, files that came out of the machine that produced the data that were used for allele scoring and interpretation). [xiv]

(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. [xv]

(e) In case of replicate samples, one genotype entry can be computed and stored in case the DNA profiles of the replicates match. In case of non-matching replicates, 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 references from the variant call file. Frequencies could also be used for heterogeneous varieties. [xv]

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

(g) Easy to share data, (e.g. API). [xv]

4.2 Requirements of the plant material [xvi]

The source and type of the material and how many samples ~~need~~ [ii] to be ~~analyzed~~ stored and shared in the database [ii] are the main issues with regard to the material to be analyzed.

4.2.1 Source of plant material

The plant material to be analyzed should be an authentic, representative sample of the variety and, ~~where~~ when [ii] 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. 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 ~~standardized and~~ [ii] 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) [ii]

It is essential that the samples taken for analysis are representative of the variety and well documented [ii]. With regard to being representative of the variety, consideration should be given to the features of propagation (see the General Introduction). ~~The size of the sample should be determined taking into account suitable statistical procedures.~~

4.2.4 DNA reference sample

~~It is recommended that~~ A DNA reference sample collection ~~should~~ may be created from the plant material sampled ~~according to sections 4.1 to 4.3. This has the benefit that the DNA reference samples can be stored and supplied to other laboratories.~~ [xvii] The method for sampling should follow recommended procedures and DNA extraction should fit some quality criteria. Both need to be documented. [ii]

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. [xvii]

~~*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. As 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 by more than one laboratory. 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.~~ [xviii]

~~*5.4 Scoring of molecular data*~~

~~A protocol for allele/band scoring should be developed in conjunction with the evaluation phase.~~

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 approached 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. [xix]

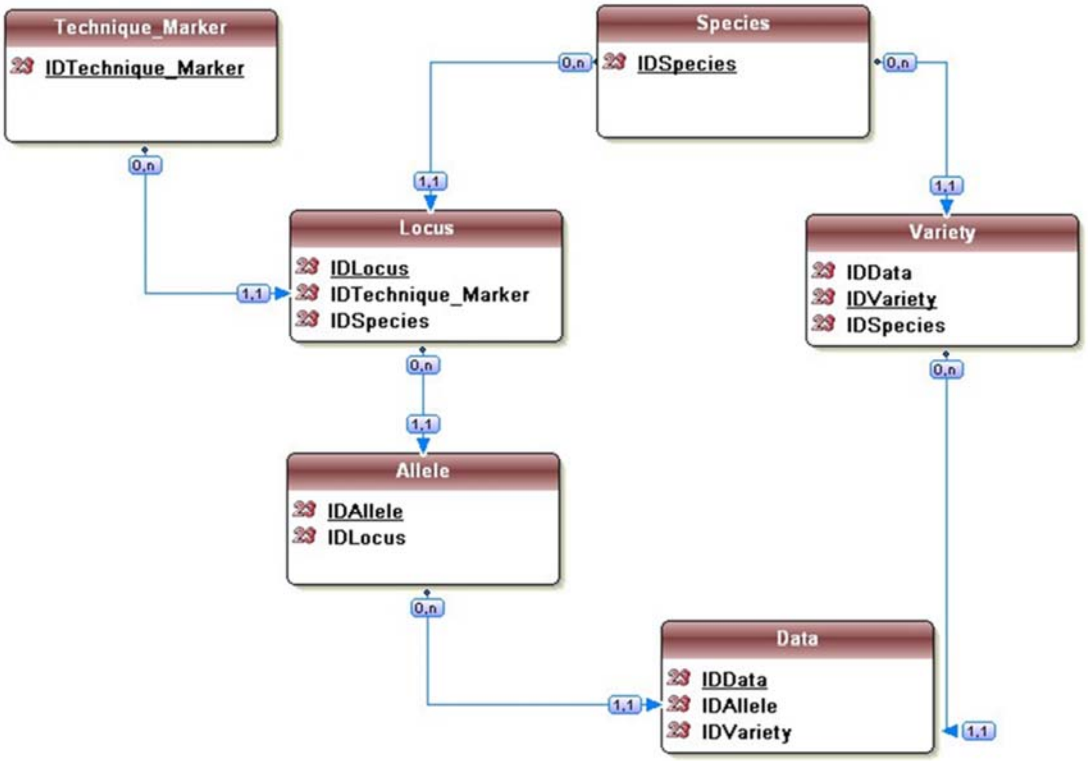
~~6. Databases~~

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. For molecular data obtained using next generation sequencing (NGS), the variant call file standard VCFv4.2 can be used.

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; ~~Technique~~ Marker detection method [ii]; 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. [xx]



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

(b) Reference genome position / 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 ~~indicates~~ [xxi] name or code of the locus for the species concerned can be used, e.g. gwm 149, A2, etc.

(c) ~~Allele code~~ 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 [xxii] indicates the name or code of the allele of a given locus for the species concerned, e.g. 1, 123, etc.

(d) Allele depths / 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, [xxiii] 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: [xxiv] the variety is the object for which the data have been obtained. ~~Grouping~~Type of variety: e.g. Inbred Line or Hybrid [xxiv]

(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.

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 should be addressed at the beginning of any work.

~~*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.~~

5. Data Exchange [xxv]

5.1 Data exchange scenarios

For cooperation purposes, the data model should allow different types of scenarios including [ii] 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.

~~Scenario 1 [xxiv]: exchange of data produced from a standardized set of markers for a specific crop~~ [ii]

~~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}~~ [xxiv]

~~For example, to obtain marker set information for maize using SSR and CE method, the following URL should be accessed:~~ [xxiv]

~~https://office.org/locus?upov_code=ZEAAA_MAY&type=SSR&method=CE~~ [xxiv]

~~The result would be:~~

~~{"techniqueid": "CN_SSR_ZEAA_MAY_CE_V_1",~~
~~["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": […]~~

~~]}[xxiv]~~

~~Scenario 2:[xxiv] search and view data of selected varieties generated from the same standardized set of markers[ii]~~

~~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}[xxiv]~~

~~For example,~~

~~https://office.org/variety?id=XU_30201800000140 &techniqueid= CN_SSR_ZEAA_MAY_CE_V_1[xxiv]~~

~~The result would be:~~

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

5.2 Data transfer methods

5.2.1 Fingerprint data transmission contains a variety of information, such as loci, samples, DNA, fingerprint data and fingerprint profiles. Commonly used data formats include: zip, csv, json, xml, and their respective characteristics are as follows:

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

(3) json and xml formats can contain more complex character data information and more redundant information, but the two formats' readability is very good.

5.2.2 The actual method of data transmission needs to be determined by the content of the transmission. A zip format is generally used to provide a format that contains transfer service of loci, samples, DNA, fingerprint data, and fingerprints spectrum. This method can be used to migrate data between systems; alternatively, csv, json or xml can be used to provide a transfer service that includes a basic fingerprint. The data transfer service also enables query and search functions. Therefore, it is recommended that the data transfer method be determined as needed to provide a better data transfer experience. Technical details on data transfer methods are described in the Annex: Data exchange scenarios and data transfer methods.

6. Summary [ii]

The following is a summary of the approach recommended for high quality DNA profiling of varieties including the selection and use of molecular markers ~~to construct central~~as well as the construction of shared and sustainable molecular 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.~~

~~Pig-tailing~~

~~In SSR analysis, “pig-tailing” is the addition of a short specific 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.~~

C. LIST OF ACRONYMS [xxvi]

BAM	Binary Alignment Map
BCF	Binary Call Format
CRAM	Compressed Reference-oriented Alignment Map
MNP	Multiple Nucleotide Polymorphism
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 [ii]

ENDNOTES

[i] See document BMT/17/25 “Report”, paragraph 16.

[ii] Joint proposals by the European Union, France and the Netherlands received on September 16, 2019

[iii] See document BMT/17/25 “Report”, paragraph 17.

[iv] See document BMT/17/25 “Report”, paragraph 20.

[v] See document BMT/17/25 “Report”, paragraph 21.

[vi] See document BMT/17/25 “Report”, paragraph 22.

[vii] See document BMT/17/25 “Report”, paragraph 23.

[viii] See document BMT/17/25 “Report”, paragraph 24.

[ix] See document BMT/17/25 “Report”, paragraph 26.

[x] See document BMT/17/25 “Report”, paragraph 36.

[xi] See document BMT/17/25 “Report”, paragraph 34.

[xii] See document BMT/17/25 “Report”, paragraph 35.

[xiii] See document BMT/17/25 “Report”, paragraph 37.

[xiv] See document BMT/17/25 “Report”, paragraph 35.

[xv] See document BMT/17/25 “Report”, paragraph 38.

[xvi] See document BMT/17/25 “Report”, paragraph 31.

[xvii] See document BMT/17/25 “Report”, paragraph 32.

[xviii] See document BMT/17/25 “Report”, paragraph 36.

[xix] See document BMT/17/25 “Report”, paragraph 39.

[xx] See document BMT/17/25 “Report”, paragraph 41.

[xxi] See document BMT/17/25 “Report”, paragraph 43.

[xxii] See document BMT/17/25 “Report”, paragraph 44.

[xxiii] See document BMT/17/25 “Report”, paragraph 45.

[xxiv] Proposals from the Office of the Union (see document BMT/17/10 “Review of document UPOV/INF/17 “Guidelines for DNA-Profiling: Molecular Marker Selection and Database Construction (‘BMT Guidelines’)”, paragraph 9).

[xxv] See document BMT/17/25 “Report”, paragraph 38.

[xxvi] See document BMT/17/25 “Report”, paragraph 49.

~~[End of document]~~ [Annex follows]

ANNEX

DATA EXCHANGE SCENARIOS AND DATA 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",
["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 ",
"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]