Marker-assisted Backcross Breeding: A Case-Study in Genotype Building Theory

Frédéric HOSPITAL

INRA
Station de Génétique Végétale
Ferme du Moulon
91190 Gif Sur Yvette --- France

Introduction

I wish here to provide an overview of some past and more recent results on marker-assisted backcross breeding theory, and discuss the general consequences for marker-assisted selection and genotype building. Backcross breeding (MAB) is a well-known procedure for the introgression of a *target gene* from a *donor line* into the genomic background of a *recipient line*. The objective is to reduce the *donor genome content* (DGC) of the progenies by repeated back-crosses to the recipient line. Genotype building (GB) terms here the use of markers to design new genotypes combining favourable alleles previously detected at a (possibly large) number of loci, in (possibly many) different parental lines. Here, the genomic background in which those alleles are combined cannot, in general, be controlled because the genes are too numerous. The theory in this domain remains largely unexplored, and few results are available. For example, de Koning and Weller (1994), and Dekkers and van Arendonk (1998) have considered the optimization of marker-assisted selection for identified quantitative trait loci (QTL) plus a possible 'polygenic' background controlling the rest of the genetic variation not explained by the identified QTL. These analyses are restricted to one or two identified QTL. Also, van Berloo and Stam (1998) and Charmet *et al.* (1999) have considered a larger set of identified QTL, each controlled by flanking markers, and studied selection of recombinant inbred lines or doubled haploids based on flanking markers to produce the best hybrid. This analysis is restricted to selection among inbreds for one or two generations only. Hospital *et al.* (2000) studied selection on marker pairs flanking 50 QTL identified in an F_2 population. With a 'QTL complementation strategy' selection of 3-5 individuals among a total of 200 for 10 generations increases the frequency of favourable alleles at the 50 QTL up to 100% when markers are located exactly on the QTL, but only to 92% when marker-QTL distance is 5 cM. The authors conclude that the efficiency of marker-based selection is bounded by the recombinations taking place between the markers and the QTL. Hence, one has to accelerate the response to selection to fix favourable QTL alleles before marker-QTL linkage disequilibrium vanishes. The main limitation identified is the fact that selected individuals are mated at random: the authors suggest that pairwise matings should increase the efficiency of selection. But, the theory in this domain remains unexplored.

Marker-assisted backcross is of great practical interest in applied breeding schemes either to manipulate 'classical' genes between elite lines or from genetic resources, or to manipulate transgenic constructions. From a theoretical standpoint, it is a 'simple' example of marker-based selection: in general, only two alleles are segregating, and the gametic phase is known because only one chromosome of each pair is issued from effective recombination (the chromosome from the gamete produced by the backcrossed parent). It is then also an

appropriate case-study to investigate how selection and recombination work together to make it work better in any type of marker-assisted selection programme.

In backcross breeding, markers can be used to: i) control the target gene (*foreground selection*) if needed: Melchinger (1990) discussed the optimal scheme to obtain a minimum number of individuals carrying a target gene of known location ; Hospital and Charcosset (1997) discussed the optimal number and positions of markers to control a QTL (target gene of uncertain location); and/or ii) control the genetic background (*background selection*). The objective of background selection is to accelerate the return to recipient genome outside the target gene, by selection of the recipient allele at markers located either on the *carrier chromosome* (the chromosome carrying the target gene) and/or on *non-carrier chromosomes* (the other chromosomes). Background selection has already been shown to be efficient by previous theoretical works (*e.g.,* Hillel *et al.*, 1990; Hospital *et al.*, 1992; Groen and Smith, 1995; Visscher *et al.*, 1996), and experimental works (*e.g.*, Ragot *et al.*, 1995). I wish here to focus on recent theoretical developments achieved by our group on two aspects of background selection: the reduction of linkage drag around the target gene, and the estimation of recipient genome content in backcross progenies.

In any case, one must keep in mind that selection on markers in backcross programs is considered efficient if it permits a return to recipient genome outside the target gene faster than the normal return rate when no selection on markers is applied (donor genome content halves at each generation). Hence, the efficiency of marker-assisted selection should always be compared with this normal rate as a reference.

The reduction of linkage drag in marker-assisted backcross programs

The carrier chromosome deserves special consideration in backcross programs because, due to selection for the donor allele at the target locus in each generation, the rate of return to recipient genotype on this chromosome is slower than on non-carrier chromosomes. Stam and Zeven (1981) provided an equation to calculate this rate of return when no selection on markers is applied. Based on a numerical comparison of these results to the known rate of return on non-carrier chromosomes (donor genome content halved each generation), Young and Tanksley (1989a) pointed out that the donor genes on the carrier chromosome were the most difficult to eliminate, and could persist in the progenies long after the donor genome content on non-carrier chromosomes has returned to approximately zero if no selection on markers was applied. They provided an impressive experimental proof of this statement, based on the *a posteriori* genotyping of a collection of tomato varieties previously introgressed for a resistance gene.

Size of intact donor chromosome segments around the target gene

The *intact donor segment* is in any BC generation the chromosome segment of donor origin containing the target locus, which has remained unaltered by crossovers since the original cross between the donor and recipient parents. Hanson (1959) first provided the theoretical expression for the expected length of this intact segment. This was later revisited by Naveira and Barbadilla (1992), who also provided the corresponding variance. It is important to note that Stam and Zeven (1981) provided the *total* proportion of donor alleles on the carrier chromosome either on the intact segment or on other non-contiguous blocks of genes elsewhere on the carrier chromosome, which is a different measure of linkage drag. In fact, comparing numerically the proportion of donor alleles on the intact segment with the total proportion shows that the vast majority of unwanted donor alleles are located on the intact

donor segment in advanced BC generations. Hence, I will focus here only on the intact segment as a measure of linkage drag.

Hospital (2001) computed the mean and variance of the length of the intact donor segment around the target gene, when background selection is applied on two markers flanking the gene, one on each side (*i.e.*, size of segment amongst *ideotypes:* individuals that are heterozygous at the target locus, and homozygous for the recipient allele at both flanking markers) in any BC generation. The numerical results indicate that the expected length of donor segment on each side of the target gene is approximately half of the distance between the gene and the flanking marker in BC1, but the length at more advanced BC generations depends on the marker distance. For distant markers (more than 30 cM), the expected length of donor segment decreases in advanced BC generations, because recombination events accumulate between the target gene and the marker during successive meioses. This is no longer the case for shorter markers distances: for markers at 20 cM from the target gene or closer, the expected size of donor segment in advanced BC generation is approximately the same as the expected size in BC1. In this case, recombination events are rare and do not accumulate: in general, the genotypes selected experienced only one crossover, the one that permitted the flanking marker to return to recipient genotype. The basic conclusion is that selecting for distant markers over several successive backcross generations cannot provide a better reduction of linkage drag than using close markers. Using very close markers is the only way to reduce linkage drag substantially.

Optimal population sizes

The above results refer to the length of donor segment in individual genotypes homozygous for the recipient allele at both flanking markers (*double recombinants*) but say nothing about the probability to obtain such genotypes. In a classical situation in plant breeding, where, among a whole population, a single individual can be selected and backcrossed to produce the population at the next generation, such probability obviously depends on population sizes. Obviously, using close markers as recommended above probably implies screening large populations which generates large genotyping costs. It is, thus, important to optimize population sizes, *i.e.,* determine the minimal population sizes (and genotyping effort) necessary to obtain the desired genotypes. Although it is intuitive that, for close flanking markers, double recombinant genotypes are highly unlikely to be obtained in one single generation (BC1) so that at least two BC generations should be performed (Young and Tanksley, 1989a), the underlying mathematics have been worked out only recently. A first solution was derived by Hospital and Charcosset (1997). This result was used by Frisch *et al.* (1999) with numerical applications in the context of single-generation optimization (population size is optimized to permit the selection of a double recombinant genotype at generation $t+1$, given that the genotype selected at generation t is known), whereas Hospital (2001) showed that a better optimization is obtained when considering all the planned generations simultaneously. The best optimization strategy is to i) determine the maximal number of BC generations that could be performed in a breeding program; ii) optimize simultaneously the population sizes at each of those previously defined generations before the programme is started; and iii) refine the optimization at each generation, when the genotype of the selected individual is known. This requires some numerical computation. A computer programme (*popmin*) that performs the corresponding numerical calculations easily was designed (Hospital and Decoux, submitted) and is freely available at http://moulon.inra.fr/~fred/programs. The results indicate that optimal population sizes should not be the same at each BC generation (using larger population sizes in advanced generations than in early generations reduces the overall number of individuals genotyped during the breeding scheme), as pointed out by Hospital and Charcosset (1997). More

generally, the results indicate that a drastic reduction of linkage drag can be obtained at reasonable costs by performing more than two BC generations. For example, for flanking markers as close as 2 cM on each side of the target gene, the minimum number of individuals that should be genotyped to obtain a double recombinant in BC1 is about 24,000. The same result can be obtained over two generations (BC2 strategy) by genotyping 290 individuals in BC1, and 500 in BC2. Finally, over three generations (BC3 strategy), the optimal population sizes are 120 individuals in BC1, 170 in BC2, and 370 in BC3. In all three strategies, the probability to obtain a double recombinant for the flanking markers by the end of the breeding programme is above 99%. In the BC3 strategy, the probability to obtain a double recombinant in BC2 is about 75%. In case this happens, the programme is obviously not pursued until BC3 (unless for other reasons not considered here). Hence, planning to perform a maximum of three BC generations (BC3 strategy) permits in 75% of the cases to obtain a double recombinant in BC2 with genotyping a total of only 290 individuals, which is much less than the 790 individuals necessary with the BC2 strategy. With the BC3 strategy, only in 25% of the cases should the programme be really conducted until generation BC3. Hence, averaging over all possibilities, the mean number of individuals that need to be genotyped to obtain a double recombinant with the BC3 strategy is only about 380, to be compared with an average of about 760 with the BC2 strategy. Hence, planning at the beginning of the programme to perform more than two BC generations is always a better strategy to optimize the costs of genotyping (unless a rapid success is really mandatory). This is equivalent to fixing a not-too-low risk of failure per generation (risk of not obtaining a double recombinant at that generation), in particular in early BC generations, which is converse to what was advocated by Frisch *et al.* (1999). Obviously, the strategy and number of individuals to be genotyped should be reconsidered at each generation once the genotype of the individual selected is known. This is also possible using our computer programme *popmin*. In conclusion, planning to perform three or more BC generations and/or increasing the risk per generation has two main advantages: First, for given and affordable population sizes (genotyping effort), it permits a more drastic reduction of linkage drag. This is particularly useful for introgression of genes from exotic genetic resources that main contain undesirable genes surrounding the gene of interest, for the manipulation of transgenic constructions (GMOs) when the introgression of the construction only is desired and close markers or better sequences of flanking regions are available, and/or for the derivation of near-isogenic lines (NIL) or congenic lines for the identification and validation of quantitative trait loci. Second, planning to perform more than two BC generations increases the probability of success (obtaining a double recombinant) in advanced BC generations. The optimal population sizes above are defined such that at least one double recombinant is obtained with a given risk. It is then likely that on average more than one is obtained. Background selection, for markers on non carrier chromosomes, is then possible among those double recombinants. This permits a better reduction of donor genome content on other chromosomes. Moreover, background selection on non-carrier chromosomes is more efficient in advanced backcross generations (Hospital *et al.*, 1992).

Background selection on non-carrier chromosomes: estimation of donor genome content.

Computation of multilocus genotype frequencies in complex pedigrees

This section is not just related to marker-assisted backcross breeding, though two applications in this field are given in the following sections. However, I want to mention these results because it can prove useful in various areas of MAS and genotype building theory, as well as for QTL detection.

Computing expected genotype frequencies at several loci (three or more) and/or in complex breeding schemes (backcrossing, hybrid mating, random mating, selfing, full-sib mating, or any combination of these) is sometimes necessary in plant breeding. Actually, it is more and more frequent when using marker information, because many theoretical calculations are based on the probabilities of the different possible genotypes at markers (*e.g.*, in QTL detection), or because one wishes to predict the probability of obtaining a particular genotype at markers or loci of interest (see an example above for the reduction of linkage drag in BC). However, such calculations are tedious and barely amenable by hand. Hospital *et al.* (1996) have proposed a general algorithm to derive such probabilities automatically by recursion, and provided the corresponding *Mathematica* notebooks (http://moulon.inra.fr/~fred/programs). These recursions were implemented in a general programme (*mdm*) performing numerical and more powerful calculations by Servin *et al.* (submitted), also available at the above Web page. The programme *mdm* has various applications in plant and animal genetics. Two exemples are provided below.

Precision graphical genotypes

To estimate the genomic composition of individuals using markers, the most basic estimate of donor genome content (DGC) could be to score the genotype at the markers, and then estimate DGC from the ratio of markers heterozygous for the donor allele over the total number of markers scores. This is a crude estimate that has the major drawback of being highly dependent upon the placement of markers along the genome. If markers are evenly spread and not too far apart from each other, the estimate is not correct (see below) but could be accepted. However, it is self evident that if markers are not evenly distributed (the real situation), weighting them equally is clearly not the best solution.

A first attempt to provide a better estimate of DGC by taking the marker locations into account was made by Young and Tanksley (1989b), who introduced the concept of *graphical genotypes*, to 'portray the parental origin and allelic composition throughout the genome'. This takes into account distances between markers in the sense that a chromosomal segment flanked by two markers of donor type (DD) is considered as 100% donor type, a chromosomal segment flanked by two markers of recipient type (RR) is considered as 0% donor type, and a chromosomal segment flanked by one marker of donor type and one marker of recipient type (DR) is considered as 50% donor type.

Using the programme *mdm*, it is possible to compute, at any point of a segment flanked by two markers, the probability of being of donor type, given the genotypes at the markers and their locations. Averaging over all possible positions between the two markers provides an estimate of DGC: *precision graphical* genotype (PGG). This shows that the estimate of Young and Tanksley (1989b) is not always correct: In DD segments, DGC is below 100% due to possible double crossovers between the markers. This error is minimal in BC1 and increases in more advanced BC generations. In RR segments, DGC is above 0% due to possible double recombinations between the markers. This error is maximal in BC1 and decreases in more advanced BC generations. However, the errors on either DD or RR segments are numerically not very important. In DR segment, DGC is exactly 50% in BC1 but decreases to below 50% in advanced BC generations. Paradoxically, although the estimate of Young and Tanksley on DR segment is correct in BC1, it is for the same segments that the error is quantitatively the most important in advanced BC generations. As the general trend in backcrossing is to have more and more markers of recipient type in advanced BC generations, even with no selection on the markers, it is expected that many segments are of DR type, hence the overall error might be important.

Extending these results using *mdm*, Servin (*in prep.*) has shown that, when estimating the DGC in a chromosomal segment flanked by two markers at a given generation, not only

the genotypes of the two markers at that generation are informative. In fact, the genotypes of the two markers at previous generations also matter, and so do the genotypes of non-flanking markers ('second' markers on the 'left' or on the 'right' of the segment, 'third' markers, and so on…). Taking this additional information into account permits in some cases to gain in precision on the estimate of the most probable genotype at any point in the segment. This is useful for graphical genotypes and DGC estimates, but also for any purpose where this type of calculation is necessary, probably the most important one being QTL detection. Using simulation, it was shown that the correlation between the 'true' DGC and its estimate by *mdm* is very good (Servin, *in prep*.). The programme *mdm* can be included as a subroutine in any programme performing such calculation (*e.g.*, QTL detection programs) and should permit a gain in the precision of the corresponding estimates. However, the amount of this gain remains to be quantified and deserves more work.

Application to maize data

Precision graphical genotypes derived using *mdm* were applied to experimental data regarding marker-assisted introgression of three favourable QTL alleles between maize elite lines (Bouchez *et al.*, in prep.). Three QTL were detected in a recombinant inbred line population. The favourable quantitative trait alleles (QTA) at those three loci originating from the first parental lines were introgressed into the genomic background of the second parent through three crosses to the second parent (*i.e.*, one non-segregating cross followed by two backcrosses, followed by one generation of selfing to fix the QTA in homozygous state. This experiment shows that marker-assisted backcrossing can be used to manipulate QTA between elite lines, although the validation of QTL effects in introgressed progenies appears easier for simple traits (*e.g.,* earliness) than for more complex traits (*e.g.,* yield), most probably because of stronger genotype-by-environment interactions. In any case, the experiment is amongst the few public experimental demonstrations of the efficiency of marker-based selection in backcross programs. In addition, the complex pedigree corresponding to this experiment was a challenging opportunity to apply the method of *precision graphical genotypes*, using *mdm* to estimate the genome contents of the products. The results show that with only about 200 individuals genotyped per generation, and a total of 15 markers on non-carrier chromosomes, the return to the recipient parent is close to 100% after two BC and one selfing generations. Chromosomal segments containing the three QTL were efficiently controlled by three markers per segment. However, the small population sizes did not permit a drastic reduction of linkage drag, which was not especially desired here because of the uncertainty about QTL locations. Comparing *precision graphical genotypes* to the approximation of Young and Tanksley described above indicates that the difference in the estimates can be important, up to ±8% genome content in some cases. The sign of the difference may vary from one chromosome to another, indicating that the error is probably more important qualitatively than quantitatively. The error is particularly important for chromosomal segments flanked by markers of different genotypes, and in advanced BC generations as expected. In particular, *mdm* can predict possible residual heterozygosity in the final material where the other approximation obviously cannot.

One possibility is to use *precision graphical genotypes* to provide a better estimate of genome contents for a set of known markers. Conversely, since the estimate provided is more accurate, this should help reduce the number of markers genotyped, and hence reduce the experimental costs. This remains under development.

References

Charmet, G., Robert, N., Perretant, M.R., Gay, G., Sourdille, P., Groos, C., Bernard, S. and Bernard, M. (1999) Marker-assisted recurrent selection for cumulating additive and

interactive QTLs in recombinant inbred lines. *Theoretical and Applied Genetics* 99, 1143-1148.
Dekkers, J.C.M. and van Arendonk, J.A.M. (1998) Optimizing selection for quantitative traits with information on an identified locus in outbred populations. *Genetical Research* 71, 257-275.
Frisch, M., Bohn, M. and Melchinger, A.E. (1999) Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Science* 39, 967-975.
Groen, A.F., and Smith, C. (1995) A stochastic simulation study on the efficiency of marker-assisted introgression in livestock. *Journal of Animal Breeding and Genetics* 112, 161-170.
Hanson, W.D. (1959) Early generation analysis of lengths of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. *Genetics* 44, 833-837.
Hillel, J., Schaap, T., Haberfeld, A., Jeffreys, A .J., Plotzky, Y., Cahaner, A., and Lavi, U. (1990) DNA fingerprint applied to gene introgression breeding programs. *Genetics* 124, 783-789.
Hospital, F. (2001) Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. *Genetics* (in press).
Hospital, F. and Charcosset, A. (1997) Marker-assisted introgression of quantitative trait loci. *Genetics* 147, 1469-1485.
Hospital, F., Chevalet, C. and Mulsant, P. (1992) Using markers in gene introgression breeding programs. *Genetics* 132, 1199-1210.
Hospital, F., Dillmann, C. and Melchinger, A.E. (1996) A general algorithm to compute multilocus genotype frequencies under various mating systems. *Computer Applications in the Biosciences*. 12, 455-462.
Hospital, F., Goldringer, I. and Openshaw, S.J. (2000) Efficient marker-based recurrent selection for multiple quantitative trait loci. *Genetical Research* 75, 357-368.
de Koning, G.J. and Weller, J.I. (1994) Efficiency of direct selection on quantitative trait loci for a two-trait breeding objective. *Theoretical and Applied Genetics* 88, 669-677.
Melchinger, A.E. (1990) Use of molecular markers in breeding for oligogenic disease resistance. *Plant Breeding* 104, 1-19.
Naveira, H. and Barbadilla, A. (1992) The theoretical distribution of lengths of intact chromosome segments around a locus held heterozygous with backcrossing in a diploid species. *Genetics* 130, 205-209.
Ragot, M., Biasiolli, M., Delbut, M.F., Dell'orco, A., Malgarini, L. *et al.* (1995) Marker-assisted backcrossing: a practical example. In: *Techniques et utilisations des marqueurs moléculaires*. (Les Colloques, no 72)} Ed. INRA, Paris, pp. 45--56.
Stam, P., and Zeven, A.C. (1981) The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica* 30, 227-238.
van Berloo, R. and Stam, P. (1998) Marker-assisted selection in autogamous RIL populations: a simulation study. *Theoretical and Applied Genetics* 96, 147-154.
Visscher, P.M., Haley, C.S. and Thompson, R. (1996) Marker-assisted introgression in backcross breeding programs. *Genetics* 144, 1923-1932.
Young, N.D. and Tanksley, S.D. (1989a) RFLP analysis of the size of chromosomal segments retained around the *tm-2* locus of tomato during backcross breeding. *Theoretical and Applied Genetics* 77, 353-359.
Young, N.D. and Tanksley, S.D. (1989b) Restriction fragment length polymorphism maps and the concept of graphical genotypes. *Theoretical and Applied Genetics* 77, 95-101.