E

BMT/8/13
**ORIGINAL:** English
**DATE:** August 13, 2003

**INTERNATIONAL UNION FOR THE PROTECTION OF NEW VARIETIES OF PLANTS**
GENEVA

# WORKING GROUP ON BIOCHEMICAL AND MOLECULAR TECHNIQUES AND DNA-PROFILING IN PARTICULAR

## Eighth Session
## Tsukuba, Japan, September 3 to 5, 2003

MARKER SELECTION AND DATABASE CONSTRUCTION FOR VARIETY CHARACTERIZATION

*Document prepared by experts from the Netherlands and the United Kingdom*

# MARKER SELECTION AND DATABASE CONSTRUCTION FOR VARIETY CHARACTERIZATION

Ben Vosman[1]*, James C. Reeves[2] & Robert J. Cooke[2]

[1] Plant Research International b.v., Wageningen, The Netherlands
[2] NIAB, Cambridge, UK
*author for correspondence: e-mail Ben.Vosman@wur.nl

## 1. Introduction

Within the context of variety registration and related activities, several possible applications of molecular markers can be envisaged. Some of these have been discussed during the BMT7 meeting, and as a result, Options 1 through 3 for the use of markers in the DUS testing process were produced. Discussions regarding these options are still ongoing. There are, however, other applications of markers, such as their use for quality assurance within DUS testing stations or other laboratories, or for tracing infringements of PBR or other authentication issues, i.e. for variety identification purposes, that are not subject to further discussion.

In all of these situations one would like to build databases containing the molecular profiles of existing varieties and to have the possibility of adding profiles of new varieties that will be produced in the future, possibly in different laboratories using different technologies. This sets high demands, not only on the quality of the markers, but also on the necessity of reproducing data using these markers in situations where equipment and/or reaction chemicals might change. In addition, specific precautions need to be taken regarding the quality of data entered into a database. In this paper we discuss these and related aspects of the application of molecular markers in variety characterization and the production of useful databases. The paper follows on from UPOV documents BMT6/12, BMT-TWV Tomato/1/1 and BMT7/16, which describe different aspects of the EU co-funded project '*Molecular Markers For Variety Testing*'.

## 2. Selection of a Molecular Marker System – General Considerations

The most important criteria for choosing a method are (i) reproducibility of data production, both between laboratories and over time, (ii) discrimination power, and (iii) the possibilities for robust databasing. In addition, although the cost aspects are an important consideration, at this stage we do not feel it is appropriate or helpful to consider cost issues, as the factors involved are different for each laboratory.

It is well known that technology development does not stand still, and as improvements in technology and new equipment become available, it is extremely important for the continued sustainability of databases that the data produced are independent of the equipment used to produce them. This is, for example, the case with DNA sequencing data. Twenty years ago radioactively labelled primers and sequencing gels were used to produce such data whereas nowadays this is done using fluorescent dyes followed by separation on high throughput, largely automated, capillary gel electrophoresis systems. Undoubtedly there will be other techniques used in future. 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 DNA microsatellite (simple sequence repeat, SSR) data, as

shown previously (Vosman *et al.* 2001), if attention is paid to certain specific factors (see below). This reproducibility in time and in space is extremely important in the construction, future operation and longevity of databases. Only in this way can a centrally maintained database, populated with verified data from a range of sources, be constructed in a cost-effective way such that the significant investment required in its establishment is only made once.

In our view, given these constraints, at the moment this effectively rules out any of the available multi-locus techniques and confines the discussion to well defined and researched SSRs (microsatellites) and for the future to sequencing information (i.e. single nucleotide polymorphisms, SNPs).

3. Selection of a Molecular Marker System – Specific Criteria

Some criteria for choosing a specific marker or set of markers are general and apply to all marker systems. These include for instance (i) distribution of the markers throughout the genome (i.e. map position) – not essential, but highly desirable, (ii) avoidance of markers with "null" alleles, (iii) a useful level of polymorphism, (iv) reproducibility of scoring between laboratories. Other criteria are much more marker system specific. For effective microsatellite analysis, selecting high quality markers is essential (this is discussed extensively by Bredemeijer *et al.* 1998, 2002 and Vosman *et al*. 2001). This includes consideration of the degree of "stuttering" and other artifacts (e.g. "echo" bands in some systems), the size of the amplification product, effective separation between the various alleles in all detection systems, reliable and repeatable scoring of the alleles/patterns and the level of polymorphism (number of alleles detected) between varieties (Vosman *et al.* 2001), which requires analysis of a substantial number of varieties. For scoring microsatellites in different laboratories and using different detection platforms, it is crucial that reference alleles (varieties) are defined and included in all analyses (Vosman *et al.* 2001; Bredemeijer *et al.* 2002). These reference alleles are necessary as molecular weight standards behave differently in the various detection systems and therefore are not appropriate for allele identification.

For the future, we envisage much greater use of SNP analysis for variety-related applications, including DUS testing. Single nucleotide polymorphisms can be detected via DNA sequencing, a routine technique which generally shows very high levels of repeatability and reproducibility in time and between laboratories. Detection of specific SNPs is currently carried out with a range of techniques, many of which are not yet routine. By their nature, SNPs have effectively two allelic states, which makes scoring them relatively straightforward and reliable and should reduce or remove many of the problems noted above. On the other hand, it means that a large number of markers therefore need to be analyzed, either singly or in multiplexes, to allow the unique identification of a particular genotype.

4. Material to be Analyzed

It is very clear from our previous work in this area that in addition to the technical issues regarding markers outlined above, it is extremely important that an agreed approach is taken to the biological material to be analyzed in the production of databases. There are two issues – the source of the material, and how many samples need to be analyzed.

It is essential that as far as possible, the plant material is obtained from a reliable source and ideally, especially for DUS purposes, we thus suggest that the material submitted by the applicant for registration purposes is used. We have evidence (Roder *et al.* 2002) that

samples bearing the same variety name can have very different genotypes. Seed lots are not always what they are supposed to be (Cooke *et al.* 2003) and commercially obtained samples may contain interfering levels of mechanical admixture. We would also like to propose the creation of a DNA reference sample collection, produced from the material sent in for DUS testing. This could be stored and supplied to others more easily than live material.

In principle, for vegetatively propagated material a single sample could be analyzed, as all individuals could be assumed to be identical and in any case mutations are usually not detected using the current molecular tools. Even so, it might be advisable to analyze at least duplicate accessions. For varieties of seed propagated inbreeding species (or hybrids produced from inbred lines) the situation is different. We have shown in previous research that in some cases, such varieties are not homozygous at all SSR loci (Bredemeijer *et al.* 2002; Röder *et al.* 2002; Cooke *et al.* 2003) and are thus "non-uniform". There are different causes for this non-uniformity – high levels are usually due to non-homogenous parental lines and/or mixing of seed lots, whereas lower levels tend to result from residual heterozygosity, cross-pollinations and accidental admixture. Such information on non-uniformity is very important for effective database construction and use. Although the above papers make some arguments for the use of bulked samples for variety evaluation and there are statistical approaches that can be useful, the best option at the moment is to analyze individual seeds. For seed propagated cross pollinating species, this is realistically the only option, as differences between varieties are the result of differences in allele (or band) frequencies, rather than the presence or absence of alleles/bands. However, in the future, methods based on SNP analysis will allow the estimation of allele frequencies in bulked seed samples.

5. Standardization of Analytical Protocols

In principle, one would like to standardize as few thing as possible. Methods used for genotyping and the construction of databases should be technically simple to perform and robust, allowing easy and indisputable scoring of marker profiles. However, some level of standardization is inevitable, such as in the selection of markers, reference alleles and allele calling (as discussed above). Regarding analytical protocols, it is not essential to standardize these, but rather to agree on certain quality criteria (such as for DNA quality, polymerase and other chemicals used). We suggest that in order to select suitable markers and laboratory protocols for a given species, there should be a preliminary evaluation phase, in which more than one laboratory is involved. Any marker which causes difficulties in any laboratory should be rejected for subsequent use. This phase could also identify an initial strategy for allele scoring for each marker and the selection of standard alleles. Ideally, 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, analyzed in different laboratories. This approach is also very effective in spotting sampling errors, or those due to heterogeneities within the samples (Vosman *et al.* 2001), and eliminates possible laboratory artifacts (systematic errors). 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.

A useful exercise to carry out when the first phase of the database is complete is 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.

6. Suggested Approach to SSR Profiling - Summary

The following is a summary of our suggested approach to the selection and use of SSRs to construct central databases of molecular profiles of varieties that are sustainable (i.e. can be populated in the future with data from a range of sources, independent of the technology used). The principles of the approach would also be suitable for use with markers such as SNPs.

6.1. Agree on quality issues (DNA, chemicals, enzymes, primers, etc.)
6.2. Verify source of the plant material used – for preference, use the DUS submissions
6.3. Agree which SSRs are to be in a preliminary collaborative phase (involving more than one laboratory and detection platform)– reject any that give problems to anyone
6.4. Conduct a preliminary collaborative analysis of a small number of varieties (c. 20) – all analyzed in different laboratories/different platforms. This will provide a strategy for allele scoring and also suggest suitable reference alleles/varieties.
6.5. Analyze a larger no of varieties in different laboratories/different platforms – use duplicate samples, and exchange samples/DNA extracts if problems occur. (This exercise will also provide information on sample heterogeneity).
6.6. Use reference varieties/alleles in all analyses
6.7. Verify all stages (including data entry) – automate as much as possible
6.8. Conduct a 'blind test' in different laboratories using the database
6.9. Adopt the same procedures for adding new data

7. Intellectual Property Issues

It would be logical, especially for DUS testing for Plant Breeders' Rights, to use methods and materials that are all publicly available. However, for obtaining high quality markers a large investment is necessary and consequently these markers are often kept confidential by the developers. It may thus be necessary to purchase or otherwise acquire access to markers that are not in the public domain. Access to markers requires detailed negotiation with their owners; depending on their interests, this may, for example, involve preferential access to data generated from use of the markers.

Similar considerations apply to ownership of the database and access to it. The number and interests of contributors to the database will influence such considerations. These contributors might include plant breeders donating the germplasm (and the legal context under which this is done may have an influence), the members of any consortium developing the database, the funders of such work and owners of any IP used in its construction. Again these interests can be reconciled through negotiation in what has become the usual manner.

A basic factor is the need, for the time being, for the purchase of a PCR licence in order to use the fundamental technology around which the techniques above are built. The PCR licence is a defined cost.

Thus it is clear that IP-related issues can add complexity, but are of themselves no insurmountable barrier to the adoption of DNA marker technologies for variety characterization. In fact, it has become usual to address IP issues at the start of many research contracts and contemporary researchers often have to deal with such considerations as a matter of routine. These kinds of negotiations are now commonplace and are an accepted part of the exploitation of novel technologies.

8. Literature:

Bredemeijer, G.M.M., P. Arens, D. Wouters, D. Vissser & B. Vosman (1998) The use of semi-automated fluorescent microsatellite analysis for tomato cultivar identification. Theor. Appl. Genet. 97:584-590.

Bredemeijer, G.M.M., Cooke, R.J., Ganal, M.W., Peeters, R., Isaac, P., Noordijk, Y., Rendell, S., Jackson, J., Röder, M.S., Wendehake, K., Dijcks, M., Amelaine, M., Wickaert, R., Bertrand, L. and Vosman, B. (2002) Construction and testing of a microsatellite database containing more than 500 tomato varieties. Theor. Appl. Genet. 105:1019-1026.

Cooke, R.J., G.M.M. Bredemeijer, M.W. Ganal, R. Peeters, P. Isaac, S. Rendell, J. Jackson, M.S. Röder, V. Korzun, K. Wendehake, T. Areshchenkova, M. Dijcks, D. Laborie, L. Bertrand & B. Vosman (2003) Assessment of The Uniformity of Wheat and Tomato Varieties at DNA Microsatellite Loci. Euphytica (in press).

Röder, M.S., Wendehake, K., Korzun, V., Bredemeijer, G., Laborie, D., Bertrand, L., Isaac, P., Rendell, S., Jackson, J., Cooke, R.J., Vosman, B. & Ganal, M.W. (2002) Construction and analysis of a microsatellite-based database of European wheat varieties. Theor. Appl. Genet. 106:67-73.

Vosman, B., Cooke, R., Ganal, M., Peeters, R., Isaac, P., Röder, M., Jackson, J., Rendell, S., Dijcks, M., Kleyn, Y., Visser, D., Wendehake, K., Areshchenkova, T., Korzun, V., Amelaine, M., Wickaert, V. and Bredemeijer, G., 2000. Standardisation of molecular marker systems for variety testing. Document for UPOV Working Group on Biochemical and Molecular Techniques and DNA profiling in particular (BMT/6/12).

Vosman, B., Cooke, R., Ganal, M., Peeters, R., Isaac, P. and Bredemeijer, G., 2001. Standardization and application of microsatellite markers for variety identification in tomato and wheat. Acta Hort. 546: 307-316.

Vosman, B. and Bredemeijer, G., 2001. Construction of the STMS database for tomato. Document for UPOV Working Group on Biochemical and Molecular Techniques and DNA profiling in particular (BMT-TWV Tomato/1/1).

Vosman, B and G. Bredemeijer (2001) Standardisation of STMS databases for tomato and wheat. Document for UPOV Working Group on Biochemical and Molecular Techniques and DNA-profiling in particular (BMT/7/16).

[End of document]