Optimal Positioning of Markers to Control Genetic Background in Marker-Assisted Backcrossing

B. Servin and F. Hospital

Molecular markers are commonly used in backcross breeding programs in plants. As genetic maps contain more and more markers, it is of interest to determine which markers are to be used for selection. Here we describe how one can compute an optimal positioning of markers resulting in a maximization of the expected proportion of recipient genome. This criterion allows us to take selection into account and to produce relevant results regarding the final efficiency of background selection in backcross programs.

Molecular markers have proven very useful in improving backcross breeding schemes. Particularly, markers allow us to estimate the genomic composition of individuals, and selection on markers can speed up the recipient genome recovery on noncarrier chromosomes (background selection). Several studies have shown that few markers (typically 2–4 markers/Morgan) are necessary to control genetic background in marker-assisted backcrossing (Hospital et al. 1992; Visscher et al. 1996). Yet more than 2–4 markers/Morgan are generally available. If we assume that all markers on the genetic map have the same technical benefits (codominance, polymorphism between parents), the choice of the markers to use for background selection has to be made according to their positions on the genetic map.

Few studies have evaluated the optimal positioning of markers to improve backcross selection efficiency. Hospital et al. (1992) showed the impact of marker positions on background selection efficiency based on simulation studies, and determined roughly the optimal positioning of two markers on a chromosome of 100 cM. Visscher (1996) computed the optimal positioning of markers, defined as the position of the last marker with a marker (herein called TM segments), of size (L−2d)/m, spaced along the chromosome, with a step of 0.1 cM. A smaller step was tried and is the recombination rate between telomere \( T_1 \) and the first marker \( M_1 \). \( d \) is also the distance between the last marker \( M_m \) and the second telomere \( T_2 \). For \( m > 2 \), the other markers \( M_2 \) to \( M_{m−1} \) are equally spaced in the segment \([M_m,M_{m−1}]\) as was also done by Visscher (1996), who used the same parameter \( d \).

Hence the chromosome is composed of two segments delimited by a telomere and a marker (herein called TM segments), of size \( d \), and \((m−1)\) segments delimited by two successive markers (herein called MM segments), of size \((L−2d)/(m−1)\).

The closed form of \( \Pi \) for two markers at generation \( BC_1 \), can be obtained by analytical derivations (see appendix):

\[
\Pi = \frac{1}{L} \int_0^L \frac{1}{P(X|d)} \ dx,
\]

where \( r_{TM} \) is the recombination rate between \( T_1 \) and \( M_1 \), and \( r_{MM} \) is the recombination rate between \( M_1 \) and \( M_2 \). For \( m > 2 \), the other markers \( M_2 \) to \( M_{m−1} \) are equally spaced in the segment \([M_m,M_{m−1}]\), as shown in Figure 1.

Methods

Throughout the article, we will assume that recombination takes place without interference and will use Haldane’s mapping function to compute genetic distances from recombination rates. Since only one chromosome of each pair is segregating in backcrossing, the analytical derivations and numerical applications are related to the segregating chromosome throughout the article. The criterion computed (II) is the expected proportion of recipient genome, given that all markers are of recipient type, which obviously can be addressed on a per chromosome basis as

\[
\Pi = \frac{1}{L} \int_0^L \frac{1}{P(X|d)} \ dx,
\]

where \( L \) is the chromosome length and \( P(X|d) \) is the probability that a locus \( X \) at position \( x \) on the chromosome is of homozygous recipient type given that all markers are of homozygous recipient type on this chromosome. The value of \( \Pi \) thus depends on the number and the positioning of markers, and on the backcross generation at which all markers are of homozygous recipient type (i.e., the last generation of background selection). For a given number of markers \( m \), the positioning of markers on a chromosome is described by a single parameter \( d \) (see Figure 1), the distance between the first telomere \( T_1 \) and the first marker \( M_1 \). For \( m > 2 \), the other markers \( M_2 \) to \( M_{m−1} \) are equally spaced in the segment \([M_m,M_{m−1}]\).

The closed form of II for two markers at generation \( BC_1 \), can be obtained by analytical derivations (see appendix):

\[
\Pi = \frac{1}{L} \int_0^L \frac{1}{P(X|d)} \ dx,
\]

where \( r_{TM} \) is the recombination rate between \( T_1 \) and \( M_1 \), and \( r_{MM} \) is the recombination rate between \( M_1 \) and \( M_2 \). For \( m > 2 \), the other markers \( M_2 \) to \( M_{m−1} \) are equally spaced in the segment \([M_m,M_{m−1}]\), as shown in Figure 1.

The closed form of II for two markers at generation \( BC_1 \), can be obtained by analytical derivations (see appendix):

\[
II = \frac{1}{L} \int_0^L \frac{1}{P(X|d)} \ dx,
\]

where \( r_{TM} \) is the recombination rate between \( T_1 \) and \( M_1 \), and \( r_{MM} \) is the recombination rate between \( M_1 \) and \( M_2 \). For \( m > 2 \), the other markers \( M_2 \) to \( M_{m−1} \) are equally spaced in the segment \([M_m,M_{m−1}]\), as shown in Figure 1.

The closed form of II for two markers at generation \( BC_1 \), can be obtained by analytical derivations (see appendix):

\[
II = \frac{1}{L} \int_0^L \frac{1}{P(X|d)} \ dx,
\]

where \( r_{TM} \) is the recombination rate between \( T_1 \) and \( M_1 \), and \( r_{MM} \) is the recombination rate between \( M_1 \) and \( M_2 \). For \( m > 2 \), the other markers \( M_2 \) to \( M_{m−1} \) are equally spaced in the segment \([M_m,M_{m−1}]\), as shown in Figure 1.
ers and for different positionings of the markers.

The closer a locus at position \( x \) is to a marker, the higher is \( P(X|L) \). When \( d \) increases, the size of the MM segments decreases and \( P(X|L) \) at any locus on MM segments increases. Conversely, when \( d \) increases, the size of TM segments increases, and \( P(X|L) \) decreases at any locus on TM segments. As \( \Pi \) is a linear combination of these probabilities, it presents a maximum (noted \( \Pi^* \)) for an optimal value of \( d \) (noted \( d^* \)), giving the optimal positioning of the markers.

Figure 2 illustrates variations in the genomic composition as a function of \( d \) for a chromosome of 100 cM controlled by two markers of fully recipient genotype at generation BC\(_1\). As explained above, the proportion of recipient genome on MM segments (dotted lines) increases with \( d \) and the proportion of recipient genome on TM segments (scattered lines) decreases as \( d \) increases. Finally, \( \Pi \) presents a maximum of coordinates \( (d^*, \Pi^*) \) indicated by the dot in Figure 2. Qualitatively, similar results are obtained with any number of markers and at any backcross generation.

Results and Discussion

Optimal Positioning

Table 1 shows the optimal positioning \( (d^*) \) of two to four markers on a 100 cM chromosome in backcross generations BC\(_1\)–BC\(_3\), as well as the corresponding \( \Pi^* \). The theoretical proportion of recipient genome without selection on markers is also recalled as a comparison \( (\pi) \). Finally, Table 1 recalls the optimal positioning of Visscher (1996), expressed as corresponding \( d \) values.

It is seen from Table 1 that optimal \( d^* \) values slightly increase with the backcross generations. This can be interpreted as follows. In BC\(_1\), the optimal length of MM segments, \( (L - 2d^*)/(m - 1) \), is much larger than the length of TM segments \( (d^*) \), because segments flanked by two selected markers are better controlled than segments flanked by only one marker. But, as meioses accumulate in advanced backcross generations, the apparent recombination rate between markers increases, and MM segments tend to be not better controlled than TM segments. The optimal size of MM segments relative to TM segments then need to be reduced compared to its value in BC\(_1\).

The variation of \( d^* \) is more important between generation BC\(_1\) and BC\(_2\) than between more advanced generations (e.g., BC\(_2\) and BC\(_3\)) as seen in Table 1. Indeed, as only one generation of recombination has taken place in BC\(_1\), it is very likely that the MM segments are introgressed as a whole and are fully recipient, because the probability that double recombination events occurred between the markers is very low, except for very large MM segments. In later backcross generations, loss of control on MM segments is due not only to (rare) double recombinations between the markers, but also to (more frequent) single recombinations between the markers occurring twice in different generations. Thus, the apparent recombination rate between markers increases faster between generation BC\(_1\) and BC\(_2\) than between two other backcross generations.

Suboptimal Positioning

Even with dense genetic maps, it can be hard to find markers exactly positioned at optimal positions \( d^* \). In this case, it is of interest to know the impact on genome content of using markers not exactly placed in the optimal positions described above (suboptimal positioning of markers). The last two columns of Table 1 show the positions \( (d^* - d^*) \) and \( (d^* + d^*) \) defining an interval in which \( \Pi^* - 1% \leq \Pi \leq \Pi^* \).

Using more markers on a chromosome leads to better control of the return to the recipient genome because the regions controlled by the markers overlap. Thus
better control of the genomic background can be achieved either by using more markers, that can be suboptimally placed, or by using fewer markers, optimally placed. For example, in generation BC2, using four suboptimally placed markers leads to an expected II of 97.6%, which can be obtained with three well-placed markers (II* = 97.6%).

For a given number of markers, the impact of suboptimal positioning of markers is less important when the backcross generation is more advanced. Indeed, even when no selection on markers is performed, the recurrent genome content increases, due to backcrossing. Thus for a given number of markers, the same value of II can be reached either by optimally placing markers and performing fewer backcross generations or by suboptimally placing the markers and performing more backcross generations. For example, a chromosome of fully recurrent genotype at three markers will present an expected proportion of recurrent genome of II = II* = 97.6% at generation BC2, if markers are optimally placed. If markers are suboptimally placed, the same return to the recipient genome will be obtained at generation BC5 (II = 97.3%).

Optimization Criterion

We found that optimal positions of two markers on a chromosome of 100 cM are about 20 cM from the telomeres (from 18.6 cM in BC1 to 22.9 cM in BC5 as recalled from Table 1). These results slightly differ from those in Visscher (1996), where the optimal marker positions are around 28 cM from the telomeres in the same conditions. Generally the positioning described in Visscher (1996) is farther from the telomeres than ours. The main difference is that optimal d* values here are given conditional on the success of selection, whereas the values given by Visscher (1996) are obtained assuming no selection on markers. This could explain the difference between our respective results. Conversely, our results for two markers fit well those of Hospital et al. (1992), obtained by simulations that take fully into account selection on markers. In fact, they found that the optimal positioning of two markers on a 100 cM chromosome was roughly at 20 cM from the telomeres. This argues for a better relevance of our optimization criterion II to predict marker positions that maximize the response to selection (i.e., the return to recipient parent genome) compared to the one used by Visscher (1996), because II takes selection into account.

The optimal positioning given in Visscher (1996) is based on the linear prediction of the proportion of recipient genome in a population composed of individuals presenting every possible genotype at markers. However, the relationship between the proportion of recipient genome and the possible genotypes at markers is linear only in BC1. Indeed, using such a linear predictor in BC1 leads to results very close to the ones obtained using our estimate II. For more advanced backcross generations, the relationship is no longer linear, and a linear predictor is not a good estimate of the proportion of recipient genome.

Efficiency of Background Selection

The proportion of recipient genome (II*) obtained for the optimal positioning is high compared to the theoretical values when no selection on markers is performed (π), as shown in Table 1. For example, a noncarrier chromosome presenting three optimally placed markers that are of recipient type in BC1 will have 99.2% of recipient genome. Without selection on markers, the same return to the recipient genome would be obtained only in BC5. Thus when all markers are of recipient type, it is expected that most of the genome is of recipient type. This confirms previous studies by showing that few markers can efficiently control large chromosomal regions (Hospital et al. 1992; Visscher et al. 1996).

Although the criterion we used to infer optimal positioning is based on the success of selection on markers, our study does not allow us to predict the efficiency of selection on markers, but previous studies have shown that it is very efficient. For example, Hospital et al. (1992) considered background selection on two markers per chromosome for 10 noncarrier chromosomes of 100 cM. They showed that, selecting down to a proportion of 10% individuals at each generation, homozygous recipient genotypes at all markers can be obtained as early as BC5. In the case of fewer noncarrier chromosomes and/or higher selection intensity, background selection may succeed in only one or two generations.

Our method could be extended to background selection on carrier chromosomes, but the optimal positioning will then depend on the position of the target gene (or of markers controlling it) and on the positions of markers used to reduce the linkage drag around the target gene.

Appendix: Analytical Derivation of II for Two Markers in Generation BC1

We consider a chromosome controlled by two markers (M1 and M2) positioned as explained in Figure 1. We denote rMM12, the recombination rate between M1 and M2, and rTM the recombination rate between T and M1. As the distance between T and M1 is also d, rTM is also the recombination rate between T and M2. We assume that recombination rates are related to genetic distances by Haldane’s mapping function, and thus rMM12 = $\frac{1}{2}(1 - e^{-2d^2})$ and $r_{TM} = \frac{1}{2}(1 - e^{-\lambda d})$. We also consider an unmarked locus, noted X, placed at position x on the chromosome.

As recalled from Equation (1) in the method section,

$$II = 100 \int_0^L \frac{1}{L} P(X|\omega) \, dx$$

$$= 100 \int_0^L \frac{1}{L} \frac{P(X \cap M)}{P(M)} \, dx, \quad (A.1)$$

where $P(X \cap M)$ is the probability to have the three loci X, M1, and M2 of homozygous recipient genotype, and $P(M)$ is the probability to have both markers M1 and M2 of homozygous recipient genotype.

As markers are placed symmetrically to the center of the chromosome, and as only $P(X \cap M)$ is a function of x, Equation (A.1) can be rewritten as

$$II = 100 \alpha(d, L) \int_0^{L/2} P(X \cap M) \, dx, \quad (A.2)$$

where $\alpha(d, L) = 2/LP(M)$ and $P(M) = \frac{1}{2}(1 - r_{TM})$.

$P(X \cap M)$ has to be divided into two parts to compute Equation (A.2), depending on the relative positions of X and M1:

$$P(X \cap M) = \begin{cases} P_{M1}(x, d, L) & \text{when } x \in [0, d] \\ P_{M2}(x, d, L) & \text{when } x \in [d, L/2]. \end{cases}$$

Let $r_1$ denote the recombination rate between X and M1, and $r_2$ the recombination rate between X and M2. Using Haldane’s mapping function we have

$$\begin{cases} r_1 = \frac{1}{2}(1 - e^{-2d^2}) \\ r_2 = \frac{1}{2}(1 - e^{-\lambda d}). \end{cases}$$
Computing $\frac{1}{2}P_{TM}(x, d, L) \, dx$

As $x$ is on the TM segment, $P_{TM}(x, d, L) = \frac{1}{2}(1 - r_1 - d - r_2)$. In this case, $r_1 = \frac{1}{2}(1 - e^{-2d-x})$. Developing $r_1$ and $r_2$ as functions of $x$, we obtain

$$P_{TM}(x, d, L) = \frac{1}{2} \left( 1 - r_{M,M} \right) + \frac{1}{4} \left( e^{-2d} + e^{-2d-x} \right) e^{2x}.$$  \hfill (A.3)

Integrating Equation (A.3) gives

$$\int_{0}^{d} P_{TM}(x, d, L) = \frac{1}{4} \left( 1 - r_{M,M} \right) (d + r_{TM}) .$$  \hfill (A.4)

Computing $\frac{1}{2}P_{MM}(x, d, L) \, dx$

As $x$ is on the MM segment, $P_{MM}(x, d, L) = \frac{1}{2}(1 - r_{M,M} - r_{M,R})$. In this case, $r_1 = \frac{1}{2}(1 - e^{-2d-x})$. Developing $r_1$ and $r_2$ as functions of $x$, we obtain

$$P_{MM}(x, d, L) = \frac{1}{2} \left( 1 - r_{M,M} \right) + \frac{1}{4} \left( e^{2d} + e^{2d-x} \right) e^{2x}.$$  \hfill (A.5)

Integrating Equation (A.5) gives

$$\int_{0}^{L/2} P_{TM}(x, d, L) = \frac{1}{2} \left( 1 - r_{M,M} \right) (L - 2d) + \frac{1}{4} r_{M,M}.$$  \hfill (A.6)

Obtaining $II$

Finally, from Equation (A.2),

$$II = 100\alpha(d, L) \left( \int_{0}^{d} P_{TM}(x, d, L) \, dx ight)$$

$$+ \int_{0}^{L/2} P_{MM}(x, d, L) \, dx$$

$$\hspace{1cm} (A.7)$$

$$II = 100 \left( 1 - \frac{1}{L} \left( 2r_{TM} + \frac{r_{M,M}}{1 - r_{M,M}} \right) \right) .$$  \hfill (A.8)

From the Station de Génétique Végétale, INRA/UPS/IN-APG, Ferme du Moulon, 91190 GI sur Yvette, France. The authors wish to thank P.M. Visscher and one anonymous referee for their helpful comments. Address correspondence to Bertrand Servin at the above address or e-mail: servin@moulon.inra.fr.

(C) 2002 The American Genetic Association

Polymorphic Microsatellites in Antirrhinum (Scrophulariaceae), a Genus With Low Levels of Nuclear Sequence Variability

D. Zwettler, C. P. Vieira, and C. Schlötterer

In Antirrhinum, reproductive systems range from self-compatible to self-incompatible, but the actual outcrossing rates of self-compatible populations are not known. Thus the extent to which levels of variability and inbreeding differ among Antirrhinum populations is not known. In order to address this issue we isolated nine Antirrhinum nuclear microsatellite loci. In contrast to several nuclear genes that show low levels of sequence variation, six of the microsatellite loci indicate high levels of variability within and between Antirrhinum species. The highly self-compatible Antirrhinum majus ssp. cirrhteum population has high levels of variability and no significant deviation from Hardy–Weinberg equilibrium, suggesting substantial rates of outcrossing.

The mating system in plants is determined by many factors, including features of the reproductive system, such as self-incompatibility mechanisms and protandry (i.e., the amount of time separating anther dehiscence and the start of stigma exertion) in hermaphroditic species, pollinator behavior, selective abortion by maternal regulation of seed quality, flowering pheno-ology (i.e., variation in floral display and structure), and population density (Shaanker et al. 1988; Marshall and Folsom 1991). The mating system affects the distribution of genetic variability, both within and between populations. For several reasons, highly inbreeding populations are expected to have low levels of variability relative to closely related outcrossing populations.

Inbreeding reduces the effective population size (Pollak 1987) and lowers effective rates of recombination due to the rarity of heterozygous individuals. Reduced recombination is associated with an increased effect of adaptive gene substitutions on neutral variability at linked sites (i.e., hitchhiking; Maynard Smith and Haigh 1974) and an increased effect of selection against deleterious alleles on neutral variation at linked sites (i.e., background selection; Charlesworth et al. 1993). Both processes tend to reduce neutral variability (reviewed in Charlesworth and Charlesworth 1998). Also, polymorphisms maintained by overdominance in outcrossing populations tend to be lost under inbreeding (Charlesworth and Charlesworth 1995; Kimura and Ohta 1971). In addition to these nonneutral effects, population structure has also been suspected to affect inbreeders. When selling species are more likely to occur in metapopulations with high rates of extinction, this will also contribute to lower levels of variability in selling populations (Barton and Whitlock 1997; Wade and McCauley 1988).

These theoretical predictions have been verified to a large extent by allozyme data, which consistently show higher levels of within-population variability in outcrossing than in selfing populations (Brown 1979; Hamrick and Godt 1990, 1996; Schoen and Brown 1991). While sequence variation data are still scarce, the available reports show the expected pattern of reduced diversity in inbreeders (Awadalla and Ritland 1997; Dvorak et al. 1998; Liu et al. 1998, 1999; Stephan and Langley 1998; Savolainen et al. 2000).

Recently several populations and species of Antirrhinum were characterized for their percentage of autogamy and self-fertility, and large variation was observed (Vieira 2000). However, the actual outcrossing rate is not known for self-compatible populations. In a recent attempt to correlate sequence variability with mating system, nuclear genes of the cycloidea and fili gene families were sequenced (Vieira and Charlesworth 2001a; Vieira et al. 1999). The low levels of sequence polymorphism observed in these studies made it difficult to correlate sequence variation with reproductive system. Furthermore,

References


Received April 23, 2001
Accepted December 31, 2001

Corresponding Editor: Leif Andersson