Marker Development in Ornamental Plants

A.W. van Heusden and P. Arens[a]
Plant Breeding, Wageningen UR
P.O. Box 16, 6700 AA Wageningen
The Netherlands

Keywords: SSR markers, variety identification, ploidy level, next generation sequencing

Abstract

Development of markers for a new crop or development of additional markers for a crop where markers have been developed in the past raises the question of the intended use of the markers. Depending on the different objectives in mind one marker type may be better suited then another. In general one can think of two main objectives for the use of markers; variety identification and breeding applications. In view of recent developments in molecular genetics, and sequencing technologies in particular, within the 23rd International Eucarpia Symposium Section Ornamentals a workshop was devoted on molecular markers and their use in ornamentals. Within this paper an overview will be presented on the development of markers for identification of ornamental crops and on the importance of the new developments in marker and sequence technology for the use of markers in ornamental breeding.

DEVELOPMENT OF MARKERS FOR VARIETY IDENTIFICATION

Molecular markers can be used for variety identification with different objectives in mind such as chain management, detection of infringements, and the use in Plant Breeders Rights (PBR) applications such as reference variety selection for Distinction Uniformity and Stability (DUS) testing in crops for which many varieties are known or within the Essentially Derived Variety (EDV) framework. For variety identification in the chain two situations can be envisaged, direct or indirect comparison. In the direct comparison a sample's identity is checked to a known variety and samples are analysed simultaneously which consequently needs a large reference collection maintenance. Basically any marker system (e.g. AFLP, NBS-profiling, SNPs or SSRs) is suitable for such a research question as long as it's reproducible. In an indirect comparison, reference material is not available and/or an unknown sample has to be identified based on identity to a known variety in a data base. In such a case, data base information has to be available by genotyping many varieties. This puts high demands on the reproducibility, reliability and scorability of markers in time. Best suitable markers are then SSR and SNP markers. Because the maintenance of a large reference collection is too expensive, indirect comparison and therefore database building is preferred for most applications in PBR.

With respect to the choice between SNP and SSR markers to use as preferred marker for data base building, there are a number of considerations to take into account. SNP markers are relatively easy to develop (e.g. using new generation sequencing). Furthermore, SNP markers are very amendable for high throughput screening but with the low samples sizes usually involved in variety identification platform choice is limited and costs may be relatively high. The main disadvantage of SNP is that they are bi-allelic markers therefore their information content is low. SSR markers have the advantage that they are multi-allelic and thus have a high information content. Development of SSR markers is quite expensive although in his presentation at the Eucarpia symposium Thomas Debener (not published) showed that with 454 sequencing of mRNA pools of differently challenged leaves already 900 SSR motifs could be detected and the power of next generation sequencing also here can reduce time and effort. What remains is that for variety identification high restrictions are put on the scorability of markers (in contrast to

[a] paul.arens@wur.nl

breeding purposes) and many SSR markers need to be tested for this. Multiplexing of SSR markers is possible but limited by the availability of different dyes (4 colours for markers on commonly used analysis platform) and by the allele size range of markers when combining markers with the same dye labelling.

In ornamental plants higher ploidy levels and aneuploïdy are very common and this is a major drawback because measurement of dosage in random sets of varieties is often not possible (Esselink et al., 2003, 2004). One approach is to take the presence or absence of each allele as a dominant marker (Esselink et al., 2003). In view of this, for most ornamental crops SSR markers are preferable over SNP markers because of their high information content (multi-allelic) as markers for database building. Therefore, the focus here will be on the development and use of SSR markers in variety identification. Starting with variety identification the first hurdle that needs to be taken is obtained sufficient markers of good quality. SSR markers (and SNP) can be found for some species in literature (e.g. rose; Hibrand-Saint Oyant et al., 2008; Zhang et al., 2006) but for most species there are no markers or insufficient markers available in publications. Another potential source is public DNA databases (e.g. EMBL) that contain sequences in which SSRs (or SNPs) can be found especially for species where large EST collections have been donated. Despite these possibilities for most ornamental species however markers still need to be developed. Most SSR markers are retrieved from genomic DNA using one of the available microsatellite retrieval methods (Zane et al., 2002) although sequencing of EST libraries is also possible as mentioned above. With the advent of the next generation sequencing techniques this approach also provides sufficient markers and the large scale sequencing has the advantage that sequences of different genotypes can be made and compared so that polymorphisms can be detected in this stage already. For breeding purposes often SSR markers with a di-nucleotide motif are retrieved because these motifs are more abundant and in general more polymorphic but the disadvantage is that for variety identification scoring is more often problematic due to so-called stutter bands due to polymerase slippage during PCR. Tri-nucleotide motifs, although less abundant, in general show much less stuttering and alleles are more separated (3 bp difference) and are easier to score. Therefore, for identification SSR markers based on a tri-nucleotide motif are preferred. After retrieval of SSR markers these have to be tested on a small but carefully ensembled set of samples for appropriateness for variety identification. Only the best markers i.e. markers without stutters, artefacts or duplicated loci and that are polymorphic can be used for database building. As a rule of thumb 1 in 10 SSR markers is suitable for identification purposes. The most informative markers are those with equal allele frequencies that distinguish samples in a large number of equally sized groups. Once a set of markers has been identified that meet the quality standards filling of the database can commence. For species where different ploidy level can be expected (e.g. carnation where varieties can be diploid, triploid, tetraploid, pentaploid and even hexaploid) flow cytometry analysis can help in interpretation but in such material also aneuploïdy can be expected which makes assessment of allele dosage complicated and the number of expected alleles uncertain. Therefore, in polyploids, scoring is done dominantly and only presence and absence of alleles can be scored. Furthermore, checking of samples with an unexpected number of alleles should be done by repeating PCR and analysis. Similarly, all failed samples should be repeated and a considerable part (and preferably all) of the samples should be analyzed twice so that mistakes in the analysis (mostly differences in scoring small peaks) can be assessed and an experimental error estimate (threshold) can be calculated (Fig. 1). This threshold avoids two samples being considered to be different on the basis of technical errors. Using the band presence/absence data, similarity between pair wise combinations of samples can be calculated and a tree can be constructed to visualize the resulting relationships between samples. After analysis and identification of those samples that cannot be distinguished from each other using the threshold value for experimental error the consequences of using this threshold can be evaluated by adding cultivar names to the samples and using knowledge on relationships (e.g. variety sports). For a good example of the potential of

variety identification using molecular markers in ornamentals see Smulders et al. (2009).

DEVELOPMENTS IN MARKER TECHNOLOGIES

Sequence technology and high-throughput genotyping are developing in a tremendous speed, this development will probably replace SSR markers as the markers of choice in ornamentals. In the past three years, the emergence of massively parallel sequencing technologies has dramatically reduced time and costs for sequencing (Fig. 1). All these developments will continue and sequencing will become cheaper and cheaper. For ornamentals where no complete sequences are known (yet!) the 454 Life Sciences option is the best option, one million reads of 400 base pairs can give a wealth of SNP markers. The SNP markers can be used to follow the alleles in which they reside. To obtain the SNPs, sequences of two or more cultivars must be compared, in the same analysis different SNPs/alleles within and between cultivars can be found. Since the reliability of 454 sequencing is slightly lower it is necessary to have a redundancy of the sequences to be analyzed (Fig. 2). There will be no redundancy of sequences if the complete genome is targeted (especially with large genome ornamentals like lily) therefore it is needed to do a complexity reduction of the genome before sequencing. There are several ways to achieve that such as sequencing cDNA or using selective bases (CRoPs-technology Keygene, http://www.keygene.com/keygene/techs-apps/technologies_crops.php). The advantage of cDNA is that directly expressed genes are targeted. To avoid abundant sequences in the cDNA a normalization procedure is necessary. To analyze the data, software programs have been developed (e.g. QualitySNP; Tang et al., 2006) that will list all true SNPs. In heterozygous ornamentals the frequency of SNPs will be high. With a high level of SNPs it is possible to identify haplotypes by analyzing more SNPs in a single sequence or SNPs in completely linked cDNA sequences. Difference in haplotypes can make a SNP assay multi-allelic in stead of bi-allelic. In tetraploid roses this will allow occasionally a better distinction of the four different alleles. All these applications will need good user friendly software to analyze the available sequence information.

After SNP discovery several methods are available for genotyping, two of the methods are the Illumina GoldenGate assay and the Illumina Infinium array (http://www.illumina.com/technology/goldengate_genotyping_assay.ilmn; http://www.illumina.com/applications.ilmn). The Illumina GoldenGate assay is capable of multiplexing from 96 to 1,536 SNPs in a single reaction over a 3-day period. It needs preferably two stretches of 50 base pairs flanking the SNP position and it can only handle base pair substitutions and not indels. It has been demonstrated that the Illumina GoldenGate assay could be used for SNP genotyping of homozygous tetraploid and hexaploid wheat lines (Akhunov et al., 2009). Quantitative scoring of SNP markers is still not reported but is expected to be possible for many of the SNPs. The Illumina Infinium array can handle a virtually unlimited number of SNPs in one run and in 80% of the SNPs only one flanking stretch of 50 base pairs has to be free of sequence differences. A disadvantage of the Infinium array is that a minimum of 1000 arrays with over 3000 SNPs has to be ordered which makes it an expensive investment (~150.000 Euro). However, a consortium might raise enough money for SNP discovery and detection and individual arrays by the partners can be used for a reasonable price (~150 Euro). A potential problem by testing unknown germplasm is that it is not known what percentage of the SNPs will not work due to polymorphisms in the flanking regions of the SNP positions in the unknown germplasm. What genotype platform to use depends on the number of the SNPs necessary for either the variety identification or for the genetic studies. After the identification of closely linked markers in genetic studies SNPs can directly be used as an extra tool in breeding and selection. For single marker experiments in marker assisted selection a number of relatively cheap techniques are available (e.g. Invader or High Melting Curve Analysis and gel based systems).

CONCLUSION

The development of next generation sequencing and genotyping will also have a large impact on ornamental research, for this it will be necessary to develop good interpretation software and accessible databases especially adapted for use in ornamentals. Good co-operations and communication between bioinformaticists, researchers and breeders will be of utmost importance. Consortia between different stakeholders and the government might be necessary to cover the initial costs of sequencing and genotyping.

Literature Cited

Esselink, D., Smulders, M.J.M. and Vosman, B. 2003. Identification of cut-rose (*Rosa hybrida*) and rootstock varieties using robust Sequence Tagged Microsatellite markers. Theor. Appl. Genet. 106:277-286.

Esselink, G.D., Nybom, H. and Vosman, B. 2004. Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting—peak ratios) method. Theor. Appl. Genet. 109:402-408.

Akhunov, E., Nicolet, C. and Dvorak, J. 2009. Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theoretical and Applied Genetics 119(3):507-517.

Hibrand-Saint Oyant, L., Crespel, L., Rajapakse, S., Zhang, L. and Foucher, F. 2008. Genetic linkage maps of rose constructed with new microsatellite markers and locating QTL controlling flowering traits. Tree Genet Genomes 4:11-23.

Smulders, M.J.M., Esselink, D., Voorrips, R.E. and Vosman, B. 2009. Analysis of a database of DNA profiles of 734 Hybrid Tea Rose varieties. Acta Hort. 836:169-174.

Tang, J.F., Vosman, B., Voorrips, R.E., Van der Linden, C.G. and Leunissen, J.A.M. 2006. QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics 7:438.

Zane, L., Bargelloni, L. and Patarnello, T. 2002. Strategies for microsatellite isolation: a review. Molecular Ecology 11:1-16.

Zhang, L.H., Byrne, D.H., Ballard, R.E. and Rajapakse, S. 2006. Microsatellite marker development in rose and its application in tetraploid mapping. J. Am. Soc. Hort. Sci. 131(3):380-387.

Tables

Table 1. Overview of three of the most used next sequencing methods. SOLiD sequencing gives very reliable sequences but with the lowest read length. (http://www.454.com/products-solutions/system-features.asp; http://www.illumina.com/technology/sequencing_technology.ilmn; http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/overviewofsolidsystem/index.htm).

System	Reads	Average length of reads	Total base pairs
454 Life Sciences GS FLX Titanium	$> 10^6$	400	0.4×10^9
Illumina Solexa genotyping	$> 400\times10^6$	75	30×10^9
SOLiD sequencing	$> 800\times10^6$	50	40×10^9

Figures

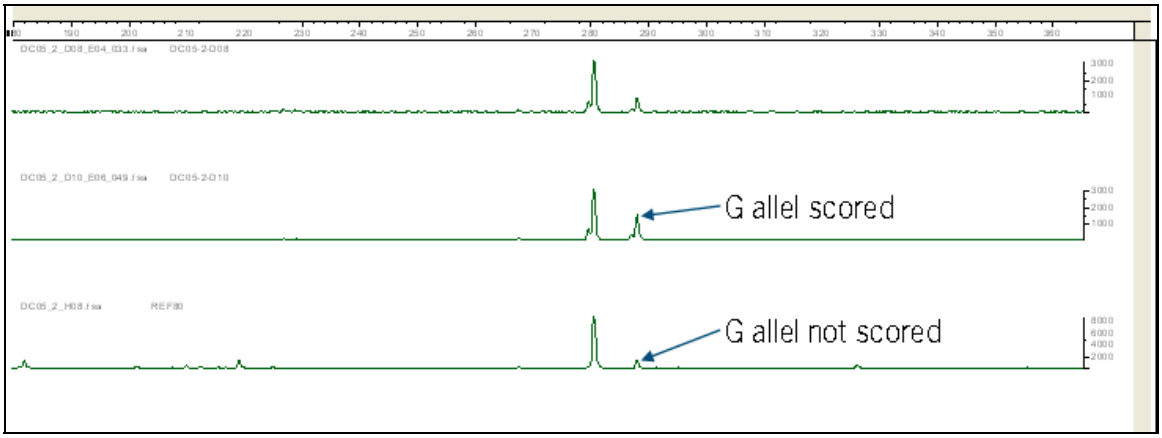

Fig. 1. Example of differences in scoring due to experimental errors (lower line peak not accessed because it is below the 400 threshold in the analyzer).

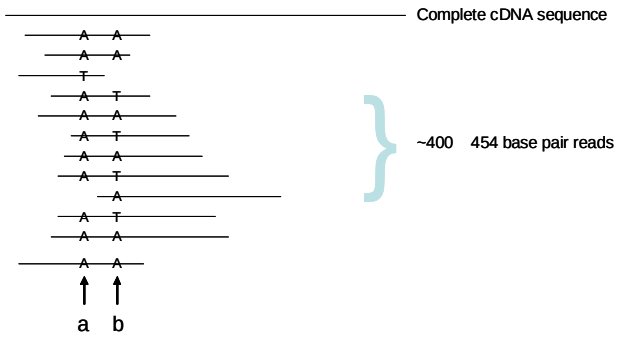

Fig. 2. Example of SNP detection. The difference in a is considered a sequence error. The difference in b as a true SNP (in more than one sequence independently found).