# Marker-assisted introgression: speed at any cost?

**Piet Stam**
*Wageningen University, Department of Plant Sciences, Laboratory of Plant Breeding, PO Box 386, 6700 AJ Wageningen, The Netherlands.*
*Contact: Piet.Stam@wur.nl, (+31) 317 842841.*

**Abstract**: Marker-assisted introgression currently represents the most widely spread application of DNA markers as an aid to selection in plant breeding. This paper discusses some quantitative aspects of this marker-guided introgression, relating to optimisation of an introgression program. Special attention is given to the effect of genomic map length, population size and duration of a backcross program on the attainable rate of donor genome substitution.

**Keywords**: marker-assisted breeding, introgression, optimisation, genome size, genome substitution rate.

## Introduction

In marker-assisted breeding the plant breeder takes advantage of the association between agronomic traits and allelic variants of genetic markers, mostly molecular markers. Usually these associations are the result of genetic linkage between markers and gene loci underlying the trait(s) of interest. Associations of this kind, also known as *linkage disequilibrium*, arise in experimental populations used for linkage mapping, such as backcrosses (BC), F2's, recombinant inbred lines (RILs) or doubled haploids (DH). In cross-fertilizing plant species a mapping population usually consists of a large full sub family resulting from a cross between single plants of divergent genotype.

Before a breeder can utilize linkage-based associations between traits and markers, the associations have to be assessed with a certain degree of accuracy, such that it can be safely relied on, and thus marker genotypes can be used as indicators or predictors of trait genotypes and phenotypes. For monogenic traits with a clear qualitative contrast between genotypes, such as a single gene-based disease resistance, the assessment of association is straightforward: mapping a monogenic trait goes along with the mapping of markers. For qualitative, polygenic traits, however, a reliable assessment of trait-marker association requires large-scale field experiments as well as statistical techniques, known as QTL mapping.

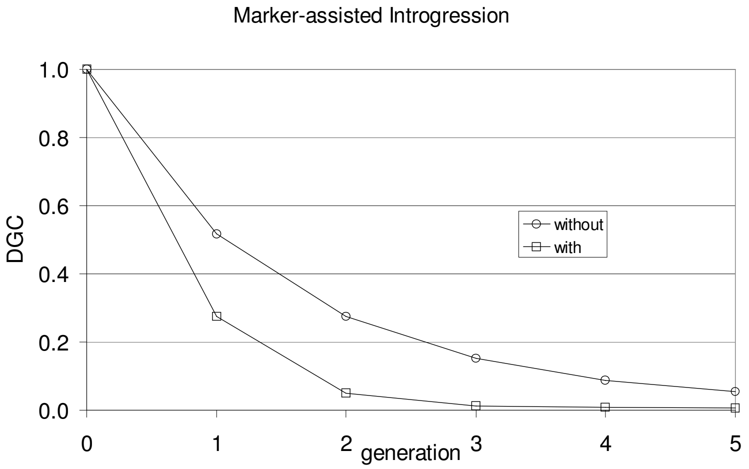
The general idea behind marker-assisted breeding is as follows. Once marker-trait associations have reliably been assessed, the breeder is able to monitor the transmission of trait genes via closely linked markers, thus enabling 'genotype building', that is the construction of desired genotypes by deliberate crossing and selection, using the marker genotype as a selection criterion.

This paper focuses on some quantitative aspects of the most widely applied form of marker-assisted breeding, *i.e.* the transfer of single or multiple genes from a (or several) donor genotypes into elite breeding material by repeated backcrossing, often referred to as *marker-guided introgression*. Although most of the ideas about the optimal design of introgression programs discussed below are not new, their relevance to applied plant breeding justifies an attempt to further dissemination into the plant breeding community.

## Marker-assisted introgression

When transferring a single gene from a donor into the genetic background of a recurrent parent by repeated backcrossing, genetic linkage will cause fragments of the donor genome surrounding the target gene to be 'dragged' along. Small donor genome fragments, not linked to the target gene, may also end up in the recipient's genetic background. Even after five or six BC generations the donor genome segment surrounding the target gene may be of considerable size (Stam and Zeven, 1981). Selection based on markers that distinguish between donor and recurrent parent genome may considerably accelerate the recovery of recurrent parent genome. Fig. 1 shows an example of the decrease of donor genome content (DGC) in heterozygous condition in successive BC generations with and without background marker selection. It demonstrates that with background selection approximately the same level of DGC reduction can be reached in half as many generations.

Figure 1. Decrease of DGC in heterozygous condition with and without background marker selection in successive BC generations. Results are based on 5000 replicate simulation runs.



The basic principle of background selection (as opposed to 'foreground selection' on the target gene) is that in any given BC generation the actual DGC varies around the theoretical mean value. In BC1 generation, for example, the expected DGC is 0.50, but in individual BC1 plants it may vary from 0.25to 0.75. Being aware of this, as well as of the influence of population size, genome size (total map length), a number of questions can be raised that are relevant to the design of an introgression-breeding program. Among these are the following.

1. What amount of variation in DGC is to be expected in generations BC1, BC2, etc?
2. To what extent does this depend on the number of chromosomes and their genetic map length?
3. What population size is required to guarantee, with 90% certainty that a least one individual in a BC1 has a DGC of less than *e.g.* 0.30?
4. If markers are flanking the target gene, what are the optimal population sizes in successive generations to ensure that the donor segment dragged along with the target gene is smaller than the segment bracketed by these flanking markers?
5. Does it pay to increase the number of markers for background selection? If so, to what extent does this depend on population sizes used and/or genome size?

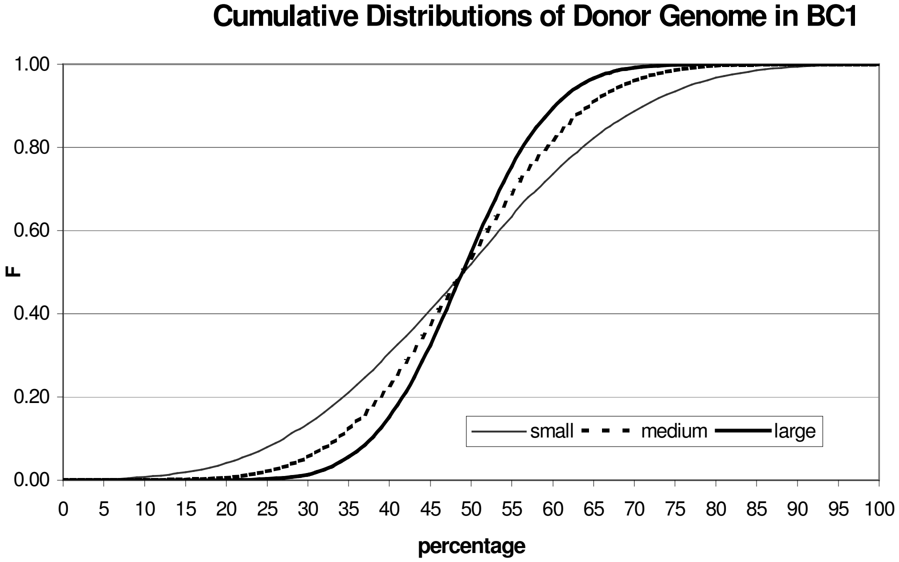
6. If a certain pre-set goal, *e.g.* less than 0.05 DGC, is to be achieved in a given number of generations, should population size in successive generations be constant or is it better to vary population size over generations?
7. If the number of generations is not a limiting factor, but the total number of plants to be genotyped is, what is then the optimal distribution of plant numbers over generations?
8. Do the same guidelines for optimal transfer of a single target gene also hold for the transfer of multiple genes?

Several of the above issues have been addressed by a number of authors, using an analytical approach, numerical methods, computer simulations or a combination thereof (Hospital, 2002; Hospital and Charcosset, 1997; Hospital *et al*., 1992; Frisch *et al*. 1997; Van Berloo *et al*., 2001). This paper focuses on the questions listed above, partly based on results of the cited papers, partly on new simulation results.

***Total genome map length***

The influence of total genomic map length on the distribution of DGC in BC generations has received little attention. Yet, it has important consequences for the attainable rate of donor genome substitution. The distribution of DGC in a BC1 generation is shown in Fig. 2 for a genome of small, medium and large size. The important feature to be observed here is that the variance in DGC decreases as genome size increases. In the remainder of this paper the following model genomes will be used (haploid number of chromosomes x map length (centimorgans): Small: 5 x 100 cM; Medium: 10 x 100 cM; Large: 15 x 150 cM.

Figure 2. The cumulative distribution (F) of donor genome content (DGC) in a BC1 generation for a small, medium and large genome. Results based on 50,000 replicate simulation runs.



From the tabulated cumulative distribution of Fig. 2 the probability of less than a given DGC can be read. For example, the probability that DGC is less than 0.35 equals 0.21, 0.12 and 0.06 for the small, medium and large genome respectively. From these probabilities one can calculate the population size required to ensure that with *e.g.* 90% certainty at least one plant will occur with less than a given DGC. Let the threshold DGC be $x$ and let the corresponding probability be $p_x$. Then the required minimum population size $N$ satisfies the equation

$$1-(1-p_x)^N > P_c,$$

where $P_c$ is the pre-set level of certainty.

Table 1 gives, for the three genome sizes, the population size required to find with 90% probability at least one or at least two plants with less than a given DGC in a BC1 generation.

Table 1. Population sizes required in a BC1 to obtain (probability 0.90) at least one or at least two plants with less than a certain donor genome content (DGC) with a small, medium or large genome.

| | at least one | | | at least two | | |
|---|---|---|---|---|---|---|
| DGC | small | medium | large | small | medium | large |
| <0.45 | 5 | 5 | 6 | 8 | 9 | 11 |
| <0.40 | 7 | 9 | 14 | 14 | 16 | 25 |
| <0.35 | 10 | 18 | 40 | 17 | 31 | 68 |
| <0.30 | 16 | 41 | 169 | 28 | 69 | 285 |
| <0.25 | 28 | 111 | 822 | 48 | 187 | >1000 |

Table 1 tells the importance of genome size. For example, for DGC to be less than 0.30 in at least one plant, a large genome requires approximately a tenfold larger population size as compared to a small genome. We infer from these simple calculations that the price to be paid for a rapid decline of DGC in a large genome is twofold:

1. A large genome requires more markers (more marker data points per plant), and
2. A large genome requires larger population sizes to attain a given rate of donor genome substitution.

***Background selection at carrier chromosome***

Donor genome substitution is, of course, most important, and at the same time most difficult, for chromosomes that carry the target gene(s) to be introgressed. In view of the need to remove unwanted linked sequences, the following question, addressed by Hospital (2001) and Frisch *et al.* (1999a) is relevant. Suppose that the target gene is flanked by two markers at map distances *d1* and *d2* which can be used for background selection. Within a given number of generations the introgressed segment must be smaller than the segment covered by *d1* – *d2*. Then, given a pre-set probability of reaching this goal (with a 99% success rate), what are the optimal population sizes in successive BC generations?

The answer to this question has been given by Hospital and can readily be obtained with the software package POPMIN (Hospital and Decoux, 2001, http://moulon.inra.fr/~fred/programs). An example of POPMIN results, illustrating several features is given in Table 2.

The main features illustrated by Table 2 are the following.

1. The smaller the segment bracketed by the markers, the more plants are required for analysis. This is because rare recombinants are less likely to occur in smaller populations.
2. Population size should increase as generations proceed. This is because two-sided detachment is in most cases a two-stage process. If no detachment occurs at either side in the first generation, more plants are required in the next generation(s).
3. Allowing more generations to achieve the goal requires, not unexpectedly, fewer plants to be grown and genotyped. There is a trade-off between speed and total size of the introgression program.

An additional feature of the POPMIN software is that it allows the user to specify the initial genotype at both markers and the target locus. So if a single individual *D1/d1 – T/t - d2/d2* (recombination at one side) in BC1 has been obtained, the user can optimise population sizes in BC2, BC3 etc., given this initial condition. This then enables further optimisation of the program. Conversely, if not a single recombinant has been obtained in BC1, an increase of the originally planned population sizes in generations thereafter is needed.

Table 2. Optimum population sizes required in successive BC generations, to achieve with 99% certainty (risk = 0.01), that two markers flanking the target gene become detached in at least one plant of the last BC generation. Σ*N*: accumulated number of plants. (Σ*N*): average accumulated number of plants; this is less than Σ*N* because with a certain probability the goal may be reached before the final generation. Figures indicated in 'configuration' column are distances in centimorgans. *T*: target locus; *d1, d2*: flanking markers.

| configuration | generations | BC1 | BC2 | BC3 | Σ*N* | (Σ*N*) |
|---|---|---|---|---|---|---|
| *d1*-10-*T*-10-*d2* | 2 | 62 | 100 | - | 162 | (137) |
| | 3 | 25 | 36 | 76 | 137 | (74) |
| *d1*-5-*T*-5-*d2* | 2 | 118 | 200 | - | 318 | (289) |
| | 3 | 48 | 70 | 149 | 267 | (149) |

***Total background selection***

The question arises over the number of markers (per chromosome) that should be used for background selection, and how this depends on genome and/or population sizes used. Several authors (see *e.g.* Hospital and Charcosset, 1997; Frisch *et al.*, 1999) have shown that in a moderately sized population from which the 'most promising' plant is selected for further backcrossing, an increase in the number of markers per chromosome beyond two is hardly rewarding (see Table 3).

An increase from 1 to 8 markers reduces DGC in relative sense (from 0.13 to 0.07 in BC2), but the absolute effect is limited. However, when rapid progress requires using larger population sizes, especially in case of a large genome (when larger population sizes are required anyway), the situation is different (Table 4).

Table 3. Average decrease of DGC in a backcross program with a medium genome size and single target gene. Each chromosome has 1, 2, or 8 markers, uniformly distributed over the chromosome. One plant out of 50 is selected in each generation for backcrossing. The selected plant satisfies the following conditions: (1) it carries the target allele and (2) it has the smallest number of markers of donor signature. Results based on 5000 replicate simulation runs.

| no. of markers | BC1 | BC2 | BC3 |
|---|---|---|---|
| 1 | 0.34 | 0.13 | 0.07 |
| 2 | 0.31 | 0.09 | 0.04 |
| 8 | 0.30 | 0.07 | 0.02 |

Table 4. Average DGC attained in BC2 for small and large genome sizes with various population sizes and number of markers per chromosome.

| no. of markers | pop. size | small | large |
|---|---|---|---|
| 2 | 50 | 0.082 | 0.121 |
|  | 200 | 0.079 | 0.095 |
|  | 400 | 0.078 | 0.088 |
| 8 | 50 | 0.040 | 0.100 |
|  | 200 | 0.021 | 0.067 |
|  | 400 | 0.019 | 0.055 |

The general conclusions to be drawn from the above are the following.

1. For a small genome with few markers per chromosome, increasing the population size makes little sense. When many markers are available, however, an increase of population size does reduce DGC, but hardly so beyond *N*=200.
2. For a large genome, increasing the population size is beneficial, irrespective of the number of markers per chromosome.

Qualitatively, these results are obvious: with increasing genome size more independent recombination events are required to attain a given reduction in DGC, which in turn demand larger populations for their discovery.

***Multiple gene transfer***

Several authors have considered optimisation aspects of multiple gene transfer by repeated backcrossing (Van Berloo *et al.*, 2001; Frisch and Melchinger, 2001; Hospital, 2002). It is clear that, roughly speaking, the effects of population size, genome size and total number of markers on the efficiency of recurrent parent genome recovery are similar to those for single gene transfer. As an example Table 5 shows the effect of population size for the introgession of three target genes.

Table 5. Average DGC in BC2 and BC3 in an example of the simultaneous introgression of three target genes in a genome of medium size, using eight markers per chromosome for background selection. A single plant was selected for further backcrossing, carrying the three target alleles and having the smallest number of markers of donor signature. Averages based on 1000 replicate simulation runs.

| pop. size | BC2 | BC3 |
|---|---|---|
| 50 | 0.18 | 0.09 |
| 100 | 0.14 | 0.06 |
| 200 | 0.11 | 0.04 |
| 400 | 0.09 | 0.03 |

With multiple targets an increase of population size enhances the efficiency (Table 5). A comparison with Table 4 demonstrates that for a given reduction of DGC, multiple targeting requires a larger population size.

Also relevant is the question whether a given total number of plants should be distributed over two or three generations. The number of target genes does affect the answer. Table 6 shows that three BC generations of 300 plants is more effective than two BC generations of 450 plants.

Table 6. Comparison of average DGC attained with a total of 900 plants, distributed over two or three BC generations with medium genome size and eight selection markers per chromosome.

| number of generations | pop. size | 1 target gene | 2 target genes | 3 target genes |
|---|---|---|---|---|
| 2 | 450 | 0.023 | 0.048 | 0.083 |
| 3 | 300 | 0.010 | 0.019 | 0.036 |

## Discussion

The results presented here focus on some of the main issues in marker-guided introgression, *i.e.* population sizes and number of selection markers in relation to chromosome number and genomic map length. Genome size has so far received little attention in the subject literature; the above results, however, indicate that it is an important factor in planning the total size and duration of a successful introgression program. Larger genomes require larger populations as well as more markers to be scored.

Various authors have discussed refinements and sophistications of the rough foreground and background selection regimes applied in the simulations of this paper. For example, Hospital (2002) considered background selection on carrier chromosomes to be more important than on non-carriers and thus assigned different weights to carrier and non-carrier markers. Based on the same idea, Frisch and Melchinger (2001b) considered 'multi-stage' selection of markers: after selection on the target gene(s), one selects plants based on carrier markers and finally, from the obtained subset, one selects on the basis of non-carrier markers. These authors also considered several other variants to this multi-stage selection.

A complication arising with multiple QTL transfer is the uncertainty about the exact location of QTLs. Hospital and Charcosset (1997) investigated the optimal location of markers to be used in foreground selection.

Van Berloo *et al*. (2001) studied the minimum population sizes required to attain a given reduction of DGC in a particular BC generation, for the specific genomic configuration of barley and its marker linkage map.

From a practical point of view the following type of question is relevant. Given a total genetic map length and (haploid) chromosome number, and given the foreground and background selection strategy, what  population sizes are required to guarantee with 90% certainty that at least one plant with DGC less that 0.05 can be recovered in BC3? Unfortunately, this type of question cannot be answered by analytical means. At this moment the only way out is to simulate thae backcross program according to the envisaged strategies to assess their efficiency. The guidelines presented here will then be helpful for the design of an optimal strategy.

The question about what final level of DGC is 'acceptable' cannot easily be answered in general terms. When only relying on estimated DGC, based on markers, one still runs risk that after finalization a tiny donor fragment contains a few 'wild type' genes that confer an undesirable trait. Especially in a rapid cycling introgression program that hardly allows phenotypic selection for general agronomic performance, undesired donor traits may unexpectedly turn up despite an expensive and theoretically powerful backcross scheme.

Therefore, it is best not to fully rely on DNA fingerprints but, if possible, apply phenotypic selection as well in introgression programs.

## References

Berloo van, R., Aalbers, H., Werkman, A. and Niks, R.E. 2001. Resistance QTL confirmed through development of QTL-NILs for barley leaf rust resistance. Molecular Breeding 8:187-195.

Frisch, M. and Melchinger, A.E. 2001a. The length of the intact donor chromosome segment around a target gene in marker-assisted backcrossing. Genetics 157:1343-1356.

Frisch, M. and Melchinger, A.E. 2001b. Marker-assisted backcrossing for simultaneous introgression of two genes. Crop Science 41:1716-1725.

Frisch, M., Bohn, M. and Melchinger, A.E. 1999a. Comparison of selection strategies for marker-assisted backcrossing of a gene. Crop Science 39:1295-1301.

Frisch, M., Bohn, M. and Melchinger, A.E. 1999b. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. Crop Science 39:967-975.

Hospital, F. 2001. Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. Genetics 158:1363-1379.

Hospital, F. 2002. Marker-assisted backcross breeding: a case study in genotype building theory. In: (M.S. Kang, ed.) Quantitative genetics, genomics and plant breeding. CAB International, New York and Oxon.

Hospital, F. and Charcosset, M. 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., Goldringer, I. and Openshaw, S.J. 2000. Efficient marker-based recurrent selection for multiple quantitative trait loci. Genetical Research 75:357-368.

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.