UPOV
International Union for the Protection of New Varieties of Plants

E

Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular

BMT/17/10

Seventeenth Session
Montevideo, Uruguay, September 10 to 13, 2018

Original: English
Date: August 29, 2018

REVIEW OF DOCUMENT UPOV/INF/17 "GUIDELINES FOR DNA-PROFILING: MOLECULAR MARKER SELECTION AND DATABASE CONSTRUCTION ('BMT GUIDELINES')"

Document prepared by the Office of the Union

Disclaimer: this document does not represent UPOV policies or guidance

EXECUTIVE SUMMARY

1. The purpose of this document is to present background information concerning the review of document UPOV/INF/17 "Guidelines for DNA-Profiling: Molecular Marker Selection and Database Construction ('BMT Guidelines')".

2. The BMT is invited to consider document UPOV/INF/17/2 Draft 1 as a basis for a revision of document UPOV/INF/17.

3. The following abbreviations are used in this document:

BMT: Working Group on Biochemical and Molecular Techniques, and DNA-Profiling in Particular
TC: Technical Committee

4. The structure of this document is as follows:

BACKGROUND

5. The BMT, at its sixteenth session, considered documents BMT/16/4 'Review of document UPOV/INF/17 "Guidelines for DNA-profiling: Molecular Marker Selection and Database Construction ('BMT Guidelines')" ' and BMT/16/5 "Standards for Databases containing Molecular Information" and received a presentation by the Office of the Union, on "Standards for databases containing molecular information", a copy of which is reproduced in document BMT/16/5 Add. (see document BMT/16/29 "Report", paragraphs 44 and 45).

6. The BMT agreed to invite members and observers to provide comments on document UPOV/INF/17 “Guidelines for DNA-profiling: Molecular Marker Selection and Database Construction (‘BMT Guidelines’)”. The comments would be compiled by the Office of the Union in a document that would form the basis of a review of document UPOV/INF/17 by the BMT at its seventeenth session. The BMT further agreed to propose to introduce a new chapter concerning cooperation in the exchange of data and construction of databases in document UPOV/INF/17 on the basis of document BMT/16/5.

7. On February 15, 2018, Circular E-18/004 was issued to designated persons of UPOV members in the Technical Committee and the BMT inviting members and observers of the BMT to provide comments on document UPOV/INF/17 “Guidelines for DNA-profiling: Molecular Marker Selection and Database Construction (‘BMT Guidelines’)” by June 15, 2018.

8. The Office of the Union received comments from Argentina, Ecuador and Spain, joint comments from the European Union, France and the Netherlands, and comments from the European Seed Association (ESA), which are reproduced in the Annexes to this document as follows:

- Annex I Comments from Argentina to UPOV circular E-18/004
- Annex II Comments from Ecuador to UPOV circular E-18/004
- Annex III Comments from Spain to UPOV circular E-18/004
- Annex IV Joint Comments from the European Union, France and the Netherlands to UPOV circular E-18/004
- Annex V Comments from the European Seed Association (ESA) to UPOV circular E-18/004

9. Document UPOV/INF/17/2 Draft 1 has been prepared on the basis of the comments received from Argentina, Ecuador, Spain, the joint comments received from the European Union, France and the Netherlands, and ESA in response to circular E-18/004, and a proposed new chapter concerning cooperation in the exchange of data and construction of databases in document UPOV/INF/17 as requested by the BMT at its sixteenth session.

10. The BMT is invited to consider document UPOV/INF/17/2 Draft 1 as a basis for a revision of document UPOV/INF/17.

[Annexes follow]

ANNEX I

COMMENTS FROM ARGENTINA TO UPOV CIRCULAR E-18/004

In regards to the email below, I send you some comments on the UPOV/INF/17/1 document.

1. page 3: first line, I would delete "for developing harmonized metodologies" and put "to standardize criteria for the use of DNA based markers".

2. page 3, point 1.2, I would add "real time based techniques and next generation sequencing".

3. page 3, point 1.5, I would delete "for the future".

4. page 4, point 2.2.1.2, I would add NGS as a technique for SSR scoring.

5. page 5, point 2.2.2, I would delete the sentence "However...yet routine".

6. page 8, point 6.2, I would change the word "technique" for "scoring system". If so, we should change also point 6.3.1 a). I think when saying scoring system we add a new quality to the data base. A scoring system would be RT PCR, sesquencing, capilar electophoresis, others.

Just one note about the Data base: nowadays most labs would use SNP markers, and the data come in an Excel file and then it is analyzed using biostatistical tools, wich are common among labs, I do not see the need of a table like 6.4.2.

[…]

[Annex II follows]

ANNEX II

COMMENTS FROM ECUADOR TO UPOV CIRCULAR E-18/004

Reference: UPOV/INF/17/1. Date: May 7 2017.

The purpose of the GUIDELINES FOR DNA-PROFILING: MOLECULAR MARKER SELECTION AND DATABASE CONSTRUCTION ("BMT GUIDLEINES") 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 plant varieties, possibly produced in different laboratories using different technologies."

Observations: What will the practical use of these guidelines be in terms of breeders' rights? If the guidelines are being considered for DUS tests, will the field phase, where a gene is expressed in a particular environment, be omitted, and if so, what will happen with genes which are hidden (DUS in the field)?

In addition, the aim is to set a high standard for the quality of the markers and fuel the desire to generate reproducible data using these markers in situations where equipment and/or chemical reagents might change. Specific precautions must be taken to guarantee the quality of database entries.

Selection of a Molecular Marking Methodology

(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;
(e) accessibility of methodology.

Observations: Will the database be shared, given that there is a large amount of information to be compiled? Ecuador is highly limited in terms of access to equipment and materials, and the cost of gathering basic information on each botanical variety and taxon is high.

Selection of Molecular Marking

The use of microsatellite markers.

Observations: It is undesirable that the sets of varieties must be defined and included in all analyses and using the same methodology. Also, equipment and the suppliers of materials must be the same to avoid obtaining varying results.

Access to Technology

Observations: 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. Markers and other methods and materials may be covered by intellectual property rights.

Material to be Analyzed

Observations: What criteria are used to guarantee an authentic, representative sample of the variety?
What is meant by "*where possible, [the plant material] should be obtained from the sample of the variety used for examination for the purposes of plant breeders' rights or for official registration*"?

What is the protocol for obtaining the sample and how is permission obtained to take the sample?

Will a bank of germplasm solely from plant material be created? What authority will be the custodian of that information or will the samples be kept by each relevant authority?

The document recommended in the general introduction that the size of the representative samples for determining the number of plants is for plants used in the open field and not for determining the number of plants for sequencing.

Observations: While it is understood that each relevant authority must establish a collection of germplasm, in the case of Ecuador, which does not have a collection of living material, it is more feasible to create living material than to have and maintain a collection of DNA reference samples.

The cost of the technology, equipment and materials is high and specialized technical equipment is required. If samples are sent to a general authority like UPOV, foreign germplasm banks used by developed countries will continue to grow; a situation that may affect the food sovereignty and security of countries.

A protocol for collection, maintenance, analysis, quality control and evaluation of samples should be standardized.

In summary, the aim of this proposal is to establish distinctions. It does not mention how homogeneity and stability will be verified in the event it is established as DUS guidelines. It does not take account of the costs entailed for countries for equipment, radioactive substances and software, nor does it detail the cooperation mechanism and under what legal guidelines these techniques may be used. It does not specify who will be the general administrator of the database and how the database will be shared with member countries. There is plenty of room for discussion aside from the techniques and methodologies to be used.

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. However, what will the practical use of these guidelines be in terms of breeders' rights? If the guidelines are being considered for DUS tests, will the field phase, where a gene is expressed in a particular environment, be omitted and if so, what will happen with genes which are hidden (DUS in the field)?

National Directorate of New Varieties of Plants

[Annex III follows]

ANNEX III

COMMENTS FROM SPAIN TO UPOV CIRCULAR E-18/004

Here we send our comments to the document UPOV/INF/17.

The document still applies in all the related to SSRs, but it should be broadened to other techniques, mainly SNPs but also those derived from NGS, as their use is now much more extended that at the moment of the writing of the document, in 2010.

Some of the principles can be extended to any other DNA based tecnology (point 3, acces to the technology; 4, material to be analyzed; 5, standardization of analytical protocols, excepting quality criteria, only useful for technologies base on PCR; or 6, databases).

We understand that to do this task it would be useful to delegate in a small working group, that could prepare a review of the new techniques that must be included.

[…]

[Annex IV follows]

ANNEX IV

JOINT COMMENTS FROM THE EUROPEAN UNION, FRANCE AND THE NETHERLANDS TO UPOV CIRCULAR E-18/004

DRAFT PROPOSAL FOR REVISION OF DOC UPOV/INF/17/1

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

prepared by Naktuinbouw (NL), GEVES (FR) and CPVO
(Proposed texts are highlighted.)

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 plant varieties, possibly produced in different laboratories using different technologies. In addition, the aim is to set high demands on the quality of the 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

Molecular markers sets and subsequently databases developmental process can be subdivided into 6 different phases:

1. Selection of molecular markers
2. Selection of detection method
3. Evaluation of the selected markers set and detection method (fit for purpose validation of the marker set and technological validation of the method)
4. Harmonization and validation of the method
5. Construction of the database.
6. Management of the database

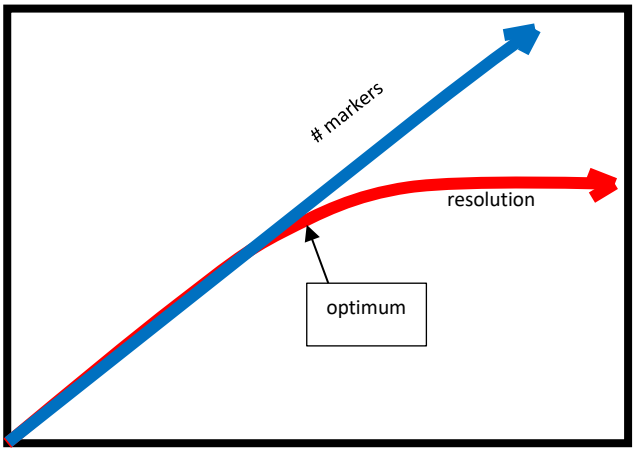
This document describes these different phases in more detail. It is considered that these phases are independent on the stage of development of genotyping technologies and future improvements in high-throughput sequencing.

1. Phase 1: SELECTION OF MOLECULAR MARKERS

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. A balance between the number of markers and the resolution or discriminative power should be reached (Cf. figure). The appropriate number of markers should be defined to reach the required discriminative power (optimum). This number is marker-type specific, species-dependent and can also be dependent on the purpose of the genetic analysis. The quality of the markers used is also to take into account: the more high quality markers are used, the more 'error-tolerant' the method will be (e.g. the impact of a single false allele score has a limited effect when the number of markers is high). However, a high number of markers is not a guarantee for a better analysis. Markers with a high error-rate are better left out, since they could hamper the quality of the analysis. Thus, the optimum number of markers should be determined also in respect to error-rate.



b) Repeatability within, and reproducibility between, laboratories in terms of scoring data

c) Good coverage of the genome. Knowing the position of the selected markers on the genome (i.e. map position) is not essential but allow avoiding the selection of markers that may be linked together.

d) Markers publically accessible, commercial, and/or newly developed depending on group/crop/species.

- Derive molecular markers from reliable public resources (e.g. peer reviewed publications and public databases NCBI, EMBL). This is the easiest and cheapest approach.
- Commercial molecular marker sets to screen and select a suitable marker set from there (e.g. species-specific chips and arrays). A special attention should be paid to this source of markers as information on position, sequences… for each marker will highly depend on the providers and some might be missing.
- Generate own sequence data. Sequence data will be generated from the selected varieties that fulfil the requirements mentioned above. However, in general to reduce time and cost, only a subset of varieties will be chosen to detect polymorphism and select the different markers. There are several options to obtain sequence information ranging from sequencing just a part of the genome to sequencing complete genomes of the selected varieties. The investigation needed in money and effort is dependent on the complexity of the genome of the particular species and the sequence data available in the public domain (e.g. reference genome). Many species contain a high level of genetic (botanic) diversity. There is no need to obtain sequence data of the complete genome; the data for only a small part of the genome may be sufficient to develop a suitable marker set for genotyping, depending on the application (e.g. genome-wide sequence capture, transcriptome sequencing (only the coding part of the genome). Presence or absence of a published reference genome is relevant to allow position determination of markers. It will be a matter of time before for a reference genome is available for all economically valuable plant species.

e) 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 markers are preferred over complex marker profiles that are sensitive to interpretation. Clear black and white answers also allows for easier harmonization.

g) Co-dominant markers are preferred over dominant markers as they have a higher discriminative power.

h) Avoidance of linkage disequilibrium

i) 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.

j) Markers located in coding and/or in non-coding regions and potentially epigenetic markers.

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

Flexibility and adaptability of a marker set

The discriminatory power of a marker set needs to be regularly assessed due to the evolution of the variety collections. Markers may need to be added or discarded depending on the modification of the genetics of varieties. In addition, New Breeding Techniques (NBT) and their resulting products may also require to use

specific markers to detect the edited sites in the genome (e.g. additional characteristics could be evaluated these markers provided that a direct correlation between the edited sites and the phenotype has been established).

Requirements on the molecular profiles

- Markers scattered all along the genome are used for the evaluation of distances/similarities between varieties through molecular distances and/or allelic frequencies. Application of this markers set is an assessment of the 'genetic background'.
- In addition, markers that correlate with defined morphological qualitative traits, fulfilling the UPOV model 1 criteria, can complement the genetic description.

2. Phase 2: SELECTION OF THE DETECTION METHOD

As a prerequisite, whatever the source of material, the method for sampling and DNA extraction should be standardized and documented.

Genotyping methods - general criteria

Important criteria for choosing a genotyping methods that generate high quality molecular data are:

- Mandatory criteria:

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) Interpretation of the data produced is independent of the equipment

- Optional criteria

a) Possibilities for databasing
b) Accessibility of methodology
c) Suitable for automation
d) Suitable for multiplexing
e) Applicable for both diploid species and polyploidy species
f) Cost effective; costs, number of samples and number of markers are in balance

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 technology and equipment used to produce them. 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.

Recommendations for the choice of the method:

- Methods that are simple to perform (limited steps in the protocol) are preferred over methods with a complex protocol that are time and labour consuming.
- Methods that allow easy, objective and indisputable scoring of marker profiles are preferred over methods that produce complex marker profiles that are sensitive for interpretation (e.g. wide range of intensities of the bands).
- Methods that are robust, not sensitive to subtle changes in the protocol or condition, but stable performance in time and conditions are preferred over methods that are sensitive to environmental conditions that are difficult to control.
- Methods that are flexible (vary in the number of samples or the number of markers) are preferred over methods that have a fixed set-up.
- Methods that are open source are preferred over methods that are completely or partly protected by IP rights or by confidential information.
- Methods that are independent of a specific machine or specific chemistry or specific supplier are preferred over methods that require a specific machine, chemistry or supplier that have a monopoly in the market. Methods without dependence on particular partners or products are preferred.
- Methods that detect molecular markers in a co-dominant way are preferred over methods that detect markers in a dominant way.
- Methods that allow multiplexing are preferred over methods that detect only one marker in one assay.
- Methods that are suitable for automation are preferred.

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 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.

Future perspective on technological development

Genotyping methods develop very fast and new technologies will keep being discovered. High-Throughput sequencing of short reads and now massive sequencing of long reads by nanopore sequencing enable the production of more and more data for a decreasing price per datapoint. As a consequence, the methods for marker set detection will alter in the future and shift from single sample endpoint methods towards whole genome sequences approaches. Irrespective of the technology used to detect the defined marker set, the genotype of a particular variety should not be affected. Both SSR markers and SNP/INDEL markers can be detected by High-Throughput Sequencing. In the (near) future, it could be cost effective to just sequence the whole genome of a plant. Even if all data produced will not be used (depending on the application), if the cost of a whole sequence become cheaper than single end point methods it may become the default method. However genotyping error of this technology need to be evaluated carefully before use.

Strategy	Reference genome	Present cost	Ease of use
Genome reduction - NGS	yes	€	+++
Genome reduction - NGS	no	€€	++
Whole genome - NGS	Yes	€€€	++
Whole genome - NGS	no	€€€€	+

3. Phase 3: EVALUATION OF THE SELECTED MARKER SET AND DETECTION METHOD (fit for purpose validation of the marker set and technological validation of the method)

General requirements for molecular marker set development

1. Selection of the varieties - defining the genetic width of the marker set
The selection of the varieties on which the molecular markers are developed is crucial. An appropriate number of varieties, based on the genetic variability within the species and type of variety concerned, should be selected. The selected varieties should be well characterized (morphologically) and true-to-type. The choice of varieties should reflect the maximum range of diversity within the group/crop/species/type - representative sampling of the particular group/crop/species/type must be guaranteed. In addition, some genetically very similar varieties or lines, some parents and offspring, genetically close but morphologically distinct varieties, some morphologically close varieties with different pedigree should be included, to enable to ‘measure’ the level of discriminative capacity of the markers and to determine the ‘suitability’ of the marker set.

2. Generation of molecular data of selected varieties – defining the genetic depth of the marker set
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.
There are several ways to collect the data on the genetic diversity within the particular group/crop/species/type for which a marker set is to be developed.

4. Phase 4: HARMONIZATION AND VALIDATION OF THE MARKER SET AND METHOD

Harmonisation and validation – general criteria

In order to select suitable markers and produce acceptable laboratory protocols for a given species, a harmonisation process 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 will involve the validation of genotyping methods and markers from which a defined set will be selected. This selection is based on performance: markers and methods should be robust and give rise to consistent results and DNA profiles in different laboratories using different equipment and chemistry. The consistence of the

markers and methods in different laboratories is evaluated in the harmonisation process. The final choice of a number to be validated 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 analysed, scored and recorded in different laboratories, potentially using different types of equipment and different sources of chemical reagents, etc.

Performance criteria

It is needed to determine whether the selected marker set is suitable (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: This can be determined by testing a defined collection of varieties – test set. For example variety pairs that are derived/excepted mutants, samples from the same variety but maintained in different places during time. Variety pairs with a known very close relation. Known pedigree. Diversity statistics such as Polymorphism Information Content (PIC)-values, expected heterozygosity (He), Effective Multiplex Ratio (EMR), Marker Index (MI) and/or Resolving power (Rp) can be calculated to illustrate the informativeness of a marker or marker set. The number of markers used should be an excess (exhausted number of markers). The minimal number of marker should be assessed and define so that analysis with a random selection of markers should not lead to different conclusions.
b) Reproducibility: Once chosen for a particular technology and DNA marker set this should repetitively reveal the same DNA profile for a variety. This is inevitable essential within one laboratory and between laboratories, especially when the DNA profiles are stored in databases.
c) Repeatability
d) Robustness
e) Error-rate: Every technology and every machine or platform has its imperfections and deficiencies. It is crucially important to be able to distinguish the technically induced variation from the real genetic diversity. This can be determined by the analysis of replica and/or duplicate samples (several DNA samples derived from the same plant material, so, identical DNA information, that proceed through laboratory process in parallel). Any deviation in the DNA profile must be a technical error.

Consistence criteria - harmonization of markers and methods in different laboratories

a) Well defined collections of varieties/samples to be applied as a reference that represent all alleles. 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 validation and harmonisation. 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. For example variety pairs that are derived/excepted mutants, samples from the same variety but maintained in different places during time. Variety pairs with a known very close relation. Known pedigree.
b) Duplicate samples (e.g. different sub-samples of seed from the same variety), analysed by more than one laboratory. Taking into account the data obtained from these samples may help reduce scoring errors and reinforce the reliability of the data stored in the databases. 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 artefacts.
c) Blind samples
d) Agreements on the scoring of molecular data. A protocol for allele/band scoring should be developed. 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 analysed, *i.e.* are not polymorphic in a particular variety collection).

5. Phase 5: CONSTRUCTION OF A SPECIES-SPECIFIC DATABASE

A database and the data that is stored in a shared database and how it is stored in a database reflects the process of producing the data. The database should store 1) the end results, e.g. the genotype as well as

how it was derived both in terms of 2) sequencing library preparation and 3) the computational steps for deriving a genotype.

Requirements of a database

- The database architecture should be flexible, e.g. allow for storing both flat files as well as compressed archives
- Contains different tables, separate tables and entries are required for library prep (the wet-lab work), data processing and the genotyping scores
- 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, xx files that came out of the machine that produced the data that were used for Allele scoring and interpretation)
- For sequencing data, variant call files in VCF or BCF format corresponding to the standard version 4.2 or higher should be used. 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 competent bioinformatician can repeat the analysis.
- In case of replicate samples, one consensus genotype entry can be computed and stored in case the genotypes 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.
- The database should validate the VCF and or BCF data against relevant specifications.
- The database should have a web front-end that enables easy uploading, downloading and interactive exploration of the data. The systems for storing, analysing and interpreting the data should be build and function separately yet function well in concert.
- Easy to share data, an API is recommended

Requirements of the plant material

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

1. Source of plant material

The plant material to be analysed 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 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.

2. Type of plant material

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

3. Sample size

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

A DNA reference sample collection may be created from the plant material sampled. The DNA samples should then be stored in such a way as to prevent degradation (e.g. storing it at -80C). The transfer of DNA reference samples to other laboratories will be submitted to the agreement of the owners of the varieties.

Data processing

The pipeline for processing the data should keep a detailed log of:

- Type and versions of tools,
- Command line used for the tool,
- Reproducibility counts,
- Open source tools are preferred,
- Sharing is encouraged,
- Raw alignment data (bam or CRAM files) should be stored where possible,
- Multi-sample VCF files are not suitable, one VCF file per cultivar must be present,
- If VCF files are stored, all positions (both variants & non-variants) and their depth should be stored,
- Both heuristic and probabilistic approached should be considered and compared for genotyping methods,
- Databases should facilitate input and output of genotype call data in standardized format (VCF or BCF),
- The data processing pipeline should result in a detailed log file which should be stored in conjunction to the variant call data,
- If possible, raw data should be stored so that data processing can be repeated with new or updated tools,
- A p-value or uncertainty for a given allele should be stored.

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 is recommended (https://samtools.github.io/hts-specs/VCFv4.2.pdf).

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. For variants obtained from sequencing data, storing VCF files in a relational or noSQL database is recommended. In this case, each database record for a variant have a defined genome version, chromosome, position, reference allele

Data Dictionary

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) 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 name or code of the locus for the species concerned can be used, *e.g. gwm 149, A2, etc.*

(c) Genotype:

For SNP genotypes, the allele composition of the SNP or MNP should be given, *e.g. A/T or A/A*

For other technique, 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 / 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: 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.

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. Phase 6: DATABASE MANAGEMENT

The effective management and updating of the database on the long term requires that appropriate agreements between partners are signed at the start of the creation process. These agreements should cover general principles (defining precisely the ownership of the materials and data, conditions of access and use, confidentiality, etc.) and technical principles (describing the types of data, identifiers, role of the partners, rules and planning of updating, etc.). The conditions under which the database could be open to additional partners wishing to contribute to its feeding after it was built needs also to be clearly established.

C. DEFINITIONS

Locus: a position on a chromosome/ a set of homologous chromosomes

Allele: a variant of a locus

Polymorphism: alleles at a particular locus that are different between individual organisms

Marker (genetic marker/DNA marker/ molecular marker): a single piece of DNA or a set of pieces of DNA that mark one or more specific alleles and can be detected through a single assay.

Genotype: the genetic constitution of an individual organism

Genotyping: the process of elucidating the genotype of an individual organism with a biological assay.

DNA profile/DNA fingerprint: a unique pattern of molecular markers, specific for an individual organism and representative for the genotype of this individual organism

Dominant marker: A marker that can mark only one of the possible alleles as either present or absent through a single assay

Co-Dominant marker: A marker that can mark different alleles through a single assay

Locus-specific marker: The position of the marker on the chromosome is exactly known. Prior knowledge on the adjacent DNA sequence is needed to develop the locus-specific assay

Random marker: The position of the marker on the chromosome is NOT known. Prior knowledge on the adjacent DNA sequence of the marker locus is NOT required

Heterozygosity: The state of an individual in which a locus carries at least two alleles

Homozygosity: A state of an individual in which a locus carries only one single allele in the number of copies equal to the ploidy level.

D: 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. 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.

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. 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.

[Annex V follows]

ANNEX V

COMMENTS FROM THE EUROPEAN SEED ASSOCIATION (ESA) TO UPOV CIRCULAR E-18/004

[…]

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, 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.

[…]

Comments by ESA

Several of the marker systems described in the document are already quite "old-fashioned". We are not really in favour of setting up marker databases on the basis of SSR's. The SSR technique has indeed some advantages, but it is an expensive technique with low throughput and it is highly dependent from the equipment that is used.

The use of SNP-markers has a preference, but it will be difficult to choose a SNP technique that all parties involved will apply. The choice for a specific SNP technology also strongly depends from the number of markers that needs to be analysed. For genotyping of bigger number of markers, the use of SNP-chips or sequence-based genotyping technology could be more appropriate, but no reference is made to these technologies at all.

Is this [“but their use in DNA profiling of plant varieties has not yet been explored”] still valid?

[…]

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;
(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; and
(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.

[…]

Comments by ESA

What does "useful" mean? Is there meanwhile any data available that can be used to frame the term in a more precise way?

[…]

2.2 Criteria for specific types of molecular markers

2.2.1 Microsatellite Markers

[…]

Comments by ESA

We are not really in favour of setting up marker databases on the basis of SSR's. The SSR technique has indeed some advantages, but it is an expensive technique with low throughput and it is highly dependent from the equipment that is used. Furthermore a technique like SSR – but also the others – has the weakness when used in polyploid species that they have to be able to distinguish between the genomes.

[…]

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 can be 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. 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.

[…]

Comments by ESA

Is this [“are not yet routine”] still valid?

The choice for a specific SNP technology also strongly depends from the number of markers that needs to be analysed. For genotyping of bigger number of markers, the use of SNP-chips or sequence-based genotyping technology could be more appropriate, but no reference is made to these technologies at all. The choice for a technology also depends on the purpose of the database (management of reference collection f.e.).

What about an extra paragraph about genotyping by sequencing? Since more and more reference genomes are available this technology might become more important and it might be useful to address the different sequencing methods and their suitability ....

The only system/technology that makes any sense is to use full DNA sequences, as anything less would result in being selective, i.e. only really screen/analyse the parts of the genome where the chosen markers are placed. And then how to decide on which parts of the genome is important?

GENERAL COMMENTS TO DOCUMENT UPOV/INF/17

- in the last BMT meeting it became clear that a lot of national PVP offices are in the process of setting up databases. For future collaboration it is essential that data compatibility is guaranteed and exchange is possible and that interfaces are created. One can also think of UPOV providing a common entry point for access....

- There might be also data that require access restrictions. Agreement on standards on how to set up these restrictions might be useful

[End of Annex V and of document]