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 Add.

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

Original: English
Date: September 10, 2018

ADDENDUM TO 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

The Annex to this document contains an extract of document UPOV/INF/17/2 Draft 1 "Guidelines for DNA-Profiling: Molecular Marker Selection and Database Construction ('BMT Guidelines')", with comments from the European Seed Association (ESA) presented in boxes highlighted in light orange and text to which those comments apply highlighted in yellow, to be presented at the seventeenth session of the Working Group on Biochemical and Molecular Techniques and DNA-Profiling in Particular (BMT).

[Annex follows]

ANNEX

COMMENTS FROM THE EUROPEAN SEED ASSOCIATION (ESA) TO DOCUMENT UPOV/INF/17/2 DRAFT 1

[…]

A. INTRODUCTION

The purpose of this document (BMT Guidelines) is to provide guidance ~~for developing harmonized methodologies~~ to standardize criteria for the use of DNA based markers 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.

[…]

Comments by ESA

It might be useful to specify the range of applications by referring to respective UPOV documents that elaborate on recommended or approved applications within the PVP system.

[…]

Joint comments from the European Union, France and the Netherlands [ii]

To add new sections 1.2 and 1.3 as follows:

1.2 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).

1.3 Requirements on the molecular profiles

1.3.1 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'.
1.3.2 In addition, markers that correlate with defined morphological qualitative traits, fulfilling the UPOV model 1 criteria, can complement the genetic description.

[…]

Comments by ESA

What exactly is meant by New Breeding Techniques? This term covers a lot of different techniques and NBT is not identical to genome editing and genome editing is not identical to targeted mutagenesis .... Why is there a need in the PVP system to have specific markers for detection of edited sites? There is scientific consensus that genetic identification of a genetic change as such is possible, but there is no way to distinguish e.g. a natural mutation from one that was introduced by genome editing. This means a genetic identification method will detect a genetic change, but not a method with which the change was introduced....

[…]

Joint comments from the European Union, France and the Netherlands [ii]

To add a new section 2.1 “Genotyping methods - general criteria” with the following subsections 2.1.1 and 2.1.2:

2.1.1 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.

2.1.2 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 is 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.

[…]

Comments by ESA

(a) and (b):

we would suggest to make this mandatory

(e):

Why?

[…]

Joint comments from the European Union, France and the Netherlands [ii]

To add a new section 2.2 "Recommendations for the choice of the method" with the following texts:

(a) 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.
(b) 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).
(c) 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.
(d) 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.
(e) Methods that are open source are preferred over methods that are completely or partly protected by IP rights or by confidential information.
(f) 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.
(g) Methods that detect molecular markers in a co-dominant way are preferred over methods that detect markers in a dominant way.
(h) Methods that allow multiplexing are preferred over methods that detect only one marker in one assay.
(i) Methods that are suitable for automation are preferred.

[…]

Comments by ESA

(f):
yes, we agree, but this should not affect standardization!

[…]

Joint comments from the European Union, France and the Netherlands [ii]

To add a new section 2.4 "Future perspective on technological development" with the following texts and table:

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	€€€€	+

[…]

Comments by ESA

what is meant by this?

[…]

Comments by Ecuador [iv]

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.

[…]

Comments by ESA

This is a very important issue! that ISTA protocols are meant to obtain authentic and representative samples in an objective of seed testing including identity and purity checks. On adventitious presence, ISTA sampling methods are also named in the EU Commission recommendation 2004/787 on technical guidelines and detection of GMOs.

[…]

Joint comments from the European Union, France and the Netherlands [ii]

To delete current section 5.3 and replace with a new section 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)” with the following subsections:

3.1 General requirements for molecular marker set development

3.1.1. 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.

3.1.2. 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.

[…]

Comments by ESA

Are primers always needed? What about NGS?.

[…]

6. Databases

Joint comments from the European Union, France and the Netherlands [ii]

To delete current section 6 and replace with a new section 5 “Phase 5: CONSTRUCTION OF A SPECIES-SPECIFIC DATABASE” with the following texts and subsection 5.1:

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.

5.1 Requirements of a database

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

(b) Contains different tables, separate tables and entries are required for library prep (the wet-lab work), data processing and the genotyping scores.

(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, xx files that came out of the machine that produced the data that were used for Allele scoring and interpretation).

(d) 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.

(e) 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.

(f) The database should validate the VCF and or BCF data against relevant specifications.

(g) 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.

(h) Easy to share data, an API is recommended.

Joint comments from the European Union, France and the Netherlands [ii]

To add a new section 5.3 “Data processing” with the following texts:

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

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

[…]

Comments by ESA

I do not really understand what is meant here....

[…]

6.1 Type of database

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

Joint comments from the European Union, France and the Netherlands [ii]

To renumber section 6.1. for a new section 5.4 and to add the following sentence to the end of the current texts:

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).

[…]

Comments by ESA

see previous comment on specific aspects.

[…]

Joint comments from the European Union, France and the Netherlands [ii]

To add a new section 6 “Phase 4: DATABASE MANAGEMENT” with the following texts:

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.

[…]

Comments by ESA

roles and associated data access/ confidentiality - meaning individualized access to data according to role

[End of Annex and of document]