Size of Donor Chromosome Segments Around Introgressed Loci and Reduction of Linkage Drag in Marker-Assisted Backcross Programs

Frédéric Hospital

Station de Génétique Végétale, INRA/UPS/INAPG, 91190 Gif sur Yvette, France

Manuscript received April 18, 2000
Accepted for publication April 30, 2001

ABSTRACT

This article investigates the efficiency of marker-assisted selection in reducing the length of the donor chromosome segment retained around a locus held heterozygous by backcrossing. First, the efficiency of marker-assisted selection is evaluated from the length of the donor segment in backcrossed individuals that are (double) recombinants for two markers flanking the introgressed gene on each side. Analytical expressions for the probability density function, the mean, and the variance of this length are given for any number of backcross generations, as well as numerical applications. For a given marker distance, the number of backcross generations performed has little impact on the reduction of donor segment length, except for distant markers. In practical situations, the most important parameter is the distance between the introgressed gene and the flanking markers, which should be chosen to be as closely linked as possible to the introgressed gene. Second, the minimal population sizes required to obtain double recombinants for such closely linked markers are computed and optimized in the context of a multigeneration backcross program. The results indicate that it is generally more profitable to allow for three or more successive backcross generations rather than to favor recombinations in early generations.

IN a backcross breeding program aimed at introgressing a gene from a "donor" line into the genomic background of a "recipient" line, molecular markers could be used to assess the presence of the introgressed gene ("foreground selection"; Tanksley 1983; Melchinger 1990; Hospital and Charcosset 1997) and/or to accelerate the return to the recipient parent genotype at other loci ("background selection"; Tanksley 1983; Young and Tanksley 1989b; Hillel *et al.* 1990). The efficiency of background selection was demonstrated by theoretical (*e.g.*, Hospital *et al.* 1992; Visscher *et al.* 1996) as well as experimental (*e.g.*, Ragot *et al.* 1995) results.

As emphasized by Young and Tanksley (1989b), after a few backcross generations with no background selection on markers, most of the donor genes still segregating in the population are found on the chromosome carrying the introgressed gene (carrier chromosome) and more precisely on the intact chromosomal segment of donor origin ("donor segment") dragged along around this gene (Hanson 1959; Stam and Zeven 1981; Naveira and Barbadilla 1992). Hence, an important objective of background selection should be the reduction of the length of this donor segment. Obviously, background selection on noncarrier chromosomes is also important, though this selection is more efficient in late backcross generations (Hospital *et al.* 1992).

Gene introgression through recurrent backcrossing can be used in various circumstances: (i) in plant or animal breeding to improve the agronomic value of a commercial strain by introgressing a mono- or oligogenic trait (typically, a resistance trait) from a wild relative or from another—less productive—strain; (ii) to transfer a transgenic construction from one (transformed) strain to another (nontransformed) strain; or (iii) to construct near-isogenic or congenic lines, *e.g.*, for the detection of quantitative trait loci (QTL) and/or the validation of candidate genes for such QTL. In examples (i) and (ii), a drastic reduction of the length of the donor segment surrounding the introgressed gene is important if undesirable genes are located close to the introgressed gene, as might be the case if the donor strain is a wild genetic resource. If introgression takes place between two commercial strains, then a drastic reduction of the length of the donor segment is not always important. In example (iii), such a reduction is always important. In cases where the reduction is important, one would like the donor segment remaining in the backcross progenies at the end of the program to be as short as possible. This article examines background selection against the donor segment on the carrier chromosome.

The article is divided into two parts. In the first part, I derive various statistics describing the length of the intact segment between the gene of interest and a flanking marker, as a function of markers' positions and number of backcross generations performed. These are original analytical results relevant not only to marker-

Address for correspondence: Station de Génétique Végétale, INRA/UPS/INAPG, Ferme du Moulon, 91190 Gif sur Yvette, France.
E-mail: fred@moulon.inra.fr

assisted backcross programs but also to many studies related to introgression. In the second part, I study the minimal population sizes that are necessary to obtain a selection objective, as a function of markers' positions and number of backcross generations performed. Here, selection is applied on two flanking markers, one on each side of the introgressed gene, and the selection objective is to obtain a backcross progeny that carries the donor allele at the locus of interest and recipient alleles at both flanking markers (such an individual genotype is called "double recombinant" herein, regardless of whether the two recombination events took place at the same or at different generations).

This simple selection scenario was chosen as a case study for several reasons. First, it is of direct practical interest because it is often used in real backcross breeding programs, for example, in plant breeding. Second, it is the simplest case study that permits the investigation of the effects of three main parameters: the positions of the flanking markers, the number of backcross generations performed, and the number of individuals genotyped at each generation. Third, the results derived in this context can be used to evaluate the efficiency of more complex selection scenarios (*e.g.*, selection on several flanking markers on each side of the introgressed gene, which is also addressed here).

Questions relevant to the study of the efficiency of such a breeding scheme are as follows: (i) What is the efficiency of marker-assisted selection for the reduction of the donor segment length—*i.e.*, what is the length of this segment among double recombinants; (ii) what is the best position of the flanking marker to reach a given efficiency; (iii) how many individuals should be genotyped at each generation; and (iv) how many successive backcross generations should be performed?

DEFINITIONS

I consider a backcross breeding program aimed at introgressing a gene from a "donor" line into the genomic background of a "recipient" line. The program starts from an F_1 hybrid between two homozygous parental lines (generation $t = 0$). I assume that the parental lines carry different alleles at each locus. At each backcross (BC) generation (BC_t, $t \geq 1$), only the genes carried by the chromosome inherited from the backcrossed parent are segregating. Thus, for simplicity I refer only to the haploid genotypes (haplotypes) on that chromosome and state that an individual "is of donor type" at a locus if this individual is in fact heterozygous donor/recipient at that locus or "is of recipient type" if in fact homozygous recipient/recipient at that locus.

I assume here that recombination is without interference and use the Haldane mapping function, giving the relationship between recombination rate r and corresponding map distance l as

$$r = r_{[l]} = \tfrac{1}{2}(1 - e^{-2l}), \qquad (1)$$

where l is in morgans. For convenience, I use morgans throughout in the analytical derivations and convert into centimorgans in the numerical applications (in tables and figures).

Let T be the locus of the introgressed gene (or "target" gene). I assume that T is flanked by two marker loci M_1 and M_2, one on each side. At each generation individuals carrying the donor allele at locus T can be identified. If the introgressed gene is identified unambiguously by its phenotype, or by its genotype, or by the genotype of an intragenic marker (*e.g.*, in the case of a transgene), then I define l_1 (respectively l_2) as the real distance from the introgressed locus T to the flanking marker M_1 (respectively M_2) on one side and L_1 (respectively L_2) as the real distance from the introgressed locus T to the chromosome end on the same side. If the introgressed gene is identified through markers (*viz.* "foreground selection markers," different from the "background selection markers" M_1 and M_2, and closer to the introgressed locus), then l_1 (respectively l_2) is the distance from the outermost foreground selection marker locus on one side of T to the background selection marker locus M_1 (respectively M_2) on the same side, and L_1 (respectively L_2) is the distance from this outermost foreground selection marker locus to the chromosome end on the same side. In any case, the only markers considered hereafter are the background selection markers M_1 and M_2. The efficiency of foreground selection was investigated elsewhere (Melchinger 1990; Hospital and Charcosset 1997) and is not considered here.

SIZE OF DONOR CHROMOSOME SEGMENTS AROUND INTROGRESSED LOCI UNDER MARKER-ASSISTED SELECTION

In this article, the efficiency of marker-assisted selection is evaluated by its ability to reduce the length of the intact donor chromosome segment dragged along around locus T (donor segment). Assuming no interference, recombination events on each side of the introgressed locus are independent and can be treated separately. For simplicity I consider in this section only the length of the donor segment on one side of the introgressed locus, where marker M (standing for either M_1 or M_2) is at distance l (standing for either l_1 or l_2) and the chromosome end at distance L (standing for either L_1 or L_2).

If the introgressed gene is identified unambiguously (see definitions) as is assumed here, then the total length of the donor segment is simply the sum of lengths on both sides. If the introgressed gene is identified through foreground selection markers (see definitions), then the donor genome within foreground selection should also be taken into account in the computation of total donor segment length. If foreground selection markers are located close to each other, as is generally recommended to provide a good control of the introgressed gene (Hospital and Charcosset 1997), then the total length of the intact donor segment is approximately equal to the sum of lengths on both sides plus the distance between those foreground selection markers. If foreground selection markers are farther apart, then slightly more complex calculations taking those markers into account are required. This was not done here for the sake of simplicity and generality. Thus, strictly speaking, the calculations in this section provide the length of the intact donor segment on one side of a locus (either the introgressed locus itself or the outermost foreground selection marker locus), which carries a donor allele, given that on this side another locus (the flanking marker M) at distance l carries a recipient allele. The case where the marker carries a donor allele is addressed separately. This makes the results of general interest, not only in the case of marker-assisted backcross programs but also in any case where one desires to estimate the genomic composition of chromosomal segments flanked by loci of known genotypes in backcross programs.

The expected length of the donor segment, without background selection on markers, was first derived by Hanson (1959) and clearly revisited by Naveira and Barbadilla (1992), who also provided the corresponding variance. Note, however, that another possible measure of the efficiency of selection is the *total* proportion of donor genes on the carrier chromosome (either on the intact segment linked to T or on noncontiguous blocks of genes elsewhere on the carrier chromosome). This was computed by Stam and Zeven (1981) for the case of no background selection. This measure is not considered here: I focus only on the length of the intact donor segment linked to T. But in any case, numerical comparison of the results of Naveira and Barbadilla (1992) and Stam and Zeven (1981) indicates that the vast majority of donor genes on the carrier chromosome are located on the intact segment, even without background selection.

Without background selection on the marker, let $X(t)$ be the random variable corresponding to the length of intact donor segment on one side of the introgressed locus at generation BC$_t$. From Naveira and Barbadilla (1992), the probability density function (PDF) $f_t(x)$ for $X(t)$ is

$$f_t(x) = te^{-tx}, \tag{2}$$

the mean of $X(t)$ is

$$E_X(t) = 1/t(1 - e^{-tL}), \tag{3}$$

where L is the distance to the chromosome end, and the variance of $X(t)$ is

$$\sigma_X^2(t) = 1/t^2\{1 - e^{-tL}(2tL + e^{-tL})\}. \tag{4}$$

Here, I derive similar results in the case where background selection on the flanking marker M at distance l is applied. Define the following variables:

t_1, generation (BC$_{t_1}$) at which the marker is observed to be of recipient type;
t_2, generation (BC$_{t_2}$) at which the length of the donor segment is observed;
$Y(t_1, t_2)$, random variable, length of donor segment at t_2 given that the marker is of recipient type at t_1;
$g_{t_1,t_2}(x)$, PDF of $Y(t_1, t_2)$ at x;
$Y(t_1) = Y(t_1, t_1)$;
$g_{t_1}(x) = g_{t_1,t_1}(x)$;
$Y^*(t_1, t_2)$, random variable, length of donor segment at t_2 given that the marker is of recipient type "for the first time" at t_1 (*i.e.*, the marker was of donor type for $t < t_1$);
$g^*_{t_1,t_2}(x)$, PDF of $Y^*(t_1, t_2)$ at x;
$P_M(t_1)$, probability that the marker is of recipient type at t_1;
$P^*_M(t_1)$, probability that the marker is of recipient type for the first time at t_1;
$E_Y(t_1, t_2)$ and $E_{Y^*}(t_1, t_2)$, means of the random variables Y and Y^*, respectively;
$\sigma_Y(t_1, t_2)$ and $\sigma_{Y^*}(t_1, t_2)$, standard deviations (*i.e.*, square roots of the variances) of the random variables.

In the following I refer, if need be, either to the probability that a crossover occurs or to the probability that a recombination occurs. Under the assumption of no interference, the positions of crossovers along a chromosome follow a Poisson process. The probability that a crossover occurs in an infinitely small interval of size dx is equal to dx. The probability that no crossover occurs in an interval of size l is equal to e^{-l}, etc. The probability that a recombination occurs in an interval (strictly speaking, between two edges of an interval) of size l is given by (1). An odd number of crossovers occurring in an interval provides recombination, and an even number of crossovers provides no recombination.

The intact donor segment is bounded by locus T on one end and by the location of the closest crossover on the other end. If several successive generations are considered, then the latter is the closest crossover among all crossovers that took place at different generations. The generation at which this bounding crossover took place is denoted t_{co}.

Single-generation information: I first study the distribution of the length of the intact donor segment in the most simple situation where all the information available was obtained at a single generation t_1. The corresponding random variable is $Y(t_1) = Y(t_1, t_1)$.

The probability that the marker is of recipient type at t_1 is

$$P_M(t_1) = 1 - (1 - r_{[l]})^{t_1}. \tag{5}$$

Let x be any chromosomal position between the locus T and the marker ($0 < x < l$). Extending the rationale of Naveira and Barbadilla (1992), the probability that at generation t_1 the length of the segment is x and the marker is of recipient type is decomposed as follows:

In the interval $]0, x]$, for the length of the intact segment to be x two conditions are required: (i) At a given generation t_{co} $(1 \leq t_{co} \leq t_1)$, a crossover must have occurred exactly in the infinitely small interval $]x, x + dx]$ (probability dx) and no crossover must have occurred in the interval $]0, x]$ (probability e^{-x}); (ii) at any of the remaining generations t $(1 \leq t \leq t_1;\ t \neq t_{co})$, no crossover must have occurred in the interval $]0, x]$ (probability $e^{-(t_1-1)x}$). The probability for the interval $]0, x]$ is then

$$e^{-t_1 x}dx. \tag{6}$$

In the interval $]x, l]$, the only possibility for the marker *not* to be of recipient type at generation t_1 is that a recombination occurred in the interval $]x, l]$ at exactly the same generation t_{co} as above (probability $r_{[l-x]}$), and no recombination occurred in the interval $]x, l]$ at any of the $(t_1 - 1)$ remaining generations (probability $(1 - r_{[l-x]})^{(t_1-1)}$). For the interval $]x, l]$, the probability that the marker *is* of recipient type at generation t_1 is then

$$1 - r_{[l-x]}(1 - r_{[l-x]})^{(t_1-1)}. \tag{7}$$

Another demonstration of (7) is provided in APPENDIX A.

Assuming no interference, the overall conditional probability is obtained by multiplying (6) by (7), combining for any $t_{co} \in [1, t_1]$ (*i.e.*, multiplying by t_1), and finally dividing by (5). Mathematically speaking, this probability is $\Pr(x < Y \leq x + dx)$. The PDF $g_{t_1}(x)$ for $Y(t_1)$ is then simply obtained by differentiating with respect to x (*i.e.*, dropping the term in dx):

$$g_{t_1}(x) = g_{t_1,t_1}(x) = \frac{1 - r_{[l-x]}(1 - r_{[l-x]})^{(t_1-1)}}{1 - (1 - r_{[l]})^{t_1}} t_1 e^{-t_1 x}$$

$$= \frac{2^{t_1} - (1 - e^{-2(l-x)})(1 + e^{-2(l-x)})^{(t_1-1)}}{2^{t_1} - (1 + e^{-2l})^{t_1}} t_1 e^{-t_1 x}. \tag{8}$$

This is a much simpler demonstration than that of FRISCH and MELCHINGER (2001), because here I use either probability of recombination or probability of crossover when appropriate, although the mathematical result is identical (except the one expressed in terms of hyperbolic trigonometric functions in FRISCH and MELCHINGER 2001). Moreover, I provide detailed numerical applications and discussion below.

An example of the distribution of $g_{t_1}(x)$ is given in Figure 1 for a marker located at distance $l = 20$ cM from locus T, at different BC_{t_1} generations. It is seen that the distribution of g_{t_1} is of decreasing exponential shape and skewed toward low x values. For a given marker position, both skewness (asymmetry) and kurtosis (peakedness) of the distributions increase in advanced backcross generations. For other marker positions (results not shown), the number and the effects of t_2 on the shape of the distributions are the same, except that for a given backcross generation both skewness and kurtosis are reduced for markers closer to T. Conversely both skewness and kurtosis are increased for markers farther apart from locus T.

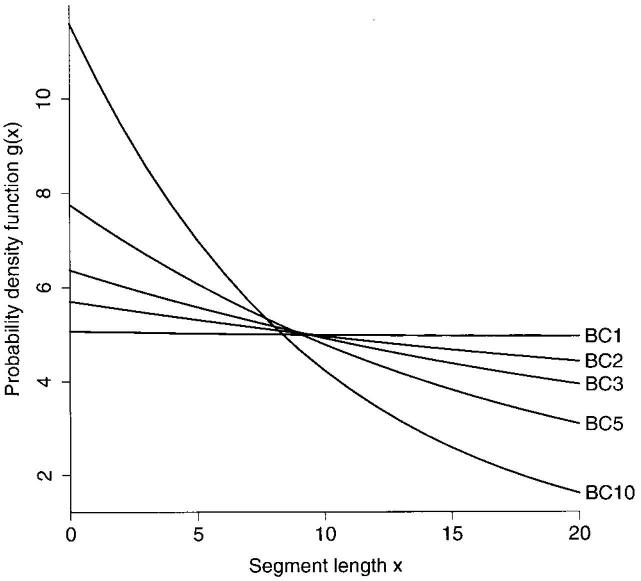

FIGURE 1.—Probability density function (PDF) of intact donor segment length on one side of locus T with single-generation information. Marker is at distance $l = 20$ cM. Abscissa: segment length x (cM). Ordinate: PDF $g_{t_1}(x)$ at different BC_{t_1} generations, with $t_1 = 1, 2, 3, 5$, or 10. See text for details.

The mean $E_Y(t_1)$ of $Y(t_1)$ is then simply obtained from (8) by integrating for x along the chromosome up to the marker position

$$E_Y(t_1) = \int_0^l x g_{t_1}(x)\,dx, \tag{9}$$

which gives

$$E_Y(t_1) = \frac{1/t_1 2^{t_1}\{1 - (1 + t_1 l)e^{-t_1 l}\} + t_1\Sigma_{k=0}^{t_1-1}\binom{t_1 - 1}{k}\{u[t_1, k+1] - u[t_1, k]\}}{\{2^{t_1} - (1 + e^{-2l})^{t_1}\}} \tag{10}$$

with

$$u[t_1, k] = \int_0^l x e^{-t_1 x} e^{-2k(l-x)}\,dx$$

$$= \begin{cases} \dfrac{1}{2} l^2 e^{-t_1 l} & \text{if } t_1 = 2k, \\[2ex] \dfrac{e^{l(t_1-2k)} - l(t_1 - 2k) - 1}{(t_1 - 2k)^2} e^{-t_1 l} & \text{otherwise.} \end{cases} \tag{11}$$

The standard deviation $\sigma_Y(t_1)$ of $Y(t_1)$ is obtained from

$$\sigma_Y^2(t_1) = \int_0^l x^2 g_{t_1}(x)\,dx - (E_Y(t_1))^2, \tag{12}$$

where $g_{t_1}(x)$ is obtained from (8) and $E_Y(t)$ from (10). A closed formula for (12) could be derived as was done above for E_Y, but the expression would be more complex and barely useful since numerical results can now be obtained directly from (12) with a mathematical software package [for example, many numerical results

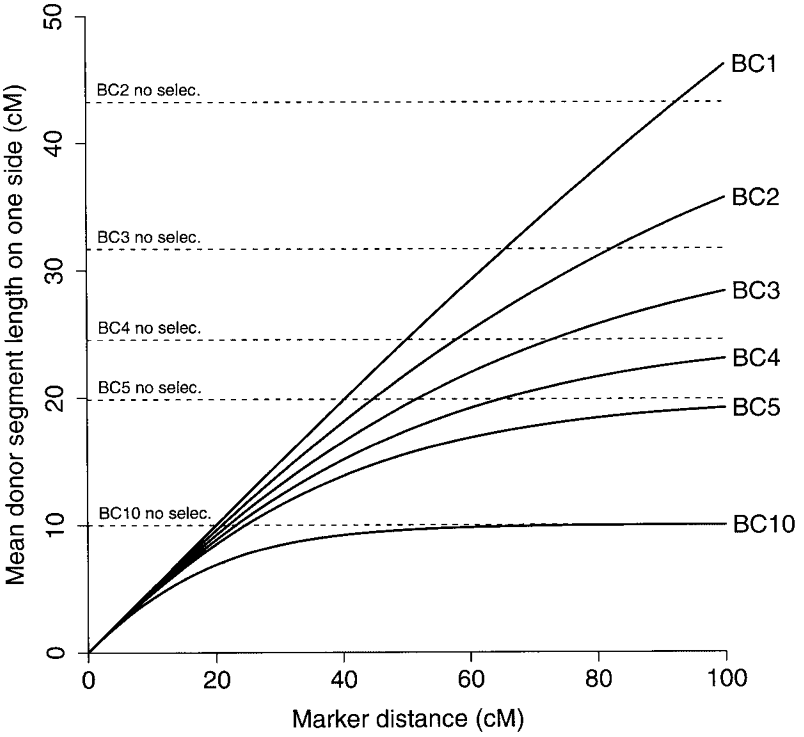

Figure 2.—The expected length of intact donor segment on one side of locus T with single-generation information. Abscissa: marker distance l (cM). Solid lines: $E_Y(t_1)$ (cM) at different BC_{t_1} generations, with t_1 = 1, 2, 3, 4, 5, or 10. Dotted lines: the expected lengths (cM) of donor segment without background selection from Equation 3 at the same generations for a chromosome end at distance 100 cM from locus T.

given in this article were obtained using *Mathematica* (Wolfram 1988)]. Other numerical results (not shown) indicate that the standard deviation σ_Y is generally of the same order as the mean E_Y, corresponding to quite large variances of donor segment length.

However, whereas the mean is always a meaningful parameter for any distribution, the standard deviation might not be the most appropriate parameter in the present case, given the shape of the PDF (Figure 1). With such distributions, parameters like quantiles are more appropriate. For example, I study the 9th decile defined as the threshold θ such that

$$\Pr(Y \leq \theta) = \int_0^\theta g(x)\,dx = 0.9. \qquad (13)$$

In the tables below, θ values were computed by solving (13) numerically.

I also computed the PDF, mean, and variance of intact segment length in the case where the marker M is of donor type. The corresponding results are given in appendix b. These results are not used in this article because I wish to focus on the optimization of marker-assisted selection, where the objective is that the marker is of recipient type. However, these results were derived for the sake of generality and could be useful in other contexts, *e.g.*, to compute "graphical genotypes" (Young and Tanksley 1989a). In the rest of this section, I refer only to the case where the marker is of recipient type.

Numerical values for the expected length of the donor segment on one side, $E_Y(t_1)$, computed using (10), are plotted in Figure 2 as a function of marker position l for different BC_{t_1} generations. Expected lengths of donor segment with no background selection (3) for the same generations are given in the graph for comparison, in the case of a chromosome end at distance 100 cM from locus T. Figure 2 shows that marker-assisted selection (solid lines) is obviously very efficient in reducing the length of the donor segment, compared to its expected value when no marker-assisted background selection is applied (dotted lines), except for markers far apart from locus T in late BC generations. Marker-assisted selection is more efficient as markers are closer to T. Note, however, that the values in Figure 2 are given for an individual that is known to be of recipient type at the marker. The probability of obtaining such an individual then depends on the population size. This is addressed later but obviously the closer the marker is to locus T, the larger the population size has to be.

The donor segment is bounded by the closest crossover position among all successive meioses. Thus, for a given marker position, the expected length of the donor segment should be smaller in advanced BC generations, because the accumulation of meioses can only reduce the segment length compared to its value in BC_1. But Figure 2 shows that the number of backcross generations (t_1) has a visible effect only for markers relatively far apart from locus T ($l \geq 20$ cM). For markers closer to T (<20 or 10 cM), the expected length of donor segment in advanced backcross generations (BC_5 or BC_{10}) is not visibly reduced below its value in BC_1, *i.e.*, $\sim l/2$. Hence, it is likely that a relatively large portion of the unwanted donor genome will remain segregating in the backcross progenies, even after several generations of marker-assisted selection.

Another way to evaluate the effect of the number of backcross generations (t_1) is to compute the minimum number of backcross generations needed to reduce the expected length of donor segment below a given threshold. This is done in Figure 3 where the threshold is expressed as a fraction c of the distance between T and the marker: minimal t_1 values such that $E_Y(t_1) \leq cl$ are given in function of l for $c = \frac{1}{2}, \frac{3}{8}, \frac{1}{4}$, or $\frac{1}{8}$. This reinforces the conclusions drawn from Figure 2: In BC$_1$, the expected segment length is approximately $l/2$ regardless of marker position. Reducing the expected segment length below $l/2$ requires unrealistically large numbers of backcross generations, unless the marker is quite far away from locus T.

I now investigate in more detail whether the accumulation of meioses could permit a better reduction of the donor segment length. If this has an important effect, it will provide an alternative to using closer markers at the expense of larger population sizes. The results in Figure 2 indicate that a moderate gain may be expected from advanced backcross generations, at least for distant markers. But, the only information considered so far in the calculations is that the marker is of recipient type at generation t_1 and the donor segment length is observed at the same generation t_1. This is not the most appropriate study of the effect of meiosis accumulation because, among the crossovers that take place between locus T and the marker, it does not permit one to distinguish between the (single) crossover that makes the marker return to recipient type and (possibly) other crossovers that would reduce donor segment length without affecting the genotype at the marker. To investigate this in more detail, slightly more complex situations must be considered.

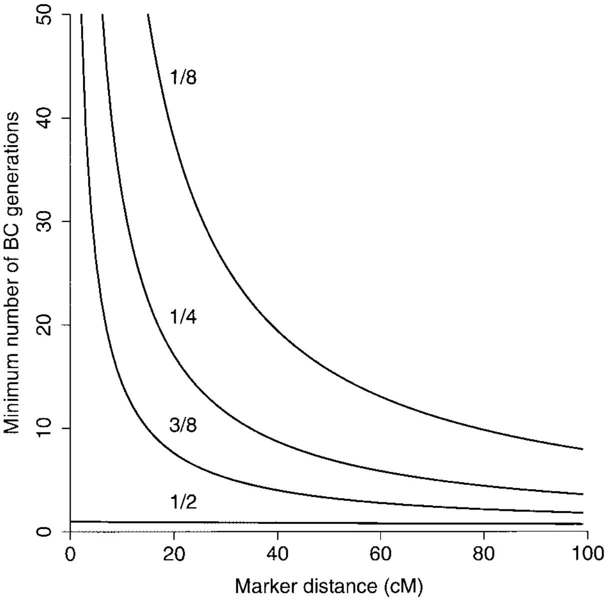

Figure 3.—Minimum number of backcross generations needed to reduce the expected length of the donor segment on one side of locus T below a given fraction of marker distance. Abscissa: marker distance l (cM). Ordinate: minimum number of backcross generations. Solid lines correspond to different threshold values expressed as a fraction of l.

The first situation considered is that when, after the (recipient) genotype of the marker has been observed at generation t_1, the backcross breeding program is nevertheless pursued until generation t_2 ($t_2 \geq t_1$), and the length of donor segment is observed at t_2. From generation t_1 to t_2, selection on the marker is no longer necessary, since the marker is then fixed for the recipient type allele. Only foreground selection for the introgressed gene is necessary. Still, additional gain in the reduction of the donor segment could be expected from crossovers taking place in this heterozygous part of the genome during meioses from t_1 to t_2. I now evaluate the amount of this possible additional gain.

The random variable corresponding to the length of donor segment at generation t_2, given that the marker was observed to be of recipient type at t_1 ($t_1 \leq t_2$), is $Y(t_1, t_2)$. The PDF g_{t_1,t_2} is calculated following the same rationale as above for g_{t_1}. Only the recombination events taking place at BC generations up to t_1 affect the marker genotype, since the marker is fixed for recipient type after t_1.

The probability that the marker is of recipient type at t_1 is given by (5). For the interval $]0, x]$, the probability is simply $e^{-t_2 x}\,dx$ similar to (6). For the interval $]x, l]$, two cases must be considered. If the crossover in the infinitely small interval $]x, x + dx]$ took place before t_1 ($1 \leq t_{co} \leq t_1$; t_1 possibilities), then the probability for the interval $]x, l]$ is the same as for g_{t_1} in (7). If the crossover in the infinitely small interval $]x, x + dx]$ took place after t_1 ($t_1 < t_{co} \leq t_2$; $t_2 - t_1$ possibilities) then, for the marker to be of recipient type at t_1, at least one recombination must have occurred in the interval $]x, l]$ at some BC generation before t_1 [probability $1 - (1 - r_{[l-x]})^{t_1}$].

Finally, the PDF g_{t_1,t_2} for $Y\,(t_1, t_2)$ is

$$g_{t_1 t_2}(x) = \frac{t_1\{1 - r_{[l-x]}(1 - r_{[l-x]})^{(t_1-1)}\} + (t_2 - t_1)\{1 - (1 - r_{[l-x]})^{(t_1)}\}}{e^{t_2 x}\{1 - (1 - r_{[l]})^{t_1}\}}. \tag{14}$$

The corresponding mean $E_Y(t_1, t_2)$, variance $\sigma^2_Y(t_1, t_2)$, and ninth decile are defined similarly to (9), (12), and (13), respectively. These were computed numerically.

Multiple-generation information: So far, the information considered in the derivations for $Y(t_1)$ or $Y(t_1, t_2)$ was that a recombination occurred between T and M at some generation $t \in [1, t_1]$, but the exact generation at which this recombination took place was considered as unknown. However, in practical situations, continuous selection on background markers is applied, as is generally recommended. In such cases, the marker genotypes are observed at every generation $t \in [1, t_1]$, to pick out recombinant individuals as soon as possible.

The exact generation at which the recombination took place between T and M is then known. To evaluate the efficiency of marker-assisted selection in such situations, different events and associated probabilities must be considered. The corresponding variables are indicated by parameters with asterisks.

The probability that the marker is of recipient type for the first time at generation t_1 (*i.e.*, the marker is of donor type for any $t < t_1$) is

$$P^*_M(t_1) = r_{[l]}(1 - r_{[l]})^{t_1-1}. \tag{15}$$

The random variable corresponding to the length of the donor segment at generation t_2, given that the marker was of recipient type for the first time at generation t_1 ($t_1 \leq t_2$), is $Y^*(t_1, t_2)$. Note that obviously Y and Y^* are identical at $t_1 = 1$. Following the same rationale as above, we need to focus only on recombination events in the chromosomal segment $]x, l]$. Three cases must be considered, t_{co} being the generation at which the crossover bounding the donor segment has occurred.

If $t_{co} < t_1$ ($t_1 - 1$ possibilities), then at generations t ($t < t_1$; $t \neq t_{co}$), for the marker to be of donor type, no recombination must have occurred in the interval $]x, l]$. At generation $t = t_{co}$, the crossover occurred in the infinitely small interval $]x, x + dx]$ so, for the marker to remain of donor type, a recombination must have occurred in the interval $]x, l]$. At generation $t = t_1$, for the marker to be of recipient type, a recombination must have occurred in the interval $]x, l]$.

If $t_{co} = t_1$ (one possibility), then at generations t ($t < t_1$; $t < t_{co}$), for the marker to be of donor type, no recombination must have occurred in the interval $]x, l]$. At generation $t = t_{co} = t_1$, the crossover occurred in the infinitely small interval $]x, x + dx]$ so, for the marker to be of recipient type, no recombination must have occurred in the interval $]x, l]$.

Finally, if $t_{co} > t_1$ ($t_2 - t_1$ possibilities), then at generations t ($t < t_1$), for the marker to be of donor type, no recombination must have occurred in the interval $]x, l]$. At generation $t = t_1$, for the marker to be of recipient type, a recombination must have occurred in the interval $]x, l]$.

Combining for all possible t_{co} values, the PDF for $Y^*(t_1, t_2)$ is then

$$g^*_{t_1,t_2}(x) = \frac{(t_1 - 1)r^2_{[l-x]}(1 - r_{[l-x]})^{(t_1-2)} + (1 - r_{[l-x]})^{(t_1)} + (t_2 - t_1)r_{[l-x]}(1 - r_{[l-x]})^{(t_1-1)}}{e^{t_2 x}r_{[l]}\ (1 - r_{[l]})^{(t_1-1)}}. \tag{16}$$

The corresponding mean $E_{Y*}(t_1, t_2)$, variance $\sigma^2_{Y*}(t_1, t_2)$, and ninth decile for $Y^*(t_1, t_2)$ are defined as before and were computed numerically. Note, however, that when the marker genotype and the length of donor segment are observed at the same generation ($t_1 = t_2$) the PDF of $Y^*(t_1) = Y^*(t_1, t_1)$ is

$$g^*_{t_1} = g^*_{t_1,t_1}, \tag{17}$$

and, in that particular case, the mean simplifies to

$$E_{Y*}(t_1) = \int_0^l x g^*_{t_1}(x)\,dx = 1/t \coth l\{1 - (\mathrm{sech}\ l)^t\}, \tag{18}$$

where coth and sech are the hyperbolic tangent and the hyperbolic secant functions, respectively.

Generally, the following relationship holds between g and g^*,

$$\sum_{k=1}^{k=t_1} P^*_M(k)g^*_{k,t_2}(x) = P_M(t_1)g_{t_1,t_2}(x), \tag{19}$$

giving

$$P^*_M(t_1)g^*_{t_1,t_2}(x) = P_M(t_1)g_{t_1,t_2}(x) - P_M(t_1 - 1)g_{t_1-1,t_2}(x). \tag{20}$$

Numerical applications: Numerical applications of the above derivations for the mean (E) and ninth decile (θ) of $Y^*(t_1, t_2)$ and $Y(t_1, t_2)$ are given in Table 1 for a marker distance of 50 cM from the introgressed gene and in Table 2 for a distance of 20 cM.

Setting the marker position at 50 cM (Table 1) is not the most realistic situation. However, it makes it easier to study the effects of the various parameters, because the results are more contrasting than for a shorter marker distance. It is thus given as an illustrative example. Results for a more realistic marker position (20 cM) are also provided (Table 2).

The results in Table 1 for multiple-generation information ($Y^*(t_1, t_2)$) highlight the effects of the two parameters: t_1, the BC generation at which the recombination occurred between the locus T and the marker, and t_2, the BC generation at which the donor segment length is observed (or the total duration of the backcross program). Note that t_1 and t_2 values shown in the tables are not continuous (1, 2, 3, 5, 10).

The results for Y^* in Table 1 at $t_1 = t_2$ indicate that the expected donor segment length is shorter in the case of a later recombination between the locus T and the marker. For example, if the marker is of recipient type for the first time in BC_1 and the donor segment is also observed in BC_1 ($t_1 = t_2 = 1$), then the expected segment length is 24.5 cM (Table 1), while if the marker is of recipient type for the first time in BC_3 only and the segment length is also observed in BC_3 ($t_1 = t_2 = 3$), then this length is 21.8 cM (Table 1). This is so because in the case of a distant marker ($l = 50$ cM in Table 1) double crossover events between T and the marker may occur at relatively high frequency before the recombination between T and the marker takes place (such double crossovers reduce donor segment length while the marker remains of donor type). Obviously, the frequency of double crossovers is much lower in the case of a shorter marker distance (*e.g.*, 20 cM, Table 2); thus values for Y^* at $t_1 = t_2$ in Table 2 are then hardly reduced with increasing t_1.

TABLE 1

Efficiency of marker-assisted selection when marker is at l = 50 cM from the introgressed gene

	Observation									
	t_2 = 1		t_2 = 2		t_2 = 3		t_2 = 5		t_2 = 10	
Recombination	E	θ	E	θ	E	θ	E	θ	E	θ
a. Multiple-generation information (Y*)										
t_1 = 1	24.5	(44.8)	20.9	(42.1)	18.1	(38.7)	14.0	(31.6)	8.6	(19.7)
t_1 = 2			23.1	(44.1)	19.8	(41.2)	15.1	(34.0)	9.0	(20.8)
t_1 = 3					21.8	(43.4)	16.3	(36.5)	9.5	(21.9)
t_1 = 5							19.5	(41.7)	10.6	(24.7)
t_1 = 10									15.1	(36.0)
b. Single-generation information (Y)										
t_1 = 1	24.5	(44.8)	20.9	(42.1)	18.1	(38.7)	14.0	(31.6)	8.6	(19.7)
t_1 = 2			21.8	(43.0)	18.8	(39.8)	14.4	(32.6)	8.7	(20.1)
t_1 = 3					19.4	(40.8)	14.8	(33.6)	8.9	(20.5)
t_1 = 5							15.6	(35.2)	9.2	(21.2)
t_1 = 10									9.6	(22.3)

Mean (E) and ninth decile (θ) of the distribution of intact donor segment length on one side of locus T observed at various BC generations t_2 given that the marker is of recipient type at generation t_1, in two situations: (a) The recombination between the introgressed gene and the marker took place exactly at t_1 (*i.e.*, the marker is of donor type at any $t < t_1$); or (b) the recombination took place at some generation $t \leq t_1$ (*i.e.*, the marker genotype at $t < t_1$ is unknown). See text for details.

However, the results in Table 1 indicate that the expected donor segment length is better reduced by crossovers occurring *after* the recombination between locus T and the marker occurred (*i.e.*, after the marker returned to recipient type). For example, if segment length is again observed in BC_3, but it is known that the marker was already of recipient type since generation BC_1 (*i.e.*, two additional BC generations were performed after the marker returned to recipient type), then the expected length of the donor segment is 18.1 cM (t_1 = 1, t_2 = 3, Table 1), compared to 21.8 cM for t_1 = 3. There is also a gain on the ninth decile θ (38.7 *vs.* 43.4 cM). In any case, the gain obtained by forcing early recombination between T and the marker is of moderate importance and would be obtained at the expense of increased population sizes (see next section).

Moreover, results for multiple-generation information (Y* values) are relevant to evaluate *a posteriori* the efficiency of selection once the BC program is completed and the genotypes of the individuals selected at

TABLE 2

Efficiency of marker-assisted selection when marker is at l = 20 cM from the introgressed gene

	Observation									
	t_2 = 1		t_2 = 2		t_2 = 3		t_2 = 5		t_2 = 10	
Recombination	E	θ	E	θ	E	θ	E	θ	E	θ
a. Multiple-generation information (Y*)										
t_1 = 1	10.0	(18.0)	9.3	(17.6)	8.8	(17.2)	7.8	(16.2)	5.9	(13.3)
t_1 = 2			9.9	(17.9)	9.2	(17.6)	8.2	(16.6)	6.2	(13.8)
t_1 = 3					9.8	(17.9)	8.6	(17.1)	6.5	(14.3)
t_1 = 5							9.6	(17.8)	7.1	(15.3)
t_1 = 10									9.1	(17.6)
b. Single-generation information (Y)										
t_1 = 1	10.0	(18.0)	9.3	(17.6)	8.8	(17.2)	7.8	(16.2)	5.9	(13.3)
t_1 = 2			9.6	(17.8)	9.0	(17.4)	7.9	(16.4)	6.1	(13.6)
t_1 = 3					9.2	(17.5)	8.1	(16.6)	6.2	(13.8)
t_1 = 5							8.5	(17.0)	6.4	(14.2)
t_1 = 10									6.9	(15.1)

Mean (E) and ninth decile (θ) of the distribution of intact donor segment length for multiple- (Y*) or single- (Y) generation information. See Table 1 or text for details.

the various generations are known. It is not relevant to the *a priori* design of a program before it is started, because it would not make sense to design a program such that the recombination between the locus T and the marker takes place, for example, in BC_3 ($t_1 = 3$) and not in BC_2 or BC_1. What makes sense is to allow a double recombinant to be selected as soon as possible, while keeping population sizes within affordable limits. To do so, one would design a program such that recombination between the locus T and the marker M takes place by a given generation. In other words, to keep with the above example one would require that recombination take place in BC_3 *or* at any previous generation. In this case, single-generation information (Y) is relevant to predict *a priori* the efficiency of such a program.

The results for single-generation information (Y) in Table 1 indicate that the reduction of donor segment length obtained by forcing early recombination between T and the marker ($t_1 < t_2$) is even less important in the context of such an *a priori* prediction. For example, for a BC program designed to last at most three BC generations, if no other information is available (*i.e.*, the recombination between T and the marker took place in BC_3 or in any earlier generation), then the expected donor segment length is 19.4 cM ($t_1 = t_2 = 3$, Table 1). If it is known that the recombination took place in BC_1, and two additional BC generations were performed afterward, then the expected donor segment length in BC_3 is 18.1 cM as before ($t_1 = 1$; $t_2 = 3$, Table 1). Hence, the gain in the reduction of donor segment length provided by forcing early recombination is only 1.3 cM on average for a marker distance of 50 cM.

Additional BC generations also have a little impact on the distribution of Y values, as indicated by θ values in Table 1: If no additional information is available, at $t_1 = t_2 = 3$, 90% of Y values are <40.8 cM, while if it is known that the recombination took place in BC_1, at $t_1 = 1$ and $t_2 = 3$, 90% of Y values are <38.7 cM; *i.e.*, the gain is only 2.1 cM.

For a more realistic marker position ($l = 20$ cM, Table 2), the numerical results indicate that the gain in the reduction of donor segment length, provided by forcing early recombination between T and M, is even smaller than for a distant marker. For example, for a program designed to last at most three BC generations, forcing the recombination between the gene and the marker to take place in BC_1 would provide a gain of only $9.2 - 8.8 = 0.4$ cM ($t_2 = 3$, Table 2).

For even closer marker positions, ($l \leq 10$ cM, results not shown), this gain tends to zero and the results for Y values at $t_1 < t_2$ are then hardly different from Y values at $t_1 = t_2$. Hence, the results for these situations can be simply taken from Figure 2.

As a conclusion to this section, in theory an additional gain in the reduction of donor segment length is always expected from allowing additional backcrosses even after a recombination between the locus T and the marker is obtained. But, the amount of this gain depends on the distance between the locus T and the marker and tends to zero for realistic (short) marker distances ($l < 20$ cM). For such short distances, the BC generation at which the recombination between the locus T and the marker takes place (t_1) has little impact on donor segment length. Moreover, at short marker distances, the total duration of the program (t_2) also has little impact on donor segment length. Hence, in such cases, the donor segment length mostly depends on the position of the marker, and not on the number of BC generations performed. Overall, it is generally more efficient to use closer markers (reduce l), than to allow additional BC generations for more distant markers. For short marker distances, reducing l has more impact on the reduction of donor segment length than increasing t_2 (as was seen from Figure 2) or increasing $t_2 - t_1$ (see Table 2). However, it is important to note that the above results on the length of donor segment were derived conditionally on obtaining a recombinant genotype for either or both markers. Thus, the probability of obtaining such a recombinant was not taken into account in the optimization of the BC scheme. Yet, the number of BC generations does affect genotype probabilities in conjunction with the population size. Hence, the number of BC generations should be optimized with respect to the population sizes needed for obtaining double recombinants. This is addressed in the following section.

MINIMAL POPULATION SIZES

In the previous section I studied the length of donor segment among genotypes that are recombinant for either M_1 or M_2. Here, I focus on the probability of obtaining an individual that is double recombinant for markers M_1 and M_2 on both sides of locus T. In that case, even when assuming no interference, recombination on both sides of T cannot be treated separately.

In one single BC generation, the probability of double recombination is easy to calculate from the product of the probabilities of single recombinations on both sides of T. But, as noted by Young and Tanksley (1989b), for close markers the probability of double recombination is much lower than the probabilities of single recombinations. Hence, the population size needed to obtain a double recombinant in one single BC generation is much larger than twice the population size needed to obtain a single recombinant on one side. On the basis of this consideration, Young and Tanksley (1989b) proposed to work on two BC generations, selecting in the first generation a single recombinant on one side for the closest marker, then selecting in the second generation a single recombinant on the other side. Young and Tanksley did not compute the corresponding population sizes. The computation in the case of two BC generations is more complex than in the case of one single generation, because it must take into account all possible successions of recombination events leading to a double recombinant in BC_2. Moreover, this computation should be extended to any number of BC generations to design a BC scheme that allows a double recombinant genotype to be obtained while minimizing the total number of individuals genotyped during the entire BC scheme.

A mathematical solution to this problem was first provided by Hospital and Charcosset (1997). This solution was derived within the complex case of introgression of a QTL and with few applications. Moreover, it does not correspond to the most realistic strategy and needs slight modifications. Here, I derive a full theoretical treatment of the problem and provide comprehensive numerical applications, which can be used for the optimization of backcross schemes in plant breeding or any backcross scheme in which it is possible to retain a single progeny for reproduction at each generation.

Let r_1 and r_2 be the recombination rates corresponding to distances l_1 and l_2, respectively. Without loss of generality, I assume hereafter $l_1 \leq l_2$. The alleles at each locus are noted "0" for donor-type allele, and "1" for recipient type. Since the genotypes carrying a donor allele at locus T are the most interesting, I define only five genotypic classes at loci M_1TM_2: $G_1 = 101$, $G_2 = 100$, $G_3 = 001$, $G_4 = 000$, and $G_5 = {*}1{*}$, the latter referring to the four possible genotypes carrying a *recipient* allele at locus T, regardless of the markers.

At each generation t a total of n_t individuals (backcross progenies) are first screened for the presence of the donor allele at locus T and then possibly for the presence of the recipient alleles at markers M_1 and/or M_2. If no carrier of the donor allele at locus T is found in the population, then the backcross scheme is interrupted (failure). If one or more carriers are found, then among those a single individual is selected on the basis of its genotype at markers in the following order of priority: (1) G_1 (double recombinant); (2) G_2 (single recombinant); (3) G_3 (single recombinant); or (4) G_4 (nonrecombinant). Note that G_2 is selected prior to G_3 because I assume $l_1 \leq l_2$. The selected individual is then backcrossed to the recipient parental line to provide the next BC generation.

If the introgressed gene is identified unambiguously (see DEFINITIONS), then the probability of transmission to a backcross progeny of the donor allele at locus T is $P = \frac{1}{2}$. This is assumed here in the numerical applications, but, for the sake of generality, I keep the literal P in the theoretical derivations. If the introgressed gene is identified by flanking markers (foreground selection markers; see DEFINITIONS), then the probability of transmission of the donor allele at locus T is $<\frac{1}{2}$ and must be calculated from the probability of transmission of those foreground selection markers (see an example in Hospital and Charcosset 1997). In that case, l_1 and l_2 are the distances from the outermost foreground selection marker on each side of T to M_1 or M_2, respectively. P is then the probability of transmission of the foreground markers and not of the introgressed gene. The probability of transmission of the introgressed gene is lower than P, depending on the position of the foreground markers, and must be calculated conditionally to the presence of those markers (see Melchinger 1990; Hospital and Charcosset 1997).

For given markers' positions and total duration of the breeding scheme, the total cost of the experiment, which we want to minimize, depends directly on population sizes at each generation. In this context, optimal population sizes are then simply the minimal population sizes necessary to achieve the experiment successfully. I now calculate such minimal population sizes.

Analytical derivations: The recursion equations of Hospital and Charcosset (1997, Equations A.16 to A.21) were derived under the following strategy (strategy A): A backcross breeding program is intended to be performed during a total of t_2 generations. The final probability that the individual selected at generation t_2 is a double recombinant (G_1) is computed by recursion. Then, population sizes n_t at generations $t \leq t_2$ are computed such that this final probability is above a given threshold (99%). These calculations need slight improvements.

The recursions of Hospital and Charcosset (1997, Equations A.17 and A.18) consider the complex case where, if no double recombinant is found at a given generation, but at least one single recombinant is present for each side of the introgressed locus, then the selected individual is chosen at random among those single recombinants. Here, I need only the simpler case where among single recombinants the selection of the recombinant for the closest marker is favored (*i.e.*, M_1 assuming $l_1 \leq l_2$). Hence, under strategy A, the recursions should be computed as follows.

Let $\mathbf{h_t}$ be the column vector of the frequencies at generation t of the five genotypic classes G_i defined above. These frequencies are given by

$$\mathbf{h_t} = \mathbf{H} \cdot \mathbf{h_{t-1}}, \tag{21}$$

with

$$\mathbf{H} = \begin{bmatrix} P & r_2P & r_1P & r_1r_2P & 0 \\ 0 & (1-r_2)P & 0 & r_1(1-r_2)P & 0 \\ 0 & 0 & (1-r_1)P & (1-r_1)r_2P & 0 \\ 0 & 0 & 0 & (1-r_1)(1-r_2)P & 0 \\ 1-P & 1-P & 1-P & 1-P & 1 \end{bmatrix} \tag{22}$$

and

$$\mathbf{h_0'} = \{0, 0, 0, 1, 0\}. \tag{23}$$

Let $a_t[i]$ be the probability that with strategy A the individual selected at generation t is of genotype G_i. Let

$\mathbf{a}'_t = \{a_t[i]\}_{1\leq i\leq 5}$ be the vector of these probabilities. We have

$$\mathbf{a}_t = \mathbf{A}_t \cdot \mathbf{a}_{t-1} \tag{24}$$

with

$$\mathbf{a}_0 = \mathbf{h}_0. \tag{25}$$

The element $A_t[i, j]$ of recursion matrix $\mathbf{A}_t$ at line i and column j is obtained by

$$A_t[i, j] = \left(1 - \sum_{k=1}^{i-1} H[k, j]\right)^{n_t} - \left(1 - \sum_{k=1}^{i} H[k, j]\right)^{n_t}, \tag{26}$$

where $H[k, j]$ is the element of matrix $\mathbf{H}$ at line k and column j.

Finally,

$$\mathbf{A}_t = \begin{bmatrix} 1-(1-P)^{n_t} & 1-(1-r_2P)^{n_t} & 1-(1-r_1P)^{n_t} & 1-(1-r_1r_2P)^{n_t} & 0 \\ 0 & (1-r_2P)^{n_t}-(1-P)^{n_t} & 0 & (1-r_1r_2P)^{n_t}-(1-r_1P)^{n_t} & 0 \\ 0 & 0 & (1-r_1P)^{n_t}-(1-P)^{n_t} & (1-r_1P)^{n_t}-(1-s_{12}P)^{n_t} & 0 \\ 0 & 0 & 0 & (1-s_{12}P)^{n_t}-(1-P)^{n_t} & 0 \\ (1-P)^{n_t} & (1-P)^{n_t} & (1-P)^{n_t} & (1-P)^{n_t} & 1 \end{bmatrix} \tag{27}$$

with the notation $s_{12} = r_1 + r_2 - r_1r_2$ introduced just to save space.

In the case of constant population sizes ($n_t = n$, $\forall t$), the elements of vector $\mathbf{a}_t$ can be obtained directly as a function of n and t after transformation of matrix $\mathbf{A}_t$ to diagonal form (see an example in Visscher and Thompson 1995). Yet, Hospital and Charcosset (1997) showed that better results are obtained with variable population sizes. In that case, $\mathbf{a}_t$ can be obtained only by recursion.

Besides other possible considerations (*e.g.*, selection on noncarrier chromosomes), the results of the previous section on segment length indicate that, once the marker is of recipient type, little gain on the reduction of the donor segment is expected from performing additional backcross generations, in particular for close markers. Hence, strategy A might not be the most realistic. I then consider a slightly different strategy (strategy B), where, if a double recombinant G_1 is found at a given generation $t < t_2$, then the backcross program is interrupted (success) rather than pursued until the initially defined generation t_2. Also, the process is interrupted (failure) if no carrier of the introgressed gene is found in the population at any $t < t_2$.

For strategy B, I define a new vector of probabilities $\mathbf{b}'_t = \{b_t[i]\}_{1\leq i\leq 5}$, such that $b_t[1] = \beta_t$ is the probability that an individual of genotype G_1 is selected at generation t but not at any previous generation; β_t is then the probability of success at generation t; conversely, $b_t[5] = \gamma_t$ is the probability that no carrier of the introgressed gene is found in the population at generation t; γ_t is then the probability of failure of the BC scheme at generation t; and finally $b_t[i]$ for $2 \leq i \leq 4$ is the probability that the individual selected at generation t is of genotype G_i, given that the individual selected at previous generation is of genotype G_j ($2 \leq j \leq 4$).

Again, we have

$$\mathbf{b}_t = \mathbf{B}_t \cdot \mathbf{b}_{t-1} \tag{28}$$

with

$$\mathbf{b}_0 = \mathbf{a}_0 = \mathbf{h}_0. \tag{29}$$

The recursion matrix $\mathbf{B}_t$ is identical to $\mathbf{A}_t$, except that the first and last columns are set to 0:

$$B_t[i, j] = \begin{cases} 0 & \text{if } j = 1 \text{ or } j = 5 \\ A_t[i, j] & \text{otherwise.} \end{cases} \tag{30}$$

Note that the events corresponding to the vector of probabilities $\mathbf{b}_t$ do not constitute a complete set of events for $t_2 > 1$. Rather,

$$\left(\sum_{k=1}^{t_2} \beta_k\right) + \left(\sum_{k=1}^{t_2} \gamma_k\right) + b_{t_2}[2] + b_{t_2}[3] + b_{t_2}[4] = 1. \tag{31}$$

For strategy B, the overall probability of success at generation t_2 is

$$S_{t_2} = \sum_{t=1}^{t_2} \beta_t. \tag{32}$$

Correspondingly, the mean total number of individuals that need to be genotyped given that the BC scheme is successful at last in generation t_2 was computed as

$$\overline{n}(t_2) = 1/S_{t_2}\left(\sum_{t=1}^{t_2} \beta_t \left(\sum_{k=1}^{t} n_k\right)\right). \tag{33}$$

Following the rationale of Hospital and Charcosset (1997), optimal population sizes are then defined such that the overall probability of success S_{t_2} at generation t_2 is above a given threshold and the total number $\overline{n}$ of individuals genotyped during the BC scheme is minimal:

$$\text{AND}\begin{cases} S_{t_2} \geq 0.99 \\ \overline{n}_{t_2} & \text{minimal.} \end{cases} \tag{34}$$

Numerical applications: Optimal population sizes n_t for strategy B were computed numerically to fulfill both conditions in (34), *i.e.*, find the population sizes that minimize $\overline{n}$ while keeping above 99% the probability that a double recombinant is obtained at last in generation t_2. Such a computation is easy when population sizes are kept constant across BC generations ($n_t = n$, $\forall t$). Yet, Hospital and Charcosset (1997) showed that

TABLE 3

Minimal population sizes for a one-, two-, or three-generation backcross program

		$t_2 = 1$	$t_2 = 2$					$t_2 = 3$						
l	Population	n_1	n_1	β_1	n_2	β_2	$\overline{n}$	n_1	β_1	n_2	β_2	n_3	β_3	$\overline{n}$
1.0	cst	93,959	921	(0.04)	921	(0.95)	1,800.9	471	(0.02)	471	(0.87)	471	(0.09)	975.2
	var		579	(0.03)	996	(0.96)	1,546.9	233	(0.01)	344	(0.73)	731	(0.25)	757.0
2.0	cst	23,961	460	(0.08)	460	(0.91)	880.7	237	(0.04)	237	(0.85)	237	(0.09)	485.1
	var		290	(0.05)	499	(0.94)	761.7	117	(0.02)	171	(0.72)	370	(0.25)	377.7
3.0	cst	10,861	306	(0.12)	306	(0.87)	574.4	158	(0.06)	158	(0.83)	158	(0.09)	320.1
	var		193	(0.08)	334	(0.91)	500.5	78	(0.03)	113	(0.70)	251	(0.25)	251.5
5.0	cst	4,066	184	(0.19)	184	(0.80)	333.0	96	(0.10)	96	(0.80)	96	(0.09)	190.5
	var		118	(0.13)	200	(0.86)	292.7	48	(0.05)	70	(0.70)	149	(0.24)	150.5
10.0	cst	1,119	92	(0.32)	92	(0.68)	154.7	49	(0.18)	49	(0.72)	49	(0.08)	93.1
	var		62	(0.23)	100	(0.76)	139.3	25	(0.10)	36	(0.66)	76	(0.23)	75.1
20.0	cst	337	47	(0.47)	47	(0.52)	71.5	26	(0.30)	26	(0.62)	26	(0.07)	46.1
	var		32	(0.35)	52	(0.64)	65.4	13	(0.16)	19	(0.61)	41	(0.22)	38.1
30.0	cst	179	32	(0.56)	32	(0.43)	45.8	18	(0.37)	18	(0.55)	18	(0.07)	30.6
	var		23	(0.45)	36	(0.54)	42.7	11	(0.25)	13	(0.55)	28	(0.19)	26.2
50.0	cst	90	21	(0.66)	21	(0.33)	28.0	13	(0.49)	13	(0.45)	13	(0.06)	20.4
	var		16	(0.56)	24	(0.43)	26.4	9	(0.37)	10	(0.49)	18	(0.13)	17.6

For each distance $l = l_1 = l_2$ (cM) between the introgressed gene and the two flanking markers, two lines give the results for either constant (cst) or variable (var) population sizes. BC schemes of different total durations t_2 are considered; in each case the data give the optimal population size n_t and probability of success β_t at generations $t \leq t_2$ and the mean total number of individuals $\overline{n}$ (Equation 33).

a better reduction of $\overline{n}$ is achieved when allowing different population sizes at different BC generations. In that case, finding the set of values $\{n_t\}_{t\in[1,t_2]}$ that satisfy (34) is more difficult, in particular for $t_2 > 2$. A computer program (F. Hospital and G. Decoux, unpublished data) was designed for the numerical optimization of population sizes in this case, using the "simulated annealing" algorithm (Press *et al.* 1992). Examples of numerical results for some marker distances $l = l_1 = l_2$ are given in Table 3 for a BC breeding scheme intended to last at most $t_2 = 1$, 2, or 3 generations and in Table 4 for $t_2 = 5$. Note, however, that our freely distributed program makes it easy to compute population sizes in any other situations and possibly with an optimization criterion different from (34).

The definition of $\overline{n}$ in (33) takes into account only the cases where the BC scheme is successful. With the population sizes in Tables 3 and 4, the probabilities γ_t of failure of the BC scheme at any generation t (no carrier of the introgressed gene in the population) are close to zero. Hence, the probability of nonsuccess at t_2 [$1 - S_{t_2} \simeq 1\%$ with the conditions of (34)] is mostly the probability of obtaining only a genotype G_2, G_3, or G_4 (single- or nonrecombinant) at t_2. In that case, the BC scheme has not failed, but simply needs to be pursued one or more additional BC generations.

For a single-generation program ($t_2 = 1$, Table 3), population sizes become very large for short marker distances ($l < 10$ cM), which are the most relevant since using close markers is the best way to reduce donor segment length (see previous section). It is then better to perform at least two BC generations, because the probability of success in BC$_2$ (β_2) is always higher than the probability of success in BC$_1$, except for distant markers ($l > 20$ cM). Performing two BC generations with constant population sizes permits a drastic reduction of the total number of individuals that have to be genotyped, which confirms the intuition of Young and Tanksley (1989b). Moreover, as already noted by Hospital and Charcosset (1997), using variable population sizes allows an even better reduction of $\overline{n}$ by slightly increasing β_2 with respect to β_1. In this case, population size in BC$_2$ needs to be higher than in BC$_1$. This is generally the case for any total duration of the BC scheme (t_2): Optimal values for variable population sizes should increase in advanced BC generations, except for distant markers and high t_2 values (*e.g.*, $t_2 = 10$, not shown).

Allowing BC schemes to last possibly more than two generations permits an even further reduction of the mean total number of individuals, though the gain on $\overline{n}$ is then less important than the gain from $t_2 = 1$ to $t_2 = 2$. The gain is nevertheless economically important, especially for close markers. Increasing total duration of the BC scheme t_2 from 2 to 3 with either constant or variable population sizes provides a gain of ~50% on $\overline{n}$ for $l \leq 20$ cM, which may correspond to hundreds less individuals. It is worth noting that, with constant population sizes, this gain is barely obtained at the expense of lower probability of success in early generations: Probability of success in at most two generations ($\beta_1 + \beta_2$) with constant population sizes is close to 90%

TABLE 4

Minimal population sizes for a five-generation backcross program

		$t_2 = 5$										
d	Population	n_1	β_1	n_2	β_2	n_3	β_3	n_4	β_4	n_5	β_5	$\overline{n}$
1.0	cst	245	(0.01)	245	(0.63)	245	(0.24)	245	(0.08)	245	(0.02)	602.9
	var	103	(0.01)	127	(0.30)	170	(0.34)	258	(0.24)	517	(0.10)	491.3
2.0	cst	123	(0.02)	123	(0.62)	123	(0.24)	123	(0.08)	123	(0.02)	300.9
	var	52	(0.01)	64	(0.30)	86	(0.34)	130	(0.24)	259	(0.10)	246.2
3.0	cst	83	(0.03)	83	(0.62)	83	(0.24)	83	(0.08)	83	(0.02)	201.4
	var	35	(0.01)	44	(0.30)	58	(0.34)	89	(0.24)	170	(0.10)	164.5
5.0	cst	50	(0.06)	50	(0.60)	50	(0.23)	50	(0.08)	50	(0.02)	120.2
	var	21	(0.02)	26	(0.29)	35	(0.33)	53	(0.24)	106	(0.10)	99.3
10.0	cst	26	(0.10)	26	(0.57)	26	(0.22)	26	(0.07)	26	(0.02)	60.7
	var	12	(0.05)	14	(0.31)	18	(0.31)	27	(0.22)	54	(0.09)	50.5
20.0	cst	14	(0.17)	14	(0.53)	14	(0.21)	14	(0.07)	14	(0.02)	31.2
	var	9	(0.12)	9	(0.38)	10	(0.27)	15	(0.16)	30	(0.06)	27.4
30.0	cst	11	(0.25)	11	(0.51)	11	(0.17)	11	(0.05)	11	(0.01)	22.7
	var	8	(0.19)	8	(0.43)	9	(0.24)	11	(0.10)	25	(0.04)	20.4
50.0	cst	9	(0.37)	9	(0.47)	9	(0.12)	9	(0.03)	9	(0.01)	16.4
	var	8	(0.34)	8	(0.46)	7	(0.14)	8	(0.04)	13	(0.01)	15.3

Same as Table 3 for $t_2 = 5$.

for $t_2 = 3$ (Table 3) compared with 99% for $t_2 = 2$ with much larger population sizes (about double). Again, using variable population sizes for $t_2 = 3$ permits a further reduction of $\overline{n}$ (~100 individuals for $l \leq 5$ cM). But, in this case, the reduction of $\overline{n}$ is obtained at the expense of a reduced probability of success in early generations. A decision must then be made between reduction of costs and reduction of duration of the breeding scheme, which is a matter for economical consideration not taken into account here (see also below). However, $(\beta_1 + \beta_2)$ with variable population sizes is still close to 75% for $t_2 = 3$ (Table 3).

The same tendencies are observed for even longer durations of the BC breeding scheme (*e.g.*, $t_2 = 5$, Table 4). Population sizes are always reduced, and even more reduced for variable than for constant values, except for distant markers. Note that for very distant markers and/or very long durations (*e.g.*, $t_2 = 10$, not shown), optimal population sizes are not reduced below a given threshold, because in those cases, while the probability of double recombinations increases, the probability of failure ("losing" the introgressed gene) becomes the most critical factor. Again, with increased duration of the program, the probability of success in early BC generations with either constant or variable population sizes decreases with respect to the probability of success in advanced generations. But, the important conclusion is that experimental costs are drastically reduced, even for very close markers. For $t_2 = 5$ (Table 4) and variable populations sizes, a mean total number of only <500 individuals need to be genotyped for flanking markers as close as only 1 cM on each side of the introgressed gene. Moreover, using the optimal population sizes defined in Table 4 for the same marker distance of 1 cM and applying (33) for the first three generations only (BC_1 to BC_3) shows that the mean total number of individuals is then <210, with a corresponding probability of success of 65%. For $l = 5$ cM and $t_2 = 5$, $\overline{n}$ is only 100 (Table 4).

The recursion equations of Hospital and Charcosset (1997) for the computation of minimal population sizes were used by Frisch *et al.* (1999) in a slightly different context. In the present study, population sizes were optimized *a priori* by considering the total possible duration of the backcross program. Instead, Frisch *et al.* (1999) use a sequential *a posteriori* approach, where minimal population sizes for any generation t are computed given that the genotype of the individual selected at the previous generation $(t - 1)$ is known. This is not the most relevant approach for the design of an optimal backcross program before the program is started. More generally, even for an already started BC program, it is not relevant to optimize population sizes beyond the very next generation. In particular, the present results indicate that the threshold risk of 1% per generation used by Frisch *et al.* is not optimal in the context of sequential predictions. In the present study, all BC generations are taken into account simultaneously, so there is a program-wise risk of 1%. This corresponds in fact to a higher risk per generation in early BC generations, as could be estimated from (31) and β values in Tables 3 and 4. This approach permits a better reduction of the mean total number of individuals genotyped over the entire program. Moreover, this approach even provides higher probabilities of obtaining double recombinants in early generations. This is so because, when the population size necessary to obtain a double recombinant in the next generation is considered as too large, Frisch *et al.* (1999) use the minimum number of individuals necessary to obtain a single recombinant on one side, while the present approach provides intermediate values. Hence, the present results should permit a better reduction of donor segment length at same experimental means (number of individuals genotyped).

Moreover, the present *a priori* approach provides an objective criterion to determine the number of individuals that have to be genotyped in the sequential approach, when the population size needed to obtain a double recombinant G_1 at one given generation is too large. Note that the calculation of optimal population sizes at any intermediate stage of an already started BC scheme is also possible using our computer program (F. Hospital and G. Decoux, unpublished data).

Other selection scenarios: The calculations in this section were derived within the framework of a breeding scenario where (i) only one individual is selected at each generation (on the basis of its genotype at flanking markers) and (ii) only one pair of flanking markers is considered.

Condition (i) is a limitation of the results on minimal population sizes, in particular for their application to breeding schemes where several individuals have to be selected at each generation (*e.g.*, animals with low fecundity). Note, however, that minimal population sizes here were computed so that at least one individual with the desired genotype is obtained; thus the expected number of such individuals is always above one. Also, the present results for single selections could be used as a per-individual approximation to the multiple selections case. But, this would provide only a crude approximation. Exact derivation of minimal population sizes in the true multiple selection case is more complex and was not considered here. It could be feasible using the present derivations in conjunction with Equation 10 provided with no application by Frisch *et al.* (1999).

Concerning condition (ii), it was shown previously (Hospital *et al.* 1992) that simultaneous selection on multiple embedded marker pairs on each side of locus T is more efficient than selection on any of the single marker pairs. The efficiency of selection on multiple embedded marker pairs can be investigated with the present derivations for a single marker pair using the following rationale: (i) Define a duration of the BC scheme t_2 and define a "limiting" marker pair for which it is mandatory to obtain a double recombinant (generally the outermost marker pair, *i.e.*, the pair of markers most distant from locus T). (ii) Take the minimal population sizes given in Table 3 for this limiting pair. This would ensure that at least a double recombinant for this pair is obtained by t_2. (iii) Apply (28) with these same population sizes for each inner marker pair in turn (this is easy to perform with our computer program).

As an example, this was done to produce the results in Table 5. I consider a three-generations BC program ($t_2 = 3$). Population size at each generation was fixed at 49 individuals, *i.e.*, the (constant) minimal population size for a 10-cM marker pair from Table 3. Then, the probabilities of obtaining double recombinants with the same population size of 49 were computed for closer marker pairs at distances 5, 3, 2, or 1 cM, one pair at a time.

TABLE 5

Efficiency of selection for inner marker pairs

l	β_1	β_2	β_3	$\Sigma_t\beta_t$
10.0	0.18	0.72	0.08	0.99
5.0	0.05	0.59	0.24	0.89
3.0	0.02	0.38	0.28	0.68
2.0	0.01	0.24	0.23	0.48
1.0	0.00	0.08	0.12	0.20

Probabilities (β_t) of obtaining a double-recombinant genotype at generation t for marker pairs at distances $l = l_1 = l_2$ (cM) in a three-generation backcross program ($t_2 = 3$), when population size at each BC$_t$ generation is $n_t = 49$. This approximates the efficiency of a multimarker pair strategy; see text for details.

Strictly speaking, in the case of a real multimarker pair selection, these computations would provide exact results only for the innermost marker pair (1 cM). This is so because, with selection on multiple pairs, the ranking of genotypes for any inner pair modifies the ranking of genotypes for any more distant pair compared to the ranking if this distant pair were alone. As a consequence, the probabilities in the last column of Table 5 do not sum up to one. However it provides a reasonable approximation of the efficiency of a real multimarker pair selection, and results in the last column can be interpreted as the probability of obtaining a double recombinant for the indicated pair or any other pair closer to T, which provides a conservative estimation of the efficiency of a real multipair strategy.

Obviously the probabilities in this situation are close to zero in BC$_1$. But the results in Table 5 indicate that with as few as 49 individuals per generation, which is minimal for a marker pair at 10 cM, there is still a good chance of obtaining by the end of the program (BC$_3$) double recombinants for closer marker pairs (above 48% for markers down to 2 cM of locus T) except for very close markers. This might then appear as a valid strategy. Note, however, that the cost of increased numbers of genotyping in a multimarker pair strategy should also be taken into account in an economic optimization (not given here).

CONCLUSIONS

The most important parameters for the optimization of marker-assisted selection for the reduction of linkage drag in backcross programs are the distances between the flanking markers and the introgressed gene, the population sizes, and the total duration of the BC scheme.

The efficiency of marker-assisted selection (evaluated from the expected length of the donor segment among genotypes that are double recombinant for both flanking markers) depends mostly on marker-gene distances, except for distant markers. For distant markers, increasing the duration of the breeding scheme (total number of BC generations performed) has an effect on the reduction of donor segment length, but this effect is small compared to the effect of shortening marker-gene distances. Hence, the best way to reduce donor segment length is to use flanking markers as close as possible to the introgressed gene. This has the drawback of considerably increasing the population size needed to obtain double-recombinant genotypes and thus increasing the cost of the experiment. Population size obviously depends on marker-gene distances, too, but in this case the duration of the breeding scheme has an important effect. In some situations, the effect of breeding scheme duration on population sizes might be more important than the effect of marker-gene distances. For close distances, performing at least two BC generations is critical. But, the present results indicate that planning to perform even more BC generations (three or more) might be generally valuable.

The results of the two previous sections are combined in Figure 4 as follows. I considered two flanking markers, located at same distance $l_1 = l_2$ from locus T, and a BC scheme designed to last at most t_2 generations. For the sake of simplicity, I did not identify single-recombinant genotypes and the probabilities to obtain them (though this is feasible with the above equations), but only focused on double-recombinant genotypes. Hence the efficiency of marker-assisted selection was evaluated from the expected length of the donor segment on both sides of the introgressed gene among double recombinants, averaged over all intermediate generations, and weighted by the corresponding probabilities of success:

$$\overline{Y}(t_2) = 2\left(\sum_{t=1}^{t_2} \beta_t E_Y(t)\right) \Bigg/ \left(\sum_{t=1}^{t_2} \beta_t\right). \tag{35}$$

Note that here and in Figure 4 I consider segment length on both sides of locus T, while other figures give segment length on one side only. This is done because as soon as probabilities of double recombinations are considered both sides cannot be treated separately. The mean total number of individuals that need to be genotyped given that the BC scheme is successful by generation t_2 was computed from (33).

Figure 4 highlights the effect of the planned total duration of the breeding scheme t_2. If the major effect of t_2 on the reduction of donor segment length is seen from increasing total duration from BC_1 to BC_2, there is also an important effect of increasing total duration up to BC_3 for most marker distances and up to BC_5 for very close markers (gene-marker distance is indicated by dots, see Figure 4 legend). Increasing total duration above BC_5 appears less valuable since the additional reduction of population size is then smaller and population sizes are then within reasonable limits, even for close markers. Remember, however, that the probability of success at intermediate generations $t < t_2$ is always above zero and might be important in some cases.

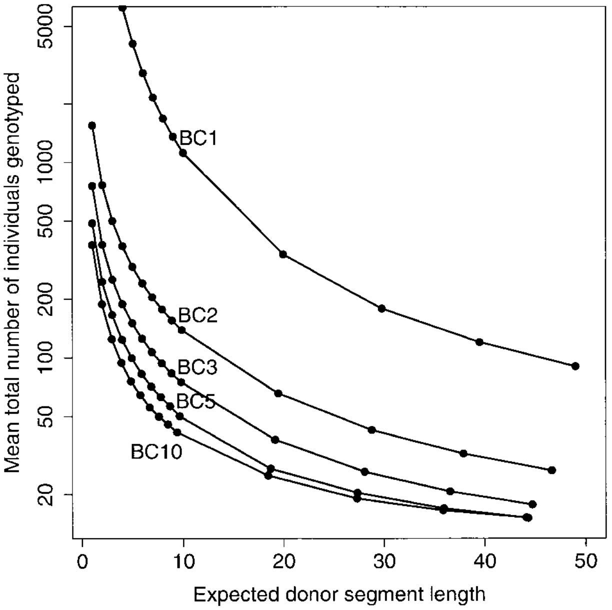

Figure 4.—Relationship between efficiency of marker-assisted selection and minimal population size. Abscissa: total length $\overline{Y}$ (cM) from Equation 35 of donor segment on both sides of locus T. Ordinate: $\overline{n}$ from Equation 33. For different BC_{t_2} generations (t_2 = 1, 2, 3, 5 or 10 from top to bottom) a parametric plot of the couple values $(\overline{Y}(t_2), \overline{n}(t_2))$ was drawn, with parameter value l = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, or 50 cM indicated by dots from left to right. Ordinate is on logarithmic scales.

Clearly, it is not possible to optimize simultaneously the three parameters l (marker distances), $\overline{n}$ (population sizes), and t_2 (total duration of the program). Rather, this study makes it possible to optimize one parameter, once decisions have been made regarding the two other parameters: If a reduction of the donor segment to a short length is not mandatory, then using markers located at ~10 cM from the introgressed gene permits a good reduction of donor segment length (compared to its expected value with no marker-assisted selection, see Figure 2) at reasonable costs (less than 200 individuals) in only two BC generations. In that case, performing additional BC generations would allow the reduction of experimental costs to very low levels (<100 or even 50 individuals). This would be seldom desired since these situations correspond mainly to the cases of introgression between commercial strains. In such cases, a fast acquisition of improved genetic material is often more desirable than a drastic reduction of cost or donor segment length. Conversely, if a drastic reduction of donor segment length is really mandatory (*e.g.*, for the construction of near-isogenic lines or congenic strains), then these results show that this may be achieved by marker-assisted selection using markers located as closely as possible to the introgressed gene, while keeping experimental efforts at a low level. But in that case optimal BC breeding schemes should generally be planned to last a possible total of three to five generations.

In general, besides considerations related to economic competition, performing several BC generations should be a better strategy, since it permits a very drastic reduction of costs. Moreover, it would also make it easier to perform marker-assisted selection on chromosomal regions not surrounding the introgressed gene (*e.g.*, selection on noncarrier chromosomes, not considered here) to reduce the overall recipient parent genome content, because this selection is also more efficient in advanced BC generations (Hospital *et al.* 1992) and because the increased frequency of double recombinants on the carrier chromosome would permit a higher selection intensity on noncarrier chromosomes.

I thank Isabelle Goldringer for helpful discussions and support; Christine Dillmann, Alain Charcosset, and André Gallais for helpful discussions; Stéphane Ballereau and Guy Decoux for their help with some of the numerical calculations; and Dan Vacso, Chris Haley, and two anonymous referees for helpful comments.

LITERATURE CITED

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

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

Hanson, W. D., 1959 Early generation analysis of lengths of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. Genetics **44:** 833–837.

Hillel, J., T. Schaap, A. Haberfeld, A. J. Jeffreys, Y. Plotzky *et al.*, 1990 DNA fingerprint applied to gene introgression breeding programs. Genetics **124:** 783–789.

Hospital, F., and A. Charcosset, 1997 Marker-assisted introgression of quantitative trait loci. Genetics **147:** 1469–1485.

Hospital, F., C. Chevalet and P. Mulsant, 1992 Using markers in gene introgression breeding programs. Genetics **132:** 1199–1210.

Melchinger, A. E., 1990 Use of molecular markers in breeding for oligogenic disease resistance. Plant Breed. **104:** 1–19.

Naveira, H., and A. Barbadilla, 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.

Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery, 1992 *Numerical Recipes in C: The Art of Scientific Computing*, Ed. 2. Cambridge University Press, Cambridge, UK.

Ragot, M., M. Biasiolli, M. F. Delbut, A. Dell'orco, L. Malgarini *et al.*, 1995 Marker-assisted backcrossing: a practical example. pp. 45–56 in *Techniques et Utilisations des Marqueurs Moléculaires. (Les Colloques, no 72)*. INRA, Paris.

Stam, P., and A. C. Zeven, 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. Mol. Biol. Rep. **1:** 3–8.

Visscher, P. M., and R. Thompson, 1995 Haplotype frequencies of linked loci in backcross populations derived from inbred lines. Heredity **75:** 644–649.

Visscher, P. M., C. S. Haley and R. Thompson, 1996 Marker assisted introgression in backcross breeding programs. Genetics **144:** 1923–1932.

Wolfram, S., 1988 *Mathematica, a System for Doing Mathematics by Computer*. Addison-Wesley, Redwood City, CA.

Young, N. D., and S. D. Tanksley, 1989a Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor. Appl. Genet. **77:** 95–101.

Young, N. D., and S. D. Tanksley, 1989b RFLP analysis of the size of chromosomal segments retained around the *tm-2* locus of tomato during backcross breeding. Theor. Appl. Genet. **77:** 353–359.

Communicating editor: C. Haley

APPENDIX A

A longer, but maybe easier, way to demonstrate Equation 7 in text is as follows. At generation t_{co}, a crossover occurred in $]x, x + dx]$; then, (i) if no recombination occurred in the interval $]x, l]$ at generation t_{co} [probability $(1 - r_{[l-x]})$], this provides a recombination between locus T and M at this generation and is sufficient for M to be of recipient type at t_1, whatever happened at other generations; (ii) if a recombination occurred in the interval $]x, l]$ at generation t_{co} (probability $r_{[l-x]}$), this provides no recombination between locus T and M at this generation. Hence, for the marker to be of recipient type at generation t_1, a recombination must have occurred in the interval $]x, l]$ in at least one of the other generations $t \neq t_{co}$ (probability $1 - (1 - r_{[l-x]})^{t_1-1}$). Finally, the probability for the interval $]x, l]$ is

$$(1 - r_{[l-x]}) + r_{[l-x]}\{1 - (1 - r_{[l-x]})^{t_1-1}\} = 1 - r_{[l-x]}(1 - r_{[l-x]})^{(t_1-1)}. \tag{A1}$$

QED

APPENDIX B: SIZE OF DONOR CHROMOSOME SEGMENT WHEN THE FLANKING MARKER IS OF DONOR TYPE

In the text of the article, I derived the PDF, mean, and variance of segment length on one side of introgressed locus T, under the condition that marker M if of recipient type. Supplementary to this, I briefly derive here the same quantities given that the marker M is of donor type. This is relevant, for example, to the estimation of "graphical genotypes" (Young and Tanksley 1989a), where both possible marker genotypes must be considered. This is also relevant to the evaluation of the efficiency of marker-assisted selection, by comparing the lengths of the intact segment when selection is successful (a recombinant was obtained and the marker is of recipient type, Equations 8, 10, and 12 in text) to lengths when selection failed (no recombinant was obtained, marker remains of donor type, Equations B1, B4, and B5 below). The results of Naveira and Barbadilla (1992; Equations 2, 3, and 4 in text) provide the lengths of intact segment averaged over both marker genotypes.

Let $Z(t_1)$ be the random variable corresponding to the length of intact donor segment on one side of locus T at generation BC_{t_1}, given that marker M at distance l is of donor type. The corresponding PDF $h_{t_1}(x)$, mean $E_Z(t_1)$, and variance $\sigma_Z^2(t_1)$ are derived as follows.

The probability that the marker is of donor type at t_1 is $\{1 - P_M(t_1)\}$ from (5).

For $l \leq x \leq L$, stating that the intact segment length is x implies that the marker is of donor type, thus the PDF is directly obtained from the results of Naveira and Barbadilla (1992), dividing (2) by $\{1 - P_M(t_1)\}$.

For $0 < x < l$, I follow the same rationale as for Equations 6 and 7 in text, that is: For the interval $]0, x]$ the probability is $e^{-t_1 x}\,dx$ as in (6). For the interval $]x, l[$, at generation t_{co} a crossover occurred in $[x, x + dx]$; then, for the marker to remain of donor type at generation t_{co} a recombination must have occurred in the interval $]x, l[$. At generations $t \neq t_{\mathrm{co}}$, no crossover occurred in $[0, x]$; then, for the marker to remain of donor type no recombination must have occurred in the interval $]x, l[$.

Finally, the PDF h_{t_1} for $Z(t_1)$ is

$$h_{t_1}(x) = h_{t_1,t_1}(x) = \begin{cases} \dfrac{r_{[l-x]}\,(1 - r_{[l-x]})^{(t_1-1)}}{(1 - r_{[l]})^{t_1}}\,t_1 e^{-t_1 x} & \text{if } 0 < x < l \\[2ex] \dfrac{1}{(1 - r_{[l]})^{t_1}}\,t_1 e^{-t_1 x} & \text{if } l \leq x \leq L \end{cases}$$

$$= \begin{cases} \dfrac{(1 - e^{-2(l-x)})(1 + e^{-2(l-x)})^{(t_1-1)}}{(1 + e^{-2l})^{t_1}}\,t_1 e^{-t_1 x} & \text{if } 0 < x < l \\[2ex] \dfrac{2^{t_1}}{(1 + e^{-2l})^{t_1}}\,t_1 e^{-t_1 x} & \text{if } l \leq x \leq L. \end{cases} \tag{B1}$$

In fact, noting that

$$X(t_1) = P_M(t_1)\,Y(t_1) + \{1 - P_M(t_1)\}Z(t_1), \tag{B2}$$

(B1) could have been computed directly from (2) and (8) as

$$h_{t_1}(x) = \begin{cases} \dfrac{f_{t_1}(x) - P_M(t_1)\,g_{t_1}(x)}{1 - P_M(t_1)} & \text{if } 0 < x < l \\[2ex] f_{t_1}(x) & \text{if } l \leq x \leq L. \end{cases} \tag{B3}$$

From (B3), the mean $E_Z(t_1)$ of Z is then simply computed from (3) and (10) as

$$E_Z(t_1) = 1/\{1 - P_M(t_1)\}E_X(t_1) - P_M(t_1)/\{1 - P_M(t_1)\}E_Y(t_1) \tag{B4}$$

and the variance $\sigma_Z^2(t_1)$ of Z is computed from (4) and (12) as

$$\sigma_Z^2(t_1) = 1/\{1 - P_M(t_1)\}^2\sigma_X^2(t_1) - P_M(t_1)^2/\{1 - P_M(t_1)\}^2\sigma_Y^2(t_1). \tag{B5}$$