

Marker-assisted breeding

Frédéric Hospital

INRA, Station de Génétique Végétale

Ferme du Moulon

91190 Gif sur Yvette, France.

1 Introduction

In contrast to past decades, when almost no markers were available and breeding was mostly based on selection on phenotype, advances in molecular genetics have been able to partially dissect the black box of quantitative traits. The use of molecular genetics rests on the ability to determine the genotype of individuals using DNA analysis (Mohan *et al.*, 1997 ; Westman and Kresovich, 1997 ; de Vienne, 2002), and results in two types of information (molecular information): identified loci are (rarely) causal mutations or (more frequently) presumed non-functional genetic markers (indirect markers). Markers can be used simply to assess the parental origin of anonymous genomic region, or to unravel the genetic architecture of the quantitative trait(s) of interest, based on evidence of empirical associations of marker genotypes with trait phenotype (QTL detection, see the chapter by Kearsy and Luo in this book). These associations can then be used for selection. Note here that whereas a causative polymorphisms give direct information about genotype for the QTL, the use of indirect markers for QTL mapping and for selection is based on existence of linkage disequilibrium (Hartl and Clark, 1989) between the marker and the QTL. Marker-assisted selection can also aim at the introduction of transgenes (see the chapters on genetic engineering in this book) into breeding populations.

Molecular information can be used in several way to make the plant or animal breeding process more efficient in so-called Marker-Assisted Selection (MAS) schemes (Dekkers and Hospital, 2002). Here, I will review some of the most promising techniques involving selection on molecular markers in order to increase the speed and/or the efficiency of plant breeding programs, *i.e.*, increase genetic gain per unit of time. Other uses of molecular information not reviewed here, though possibly

important, include parentage verification or identification, identification and characterization of genetic resources, quantification of genetic diversity, etc.

Selection decisions in breeding programs can be based on phenotypic information alone (conventional selection), molecular information alone, or a combination or both. Breeding strategies involving selection based on molecular information alone are termed ‘Genotype Building’ (GB) strategies here, because the selection phase can be reduced to a simple ‘building blocks’ problem. Based on phenotypic and/or molecular information available prior to the start of the selection program, the breeder defines the ideal genotype (ideotype) at a collection of loci (target loci), as the one that meets the selection objective. The parents originally hosting the different target genes are crossed. Then, selection consists in screening, among the different genotypes produced by recombination in one or more generations, the one(s) that is closest to the ideotype, or that permits to obtain the ideotype most rapidly, simply based on DNA analysis (marker genotypes). Finally at the end of the MAS program, phenotypic evaluation is performed in order to evaluate the agronomic value of the resulting progenies.

One case of GB widely used in plant breeding is marker-assisted introgression in backcross programs, and this case is reviewed in details below. Other possible GB schemes (marker-based population screening, recurrent selection or gene pyramiding) are reviewed more rapidly. Obviously, phenotypic information may also be used during the selection phase in addition to, or in combination with, molecular information. The corresponding techniques are also reviewed.

Compared to breeding schemes involving phenotypic evaluation at each selection step, GB schemes where selection is based solely on molecular information capitalize on a potentially important saving of time and/or experimental means, particularly in cases where phenotypic information is longer, more difficult, and/or more expensive to score than molecular information (e.g., testcross breeding for yield in maize involving progeny testing, malting quality in barley, breeding for diseases resistance in most crops). However, such a breeding strategy assumes that the effects associated to markers are sufficiently well estimated, and sustainable across agronomic conditions and genetic background, so that the realized genetic gain will meet the expectations. The risk inherent to this assumption is finally discussed in the light of some recently published results of GB experiments in plants.

2 Marker-assisted backcrossing of a single target gene

Backcross breeding 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 target locus (chromosomal location of the target gene) is kept heterozygous (donor/recipient) by selection for the donor type allele at each generation. On chromosomal locations outside of the target locus, the objective is to increase the *recipient genome content* (RGC, often expressed as the percentage of recipient alleles) of the progenies. Increasing RGC is particularly important, but difficult, on the chromosome carrying the target locus (*carrier chromosome*). Because at each generation there is selection for the donor allele at the target locus on the carrier chromosome, unwanted donor genes located in the vicinity of the target locus may be dragged along with the target gene (*linkage drag*). Hence, RGC is expected to be lower on the carrier chromosome than on *non-carrier chromosomes*.

In conventional breeding schemes, presence of the target gene is assessed phenotypically, provided this gene has a visible or measurable effect, and increase of RGC is ensured by repeatedly back-crossing progenies carrying the target gene to the recipient line. Given that donor content on non-carrier chromosomes is expected to be reduced by one half after each backcross, it is generally assumed that at least six or seven backcross generations are necessary to ensure a sufficient genomic similarity between the backcross line and the recipient line (except the target locus). In practice, the number of backcrosses performed is sometimes even more important (10 or more). Using molecular information may improve the efficiency of backcross breeding schemes in several non exclusive ways, through selection on molecular markers (marker-assisted backcrossing): i) to control the target gene (*foreground selection*) and/or ii) to control the genetic background (*background selection*). In all cases, marker assistance is expected to provide higher efficiency, reduced cost, and/or shorter duration of the backcross breeding scheme, compared to conventional methods. In addition, markers can also be used to estimate RGC in the backcross progenies. Such uses of markers are detailed below in the case of a backcross program involving one single target gene. Additional considerations related to the manipulation of more than one gene are addressed in the following section, along with other marker-assisted breeding strategies for several genes.

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, or quantitative trait loci (see other chapters). From a theoretical standpoint, it is a ‘simple’ example of marker-based selection: in general, only two alleles are segregating, and the gametic phase (parental origin of the alleles on a chromosome) 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 program.

2.1 Foreground selection

Here, we address the use of molecular markers to assess the presence of the target gene in backcross progenies.

2.1.1 Target locus is a known locus

Molecular data can be obtained at early stage as soon as DNA can be obtained (*e.g.* from leaf sample), heritability at the molecular marker level is one, and most often it is possible to find markers for which the dominance relationship is favorable. Conversely, phenotypic assay is often longer and/or more difficult and/or more costly than molecular genotyping. Classical examples are cases when phenotypic assay involves progeny testing, because the phenotype is expressed posterior to reproduction (*e.g.* grain yield, any testcross performance, malting quality in barley), because phenotypic assay is destructive, or because the target gene is recessive. In such circumstances, even when the target is a major gene (of known phenotypic effect and chromosomal location), control of the target with molecular markers may be more profitable than phenotypic assay.

When the presence of the target gene is not controlled directly through its phenotypic expression, but indirectly through the observed genotype at one or several marker(s), it is important that observed genotype at the marker provides a good control over the (true but unknown) genotype at the target. This may be assessed for example by computing the Target Control Rate (TCR) defined for any given individual as the probability that the individual is heterozygous donor/recipient at the target given that it is heterozygous donor/recipient at the marker. If we denote T^D (resp. T^R) the donor

(resp. recipient) allele at the target and M^D (resp. M^R) the donor (resp. recipient) allele at the marker, we have:

$$\begin{aligned}\text{Target Control Rate} &= \text{TCR}(\%) \\ &= \Pr \{ (T^D/T^R) \mid (M^D/M^R) \} \times 100 \\ &= \Pr \{ (T^D M^D / T^R M^R) \} / \Pr \{ (M^D/M^R) \} \times 100\end{aligned}$$

where $\Pr\{(X/Y)\}$ denotes the probability of being of genotype X/Y , and \mid denotes condition. In other words, the risk that an individual, which displays the desired genotype at the marker(s), does *not* have the desired genotype at the target is $(100 - \text{TCR})$.

In some favorable cases, it is possible to find a *direct marker* inside the target locus. Most classical example is the case when the target gene is a transgenic construct which complete DNA sequence is in general known. In such cases, recombination rate between the marker and the target locus is zero, the probability of transmission of the marker is $\frac{1}{2}$ at each backcross (BC) generation, as is the probability of transmission of the target gene, and the genotype at the target locus equals that at the marker, so $\text{TCR}=100\%$. However, such cases are rare. In general, the target has to be controlled by *indirect markers* located outside of the target locus, so that recombination rate between the target and the marker(s) is not zero.

If the target is controlled by one single marker M_1 such that the recombination rate between the target and the marker is r_1 , then after n BC generations:

$$\text{TCR}(M_1) = \{ (1/2)^n (1-r_1)^n \} / \{ (1/2)^n \} = (1-r_1)^n$$

If the target is controlled by two markers M_1 and M_2 (one on each 'side' of target locus T) such that the recombination rate between T and M_1 is r_1 , the recombination rate between T and M_2 is r_2 , and the recombination rate between M_1 and M_2 is r , then after n BC generations:

$$\text{TCR}(M_1 M_2) = \{ (1/2)^n (1-r_1)^n (1-r_2)^n \} / \{ (1/2)^n (1-r)^n \} = (1-r_1)^n (1-r_2)^n / (1-r)^n$$

--- TABLE X.1 around here ---

Tabulated numerical values for $\text{TCR}(M_1)$ or $\text{TCR}(M_1 M_2)$ are given in Table x.1 for a range of marker positions, where marker position is expressed as the distance between the target and the marker in Haldane centimorgans (cM) assuming no

interference in recombination. It is clearly seen from the table that control of the target by a single marker is not satisfactory in most cases. The marker must be as close as 1 cM to the target to keep the risk of ‘losing’ the target below 5% over five BC generations. Even with a single marker at 1 cM, the risk of losing the target is close to 10% in BC₁₀. For greater distances of a single marker, the risk becomes rapidly too high. Conversely, the table shows that control of the target by two marker (*marker bracket*) is much more satisfactory, even for larger marker distances. A bracket with each marker as far as 10 cM from the target provides approximately the same control as a single marker at 1 cM, and the control provided by any bracket closer than 10 cM is quite satisfactory. Control provided by brackets further than 10 cM is also acceptable, but only for a few BC generations. Obviously, this is because when the target is controlled by two markers a (rare) double recombination is necessary to break the bracket-target association, while a (more frequent) single recombination is sufficient to break a single marker-target association.

2.1.2 Target locus is a quantitative trait locus (QTL)

A quantitative trait locus (QTL) is a locus, or a chromosomal segment involved in the variability of a quantitative trait, which is detected by appropriate statistical methods putting in relation molecular and phenotypic information (see the corresponding chapter). By definition such target can only be controlled by indirect markers (unless it is characterized at the molecular level, which is rarely the case and necessitates a lot more additional work). Control of a target QTL in foreground selection poses additional problems, because the exact location of the target is often not known, but rather estimated with a given precision. Hence, number and chromosomal positions of the markers devoted to target control must take account of the uncertainty of the true target location. Extending a first result by Visscher *et al.* (1996), Hospital and Charcosset (1997) discussed the optimal number and positions of markers to control a QTL in the foreground selection step of a BC program. Calculations are based on the following rationale: it is assumed that QTL detection has been already performed, and has provided the expected (most likely) position of the QTL, along with a confidence interval on that position. Then, uncertainty in the target position is modeled by assuming that the true target is located somewhere around the expected QTL position, with a distribution following a normal law of mean 0 and of variance derived from the length of the confidence interval. For a given number of markers, Hospital and

Charcosset (1997) then computed the optimal positions of the markers as the ones that maximize the Target Control Ratio. For each possible position of the true target, TCR is computed as we did above for a known target (section 2.1.1), then TCR is averaged over all possible positions. Some relevant numerical results are given in Table x.2.

--- TABLE X.2 around here ---

It is seen that target control provided by optimally placed markers is very good in BC₁ and BC₃. In general, three markers optimally placed are sufficient to ensure a TCR above 99%, even for large confidence intervals (60 cM). Even fewer markers can be used to control smaller confidence intervals. However, Hospital and Charcosset (1997) showed that control by few markers (three or less) is quite sensitive to marker positions, so that in the general case where markers cannot be placed optimally, using at least three markers per QTL is recommended, except when precision on QTL location is very high.

2.1.3 Minimal population sizes

When the target is controlled indirectly via molecular markers, it is never possible to know whether an individual that carries the desired genotype at foreground selection markers (*i.e.* an individual that is heterozygous at all markers devoted to target control) also really carries the desired genotype at the target locus (*i.e.* is heterozygous at that locus) until a phenotypic assay is performed. Hence, a strategy could be to 1) choose markers number and positions based on above calculations so that these markers provide a high as possible Target Control Ratio, then 2) compute minimal populations sizes such that the risk of not obtaining at least one individual with the desired genotype at the markers is below a given threshold, say α_M . With this strategy, the risk of not obtaining an individual with the desired genotype at the target is simply $(100 - \text{TCR}) \times \alpha_M$, supposedly low enough. Computing minimal population sizes in this context is quite simple.

At each BC generation the probability that an individual has the desired marker genotype given that his backcrossed parent had the desired genotype is:

$$P_M = \frac{1}{2} (1 - r_1) \times (1 - r_2) \times \dots \times (1 - r_{m-1})$$

where m is the total number of markers and r_k is the recombination rate between the k^{th} and $(k+1)^{\text{th}}$ markers. The minimum number N_{\min} of individuals that should be

genotyped at each BC generation to obtain at least one individual with the desired genotype at the markers with risk α_M is obtained by solving the equation $(1-P_M)^N = \alpha_M$, so that:

$$N_{\min} = \ln(\alpha_M) / \ln(1-P_M)$$

where \ln denotes the naperian logarithm. Numerical values for N_{\min} are given in the last columns of Tables x.1 and x.2 for the corresponding marker positions. Minimal population sizes are quite low (between 7 and 13 individuals at each BC generation) in all cases, even for QTL located with poor precision. However one should keep in mind that: 1) This ensure that the genotype at *flanking markers* is obtained, and one should check that the corresponding TCR allows a sufficient control of the target. 2) This ensures that at least *one* individual with desired genotype is obtained ; more complex calculations to obtain several carriers of the target gene were derived by Melchinger (1990). 3) This is for the control of a single target (either known gene or QTL). For BC breeding programs aiming at introgression of several targets, minimal population sizes rapidly increase with the number of targets, and turn out to be one of the most limiting parameter. This is addressed later in this chapter (section 3).

2.2 Background selection

Whether the target is controlled directly through its phenotypic expression, or indirectly by markers (*foreground selection*, see previous section) molecular markers can always be used for *background selection*. The aim of background selection is to fasten the recovery of recipient parent genotype outside the target(s) (*genetic background*). Usually, foreground and background selection are performed in two distinct and successive steps at each BC generation, because it is not necessary to genotype the background of individuals that do not carry the target (unless such global genotyping is less expensive because of the particular molecular technique used).

In any case, we assume here that selection is in two steps, and that step 1 (foreground selection) can be achieved successfully, either using phenotype or markers, so that Target Control Ratio (see previous section) is close to 100% and step 2 (background selection) is amongst individuals that are all heterozygous for the target.

Even when no background selection is performed (random selection), the percentage of alleles inherited from the recipient parent (RGC) is expected to increase

in BC progenies, due to successive backcrosses. Hence, background selection is considered efficient only if it permits a return to recipient genome *faster* than the normal return rate when no selection on markers is applied. Then, the efficiency of marker-assisted selection should always be compared to this normal rate as a reference. We need first to recall what is the normal rate of return expected with no background selection.

2.2.1 Expected genome contents with no selection

On *non-carrier* chromosomes (chromosomes not hosting the target), the probability that any locus remains heterozygous donor/recipient after n backcrosses is $(\frac{1}{2})^n$. Hence, starting from 50% in the original F_1 , the expected RGC(%) on non-carrier chromosomes at generation BC_n with no background selection is $100 \times \{1 - (1/2)^{(n+1)}\}$. Based on this consideration, it is generally considered that at least six BC generations are necessary to insure a similarity with the recipient parent above 99%. However, it is important to note that on a *carrier chromosome* (chromosome hosting the target), the rate of return to recipient type is much slower than that.

On the carrier chromosome, selection of individuals that are heterozygous at the target locus is mandatory. Hence, because of this foreground selection, any locus that is linked to the target locus on the carrier chromosome is more likely to be heterozygous than a locus on a non-carrier chromosome. This *linkage drag* makes the RGC on the carrier chromosome with no background selection always lower than the RGC on non-carrier chromosome. There are two possible measures of linkage drag:

One can compute the *total* proportion of donor alleles on the carrier chromosome. This was provided by Stam and Zeven (1981) based on the following calculation. Given that the target locus T is heterozygous, any locus X on the carrier chromosome is heterozygous with probability $(1-r)$, where r is the recombination rate between T and X . Then, the expected proportion of loci that are heterozygous is obtained by integrating along the chromosome, assuming no interference in recombination, *i.e.* the relation between map distance d (in centimorgans, cM) and corresponding recombination rate $r[d]$ is obtained from Haldane's formula: $r[d] = \frac{1}{2} (1 - \text{Exp}[-2d/100])$, where $\text{Exp}[]$ denotes exponential.

Also, one can focus on the *intact donor segment*, *i.e.* 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. The computation similarly involves integration along the chromosome, but here the relevant probability is that of absence of crossover between T and X, not absence of recombination as in the previous case. Recall that the probability of absence of recombination between two loci is the probability of an even number (0, 2, 4, ...) of crossovers between the two loci, not only zero crossover. Obviously, the *total* proportion of donor alleles on the carrier chromosome computed by Stam and Zeven (1981) comprises the length of the intact donor segment plus other blocks of donor alleles elsewhere on the carrier chromosome.

--- Figure x.1 around here ---

The quantity of donor genes (in cM) on the different parts of the genome is shown in Figure x.1 for a genome composed of one carrier chromosome and 9 non-carrier chromosomes, each chromosome of length 200 cM. Because there are 9 non-carrier chromosomes, most of the unwanted donor genes are located on non-carrier chromosomes in early BC generations, but these genes are rapidly removed as noted above. Conversely, the quantity of donor genes on the carrier chromosome decreases much slower, so that after generation BC₆ most of the unwanted donor genes still segregating are located on the carrier chromosome. Donor fragments on the carrier chromosome represent a total of 40 cM in BC₆, 23 cM in BC₁₀, and still more than 10 cM in BC₂₀. Although these fragments represent only 1% of the total genome length (4000 cM) after generation BC₆, these may still host numerous unwanted genes, in particular if the donor is a wild genetic resource. Moreover, the variance around these expected values is important (Stam and Zeven, 1981 ; Naveira and Barbadilla, 1992). Also, it is seen from Figure x.1 that the difference between the two measures of linkage drag on the carrier chromosome is small. Hence, the vast majority of unwanted donor genes on the carrier chromosome are located on the intact donor segment surrounding the target. An impressive experimental proof of this was provided by Young and Tanksley (1989a) who genotyped *a posteriori* with RFLPs a collection of tomato varieties previously introgressed for the resistance gene at the *Tm-2* locus. The size of chromosomal segments retained around the *Tm-2* locus during backcross breeding was

very variable and sometimes quite long: one line exhibited a donor segment of 50 cM after 11 backcrosses, another 36 cM after 21 backcrosses, etc.

2.2.2 Marker-based estimate of recipient genome content

In a particular experiment, one can use molecular markers simply to estimate the Recipient Genome Contents (RGC) of backcross progenies. The most basic estimate is provided by scoring the genotype at a collection of markers over the genome, and then estimating RGC from the ratio of the number of markers homozygous for the recipient allele over the total number of markers scored. This simplest estimate does not take account of the positions of the markers. However, it is self evident that if markers are not evenly distributed along the genome (the real situation), weighting them equally is clearly not the best solution. Several solutions have been proposed to take account of markers locations. Visscher (1996) proposed to include molecular information in a BLUP-like estimate of RGC, and derived the optimal positions of markers in this context. Young and Tanksley (1989b) introduced the concept of *graphical genotypes* to ‘portray the parental origin and allelic composition throughout the genome’. For each chromosomal segment flanked by two markers, RGC is approximated based on the genotypes of the flanking markers: 100% if two markers of recipient type, 0% if two markers of donor type, and 50% if one marker of donor type and one marker of recipient type. This ignores the possible recombination events taking place between the two flanking markers. To take recombination into account, one can compute for any point of a chromosome the probability of being of recipient type, given marker genotypes, marker positions, and the breeding scheme (Servin *et al.*, 2002). In any case, the general conclusion is that few well-placed markers (two to four markers on a chromosome of 100 cM) provide adequate coverage of the genome in backcross programs (Visscher, 1996 ; Servin and Hospital, 2002)

2.2.3 Reduction of linkage drag (carrier chromosome)

As seen above (section 2.2.1), the carrier chromosome deserves special consideration in backcross programs because this chromosome returns to recipient type slower than non-carrier chromosomes, due to selection for the target gene in each generation (linkage drag). Here, I will focus only on the intact donor segment (see above) as a measure of linkage drag.

Basically, linkage drag can be reduced by performing background selection at two markers flanking the target, one on each side. Here, the objective is to select individuals that are heterozygous at the target locus, and homozygous for the recipient allele at both flanking markers (such individuals are termed *double recombinants* herein).

Hospital (2001) computed the mean and variance of the length of the intact donor segment around the target gene, for double recombinant individuals, in any BC generation. This gives the efficacy of background selection for the reduction of linkage drag. The numerical results indicate that the expected length of donor segment on each side of the target gene is approximately half the distance between the target and the flanking marker in BC_1 . The length in more advanced BC generations depends on the marker distance, but for short markers distances (20 cM from the target or less), the expected length of donor segment in advanced BC generation is not much below the length in BC_1 . For short markers distances, 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.

Obviously, selecting for flanking markers close to the target implies genotyping and screening large populations before a double recombinant genotype is obtained. In order to optimize genotyping effort (*i.e.*, the cost of the program) it is thus important to determine the minimal population sizes necessary to obtain the desired genotypes at the flanking markers. Intuitively, for close flanking markers, double recombinant genotypes are highly unlikely to be obtained in one single generation (BC_1) so that at least two BC generations should be performed, with selection for a single recombinant genotype on one side of the target in BC_1 , and a single recombinant on the other side in BC_2 (Young and Tanksley, 1989a). However, 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 (assuming that the genotype selected at generation $BC_{(n)}$ is known, population size at generation $BC_{(n+1)}$ is optimized to permit the selection of a double recombinant genotype at generation $BC_{(n+1)}$). However, Hospital (2001) showed that a better optimization is obtained when considering all the planned generations

simultaneously, because the optimal population size at each BC generation depends on the total duration of the breeding scheme.

Optimizing populations sizes over several successive generations requires some numerical calculations. A computer program (*popmin*) that performs these calculations easily was designed (Hospital and Decoux, 2002) and is freely available at <http://moulon.inra.fr/~fred/programs>. This program works as follows. The user enters a given value n for the total duration of the breeding scheme (maximal number of BC generations that could be performed) and a given risk α . Consider a marker-assisted backcross scheme involving n generations with populations sizes $N_1 \dots N_n$ at generations $BC_1 \dots BC_n$, respectively. The selection objective is here to obtain a double recombinant genotype at any generation BC_k ($k \leq n$) but, obviously, if a double recombinant is obtained at generation $k < n$, then the BC scheme is interrupted. The program computes the probabilities S_k that a double recombinant is obtained at any generation BC_k ($k \leq n$). Then, it determines optimal populations sizes $N_1 \dots N_n$ at generations $BC_1 \dots BC_n$ such that i) the risk that *no* double recombinant is obtained after n BC generations is α ($\sum_k S_k \geq 1 - \alpha$), and ii) the average number of individuals genotyped ($N^* = \sum_k N_k S_k$) is minimal.

One can run the *popmin* program to investigate any particular situations. However, the general conclusions that one should keep in mind are as follows. First, it is often preferable to genotype *more* individuals in advanced BC generations than in early BC generations (*e.g.*, for a BC scheme lasting two generations, genotype more individuals in BC_2 than in BC_1 , not the reverse). This reduces the average number of genotypings over the entire BC scheme. Second, planning to perform a total of more than two BC generations is in general recommended.

--- Table x.3 around here ---

This is exemplified in Table x.3 showing numerical results obtained with the *popmin* program, for flanking markers located at 2 cM from the target on each side, and BC schemes of different durations, with a risk $\alpha=1\%$. The minimum number of individuals that should be genotyped to obtain a double recombinant in BC_1 is about 24000, obviously far too many. The same result can be obtained over two generations (BC2 strategy) by genotyping 290 individuals in BC_1 , and 500 in BC_2 . Finally, over three generations (BC3 strategy), the optimal population sizes are 120 individuals in BC_1 , 170 in BC_2 , and 370 in BC_3 . In all three strategies, the probability to obtain a double recombinant for the flanking markers by the end of the breeding scheme is above

99%. In the BC3 strategy, the probability to obtain a double recombinant in BC₂ is about 75%. In case this happens, the BC scheme 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 program 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. Planning at the beginning of the program to perform more than two BC generations is then 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 actual genotype of the individual selected is known. This is also possible using the computer program *popmin* with a relevant option. Finally, the best optimization strategy is as follows: i) before starting a marker-assisted backcross program, investigate different scenarios to determine the maximal number of BC generations (n) that could be performed, given available genotyping means, and other possible economic considerations; ii) start the BC scheme with the optimal population size in BC₁ corresponding to the chosen scenario; and iii) refine the optimization at each following generation, once the genotype of the selected individual is known.

--- Figure x.2 around here ---

To synthesize the above results and help one make a decision in designing a marker-assisted backcross scheme in particular conditions, the *popmin* program was run for a range of parameters to provide the direct relationship between the efficiency of marker-assisted selection (expected linkage drag at the end of the breeding scheme) and the total number of individuals genotyped for flanking markers during the breeding scheme. The results are shown in Figure x.2 for breeding schemes involving a total of 1, 2, 3, 5, or 10 BC generations. This can be used as follows. One should define the upper limit for the length of linkage drag that is acceptable at the end of the experiment (*e.g.*

for introgression of a gene from a wild genetic material, or for fine-mapping of a QTL, and/or for the derivation of near-isogenic lines (NIL) or congenic lines for the identification and validation of quantitative trait loci, acceptable linkage drag should be much smaller than that for introgression between elite lines). Also, one should define the upper limit for the total number of individuals that can be genotyped during the experiment (*e.g.*, based on available molecular facilities and cost of the molecular technique used). Then, one can use Figure x.2 to determine how many BC generations should be performed in order to remain below the two defined limits. It is seen from the figure that BC schemes involving only one or two BC generations will rarely be affordable, unless large linkage drag can be accepted. Conversely, the small difference between the lines in Figure x.2 for 5 vs 10 BC generations indicates that performing more than 5 BC generations is rarely necessary, unless very low molecular cost is sought. Note however that both the x (abscissa) and y (ordinate) axis in the Figure are in logarithmic scales, so that small differences on the x axis may correspond to hundreds individuals for small linkage drag values.

The general conclusion is then that a typical marker-assisted backcross scheme should involve three to four BC generations in most cases, unless rapid success is sought for particular reasons. Planning to perform three or more BC generations has two main advantages: First, it permits a more drastic reduction of linkage drag while reducing the genotyping effort. Second, it increases the probability of success (obtaining a double recombinant) in advanced BC generations. The optimal population sizes above were 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 on non carrier chromosomes, is then possible among those double recombinants as described in the next section.

2.2.4 Selection on non-carrier chromosomes

Since the benchmark papers of Tanksley (1983) and Tanksley *et al.* (1989), numerous papers have addressed the use of markers to fasten the recovery of recipient genome on non-carrier chromosomes in BC breeding schemes (*e.g.*, Hillel *et al.*, 1990; Hospital *et al.*, 1992; Groen and Smith, 1995; Visscher *et al.*, 1996). This was also validated by experiments (see section 4). In such cases, the objective is to select individuals that are of homozygous recipient type at a collection of markers located on non-carrier chromosomes. Again, several markers are involved and it is unlikely that the

selection objective is fulfilled in one single generation (BC_1), so that selection on markers should be performed over two or more BC generations.

The general conclusions that can be drawn from these theoretical works summarize as follows. First of all, a dense coverage of the non-carrier chromosomes by molecular markers is not mandatory to increase the overall recipient genome content (unless fine-mapping of particular chromosomal regions is important). For a chromosome of 100 cM, two to four markers are sufficient. Obviously, selection on markers is most efficient if the markers are optimally positioned along the chromosomes. Such optimal positions were derived by Servin and Hospital (2002) and are recalled in Table x.4.

--- Table x.4 around here ---

However, a precise positioning of the markers on non-carrier chromosomes is again not mandatory (conversely to the case of the reduction of linkage drag on the carrier chromosome, see above). As can be seen from Table x.4 (d^*/d^{*+}) a variation of marker positions several centimorgans away from their optimal positions does not reduce much the efficacy of selection (RGC%), in particular when several markers per chromosome are used. In fact, what is important is to have at least two or three markers per chromosome, and that no chromosome is unmarked (zero marker). Given that this condition is fulfilled, the second conclusion is that selection on markers is quite efficient. In general, three or four generations of marker-assisted selection are sufficient to increase RGC on non-carrier chromosomes above 99%. Hence, the gain due to selection is of about two BC generations (RGC in BC_4 with selection is approximately the same as RGC in BC_6 with no selection on markers). This gain can be economically very valuable, for example with respect to the time necessary to release new products on the market.

Another important conclusion is that background selection is more efficient in late BC generations than in early BC generations. For example, if a BC breeding scheme is conducted over three successive BC generations, but it is wanted to genotype individuals for molecular markers at only one generation, then it is more efficient to genotype and select the individuals only in the BC_3 generation, rather than only in the BC_1 generation. This was demonstrated analytically in Hospital *et al.* (1992), and recently observed in simulations by Ribaut *et al.* (2002). This conclusion may seem

counter-intuitive, because recipient genome content is lower in BC₁, so there is ‘more to select’. However, this may be explained as follows. Suppose the series of recombination events that will take place during the breeding scheme were already drawn before the start of the program, but remained unknown to the breeder. In the BC₁ population, many chromosomal segments of donor origin are segregating. However, during the following backcross process, some of these segments will return to recipient type simply by chance. Hence, the experimental means devoted to the genotyping of these very segments is useless. Conversely, it is now clear that genotyping in the last generation the donor chromosome segments that were not previously removed by chance is more efficient. Hence, genotyping only the last generation could be a way to reduce the cost of the experiment. However, the efficiency of such a selection strategy will always remain below the efficiency of a strategy involving selection at every BC generation. In practice recombination events occur at random in an unpredictable manner, so that not all the genotyping efforts in early BC generations is useless, and the small gain provided can only increase the final efficiency of the breeding scheme.

2.2.5 Example of efficiency for a complete scheme

In order to synthesize the above conclusions, I present here the results of what could be a typical marker-assisted backcross breeding scheme. The results are given in Table x.5 and were obtained by simulation of the following strategy. At each BC generation, selection was in three steps. 1) Foreground selection: selection of all individuals that are heterozygous at the target locus (assumed controlled directly here, not by distant markers). 2) Reduction of linkage drag: selection of all individuals that are homozygous for the recipient allele at two markers flanking the target locus on each side (double recombinants) or, if no double recombinant is present in the population, selection of all individuals that are homozygous for the recipient allele at either of the two flanking markers (single recombinants). 3) Background selection on non-carrier chromosomes: selection of the one individual that is most homozygous for the recipient allele at markers on non-carrier chromosomes. Target-marker distance on the carrier chromosome was 2 cM. Each non-carrier chromosomes was controlled by three markers located at optimal positions given in Table x.4. We considered a breeding scheme involving four BC generations. Populations sizes at each generation were taken from optimal values in Table x.3.

--- Table x.5 around here ---

The results in Table x.5 confirm that marker-assisted backcrossing is expected to be quite efficient, providing a Recipient Genome Content of 99% in BC₄. Again, this represents a gain of two BC generations, because a RGC of 99% would be obtained only in BC₆ with no selection on markers. It is seen from Table x.5 that all selected markers have returned to fully homozygous recipient type in BC₄. Hence, selection on these markers would not be efficient in additional BC generations. RGC in BC₃ is already high (98%), but reduction of linkage drag is not complete at this stage, because the scheme of Table x.5 was optimized for a total of four BC generations. If maximal efficiency in only three generations were sought, then larger population sizes should be used (see Table x.3), providing a RGC in BC₃ of 98.5% (simulations not shown).

Note also that in practice the breeding strategy and the population sizes should be optimized in consideration with the particular molecular technique used for molecular assay (*e.g.*, Ribaut *et al.*, 1997)

3 Strategies for multiple target genes

In some cases, it is wanted to manipulate several genes of interest, either known genes or favorable alleles at quantitative trait loci (all termed target genes here) in a same breeding scheme, with or without controlling simultaneously the genetic background in which those target genes are introduced. Again, use of molecular markers can make such breeding schemes more efficient in various aspects. However, the underlying theories are still under development, and optimal strategies are not as well established as in the case of backcrossing for a single target. Hence, the general principles are reviewed here more briefly. The reader should refer to the cited references for more details.

3.1 Genotype building strategies for several target genes

Again, GB strategies here assume that an ideal genotype (ideotype) has been previously defined at a collection of loci. These loci may be known loci of major effects, or quantitative trait loci, but in any case it is assumed that gene effects are well estimated, and sustainable, so that the selection is only at the molecular level and simply consists in screening the products of meioses (recombination) taking place in successive generations in order to obtain the ideotype as fast as possible (*i.e.* accumulate the favorable alleles at all previously defined loci). Strategies where the selection criterion

is weighted as a function of the estimated effects of the genes considered are addressed in the next section.

3.1.1 Marker-based population screening (RIL, DH)

When several favorable genes are originally hosted by two different parents, the simplest strategy involves production of an F_2 , F_3 , or (if possible) Recombinant Inbred Lines (RIL) or Doubled-Haploid (DH) population. Then, screen the population based on molecular markers for individuals homozygous at the requested loci. In this context, van Berloo and Stam (1998) have considered a set of identified QTL, each controlled by two flanking markers, and studied selection in RIL populations based on flanking markers to produce the best hybrid. If all the genes cannot be fixed in a single step of selection, it is necessary to cross again selected individuals with incomplete, but complementary, sets of homozygous loci (Charmet *et al.*, 1999). However, such strategies are limited to small numbers of target loci, because the population size necessary to fix the target genes increases exponentially with the number of loci: for example in a RIL population, the frequency of homozygotes is $\frac{1}{2}$ for one gene at one locus, and $(\frac{1}{2})^k$ for k unlinked target loci, *i.e.* less than 1/1000 individuals for ten target loci.

3.1.2 Marker-based recurrent selection

For even more loci, recurrent selection should be used, *i.e.* a breeding scheme involving several generation of selection and random mating of the selected individuals. Hospital *et al.* (2000) studied selection on marker pairs flanking 50 QTL identified in an F_2 population. The best strategy seems to select at each generation a set of individuals that are complementary for their genotypes at flanking markers, such that each target is carried by at least two selected individuals. With this strategy, selection of 3 to 5 individuals among a total of 200 for 10 generations increases the frequency of favorable 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. In this case, the efficiency of marker-based selection is bounded by the recombination taking place between the markers and the QTL. Hence, one has to accelerate the response to selection to fix favorable 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 pair wise mating of individuals based on their marker genotypes might increase the efficiency of selection. But, the theory in this domain remains unexplored.

3.1.3 Marker-based gene pyramiding

When the target genes are originally hosted by multiple parents one can perform a marker-assisted gene pyramiding scheme, involving several initial crosses between the parents. For example, four genes (G1-G4), that are present in four different lines (L1-L4), can be combined into a single line in a two-steps procedure. In the first step, two lines that are homozygous for two target genes each (G1/G2 vs. G3/G4) are developed by crossing pairs of lines (L1×L2 vs. L3×L4), followed by selection of homozygotes among F₂, recombinant inbred line (RIL), or double haploid (DH) progeny. In the second step, such individuals are crossed to produce individuals that are homozygotes for all four target genes. Selection of homozygotes can be on the basis of linked markers. An example of experimental implementation of such strategy for the manipulation of QTL is given in section 4. This process can be expanded to more than four genes by expanding the pyramid. However, such basic scheme is limited to a small numbers of target loci. If target loci are numerous, and in particular if some loci are linked on the same chromosome, such schemes certainly deserve optimization, but again the theory in this domains remains largely unexplored.

3.1.4 Marker-assisted backcrossing for several target genes

In the above marker-based strategies for multiple targets, it was not wanted to control the genetic background in which the target genes were accumulated (*i.e.*, genomic regions outside the target loci). However, some traits that are improved through introgression by backcrossing might have an oligo- or poly-genic basis. This is for example the case for the accumulation of resistances, or for the introgression of complex traits, because in general several QTL of small or medium effects account for the variability of these traits. In such cases, it is important to control several targets *and* the genetic background in marker-assisted backcross breeding schemes.

Again, the number of individuals that must be genotyped increase exponentially with the number of target loci. Hospital and Charcosset (1997) computed such population sizes for different numbers of target QTL, and concluded that in general it is illusive to plan to manipulate more than three or four QTL simultaneously in a marker-assisted backcross program. If the targets are known loci controlled directly, not QTL controlled indirectly by markers, the maximum number of targets could be slightly higher, but should not exceed five or six.

However, when several targets are controlled simultaneously in one single backcross scheme, the number of individuals heterozygous at all targets is low in each BC generation, and this leaves little opportunity to select among those individuals for the genetic background (background selection). In such cases, introgression can be combined with gene pyramiding to decrease the number of individuals required (Hospital and Charcosset 1997, Koudandé et al. 2000). For example, if it is wanted to introgress four targets, one could first perform four parallel backcross schemes each introgressing one target in the genetic background, or two backcross schemes each introgressing two targets, then accumulate all targets in the same background by gene pyramiding (see above). This capitalizes on a higher efficiency of background selection in the separate backcross lines (provided at least two BC generations are performed), hence the final efficiency after pyramiding is higher than when the targets are controlled simultaneously in one single backcross scheme. However, the total duration of the program is then longer, because of the additional generations of pyramiding.

3.2 *Selection combining molecular and phenotypic information*

When the target genes do not account for all of the variability of the selected trait(s), as would be the case for many complex traits, the gain expected from the cumulated effects of the target genes might not be worth performing selection based solely on molecular information. In such case it could be wanted to control both the variability accounted for by the target genes (major genes or more frequently QTL with medium to low effects) and the ‘unmarked’ variability. For example, de Koning and Weller (1994), Dekkers and van Arendonk (1998), or Chakraborty et al. (2002) have considered the optimization of marker-assisted selection for identified 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 of large enough effect.

3.2.1 The marker-phenotype index

Other methods exist in order to capitalize on all the variability accounted for by the markers, including QTL of small effects. Here, we do not aim to precisely estimate the chromosomal locations and the effects of the QTL, but simply use markers to improve the prediction of the breeding value of each individual and select the best

individuals according to this value. The method is rapidly outlined here, more details can be found in Whittaker (2001).

The bases were set by the landmark paper of Lande and Thompson (1990). For a single marker, the ‘molecular score’ of an individual for use in recurrent selection is obtained as the estimate of the statistical association between marker genotype and phenotype. For multiple markers, genotype effects can be summed over markers into a single molecular score. Then Lande and Thompson (1990) derived an index for selection combining molecular and phenotypic information. Considering molecular score M and phenotype P as two correlated trait, the authors used classic selection index theory to compute the coefficient b_M and b_P of the index $I = b_M M + b_P P$, optimally weighting both types of information in order to maximize the genetic gain. Lande and Thompson concluded that the method was most efficient for low heritable traits. However, Moreau *et al.* (1998) later showed that, because low heritability also reduces the power of detection of the effects associated to markers in a finite (real) population, greatest opportunities for MAS with this method may exist for traits with moderate rather than low heritability. Efficiency of the method has been tested by simulations in several paper (*e.g.*, Gimelfarb and Lande, 1994 ; Hospital *et al.* 1997) and proved efficient. In all cases, population size (that must be large enough to allow a good estimate of marker effects) appears as the most critical factor limiting the efficiency of MAS. The current debate is whether only markers with significant effects should be taken account in the molecular score, as first proposed by Lande and Thompson (1990). Moreau *et al.* (1998) showed that increasing type I error risk could increase the efficiency of MAS, by increasing the power of detection of genes with small effects. More recently, methods were proposed that include all marker effects in the index, regardless of their statistical significance, and provide increased selection response (Meuwissen *et al.*, 2001 ; Lange and Whittaker, 2001).

One problem with this method is that QTL effects are often overestimated, as shown by both theory (Beavis 1994, Bost *et al.* 2001) and experimentation (Melchinger *et al.* 1998). Overestimation of QTL effects leads to too much emphasis on molecular scores in selection relative to phenotypic data and results in a less than optimal response to selection. Alternative statistical methods for analysis of QTL data that avoid overestimation or reduce its impact on selection response are needed (*e.g.* Fernando and Grossman 1989).

Another, perhaps more critical problem, is that here molecular costs are in addition to, not in place of, phenotypic costs, contrary to Genotype Building strategies with selection based solely on genotype. But, resources allocated in each generation to molecular assay could also be allocated to enhance conventional phenotypic selection (*e.g.*, by increasing the number of individuals tested) with more profit, because the molecular costs are still high relative to phenotypic costs (Moreau *et al.*, 2000). The economic merit of MAS could be restored by reducing the frequency of re-evaluation of marker effects (Hospital *et al.*, 1997). However, further work on the optimization of such strategies is required, and it is likely that the economically optimal use of MAS necessitates a complete re-thinking of the design of breeding schemes (see for example Ribaut and Hoisington, 1998 for a review of changes required for plant breeding programs).

3.3 Selection for hybrid performance

In theory, crosses between lines that are genetically more distant are expected to show greater heterosis. Genetic distance can be measured from differences in allele frequencies at anonymous markers spread throughout the genome. Evaluation of this concept for a large number of crops (Melchinger, 1999) shows that marker-based prediction of hybrid performance can be efficient if hybrids include crosses between lines that are related by pedigree or which trace back to common ancestral populations. On the other hand, prediction is not efficient for crosses between lines that are unrelated or that originated from different populations, because the associations (via linkage disequilibrium) between marker loci and QTL involved in heterosis are not the same in the different populations (Charcosset and Essioux 1994).

The limited ability to predict hybrid vigor in untested crosses has motivated the development of strategies to use knowledge of QTL effects to generate crosses that are predicted to create QTL genotypes with favorable non-additive effects. An example is the use of marker-based statistical methods to predict the performance of untested crosses from performance of parental lines in a limited number of test crosses (Bernardo 1994, 1999).

4 Experimental results

Few results of real MAS experiments have yet been published. Some recent results in plants are presented below by increasing level of complexity for the use of markers, the genes manipulated and/or the traits under control.

Using markers as simple marks to fasten the recovery of recipient genome background (background selection) in backcross introgression programs for the transfer of a single well identified target region (direct marker) has been nicely proved efficient by the integration of the *Bt* transgene into different maize genetic backgrounds (Ragot *et al.* 1995). This confirmed the theoretical prediction that use of markers provides a gain in time of approximately 2 BC generations. If few other results on this matter have been published, it is known that the technique is now largely used, in particular by private plant breeding companies.

Other experimental reports for the manipulation of known genes with indirect (linked) markers include ‘pyramiding’ of several major resistance genes in rice, from near-isogenic lines (NILs), each carrying only one gene, into a common background (Huang *et al.* 1997 ; Hittalmani *et al.* 2000). In all cases, control of the target genes by indirect linked markers was successful, as later checked by phenotypic assay of resistance. Huang *et al.* (1997) pyramided four genes for blight resistance into different combinations (2, 3 or 4 genes) that exhibited higher level of resistance and/or wider spectrum than the original parents. Moreover, some pyramided lines showed resistance to pathogen races to which all parents were susceptible. Results were also generally successful for Hittalmani *et al.* (2000), who pyramided 3 genes for blast resistance into different combinations. However, in this case some multiple-genes combinations did not perform any better than the single-gene one, indicating that a good knowledge of the spectrum of gene effects is necessary prior to performing the MAS program. In any case, pyramiding multiple resistance genes is a valuable step towards more durable and stable crop resistance, that could hardly be achieved without the use of marker-based selection, because epistasis and/or the masking effects of genes limit the efficiency of conventional (phenotypic) breeding methods. Moreover, use of markers not only alleviates this limit, but also provides a better understanding of these gene interactions.

Experimental results of MAS for the manipulation of QTL (not known major genes) are more contrasted. Toojinda *et al.* (1998) introgressed 2 QTL for stripe rust

resistance in barley, through 1 backcross followed by 1 haplo-diploidisation with selection on marker genotype and phenotype, into a genetic background different from the one used to map QTL. Both QTL were confirmed, and additional QTL were detected in the new background, including some resistance alleles brought by the susceptible parent. Probably those alleles were fixed in the mapping population, but this illustrates the importance of the genetic background, both for QTL detection and MAS. Han *et al.* (1997) manipulated 2 QTL for a component of malting quality in six-row barley, a trait that is very difficult and costly to work phenotypically. They screened and selected DH lines with four different strategies: i) phenotype alone, ii) marker genotype alone, iii) genotype followed by phenotype in tandem selection, or iv) genotype and phenotype combined in an index. This either on a single-trait, or a multiple-trait basis. Results were successful for one QTL, but not for the other QTL, for which tandem and combined selection based on both marker genotype and phenotype did not perform any better than selection on phenotype alone, probably because the location of the QTL was inaccurate. However, the authors point out that, even not performing any better, tandem selection provides a valuable gain in time and efforts, compared to phenotypic selection. Lawson *et al.* (1997) introgressed four target chromosomal regions containing five QTL for pest resistance (acylsugar accumulation) from wild tomato into cultivated tomato. Starting with the introgression lines of Eshed and Zamir (1995), each carrying one target region, they performed three backcrosses followed by one intermating generation to obtain progenies homozygous for the resistance alleles at the five QTL. Selection was based on both marker genotype and phenotype. The introgression of the four regions was successful at the genomic level. However, the level of acylsugar accumulation in the progenies introgressed for the five QTL was lower than expected, and in particular lower than that of the interspecific F₁ hybrid, indicating that some genetic factors (QTL) of the accumulation were missing, either lost or not controlled in the program. Shen *et al.* (2001) manipulated four QTL for drought resistance (root depth) in rice, a trait that is very difficult to manage phenotypically. Starting from DH lines, they produced a number of BC₃F₃ lines, each introgressed for one or two QTL at most, using selection on marker genotype alone, not phenotype. They re-detected and fine mapped the QTL in the progenies. Among the four QTL, one exhibited the expected effect in the progenies, one was finally revealed as a false-positive, one segment was shown to contain in fact two QTL in repulsion phase (+/-) that reduced its expression, and one segment did not exhibit the expected effect, either because the QTL was lost in the

program, or because its effect was masked by epistatic interactions. This again highlights the problems linked to the precision of the initial QTL detection with regard to the position and effect of the QTL, and the effect of possible epistatic interactions on the expression of the QTL in the progenies. Ribaut *et al.* (2002) introgressed five target regions containing QTL for drought tolerance (reduction of ASI) in maize. The results depended on the condition of the phenotypic assay of the progenies: under stress conditions (drought), the introgressed progenies exhibited a reduced ASI, while the introgression had no visible effect in the absence of stress. Zhu *et al.* (1999) screened DH lines of barley for the presence of several QTL for yield, a very complex composite trait, based on selection for marker genotype alone. They evaluated phenotypically the progenies in five environments, including four locations and two years. The results indicate that the position of the QTL were confirmed as correct in the progenies. However the effect of the QTL in the progenies were often different from the expectation with regards to magnitude and sign. Moreover, the authors detected epistatic interactions between QTL, as well as numerous GxE interactions. The authors conclude that selection for complex traits should focus on allelic combinations (based on epistatic interactions) rather than on individual QTL effects.

The above experimental results of MAS for QTL are synthesized by an experiment performed in our lab (Bouchez *et al.*, 2002). The introgression of favourable alleles at 3 QTL for 2 traits (earliness and yield) between maize elite lines by marker-assisted backcrossing showed that use of markers as simple marks to improve background selection is efficient, even with few markers, especially on non-carrier chromosomes. Foreground selection on markers to control the three target regions without the help of phenotypic assay was also efficient. However, results of the phenotypic evaluation of introgressed progenies, as well as the re-detection of QTL among those progenies depended upon the complexity of the trait under control. For the simple trait (earliness), QTL effects in the progenies were in general accordance with those expected from the original detection in the parental lines. For the more complex trait (yield) results were in general not as good as expected, and one high-yielding allele putatively detected from the low-yielding parent finally exhibited an effect opposite to the expectation. This indicates that the estimates of QTL positions appear more reliable than the estimates of their effects, in particular with regards to genotype by environment (GxE) interactions, which were found significant in the experiment.

5 Conclusions

The application of molecular genetics in breeding programs is currently bounded by the precision of the effects associated to markers, and the economic merit of marker-assisted selection.

Using marker-based selection is definitely useful to manipulate chromosomal regions and design rapidly new genotypes combining favorable regions (Genotype Building). The clearer example is marker-assisted selection in backcross breeding schemes for the introgression of one or a few target genes in a given genetic background. This is probably the implementation of MAS that is the most widely used in practice, in particular by private plant breeding companies, although the corresponding results are often not published. In this case, the target gene(s) could be major genes of well known, or well estimated effects. Marker-assisted backcrossing is also particularly useful for the introgression of transgenes. Control of the target (foreground selection) is easy because its DNA sequence is known. Moreover, it is often easier to introgress a transgene from an already genetically modified material into a new (non-modified) line, than engineer the new line.

Classical (phenotypic) selection in plant breeding is limited by the ability to estimate genetic parameters (breeding values of the individuals candidate to selection) for the traits of interest, using statistical analysis of phenotypic data (quantitative genetics). Use of marker information alleviates some of the limits of quantitative genetics selection, and provides better estimates of breeding values, by increasing the apparent heritability of the trait. However, the alleviation may be only partial, depending on the complexity of the genetic architecture of the trait. Genotype Building experiments, or selection based solely on molecular information appears restricted to simple traits that are governed by few genes of large effects, so that genetic markers capture most genetic variation for the trait, and provide precise and sustainable estimates of breeding values. Conversely, for complex traits that are governed by several genes of medium to low effects, possibly affected by environment, it appears necessary to have an accurate evaluation of QTL effects in varying environments before initiating a genotype building program. If QTL effects are not perfectly estimated and sustainable, it might be risky to perform selection based solely on markers. In such cases, selection must be on a combination of marker and phenotypic data and hence will suffer from the same limitations as conventional breeding.

Economics is the other key determinant for the application of molecular genetics in breeding programs. Cases where the economic merit of MAS is clear include situations where molecular costs are more than offset by the savings in phenotypic evaluation. Examples are the use of markers in genotype building programs and selection on traits that are costly to evaluate, but well characterized at the molecular level (e.g. oligogenic disease resistances). In other cases, the ability to select early offsets the extra costs associated with MAS. The benefits of being able to release new genetic material more quickly can be substantial, particularly in competitive markets. The economic merit of MAS becomes questionable and more difficult to evaluate in cases where MAS is expected to provide greater genetic gain at increased costs. This is particularly the case for selection on a combination of phenotype and molecular score (see section 3).

Clearly, marker-assisted selection is efficient and valuable for simple traits and/or traits for which increase of genetic gain per unit of time is of high economic return. However, the advent of marker-assisted selection for the ordinary breeding of complex traits relies on a re-thinking of breeding strategies, and on the availability of both statistical and molecular techniques that would provide precise estimates of gene effects in selected populations at low cost, which is far from being the case at present.

6 References

- Beavis, W. D. (1994) The power and deceit of QTL experiments: lessons from comparative QTL studies, pp. 250-266 in *Proc. 49th Annual Corn and Sorghum Research Conference*, ASTA, Washington D.C.
- Bernardo R. (1994) Prediction of maize single-cross performance using RFLPs and information from related hybrids. *Crop Science*, 34, 20-25.
- Bernardo R. (1999) Marker-Assisted Best Linear Unbiased Prediction of Single-Cross Performance. *Crop Science*, 39, 1277-1282
- Bost B., de Vienne D., Hospital F., Moreau L., and Dillmann C. (2001) Genetic and nongenetic bases for the L-shaped distribution of quantitative trait loci effects. *Genetics*, 157, 1773-87.
- Bouchez A., Hospital F., Causse M., Gallais A. and Charcosset A (2002) Marker assisted introgression of favorable alleles at quantitative trait loci between maize elite lines. submitted.
- Chakraborty, Moreau, and Dekkers. 2002. A general method to optimize selection on multiple identified QTL. *Genetics Selection Evolution*, (in press).
- Charcosset, A. and Essioux L. (1994) The effect of population structure on the relationship between heterosis and heterozygosity at marker loci. *Theoretical and Applied Genetics*, 89, 3336-343.
- 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.
- 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.
- Dekkers, JCM, and van Arendonk JAM. (1998). Optimum selection for quantitative traits with information on an identified locus in outbred populations. *Genetical Research*, 71, 257-275.
- Dekkers JCM and Hospital F. (2002) The use of molecular genetics in the improvement of agricultural populations. *Nature Reviews Genetics*, 3, 22-32.
- de Vienne, D. (Ed) (2002) *Molecular markers in plant genetics and biotechnology*, Oxford Publishing Ltd., in press.

- Eshed Y. and Zamir D. (1995) An introgression line population of *Lycopersicon pennellii* in the cultivated tomato enables the identification and fine mapping of yield-associated QTL. *Genetics* 141(3), 1147-62.
- Fernando RL, and Grossman M. (1989) Marker-assisted selection using best linear unbiased prediction. *Genetics Selection Evolution*, 21, 467-477
- 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.
- Gimelfarb A, and Lande R (1994) Simulation of marker-assisted selection in hybrid populations. *Genetical Research*, 63, 39-47.
- 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.
- Han F., Romagosa I., Ullrich S. E., Jones B. L., Hayes P. M. and Wesenberg D. (1997) Molecular marker-assisted selection for malting quality traits in barley. *Molecular Breeding*, 3, 427-437.
- 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.
- Hartl, D. L. Clark, A. G (1997) *Principles of population genetics*. Sinauer Associates Incorporated, Sunderland, US., Ed. 3, xiii + 542 pp.
- 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.
- Hittalmani S., Parco A., Mew T. V., Zeigler R. S. and Huang N. (2000) Fine mapping and DNA marker-assisted pyramiding of the three major genes for blast resistance in rice. *Theoretical and Applied Genetics*, 100, 1121-1128.
- 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. and Charcosset, A. (1997) Marker-assisted introgression of quantitative trait loci. *Genetics*, 147, 1469-1485.

- Hospital, F. and Decoux, G. (2002) Popmin: a program for the numerical optimization of population sizes in marker-assisted backcross programs. *The Journal of Heredity*, in press.
- Hospital, F., Chevalet, C. and Mulsant, P. (1992) Using markers in gene introgression breeding programs. *Genetics*, 132, 1199-1210.
- Hospital, F., Moreau, L., Lacoudre, F., Charcosset, A. and Gallais, A. (1997) More on the efficiency of marker assisted selection. *Theoretical and Applied Genetics*, 95, 1181-1189.
- Hospital, F., Goldringer, I. and Openshaw, S. (2000) Efficient marker-based recurrent selection for multiple quantitative trait loci. *Genetical Research*, 75, 357-368.
- Huang N., Angeles E. R., Domingo J., Magpantay G., Singh S., Zhang G., Kumaravadivel N., Bennett J., Khush G. S (1997) Pyramiding of bacterial blight resistance genes in rice: marker-assisted selection using RFLP and PCR. *Theoretical and Applied Genetics*, 95, 313-320.
- Koudande, O. D. Iraqi, F. Thomson, P. C. Teale, A. J. Arendonk, J. A. M. van (2000) Strategies to optimize marker-assisted introgression of multiple unlinked QTL. *Mammalian Genome*, 11, 145-150.
- Lande R. Thompson R. (1990) Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics*, 124(3), 743-56,
- Lange C and Whittaker J. C. (2001) On Prediction of Genetic Values in Marker-Assisted Selection. *Genetics*, 159, 1375-1381.
- Lawson D. M., Lunde C. F. and Mutschler M. A. (1997) Marker-assisted transfer of acylsugar-mediated pest resistance from the wild tomato, *Lycopersicon pennellii*, to the cultivated tomato, *Lycopersicon esculentum*. *Molecular Breeding*, 3, 307-317.
- Melchinger, A.E. (1990) Use of molecular markers in breeding for oligogenic disease resistance. *Plant Breeding*, 104, 1-19.
- Melchinger A.E. (1999) Genetic diversity and heterosis In " *The genetics and exploitation of heterosis in crops* ", J.G. Coors et S. Pandey (eds.), Crop Science Society of America, pp. 99-118
- Melchinger AE., Utz HF. , and Schon CC. (1998) Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. *Genetics*, 149(1), 383-403,

- Meuwissen THE, Hayes BJ, and Goddard ME. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics*, 157, 1819-1829.
- Mohan, M., Nair, S., Bhagwat, A., Krishna, T. G., Yano, M., Bhatia, C. R, and Sasaki, T (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. *Molecular Breeding*, 3, 87-103.
- Moreau L. Charcosset A., Hospital F. and Gallais A. (1998) Marker-assisted selection efficiency in populations of finite size. *Genetics*, 148(3), 1353-65,
- Moreau, L., Lemarié S., Charcosset A and. Gallais A, (2000) Economic efficiency of one cycle of marker-assisted selection. *Crop Science*, 40, 329-337.
- 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., Biasioli 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), INRA, Paris, France.
- Ribaut JM, and Hoisington D (1998) Marker-assisted selection: new tools and strategies. *Trends in Plant Science*, 3(6), 236-239.
- Ribaut, J. M., Hu XueYi., Hoisington, D. and Gonzalez-de-Leon, D (1997) Use of STSs and SSRs as rapid and reliable preselection tools in a marker-assisted selection-backcross scheme. *Plant Molecular Biology Reporter*, 15, 154-162.
- Ribaut, J.-M., Jiang, C., and Hoisington, D. (2002) Simulation Experiments on Efficiencies of Gene Introgression by Backcrossing. *Crop Science*, 42, 557-565
- Ribaut J.M., Banziger M., Betran J., Jiang C., Edmeades G.O., Dreher K., and Hoisington D. (2002) Use of molecular markers in plant breeding: drought tolerance improvement in tropical maize in “*Quantitative Genetics, Genomics, and Plant Breeding*”, M. S. Kang (ed), CABI Publishing, Wallingford, UK, in press.
- Servin B and Hospital F (2002) Optimal positioning of markers to control genetic background in marker assisted backcrossing. *The Journal of Heredity*, in press.
- Servin B, Dillmann C, Decoux G and Hospital F (2002) MDM : a program to compute fully informative genotype frequencies in complex breeding schemes. *The Journal of Heredity*, in press.

- Shen L., Courtois B., McNally K.L., Robin S. and Li Z. (2001) Evaluation of near-isogenic lines of rice introgressed with QTLs for root depth through marker-aided selection. *Theoretical and Applied Genetics*, 103, 75-83.
- 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.
- Tanksley, S. D., (1983) Molecular markers in plant breeding. *Plant Molecular Biology Reporter*, 1, 3-8.
- Tanksley, S.D., Young ND, Paterson AH and Bonierbale MW (1989) RFLP mapping in plant breeding: new tools for an old science, *Biotechnology*, 7, 257-264
- Toojinda T., Baird E., Booth A., Broers L., Hayes P., Powell W., Thomas W., Vivar H. and Young G. (1998) Introgression of quantitative trait loci (QTLs) determining stripe rust resistance in barley: an example of marker-assisted line development. *Theoretical and Applied Genetics*, 96, 123-131.
- 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. (1996) Proportion of the variation in genomic composition in backcrossing programs explained by genetic markers. *The Journal of Heredity*, 87, 136-138.
- Visscher, P.M., Haley, C.S. and Thompson, R. (1996) Marker-assisted introgression in backcross breeding programs. *Genetics*, 144, 1923-1932.
- Westman AL and Kresovich (1997) Use of molecular marker techniques for description of plant genetic variation. In *Biotechnology and Plant Genetic Resources: Conservation and Use*. Callow JA, Ford-Lloyd BV and Newbury HJ (Eds) CAB INTERNATIONAL, Wallingford, UK. Pp. 9-48.
- Whittaker, J. C., (2001) Marker assisted selection and introgression, pp. 673–695 in *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop and C. Cannings (eds). Wiley, New York.
- 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.

Zhu H., Briceno G., Dovel R., Hayes P. M., Liu B. H., Liu C. T., and Ullrich S. E.
(1999) Molecular breeding for grain yield in barley: an evaluation of QTL effects
in a spring barley cross. *Theoretical and Applied Genetics*, 98, 772-779.