High Throughput Marker Development and Application in Horticultural Crops

M.J.M. Smulders[a], M. Vukosavljev, A. Shahin, W.E. van de Weg and P. Arens
Wageningen UR Plant Breeding
PO Box 16, NL-6700 AA Wageningen
The Netherlands

Keywords: SSR, SNP, ploidy level, next generation sequencing, complexity reduction

Abstract

In this paper we present an overview of current developments in sequencing that offer the possibility to generate large numbers of markers in ornamental crops. The prospects of this new sequence technology for the application of markers in breeding of outcrossing and/or polyploid crops are discussed using examples in rose and lily.

MOLECULAR MARKERS IN BREEDING

Marker development in ornamentals has been lagging behind compared to marker development in large agricultural and horticultural crops. This is partly due to the fact that there are many different ornamental crops. As markers need to be developed for each of these crops separately, the costs related to development and the time needed to develop them were obstacles. In addition, morphological characteristics are important during ornamental breeding of new cultivars, and these can be assessed without the use of markers. This is slowly changing, as it becomes more important to develop cultivars that have growth characteristics suitable for specific environments, and breeding also has to focus on traits that are difficult to assess and/or are controlled by multiple loci (quantitative traits), including stem production, time to flowering, flower size, and disease resistances. For instance, *Fusarium* resistance in lily is controlled by six putative QTLs (Shahin et al., 2010). Markers for each of these QTLs would make it possible to select progeny plants that have inherited the combination of these six loci. As this can already be done in a seedling stage, this would speed up the breeding process, especially in bulbous ornamentals, which have a long juvenile phase (3 years in lily, 5 years in tulip). In the case of introgression of disease resistances from wild relatives, markers can also be used to assist selection against wild germplasm unlinked to the desired QTLs.

PROGRESS IN SEQUENCING

In the past five years, the emergence of massively parallel sequencing technologies has dramatically reduced time and costs for sequencing. These developments will continue and sequencing will become cheaper while fragment lengths will increase. Importantly, parallel sequencing can be done for many targets, or even on the complete genomic DNA, without prior knowledge of the DNA. Hence it now becomes feasible to generate large amounts of DNA sequence information for species for which little prior sequence information exists, and to mine these sequences for polymorphisms that form the basis of the development of molecular markers. These molecular markers can be applied for marker-assisted breeding (MAB) and identification of cultivars and hybridization events. It offers great opportunities for ornamental crops, for which few molecular markers are currently available.

SEQUENCING FOR MARKER DEVELOPMENT

SNP Marker Discovery

Sequence differences of a single nucleotide are called ‘single nucleotide

[a] rene.smulders@wur.nl

polymorphism' or SNP. To obtain SNPs, sequences of two strands of DNA must be compared. In inbreeding crops this means comparing the sequences from at least two different plants. Most ornamentals are outbreeding crops. That means that even within the DNA of one heterozygous diploid plant there will be many sites at which the sequence of one chromosome is different from that of the homologous chromosome. Therefore, an analysis of sequences obtained from even a single plant can be mined to find SNPs between the homologous chromosomes. An analysis of sequences from two plants (e.g., the two parents of a cross) will yield both SNPs within each of the parents and SNPs between the parents.

Many ornamental genomes are too large to completely sequence to sufficient depth with the current techniques. Besides, it is not necessary to generate so much sequence information, as 200 to 1000 SNPs will suffice to generate a genetic map of the two parents, and they are also sufficient to be able to find markers that will be associated with traits in a segregating population derived from a cross between two plants (for an association study more markers mat be needed). So we only want to sequence a small part of the genome of a species of which we do not have much information, but, importantly, we do want to sequence the same part from various plants from this species, as we want to compare the sequences we obtain and screen them for differences. Several methods have been developed for this so-called 'complexity reduction'. Probably the easiest method is to sequence not genomic DNA but cDNA made from the mRNAs that are being expressed in one or more tissues (this is commonly referred to as RNA-seq or Whole Transcriptome Shotgun Sequencing). As genes only represent a small fraction of the genome, and not all genes are being expressed in a given tissue and developmental state, this gives a strong reduction of the amount of sequences that need to be generated. Additionally, any polymorphism found will be in or close to genes. Such polymorphisms are often also useful in related species. As genes are not the most polymorphic part of the genome, one will find fewer polymorphisms than in non-coding DNA, but that also means that SNP marker development will be relatively easy as there are fewer flanking SNPs that could interfere with the development of a SNP assay. A problem is that some mRNAs are present in many copies, so either we accept that part of the sequence capacity is consumed to sequence many identical copies of a few common genes (this often includes chloroplast genes), or we take the effort to 'normalise' the contribution of all mRNAs during sample preparation, which is usually done through hybridisation, requires some effort, and does not work perfectly.

Alternative methods exist for complexity reduction of genomic DNA. They include hybridisation to a set of COS (conserved orthologous set) genes (Wu et al., 2006; Li et al., 2008), or variants thereof such as SureSelect, Motif-directed profiling (Van der Linden et al., 2006) to amplify members of a gene family, and the use of restriction enzymes as in multiplexed shotgun genotyping (MSG, Andolfatto et al., 2011) and Restriction-site Associated DNA (RAD) tag sequencing (Emerson et al., 2010). Using the latter approach, Barchi et al. (2011) recently developed SNP markers in eggplant.

To analyze the data, software programs have been developed (e.g., QualitySNP; Tang et al., 2006, 2008) that will attempt to distinguish 'true' from 'false' SNPs. False SNPs can be due to sequencing errors, but they can also be generated when two paralogous genes from the same genome are compared. This is a problem in diploid as well as polyploid species.

SSR Marker Development

The sequences that are generated can also be mined to discover SSR ('simple sequence repeat'; syn. STR or microsatellite) repeats that can be used for developing SSR markers. Analysis of expressed sequences in several species has shown that a few percent of them contain SSR regions that can be used for developing SSR markers. Although genic SSRs tend to be less polymorphic than SSR repeats in non-coding DNA, they are more easily transferable to related species, as the primers often reside in coding regions. There are several recent papers that report generating SSR markers from next generation

sequences, including Tangphatsornruang et al. (2009) in mungbean, Parida et al. (2010) in sugarcane, and Blanca et al. (2011) in *Cucurbita pepo*. Recently, Li et al. (2011) analysed the position of exonic and intronic SSR markers in poplar, and found that some chromosomes have regions in the genome in which there are no exonic SSRs. Park et al. (2010) mined ESTs in the Genome Database for *Rosaceae* (GDR) for SSR markers. Shahin et al. (in preparation) mined lily and tulip sequences for SSRs.

CHALLENGES FOR MARKER-ASSISTED BREEDING IN ORNAMENTALS

Large Genomes

Many bulbous ornamental species have huge genomes. For instance, lily has a genome size of 36 Gb, which is more than 280 times larger than the genome of *Arabidopsis*. This is not due to an increase in the number of genes but in an increase in repetitive DNA. Transposable elements, especially long terminal repeat (LTR) retrotransposons, comprise the vast majority of this repetitive DNA, and they have a significant impediment on physical mapping, genome sequencing, and map based gene isolation (Bennetzen et al., 1994). Horning et al. (2003) developed SSR markers from lily genomic DNA but could only produce six polymorphic SSR markers, which is an insufficient number for mapping purposes. It is expected that the genome size will not hamper marker development based on next generation sequencing of cDNAs.

Outcrossing

Recently, 454 pyrosequencing was used to sequence the cDNA of *Lilium* and tulip as a first trial in ornamental plants to generate SNP and SSR markers using high throughput technologies (Shahin et al., in prep.), and Illumina sequencing was used for the transcriptome of rose (Boucoiran, Gitonga et al., in prep.). As is common in lily, the cultivars sequenced are hybrids made by crossing parents from different species within a section of the genus. As a consequence, they are very heterozygous, and the genetic distances between the parental species leads to so many genetic differences between alleles that the distinction between homologous and paralogous genes sometimes is not straightforward in all cases. In the assembly of sequences of lily and tulip it was observed that, due to high divergence among sequences, orthologous genes sometimes were split up into different contigs. This may be related to the fact that all software and programs were developed and their parameter settings were optimised for analysing data of model species such as *Arabidopsis* and rice, which are selfing species. The parameters should be adjusted to deal with outcrossing, highly heterozygous species such as many ornamentals.

Polyploidy

Many domesticated agricultural, horticultural, and ornamental crops are polyploid. In polyploids each individual carries more than two alleles at a locus. This means that identical marker alleles can occur simultaneously on different homologous chromosomes. This makes practical application of molecular markers in breeding more complicated, as there can be multiple alleles at loci for important traits, and different allele dosages for alleles present. In addition, the genetic analysis may be more complicated because of various polyploid-specific phenomena: preferential pairing of different homologs, double reduction, and multivalent pairing in meiosis.

The availability of linkage maps and the application of molecular markers to segregating populations greatly facilitate genetic analyses. This poses technological and computational problems in tetraploids. For instance, Gar et al. (2011) constructed a map for tetraploid rose, but most markers used were dominant and part of the SSR alleles were scored dominantly as well. At the moment SSRs are the only codominant marker system that is sufficiently multi-allelic to provide the opportunity to distinguish multiple or all homologous chromosomes. However, for successful analysis it is necessary to involve as many alleles, or close to that number, as there are chromosomes that segregate. In case there are fewer alleles than chromosomes the only way to deduce which allele is present

in more than one copy in parents and progeny is to score alleles quantitatively.

Using quantitative scoring it is possible to extract more information, map more markers, but also to map more accurately. For genotyping tetraploid garden roses we use an ABI platform. Quantitative scoring is performed on the base of peak areas and the comparison of their ratio across parents and progeny plants (in a variant of Esselink et al., 2004; Vukosavljev et al., in prep). Hurdles include: differential amplification of alleles and stutter bands that overlap with other alleles. A specific complication that is regularly encountered in polymorphic loci in tetraploid rose is that a single plant carries two alleles of the same length but from different origin (different parents), and that these alleles have differential amplification (presumably due to a point mutation in one of the primers) or even a totally different pattern (one allele can occur with stutters, while the other one does not have stutters, possibly indicating a rearrangement of the repeat region itself).

SNP detection in polyploids is possible (see e.g., Akhunov et al., 2009), but the interpretation poses an additional challenge. More allelic states are possible, as a biallelic SNP (a/b) can be present in a particular plant in five different genotypes: aaaa, baaa, bbaa, bbba and bbbb; nulliplex to quadruplex. Voorrips et al. (2011) developed an algorithm that can interpret the genotype data and distinguish these states.

In heterozygous ornamentals the frequency of SNPs will be high. With a high level of SNPs it is possible to analyze more than one SNP in a single sequence or SNPs in completely linked cDNA sequences. Combinations of multiple SNPs may define haplotypes, which would turn a SNP assay from a bi-allelic assay into a multi-allelic assay. In polyploids this would be a large advantage, as it would allow tagging more than two homologous chromosomes at the same time. A separate analysis of multiple SNPs does not, however, automatically disclose the identity of the haplotypes they reside in. For instance, two A/C SNPs within 100 bp could be present as A-A and C-C haplotypes, but also as A-C and C-A. In polyploids all of these combinations may even co-exist in the same plant. Using sequences directly as markers may therefore be an attractive alternative, notably in polyploids, once complexity reduction methods have sufficiently improved and library preparation methods become routine. It does require the availability of bioinformatics pipelines to automatically cluster the reads and identify the haplotypes. Once these are in place, and with further reductions in sequencing costs, developments may eventually turn towards genotyping-by-sequencing for diploid species as well, whereby separate marker development becomes obsolete.

CONCLUSION

The development of next generation sequencing and genotyping will have a large impact on breeding, also in ornamentals. To achieve the potential, co-operation and communication between bioinformaticians, researchers and breeders will be of utmost importance.

ACKNOWLEDGEMENTS

Research on marker development in ornamentals is supported by TTI-Green Genetics (projects Hyperrose, Polyploids, Lily) and several ornamental breeding companies.

Literature Cited

Akhunov, E., Nicolet, C. and Dvorak, J. 2009. Single nucleotide polymorphism genotyping in polyploid wheat with the Illumina GoldenGate assay. Theor. Appl. Genet. 119:507-517.

Andolfatto, P., Davison, D., Erezyilmaz, D. et al. 2011. Multiplexed shotgun genotyping for rapid and efficient genetic mapping. Genome Res. 21:610-617.

Barchi, L., Lanteri, S., Portis, E., Acquadro, A., Valè, G., Toppino, L. and Rotino, G.L. 2011. Identification of SNP and SSR markers in eggplant using RAD tag sequencing. BMC Genomics 12:304.

Blanca, J., Canizares, J., Roig, C., Ziarsolo, P., Nuez, F. and Picó, B. 2011. Transcriptome

characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (*Cucurbitaceae*). BMC Genomics 12:104.
Bennetzen, J.L., Schrick, K., Springer, P.S., Brown, W.E. and SanMiguel, P. 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. Genome 37:565-576.
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.
Gar, O., Sargent, D.J., Tsai, C.-J., Pleban, T., Shalev, G. et al. 2011. An autotetraploid linkage map of rose (*Rosa hybrida*) validated using the strawberry (*Fragaria vesca*) genome sequence. PLoS ONE 6:e20463.
Horning, M.E., Maloney, S.C. and Webster, M.S. 2003: Isolation and characterization of variable microsatellite loci in *Lilium philadelphicum* (*Liliaceae*). Molecular Ecology Notes 3:412-413.
Li, M. et al. 2008. Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics 24:727-745.
Li, S., Yin, T., Wang, M. and Tuskan, G.A. 2011. Characterization of microsatellites in the coding regions of the *Populus* genome. Mol. Breeding 27:59-66.
Linden van der, C.G., Smulders, M.J.M. and Vosman, B. 2005. Motif-directed profiling: a glance at molecular evolution. p.291-303. In: F.T. Bakker, L.W. Chatrou, B. Gravendeel and P.B. Pelser (eds.), Plant species-level systematics: new perspectives on pattern & process. Regnum Vegetabile 143. ARG Gantner Verlag, Ruggell, Liechtenstein; Koeltz, Koenigstein, Germany.
Parida, S.K., Pandit, A., Gaikwad, K., Sharma, T.R., Srivastava, P.S., Singh, N.K. and Mohapatra, T. 2010. Functionally relevant microsatellites in sugarcane unigenes. BMC Plant Biol. 10:251.
Park, Y.H., Ahn, S.G., Choi, Y.M., Oh, H.J., Ahn, D.C., Kim, J.G., Kang, J.S., Choi, Y.W. and Jeong, B.R. 2010. Rose (*Rosa hybrida* L.) EST-derived microsatellite markers and their transferability to strawberry (*Fragaria* spp.). Scientia Horticulturae 125:733-739.
Shahin, A., Arens, P., Van Heusden, A.W., Van Der Linden, G., Van Kaauwen, M., Khan, N., Schouten, H.J., Van De Weg, W.E., Visser, R.G.F. and Van Tuyl, J.M. 2010. Genetic mapping in *Lilium*: mapping of major genes and quantitative trait loci for several ornamental traits and disease resistances. Plant Breeding 130:372-382.
Tang, J., 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.
Tang, J., Leunissen, J.A.M., Voorrips, R.E., Van der Linden, C.G. and Vosman, B. 2008. HaploSNPer: a web-based allele and SNP detection tool. BMC Genetics 9:23.
Tangphatsornruang, S., Somta, P., Uthaipaisanwong, P., Chanprasert, J., Sangsrakru, D., Seehalak, W., Sommanas, W., Tragoonrung, S. and Srinives, P. 2009. Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek). BMC Plant Biol. 9:137.
Voorrips, R.E., Gort, G. and Vosman, B. 2011. Genotype calling in tetraploid species from bi-allelic marker data using mixture models. BMC Bioinformatics 12:172.
Wu, F. et al. 2006. Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the Euasterid plant clade. Genetics 174:1407-1420.