WAGENINGEN
UNIVERSITY & RESEARCH

Using molecular markers in breeding: ornamentals catch up

Acta Horticulturae

Smulders, M.J.M.; Bourke, P.M.; Tumino, G.; Voorrips, R.E.; Maliepaard, C. et al

https://doi.org/10.17660/ActaHortic.2020.1283.8

This article is made publicly available in the institutional repository of Wageningen University and Research, under the terms of article 25fa of the Dutch Copyright Act, also known as the Amendment Taverne. This has been done with explicit consent by the author.

Article 25fa states that the author of a short scientific work funded either wholly or partially by Dutch public funds is entitled to make that work publicly available for no consideration following a reasonable period of time after the work was first published, provided that clear reference is made to the source of the first publication of the work.

This publication is distributed under The Association of Universities in the Netherlands (VSNU) 'Article 25fa implementation' project. In this project research outputs of researchers employed by Dutch Universities that comply with the legal requirements of Article 25fa of the Dutch Copyright Act are distributed online and free of cost or other barriers in institutional repositories. Research outputs are distributed six months after their first online publication in the original published version and with proper attribution to the source of the original publication.

You are permitted to download and use the publication for personal purposes. All rights remain with the author(s) and / or copyright owner(s) of this work. Any use of the publication or parts of it other than authorised under article 25fa of the Dutch Copyright act is prohibited. Wageningen University & Research and the author(s) of this publication shall not be held responsible or liable for any damages resulting from your (re)use of this publication.

For questions regarding the public availability of this article please contact openscience.library@wur.nl

Using molecular markers in breeding: ornamentals catch up

M.J.M. Smulders[a], P.M. Bourke, G. Tumino, R.E. Voorrips, C. Maliepaard and P. Arens

Plant Breeding, Wageningen University & Research, Wageningen, The Netherlands.

Abstract

Thanks to advances in next generation sequencing it is now straightforward to develop tens or even hundreds of thousands of SNP markers. Advances in genotyping technology have made it feasible to genotype progenies of crosses, panels of genotypes, or even a complete breeding program, by using arrays with tens of thousands of SNPs, or by random or targeted sequencing technologies. Recently software has been developed for dosage scoring and linkage mapping in polyploid crops. This means that advanced genetic analyses can now also be performed in many polyploid ornamentals. A DNA marker, such as a single nucleotide polymorphism (SNP), linked to a trait enables following a gene or allele during crosses and in a breeding program. Association of a SNP marker to a trait or a component of a trait may be done through QTL analysis in segregating populations, by genome-wide association analysis (GWAS) in a set of accessions, or through an analysis across a pedigree. In these analyses, a dense linkage map is a very important tool, to delineate and possibly narrow down the QTL interval and to filter away false positive SNPs. New developments in linkage mapping include paying attention to even marker coverage, and new ways to use markers to infer haplotypes. The latter is especially important in polyploids, in populations with multiple parents, or in wide panels used for association studies. In such cases multiple functional alleles may segregate simultaneously, that cannot all be tagged uniquely by single biallelic SNP markers.

Keywords: molecular marker, SNP, linkage map, QTL, GWAS, polyploid, allele dosage

INTRODUCTION

The potential for developing and using molecular markers has improved tremendously in the past decade due to advances in three fields: next generation sequencing technologies for generating large numbers of single nucleotide polymorphisms (SNP) based on genomic or transcriptomic sequences, SNP detection systems for genotyping and automated SNP calling, and methods and software to analyze these data. Software now also exists to find associations with trait phenotypes in polyploid crops and to generate tools for marker-assisted breeding (Bourke et al., 2018a, b). As a result, nowadays it is possible to genotype a large number of samples at many marker loci simultaneously with low costs enabling the generation of high-density genetic maps.

The effects of these developments on the research in ornamental breeding have been tremendous. Whereas not so long ago, research in ornamentals was mostly suffering from the lack of a sufficient number of markers, no genomic resources, and the difficulty of genetic analysis in polyploids (Arens et al., 2012), the situation in a number of major ornamental crops has changed dramatically. A good example is the current situation in rose (Smulders et al., 2019) but also for other polyploid crops including chrysanthemum (Van Geest et al., 2017a, b), *Alstroemeria* and *Phalaenopsis* (Cai et al., 2015) developments have been significant. The challenge for the coming years will be to implement the use of these tools in ornamental crops (Smulders and Arens, 2018) and also develop new strategies to deal with the high level of genetic variation in outcrossing polyploid ornamentals for use in marker-assisted breeding (MAB).

[a]E-mail: rene.smulders@wur.nl

HOW TO IMPLEMENT MARKERS?

How to employ the existing knowledge for steps in the breeding of rose in an efficient and cost-effective way? At least four questions are important: which steps of the breeding process may benefit from marker information, what is the value of the identified plus-alleles for the trait, where in the breeding germplasm do these plus-alleles occur, and how can plants carrying the alleles be selected efficiently (Smulders and Arens, 2018)?

Cameron Peace (2017) has defined a scheme of five steps for translating the output of genomics research into a routine application that is integrated in a breeding program, a situation which he coined 'DNA-informed breeding'. This term is equivalent to the term marker-assisted breeding, with two differences. First, it does not require that the reader understands what a marker is and how it works, which may be an advantage during communication with breeders, other possible users, and the general public. Second, it also includes exploitation of neutral markers, as distinct from markers that are associated with traits, for applications such as determining parentage or checking identity.

Peace (2017) recognizes the following five steps:

1. Establishing a breeder's need (or advantage) for use of DNA information for important traits;
2. Adapting tools to the local breeding situation;
3. Identifying efficient application schemes;
4. Accessing effective services in DNA-based diagnostics (this step is often outsourced, balancing cost-effectiveness with throughput capacity and time needed to obtain the results);
5. Gaining experience in conducting DNA-informed breeding.

DNA information may be used for a range or purposes. Without the intention of being complete, this includes:

- Structure of the germplasm;
- Parental selection;
- Seedling selection;
- Identity checks in the breeding program;
- Variety protection.

Also, phenomena such as meiotic behavior and segregation distortion can be studied when using large bi-parental populations (Bourke et al., 2017; Smulders and Arens, 2018).

IMPORTANCE OF A DENSE LINKAGE MAP AND A GENOME SEQUENCE

The availability of a high-density integrated genetic linkage map is important for several reasons, both in trait discovery as well in later use for selection purposes in breeding. Of course, proximity of markers to the trait gene (preferably on both sides) is important even in single dominant trait situations where for instance a resistance gene from a single unique source should be followed in crosses. In a QTL analysis, where the location of the trait gene QTL is delineated on a region of the genetic map, using too few mapped markers and/or too few offspring plants means that the QTL region is too large and the exact location remains unclear. In a GWAS analysis many false positive SNPs may occur, and as a consequence one does not look at the p-value of an individual marker but at the shape of the peak of p-values of SNPs in certain regions on the genome. Therefore, in a GWAS, having a genome may enable to position any unmapped SNP marker which otherwise is effectively ignored.

With a genome sequence one can even drill down to candidate genes (Smulders et al., 2019). In practice, the QTLs are detected with linked markers and these markers are then located on the scaffolds or pseudochromosomes to obtain the corresponding region of the genome sequence. To define the region there are some steps that might have to be developed a bit more depending on the accuracy of genome sequences and genetic resources available. With the current situation in rose having a high-quality genome sequence from di-haploids (Hibrand Saint-Oyant et al., 2018; Raymond et al., 2018), Linkage Disequilibrium (LD) can be used to define the QTL region. Using the pairwise SNP marker information over the whole GWAS set it is possible to calculate chromosome specific LDs. From here on, using the functional annotation of the genome, putative candidate genes can be identified based on

similarities with genes known to be involved in the studied process in other plants including model plants such as *Arabidopsis thaliana*. The next step is to look for differential expression or the presence of structural variants or SNPs between varieties on the extremes of the trait values. Subsequently other candidate gene validation methods (VIGS, RNAi or gene editing) may be tested for effects on the trait value depending on the particular QTL gene effect size.

GENOTYPIC INFORMATION CONTENT (GIC)

A marker can either tag a single allele, or multiple alleles simultaneously. Haplotype-specific markers are extremely useful if they are closely linked to a single allele of interest, as they can be directly used for selection purposes in a breeding program. However, in heterozygous or polyploid species, it is often the case that markers tag multiple alleles (particularly in the heterozygous and polyploid scenario). One of the consequences of this is that incomplete inheritance information is carried by these markers. In order to quantify this phenomenon, the concept of genotypic information has been proposed to highlight genomic regions where marker coverage is less informative than might appear initially. Genetic maps are often presented in terms of numbers of markers, total map length, average inter-marker distances, or maximum gap-size between successive markers. Although these statistics provide some descriptive information, they often mask the inadequacies of a genetic map by presenting an overly-optimistic picture of marker density and distribution. The GIC measure, on the other hand, is a direct representation of the amount of information (carried by all markers) on the transmittance of alleles across generations, which is directly relevant to subsequent applications like QTL detection or marker-assisted selection. Although somewhat less amenable to tabular form, the GIC can easily be visualized along chromosomes (much like a QTL detection profile) and thus provide a complete overview of the information content carried by a marker set in a particular population (Figure 1).

Through a series of simulation studies, it was found that variable GIC can influence both the power to detect QTL effects and the precision of QTL mapping (Bourke et al., 2019). In other words, simply developing high-density marker data sets may not be enough to fully saturate complex genomes from a breeding/inheritance-tracking perspective. Regions showing low GIC levels could therefore benefit from further marker development (particularly if such regions are thought to harbour interesting alleles, for example through previous studies in the literature, or based on synteny analyses).

HAPLOTYPES

So far, we have assumed that the markers we have available are biallelic: a SNP usually has only two alleles. Although in theory three or four SNP alleles may occur this is the case for only a minority of SNP positions (and such SNPs are problematic to score, at least using array technologies or single SNP marker technologies like KASP). However, if we consider a relatively short segment of DNA (which we call a haploblock) covering multiple bi-allelic SNPs, then more than two alleles of such a segment (haplotypes) may exist. For example, let us assume the haploblock covers 3 bi-allelic SNPs of which the first has alleles A and G, the second has alleles C and G and the third one A and T. Then in principle 8 different haplotypes may occur for this haploblock, for example one haplotype with bases A-C-A at the SNP positions, a second haplotype G-C-A, a third G-G-T. etc. Potentially such multi-allelic haploblocks carry more information than separate SNPs, which can be exploited in linkage mapping or QTL mapping in experimental populations, for GWAS analyses in panels or in pedigreed breeding populations. The basic idea is that the multiple alleles provide more information to discriminate founder alleles in the total population; if sufficient haplotypes can be recognized Identity-by-State (IBS) becomes a reliable indication for Identity-by-Descent (IBD), which is the basis for many QTL mapping approaches. An assumption here is that the haploblock is short enough that recombinations within the haploblock are present at a negligible rate in the population studied.

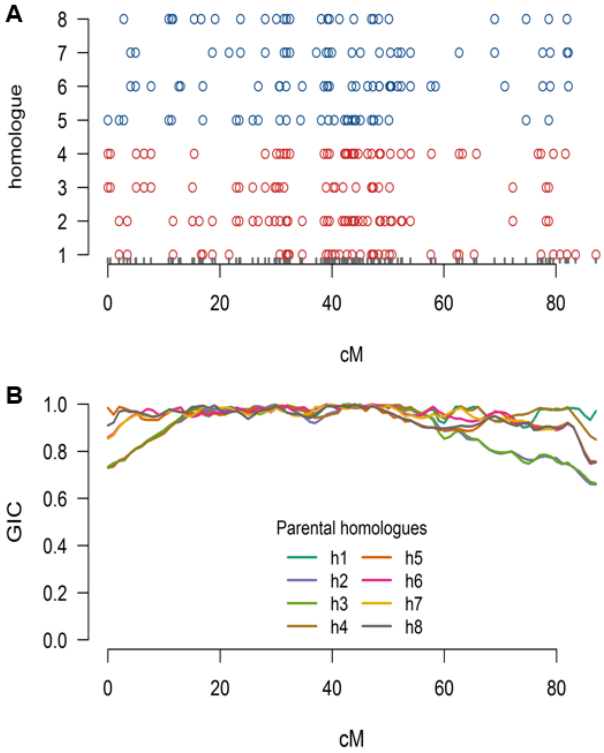

Figure 1. Effect of marker distribution on genotypic information content (GIC). A. Marker distribution across 8 parental homologs (chromosome copies from an autotetraploid cross). Maternal alleles are shown in red, paternal in blue. B. GIC values plotted for the same data set, with near-full information in the central regions but a drop-in information toward the telomeres. Homologs 2 and 3 are less well-covered, shown as near-overlapping purple and green lines that drop distinctively more toward the telomeres. Rudimentary genetic map statistics do not provide such a detailed picture of marker distribution and informativeness.

While multi-SNP, multi-allelic haploblocks have advantages in genetic analyses, inferring the haploblock genotype of an individual (i.e. the combination of haplotypes it carries at the haploblock) is more challenging than inferring its SNP genotypes at the individual SNP positions. SNP dosages are most reliably estimated using array technologies, using software tools as discussed earlier. Alternatively, they can be obtained from sequencing experiments. Similarly, haploblock genotypes may be inferred from SNP arrays or sequencing. In the case of SNP arrays, the initial genotyping results are dosages of the separate SNPs. For diploids established software tools such as Beagle (Browning and Browning, 2009) and AlphaPhase (Hickey et al., 2011) exist to infer haplotype combinations from separate SNP data. For polyploids the first tools are appearing. We are developing a haplotyping tool that takes advantage of Full-Sib populations, but which can deal with unstructured populations as well. Our group has also developed and published PopPoly (Motazedi et al., 2019), software to infer haploblock genotypes from sequence reads.

Inferring haploblock genotypes is only part of the challenge. In order to use these genotype data for linkage and QTL mapping, adaptations of existing software, or entirely new approaches are needed. Also, in these fields we are in the process of developing the necessary tools. For practical applications in breeding selection the haploblock SNPs ideally should be within such close proximity that they can be captured in single sequence reads allowing for direct haplotype assessment. The high degree of genetic variation in many polyploid ornamentals may pose opportunities for selection of markers for such applications.

ACKNOWLEDGEMENTS

This research was partially supported by KB-24-002-017 and the TKI-U Polyploid projects BO-26.03-002-001 and BO-26.03-009-004. The support of the companies participating in the Polyploid projects is gratefully acknowledged.

Literature cited

Arens, P., Bijman, P., Tang, N., Shahin, A., and Van Tuyl, J.M. (2012). Mapping of disease resistance in ornamentals: a long haul. Acta Hortic. *953*, 231–238 https://doi.org/10.17660/ActaHortic.2012.953.32.

Bourke, P.M., Arens, P., Voorrips, R.E., Esselink, G.D., Koning-Boucoiran, C.F., Van't Westende, W.P., Santos Leonardo, T., Wissink, P., Zheng, C., van Geest, G., et al. (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. Plant J. *90* (*2*), 330–343 https://doi.org/10.1111/tpj.13496. PubMed

Bourke, P.M., van Geest, G., Voorrips, R.E., Jansen, J., Kranenburg, T., Shahin, A., Visser, R.G.F., Arens, P., Smulders, M.J.M., and Maliepaard, C. (2018a). polymapR-linkage analysis and genetic map construction from F1 populations of outcrossing polyploids. Bioinformatics *34* (*20*), 3496–3502 https://doi.org/10.1093/bioinformatics/bty371. PubMed

Bourke, P.M., Voorrips, R.E., Visser, R.G.F., and Maliepaard, C. (2018b). Tools for genetic studies in experimental populations of polyploids. Front. Plant Sci. *9*, 513 https://doi.org/10.3389/fpls.2018.00513. PubMed

Bourke, P.M., Hackett, C.A., Voorrips, R.E., Visser, R.G.F., and Maliepaard, C. (2019). Quantifying the power and precision of QTL analysis in autopolyploids under bivalent and multivalent genetic models. G3 (Bethesda) *9* (*7*), 2107–2122 https://doi.org/10.1534/g3.119.400269. PubMed

Browning, B.L., and Browning, S.R. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am. J. Hum. Genet. *84* (*2*), 210–223 https://doi.org/10.1016/j.ajhg.2009.01.005. PubMed

Cai, J., Liu, X., Vanneste, K., Proost, S., Tsai, W.C., Liu, K.W., Chen, L.J., He, Y., Xu, Q., Bian, C., et al. (2015). The genome sequence of the orchid *Phalaenopsis equestris.* Nat. Genet. *47* (*1*), 65–72 https://doi.org/10.1038/ng.3149. PubMed

Hibrand Saint-Oyant, L., Ruttink, T., Hamama, L., Kirov, I., Lakhwani, D., Zhou, N.N., Bourke, P.M., Daccord, N., Leus, L., Schulz, D., et al. (2018). A high-quality genome sequence of Rosa chinensis to elucidate ornamental traits. Nat. Plants *4* (*7*), 473–484 https://doi.org/10.1038/s41477-018-0166-1. PubMed

Hickey, J.M., Kinghorn, B.P., Tier, B., Wilson, J.F., Dunstan, N., and van der Werf, J.H. (2011). A combined long-range phasing and long haplotype imputation method to impute phase for SNP genotypes. Genet. Sel. Evol. *43* (*1*), 12–24 https://doi.org/10.1186/1297-9686-43-12. PubMed

Motazedi, E., Maliepaard, C., Finkers, R., Visser, R., and de Ridder, D. (2019). Family-Based Haplotype Estimation and Allele Dosage Correction for Polyploids Using Short Sequence Reads. Front. Genet. *10*, 335 https://doi.org/10.3389/fgene.2019.00335. PubMed

Peace, C.P. (2017). DNA-informed breeding of rosaceous crops: promises, progress and prospects. Hortic. Res. *4* (*1*), 17006 https://doi.org/10.1038/hortres.2017.6. PubMed

Raymond, O., Gouzy, J., Just, J., Badouin, H., Verdenaud, M., Lemainque, A., Vergne, P., Moja, S., Choisne, N., Pont, C., et al. (2018). The *Rosa* genome provides new insights into the domestication of modern roses. Nat. Genet. *50* (*6*), 772–777 https://doi.org/10.1038/s41588-018-0110-3. PubMed

Smulders, M.J.M., and Arens, P. (2018). New developments in molecular techniques for breeding in ornamentals. In Ornamental Crops. Handbook of Plant Breeding, vol 11, J. Van Huylenbroeck, ed. (Cham: Springer) p.213–230.

Smulders, M.J.M., Arens, P., Bourke, P.M., Debener, T., Linde, M., Riek, J., Leus, L., Ruttink, T., Baudino, S., Hibrant Saint-Oyant, L., et al. (2019). In the name of the rose: a roadmap for rose research in the genome era. Hortic. Res. *6* (*1*), 65 https://doi.org/10.1038/s41438-019-0156-0. PubMed

van Geest, G., Voorrips, R.E., Esselink, D., Post, A., Visser, R.G., and Arens, P. (2017a). Conclusive evidence for hexasomic inheritance in chrysanthemum based on analysis of a 183 k SNP array. BMC Genomics *18* (*1*), 585 https://doi.org/10.1186/s12864-017-4003-0. PubMed

van Geest, G., Bourke, P.M., Voorrips, R.E., Marasek-Ciolakowska, A., Liao, Y., Post, A., van Meeteren, U., Visser, R.G.F., Maliepaard, C., and Arens, P. (2017b). An ultra-dense integrated linkage map for hexaploid chrysanthemum enables multi-allelic QTL analysis. Theor. Appl. Genet. *130* (*12*), 2527–2541 https://doi.org/10.1007/s00122-017-2974-5. PubMed