Marker-Assisted Introgression of Quantitative Trait Loci

Frédéric Hospital and Alain Charcosset

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

Manuscript received October 10, 1996
Accepted for publication August 15, 1997

ABSTRACT

The use of molecular markers for the introgression of one or several superior QTL alleles into a recipient line is investigated using analytic and simulation results. The positions of the markers devoted to the control of the genotype at the QTLs in a "foreground selection" step are optimized given the confidence interval of the QTL position. Results demonstrate that using at least three markers per QTL allows a good control over several generations. Population sizes that should be recommended for various numbers of QTLs are calculated and are used to determine the limit in the number of QTLs that can be monitored simultaneously. If "background selection" devoted to accelerate the return to the recipient parent genotype outside the QTL regions is applied, the positions of the markers devoted to the control of the QTLs have to be reconsidered. When several QTLs are monitored simultaneously, background selection among the limited number of individuals resulting from the foreground selection step accelerates the increase in genomic similarity with the recipient parent, with only limited costs. Background selection is even more efficient in a pyramidal backcross program where QTLs are first monitored one by one.

A backcross breeding program is aimed at gene introgression from a "donor" line into the genomic background of a "recipient" line. The potential utilization of molecular markers in such programs has received considerable attention in the recent past. Markers could be used to assess the presence of the introgressed gene ("foreground selection") when direct phenotypic evaluation is not possible, or too expensive, or only possible late in the development. This was proposed by Tanksley (1983), and later reviewed in various papers (see for example Melchinger 1990). Markers could also be used to accelerate the return to the recipient parent genotype at other loci ("background selection"). This was first proposed by Hillel *et al.* (1990 and also 1993) and later investigated by Hospital *et al.* (1992). In both papers, it was assumed that the introgressed gene could be detected without ambiguity, and the theoretical study was restricted to background selection only. The use of molecular markers for background selection in backcross programs has been tested experimentally and proved to be very efficient (Ragot *et al.* 1995).

Recently, Groen and Smith (1995) and Visscher *et al.* (1996) investigated both foreground and background selection. Visscher *et al.* (1996) also investigated the case when the introgressed gene is a QTL (quantitative trait locus), that is a gene whose position is not known with certainty, but only estimated. In fact, introgressing the favorable allele of a QTL by recurrent backcrossing could be a powerful mean to improve the economic value of a line, provided the expression of the gene is not reduced in the recipient genomic background. Yet, recent results show that for many traits of economic importance QTLs have rather small effects. In this case, the economic improvement resulting from the introgression of the favorable allele at a single QTL may not be competitive when compared with the improvement resulting from conventional breeding methods over the same duration. Marker-assisted introgression of superior QTL alleles could then compete with classical phenotypic selection only if several QTLs could be manipulated.

In this article, we want to investigate the potential use of molecular markers for both foreground and background selection in backcross breeding programs aimed at introgressing one to several QTLs. We will first determine the optimal number and positions of the markers needed to control the QTLs during the foreground selection step and the maximum possible number of QTLs that could be monitored simultaneously with realistic population sizes. Then, we will investigate the use of markers for background selection, and the potential efficiency of selection on the possibly small number of individuals carrying all favorable alleles.

METHODS

Foreground selection: First, we investigate the optimal use of markers to assess the presence of the desirable allele at a QTL. This is done using approximate analytic calculations where the possible effects of background selection over successive generations are not

Corresponding author: Frédéric Hospital, Station de Génétique Végétale, INRA/UPS/INA-PG, Ferme du Moulon, 91190 Gif Sur Yvette, France. E-mail: fred@moulon.inra.fr

taken into account. We consider that one QTL has been detected on a chromosome of total length L (in Morgans). Positions on this chromosome are represented using an arbitrarily oriented scale ranging from 0 to L. We assume that the most likely position of the QTL has been previously estimated using any appropriate approach (*e.g.*, LANDER and BOTSTEIN 1989; KNAPP *et al.* 1990; HALEY and KNOTT 1992) as x_l. At each backcross generation, the selection of individuals that carry the allele of interest at the QTL is based on their genotype at neighbor marker(s). We consider that this selection involves m markers, located at positions $\{x_1, \ldots, x_m\}$. The choice of these markers must allow a good control of the genotype at the QTL, to limit the risk that an individual that displays the desired genotype at the marker(s) does not carry the allele of interest at the QTL. We only consider here probabilities of genotypes on the chromosome derived from the non-recurrent parent.

The probability that a given progeny inherits the donor allele at all markers from its parent is

$$P_M = \frac{1}{2} \prod_{k=1}^{m-1} (1 - r[x_k, x_{k+1}]), \qquad (1)$$

where $r[x_k, x_{k+1}]$ is the recombination rate between loci x_k and x_{k+1}. In the case of a single marker, $P_M = {}^1/_2$. If we consider that at each generation, among a total of N progenies, only those with appropriate genotype at all the markers are selected for reproduction, the probability of obtaining at least one individual with the donor allele at all the markers after t generations is

$$P_N[t] = (1 - (1 - P_M)^N)^t, \qquad (2)$$

and the probability that a given backcross progeny carries the donor allele at all m markers at generation t is (see MELCHINGER 1990)

$$P_M[t] = P_N[t-1] P_M. \qquad (3)$$

From (2), the minimum number $N_\alpha[t]$ of individuals that should be genotyped at each generation, so that at least one individual with the desired genotype at all the markers is obtained with risk α after t generations, is

$$N_\alpha[t] = \frac{\ln\,[1 - (1-\alpha)^{1/t}]}{\ln\,[1 - P_M]}. \qquad (4)$$

We need now to evaluate the risk that an individual that displays the desired genotype at the marker(s) does not carry the allele of interest at the QTL. Let P_{MQ} be the probability that a given backcross progeny has the requested genotype at all m markers *and* at the QTL. The conditional probability $P_{Q|M}[t]$ that at generation t a given progeny has the donor allele at the QTL, given that it has the donor allele at all the markers, is then simply (from MELCHINGER 1990) as follows:

$$P_{Q|M}[t] = \frac{(P_{MQ})^t}{(P_M)^t}. \qquad (5)$$

If the actual position of the QTL is x, this probability is

$$P_{Q|M}[t] = f[x] = \frac{(1 - r[x_k, x])^t (1 - r[x, x_{k+1}])^t}{(1 - r[x_k, x_{k+1}])^t}, \qquad (6)$$

if the QTL lies between two markers, respectively, at positions x_k and x_{k+1}, and

$$P_{Q|M}[t] = f[x] = (1 - r[x, x_1])^t, \qquad (7)$$

if the QTL is controlled by a single marker at position x_1.

In the extreme case where a single marker identifies the allele to be introgressed without error, then $r[x, x_1] = 0$ and $P_{Q|M}[t] = 1$. In practice, the position of the QTL is estimated with a given error in an experiment. Thus, the actual position of the QTL is unknown. Probability $P_{Q|M}[t]$ must then be integrated over all putative positions of the QTL:

$$P_{Q|M}[t] = \int_0^L f[x] g[x]\,dx, \qquad (8)$$

where $f[x]$ is taken from (6) or (7), and where $g[x]$ is the density of probability of the true given the expected position of the QTL. When $g[x]$ follows a Gaussian distribution, this calculation was done by VISSCHER *et al.* (1996) in the case of one or two markers. An extension of their approach to any number m of markers is presented in the APPENDIX (Equations A1, A5, A6 and A7).

For any number m of markers, it is possible to determine the positions of the markers that maximize $P_{Q|M}[t]$. Then, given the marker positions, it is possible to compute P_M and N_α numerically. This was done using software package *Mathematica* (WOLFRAM 1988).

The analytical approach can be extended to the case where q unlinked QTLs are considered. In this situation the definitions of P_M, N_α and $P_{Q|M}$ can be extended as follows:

$$P_M = \prod_{l=1}^{q} P_{M(l)} \qquad (9)$$

$$N_\alpha[t] = \frac{\ln\,[1 - (1-\alpha)^{1/t}]}{\ln\,[1 - P_M]} \qquad (10)$$

$$P_{Q|M}[t] = \prod_{l=1}^{q} P_{Q|M(l)}[t], \qquad (11)$$

where P_M is the probability that a given backcross progeny carries the donor allele at all markers (controlling q QTLs), $N_\alpha[t]$ is the minimum number of individuals that should be genotyped at each generation so that at least one individual with requested genotype is obtained at generation t with risk α and $P_{Q|M}[t]$ is the conditional probability that at generation t, a given progeny has the donor allele at all q QTLs, given that it has the donor allele at all the markers. $P_{M(l)}$ and $P_{Q|M(l)}[t]$ are the individual probabilities associated with QTL l.

Background selection: The analytic approach was extended to the study of combined foreground and background selection on the carrier chromosome, when a single QTL is considered. In addition to the m markers x_i devoted to the control of the QTL as in the previous section, we consider two additional markers located on each side of the QTL on the chromosome at positions y_1 and y_2 ($0 \leq y_1 < x_1 \leq \cdots \leq x_m < y_2 \leq L$) on which background selection, *i.e.*, selection for the recipient type allele is performed. We define $P^*_{Q|M}$ as the probability of having the donor type allele at the QTL, given that we have the donor type allele at all markers x_i *and* the recipient type allele at both y_1 and y_2. The aim of background selection on the carrier chromosome is that at least the chromosomal segments $[0, y_1]$ and $[y_2, L]$ return to a 100% recipient-type genomic composition as fast as possible. Hence, we restrict the calculation of $P^*_{Q|M}$ on the segment $]y_1, y_2[$. The corresponding calculations are described in the APPENDIX (Equations A10, A12, A13 and A15). Also, we compute the minimum number of individuals that should be genotyped at generations 1 to t, so that in generation t at least one individual with the requested genotype at all markers x and y is obtained with a given risk. The requested genotype at all markers can be obtained if there is no recombination between the x_i's, as in the foreground selection case, and in addition if there is at least one recombination between y_1 and x_1, and between x_m and y_2. Hence, when more than one generation is considered, the calculation of minimal population size for background selection is slightly more complicated than for foreground selection only (Equation 4). The recursion equations for the calculation of minimal population size for background selection are derived in the APPENDIX (Equations A16, A17, A18 and A19) in the case when one and only one individual is retained after the background selection step at each generation. In these equations, the number of genotyped individuals was allowed to possibly differ at each generation and is denoted $n[u]$ ($1 \leq u \leq t$). Minimal population sizes can be derived by solving (A22) numerically.

Combined foreground and background selection with one or several QTLs was investigated through computer simulations. The model involves 10 pairs of chromosomes of length 150 cM. Each chromosome is described by 151 equally spaced loci. Two alleles per locus are considered (donor or recipient type). Crossing overs are simulated assuming no interference. At each generation the genotypes of the requested number N of backcross progenies are generated, then among those, all (N') individuals carrying the donor type allele at all markers x are selected. Finally, among the N' individuals, one individual is retained based on its genotype at background selection markers. In the simulations, background selection is based not only on markers located on the chromosome(s) carrying the QTL(s) (carrier chromosomes), but also on noncarrier chromosomes. Each background selection marker locus is given a score of 1 (if homozygous for the recipient type allele) or 0 (if heterozygous), and the scores are combined in a selection index in which different weights can be assigned to markers on carrier or noncarrier chromosomes, so that background selection is performed on the former or the latter type of markers in priority. For example, to give priority to markers on carrier chromosomes, weights were chosen such that the weight of any marker on a carrier chromosome was greater than the sum of weights of all markers on noncarrier chromosomes. The probability of having the donor type alleles at the QTLs, as well as the percentage of recipient type genome (*genomic similarity*) on both carrier and noncarrier chromosomes are estimated from the genotypes at corresponding loci. These estimations are performed three times for each generation: among the N backcross progenies before selection, among the N' individuals selected after the foreground selection step, and among the individual(s) selected after the background selection step.

CONTROL OF THE QTLs

Foreground selection only: We first investigate the case when markers are only used to assess the presence of the donor allele at the QTL(s). At each generation, selection is for the donor allele at markers.

One QTL: The calculation of $P_{Q|M}[t]$ presented in METHODS and the numerical applications below depend on the assumptions made for $g[x]$. In a given experiment, values for $g[x]$ may be derived for example from the LOD curve provided by the QTL detection program, but the predictions obtained in this framework would be restricted to particular data whereas we want here to derive general conclusions. In expectation, the distribution of $g[x]$ is likely to be Gaussian, as assumed in the APPENDIX (VISSCHER *et al.* 1996). The results presented thereafter are also obtained under this framework. Then, one needs to choose realistic values for the mean x_0 and variance σ^2 of $g[x]$. We assume here that x_0 is at the estimated position of the QTL ($x_0 = x_l$), as done by VISSCHER *et al.* To choose a value for σ we assume that, in addition to the estimated position of the QTL, a confidence interval was also provided (LANDER and BOTSTEIN 1989; DARVASI *et al.* 1993; MANGIN *et al.* 1994). Let x_{inf} and x_{sup} be, respectively, the lower and upper bounds of the confidence interval on the arbitrary scale. We compute σ *a posteriori* as the solution of the equation:

$$\int_{x_{inf}}^{x_{sup}} g[x]\,dx = 1 - \alpha_{CI}, \qquad (12)$$

where α_{CI} is the risk associated with the confidence

TABLE 1

Confidence intervals for QTL location

σ	$S_{0.01}$	$S_{0.05}$	$S_{0.10}$
0.97	5.0	3.8	3.2
1.94	10.0	7.6	6.4
2.91	15.0	11.4	9.6
3.88	20.0	15.2	12.8
5.82	30.0	22.8	19.2
7.76	40.0	30.4	25.5
9.71	50.0	38.0	31.9
11.65	60.0	45.7	38.3
13.59	70.0	53.3	44.7

σ, standard error of the normal distribution of the true given the expected position of the QTL; $S_{\alpha_{CI}}$, corresponding length of the confidence interval at α_{CI} risk level, when the expected position of the QTL is at the center of the confidence interval. Values in centiMorgans.

interval. Approximate numerical solutions of (12) are given in Table 1 for different values of α_{CI} and of $S = x_{sup} - x_{inf}$, assuming that x_l is at the center of the confidence interval. We considered that confidence interval length S was a more meaningful parameter than σ. Hence, we will use thereafter S as a parameter of the model, taking $\alpha_{CI} = 0.01$ as the reference value. The corresponding value of σ can be inferred from the length of the confidence interval S at an α_{CI} risk level, using Table 1. It is important to notice that different combinations of S and α_{CI} values corresponding to the same σ (Equation 12) give identical results, so that the results shown in Tables 2 to 6 below for $\alpha_{CI} = 0.01$ can be easily applied to a different combination (S, α_{CI}) using Table 1.

Previous developments allow computation of numerical values for P_M, N_α and $P_{Q|M}$ as function of L, t, x_0, S, m and the x_i's. We consider first that the estimated position of the QTL lies far from the ends of the chromosome, so that $g[x]$ gets close to zero at these ends. This situation is illustrated by the case when the confidence interval is at the center of the chromosome ($x_0 = L/2$, $x_{inf} = x_0 - S/2$, $x_{sup} = x_0 + S/2$). In this case, optimal marker positions for the first backcross generation ($t = 1$) are given in Table 2 for a chromosome of length 150 cM. For different lengths of the confidence interval and different numbers of markers, optimal marker positions relative to the estimated position of the QTL ($x_i - x_0$) are given along with the corresponding values of $P_{Q|M}$, P_M and N_α for $\alpha = 0.01$. The number $N_{0.01}$ represents the minimum population size requested so that at least one individual carrying the favorable allele at all markers is obtained with a probability of 0.99. Since the confidence interval is at the center of the chromosome, optimal marker positions in Table 2 are always symmetrical with respect to the center of the confidence interval, and for odd numbers of markers the central marker is always at x_0. Optimal marker positions are clearly not evenly spread over the confidence interval, and depend on S and m: the distance between neighboring markers is less for markers closer to x_0. For large values of S, marker positions remain inside the confidence interval for any number of markers, whereas for lower values of S outer markers are outside the confidence interval for large numbers of markers. These computations were performed for the first backcross generation ($t = 1$). It was checked that considering further generations ($t > 1$) only leads to minor modifications of optimal marker positions (~1 cM for the largest confidence interval at $t = 3$, data not shown). This is consistent with the results obtained by VISSCHER (1996) in a simplified case without selection.

It is seen from Table 2 that optimal marker spacing provides remarkably high probabilities $P_{Q|M}$ of having the desired genotype at the QTL given the desired genotype at the markers in almost any situation, and that the corresponding minimum population sizes are always small. For instance, for a 60-cM confidence interval, $P_{Q|M} = 0.99$ can be obtained using only three markers: one at the center of the confidence interval, and each of the two other at 20.5 cM from this center, on both sides. In the same situation, the use of only two markers 28 cM apart still allows a good control of the genotype at the QTL: $P_{Q|M} = 0.978$. For short confidence intervals, the number of markers has almost no visible effect on $P_{Q|M}$ and N_α, and even for longer confidence intervals the only marked effect is seen between $m = 1$ and $m = 2$. Hence, using at most $m = 2$ markers for a single-QTL introgression program with no background selection on the chromosome carrying the QTL seems sufficient in the first backcross generation.

The results for one QTL in the third backcross generation ($t = 3$) are reported in Table 3 ($q = 1$). Compared with the results in Table 2, minimal population sizes are only slightly increased, all $P_{Q|M}$ values are decreased, and the effect of varying the number of markers becomes more important. Using more than two markers for a single QTL is justified for large confidence intervals ($S \geq 40$ cM), or projects that will be run over several generations.

Second, we consider what happens when the estimated position of the QTL gets closer to one end of the chromosome, so that $g[x]$ differs from zero at this end. This situation is illustrated by the case when one edge of the confidence interval meets exactly one edge of the chromosome ($x_0 = S/2$, $x_{inf} = 0$, $x_{sup} = S$). In this case, optimal marker positions in the first backcross generation are given in Table 4 for the same chromosome length as in Table 2. Optimal marker positions in Table 4 are no longer symmetrical with respect to the center of the confidence interval. For odd numbers of markers (except $m = 1$), central marker position is no longer at x_0 and marker positions are moved "to the right" compared with the values shown in Table 2, the variation being more important for markers that

TABLE 2

Optimal marker positions when confidence interval is at center of chromosome

S	m	$x_1 - x_0$	$x_2 - x_0$	$x_3 - x_0$	$x_4 - x_0$	$x_5 - x_0$	$P_{Q\|M}$	P_M	$N_{0.01}$
10	1	0.0					0.985	0.500	7
	2	−3.6	+3.6				0.999	0.466	8
	3	−4.7	0.0	+4.7			1.000	0.456	8
	4	−5.4	−1.4	+1.4	+5.4		1.000	0.450	8
	5	−5.9	−2.2	0.0	+2.2	+5.9	1.000	0.445	8
20	1	0.0					0.970	0.500	7
	2	−6.2	+6.2				0.996	0.445	8
	3	−8.5	0.0	+8.5			0.998	0.425	9
	4	−9.9	−2.6	+2.6	+9.9		0.999	0.413	9
	5	−10.8	−4.1	0.0	+4.1	+10.8	1.000	0.405	9
40	1	0.0					0.944	0.500	7
	2	−10.4	+10.4				0.987	0.415	9
	3	−14.9	0.0	+14.9			0.995	0.379	10
	4	−17.7	−4.8	+4.8	+17.7		0.997	0.358	11
	5	−19.6	−7.8	0.0	+7.8	+19.6	0.998	0.345	11
60	1	0.0					0.919	0.500	7
	2	−14.0	+14.0				0.978	0.393	10
	3	−20.5	0.0	+20.5			0.990	0.346	11
	4	−24.6	−6.8	+6.8	+24.6		0.995	0.318	13
	5	−27.5	−11.2	0.0	+11.2	+27.5	0.997	0.300	13

Foreground selection only with a single QTL at $t = 1$. S, length of the confidence interval at $\alpha_{CI} = 0.01$ risk level; m, number of markers; x_0, estimated QTL position; x_i $(1 \le i \le 5)$, optimal marker positions; $P_{Q|M}$, efficiency of foreground selection; P_M, probability of inheritance; $N_{0.01}$, minimum population size. $L = 150$ cM, x_0 75 cM, $[x_{inf}, x_{sup}] = [x_0 - S/2, x_0 + S/2]$.

are "on the left" of the center of the confidence interval (*i.e.*, at positions between 0 and x_0). Although always inferior to the values in Table 2, probabilities $P_{Q|M}$ are still remarkably high in this case for almost all sets of parameters. Note that in this situation the maximum possible value for $P_{Q|M}$ is 0.995. The probability P_M of having the requested alleles at all m markers is slightly increased compared with previous situation, but minimal population size is not affected. In any case, the differences between optimal marker positions in Tables 2 and 4 are small compared with the precision and density of genetic maps for most species.

Several unlinked QTLs: The previous approach can be extended to programs aimed at introgressing simultaneously the favorable alleles at several QTLs. We will first consider the case where q QTLs are unlinked. We assume that all QTLs have confidence intervals of equal length (located at the center of the chromosome) and that each QTL is controlled by the same number m of markers.

First, we consider optimal marker positions for each QTL (taken from Table 2). Results are reported in Table 5 for the first backcross generation ($t = 1$) and in Table 3 for $t = 3$. In the case where all QTLs are identical, the number of individuals that need to be genotyped at each generation increases exponentially with the number of QTLs that are considered, the increase in N_α per QTL being approximately a factor $1/P_{M(1)}$ for q not too small (Equations 9 and 10). This is illustrated in Table 2 and 5: when considering for instance 40-cM confidence intervals and two markers per QTL, the number of individuals to be genotyped for a one generation project ($t = 1$) are 9, 25, 63, 154, 373 and 901 for one, two, three, four, five and six QTLs, respectively. If population size is set at the value $N_{0.01}$ corresponding to each set of parameters, the actual number of individuals with the requested genotype may be greater than one, but still not very important: the expected number $N_{0.01} \times P_M[t]$ of individuals carrying the requested allele at all markers varies from 3.3 to 4.6 for $t = 1$ (Table 2 and 5) and from 4.1 to 5.7 for $t = 3$ (Table 3). In such cases, few if any individuals are available for background selection at each generation. Hence, the population sizes shown in Tables 5 and 3 can be used to determine an upper bound to the number of unlinked QTLs that can be transferred simultaneously, with given experimental means. It seems illusive to work with more than four QTLs, unless very large population sizes can be considered, or the precision of the QTL location is very high.

In the first generation, the results of Table 5 indicate that for up to four QTLs, reasonably high values of $P_{Q|M}$ can be obtained with no more than two markers per QTL (and realistic population sizes): only for three or four QTLs with confidence intervals of length 60 cM is $P_{Q|M}$ reduced below 0.95 (and yet still above 0.90). This is no longer true in the third backcross generation, as can be seen from the results in Table 3: $P_{Q|M}$ values are then substantially reduced, so that using $m = 2$ markers appears sufficient only if confidence intervals no more

TABLE 3

Efficiency of foreground selection and minimum population size for a three-generations backcross program

		$q = 1$		$q = 2$		$q = 3$		$q = 4$		$q = 5$		$q = 6$	
S	m	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$
10	1	0.956	9	0.913	20	0.873	43	0.834	89	0.797	180	0.762	362
	2	0.996	10	0.991	24	0.987	54	0.983	118	0.979	256	0.975	551
	3	0.999	10	0.997	25	0.996	58	0.995	129	0.993	286	0.992	630
	4	0.999	10	0.999	26	0.998	60	0.997	137	0.997	308	0.996	687
	5	1.000	10	0.999	26	0.999	62	0.999	143	0.998	324	0.998	730
20	1	0.915	9	0.838	20	0.767	43	0.702	89	0.643	180	0.588	362
	2	0.987	10	0.974	26	0.962	62	0.949	143	0.937	324	0.925	731
	3	0.995	11	0.991	29	0.986	72	0.982	173	0.977	409	0.973	967
	4	0.998	11	0.995	31	0.993	79	0.991	194	0.989	472	0.986	1147
	5	0.999	11	0.997	32	0.996	83	0.995	209	0.993	518	0.992	1283
40	1	0.845	9	0.713	20	0.603	43	0.509	89	0.430	180	0.363	362
	2	0.964	11	0.928	31	0.895	77	0.862	190	0.831	461	0.800	1115
	3	0.985	12	0.970	37	0.956	102	0.942	273	0.928	722	0.914	1907
	4	0.992	13	0.984	42	0.976	121	0.969	343	0.961	961	0.953	2685
	5	0.995	14	0.990	46	0.985	137	0.981	402	0.976	1170	0.971	3398
60	1	0.785	9	0.616	20	0.483	43	0.379	89	0.298	180	0.234	362
	2	0.936	12	0.876	35	0.820	92	0.768	237	0.718	607	0.672	1550
	3	0.971	14	0.944	45	0.917	135	0.890	396	0.865	1148	0.840	3322
	4	0.984	15	0.968	54	0.953	174	0.938	552	0.923	1739	0.908	5467
	5	0.990	16	0.980	61	0.970	209	0.961	702	0.951	2346	0.941	7827

Foreground selection only at $t = 3$. q, number of QTLs; S, length of the confidence interval at $\alpha_{CI} = 0.01$ risk level; m, number of markers; $P_{Q|M}$, efficiency of foreground selection; $N_{0.01}$, minimum population size. $L = 150$ cM, $x_0 = 75$ cM, $[x_{inf}, x_{sup}] = [x_0 - S/2, x_0 + S/2]$.

than 20 cM long are considered. In the other situations, the number of markers has to be increased to ensure higher $P_{Q|M}$ values. However, one has to increase population size simultaneously [*e.g.*, from 92 to 174 if using $m = 4$ markers ($P_{Q|M} = 0.953$) instead of $m = 2$ markers ($P_{Q|M} = 0.82$) for three QTLs with 60-cM confidence intervals]. In this case, experimental costs increase dramatically. Thus, in general, a decision must be taken between (i) setting a lower bound to $P_{Q|M}$, and thus increasing markers number and population size or (ii) setting an upper bound for the cost of the experiment (population size and genotyping) by limiting the number of markers used to control each QTL, and hence accepting a relatively high risk of "loosing" the favourable allele for at least one of the QTLs. Experimental results relating to the latter strategy were provided by STUBER and SISCO (1992).

The values shown in Tables 5 and 3 were obtained with optimal marker spacing (with respect to $P_{Q|M}$) for each set of parameters. In particular situations, such as for example the case of $q = 4$ QTLs with confidence intervals of length $S_{0.01} = 40$ cM, it may be interesting to compare the respective sensitivity of $P_{Q|M}$ and N_α to the deviations of marker positions from the optimum. This was done by "moving" the markers using the following model for $m = 2$–5 markers. The positions of the outer markers x_1 and x_m were set at $x_0 - d/2$ and $x_0 + d/2$, respectively. For $m = 3$ or 5, the position of the central marker was set at x_0, and for $m = 4$ or 5, the position of the inner markers x_2 and x_{m-1} were optimized with respect to $P_{Q|M}$ given the positions of the outer markers. Parameter d was varied from d^* to 0, where d^* corresponds to optimal marker spacing. For each number of markers, the couples of values (N_α, $P_{Q|M}$) corresponding to each value of d are plotted in Figure 1 (dotted lines) for $t = 3$, $q = 4$, $S_{0.01} = 40$ cM, $L = 150$ cM, $x_0 = 75$ cM. The points corresponding to optimal marker spacing are identified by the diamonds for each number of markers. Note that the coordinates of these points may be slightly different from the values in Table 3, since optimal marker spacing in the figure is for $t = 3$ whereas optimal spacing at $t = 1$ was used in both Table 5 and 3, but the difference if any is very small (see below and the comments on Table 2).

It is seen from Figure 1 that the shape of the optimum in $P_{Q|M}$ depends on the number of markers and flattens when larger values of m are considered. Also, the curves corresponding to different numbers of markers all tend toward a common value for small population sizes (*i.e.*, low d values). The consequences are twofold. First, as more markers are taken into account, $P_{Q|M}$ becomes less sensitive than N_α to the variation of marker positions. Hence, it is possible to reduce population size substantially without increasing much the risk of having the requested alleles at all markers but not at all QTLs. For example, with m = three markers, reducing the distance $x_3 - x_1$ between outer markers from 29.6 cM (optimum) to 23.8 cM would reduce minimum popula-

TABLE 4

Optimal marker positions when confidence interval is at one edge of chromosome

S	m	$x_1 - x_0$	$x_2 - x_0$	$x_3 - x_0$	$x_4 - x_0$	$x_5 - x_0$	$P_{Q\|M}$	P_M	$N_{0.01}$
10	1	0.0					0.980	0.500	7
	2	−3.4	+3.6				0.994	0.467	8
	3	−4.3	+0.1	+4.8			0.995	0.457	8
	4	−4.7	−1.1	+1.5	+5.5		0.995	0.452	8
	5	−4.8	−1.8	+0.2	+2.3	+6.0	0.995	0.450	8
20	1	0.0					0.966	0.500	7
	2	−6.0	+6.2				0.991	0.446	8
	3	−8.0	+0.2	+8.6			0.994	0.426	9
	4	−8.9	−2.2	+2.8	+10.0		0.994	0.416	9
	5	−9.3	−3.6	+0.3	+4.4	+10.9	0.995	0.411	9
40	1	0.0					0.940	0.500	7
	2	−10.3	+10.5				0.983	0.415	9
	3	−14.4	+0.2	+15.0			0.990	0.381	10
	4	−16.5	−4.3	+5.0	+17.8		0.993	0.362	11
	5	−17.7	−7.1	+0.5	+8.1	+19.8	0.994	0.350	11
60	1	0.0					0.915	0.500	7
	2	−13.8	+14.1				0.973	0.393	10
	3	−20.0	+0.2	+20.7			0.986	0.347	11
	4	−23.4	−6.3	+7.1	+24.8		0.990	0.321	12
	5	−25.5	−10.4	+0.5	+11.6	+27.8	0.992	0.304	13

Foreground selection only with a single QTL at $t = 1$. S, length of the confidence interval at $\alpha_{CI} = 0.01$ risk level; m, number of markers; x_0, estimated QTL position; x_i ($1 \leq i \leq 5$), optimal marker positions; $P_{Q|M}$, efficiency of foreground selection; P_M, probability of inheritance; $N_{0.01}$, minimum population size. $L = 150$ cM, $x_0 = S/2$, $[x_{inf}, x_{sup}] = [0, S]$.

tion size from 271 (optimum) to 221 while only reducing $P_{Q|M}$ from 0.942 (optimum) to 0.932. Second, for any number of markers, adding one inner marker and slightly reducing the distance between outer markers always increases $P_{Q|M}$ without affecting population size. For example, while for $m = 2$ markers with $x_2 - x_1 = 20.6$ cM, the optimum is at $P_{Q|M} = 0.862$ and $N_\alpha = 189$, considering $m = 3$ markers with $x_3 - x_1 = 19.6$ cM and $x_2 = x_0$ would increase $P_{Q|M}$ up to 0.907 with corresponding $N_\alpha = 190$. This would only increase the cost of the experiment by the cost of genotyping one more marker per QTL for the same number of individuals. Moreover, with one more marker, it is always possible to both increase $P_{Q|M}$ and reduce the population size, or at the limit to reduce population size for the same $P_{Q|M}$. For example, considering $m = 4$ markers with $x_4 - x_1 = 23.2$ cM and $x_3 - x_2 = 7$ cM instead of $m = 3$ markers with optimal positions would reduce population size from 271 to 220 for the same $P_{Q|M}$. Since the cost of genotyping is roughly proportional to the product of population size by the number of markers, this could in some cases reduce the cost of the experiment for the same risk on QTLs control.

These considerations may alter previous conclusions drawn from Tables 5 and 3. Considering markers at positions differing from the optima defined in Table 2 could (i) allow using more than two or three markers for $q \leq 4$, while increasing foreground selection efficiency and/or reducing the cost of the experiment, or (ii) make it possible to monitor more than four QTLs with both reasonable population size and acceptable risk on foreground selection efficiency.

Linked QTLs: In the framework of the foreground selection step in a backcross program, only coupling associations have to be considered (*i.e.*, linkage between QTLs for which the favorable allele is carried by the donor parent), since the control of QTLs for which the favorable allele is carried by the recipient parent is a matter for the background selection step. Linkage between QTLs first modifies the probability that a given backcross progeny carries the donor allele at all markers (controlling q QTLs). In this situation, (9) becomes

$$P_M = P_{M(1)} \prod_{l=2}^{q} (2(1 - r_{l-1,l}) P_{M(l)}), \qquad (13)$$

where $r_{l-1,l}$ is the recombination rate between the last marker controlling QTL $l - 1$ and the first marker controlling QTL l, when markers are ordered following (1) chromosome number and (2) map positions on a same chromosome. Equation 13 shows that linkage between QTLs increases P_M, when compared with the independence situation studied previously. Thus, the number of individuals to be genotyped at each generation has to be revised downward in the case of linkage. Linkage between QTLs also modifies probability $P_{Q|M}$. It would be possible to extend the equations for P_1, $P_{2,k}$ and P_3 described in the APPENDIX to the case when two or more normal distributions $g[x]$ are mixed on the same chromosome, but it is not certain that the joint distribution of the true given the expected positions of

TABLE 5

Efficiency of foreground selection and minimum population size in the first backcross generation

		$q = 2$		$q = 3$		$q = 4$		$q = 5$		$q = 6$	
S	m	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$	$P_{Q\|M}$	$N_{0.01}$
10	1	0.970	17	0.955	35	0.941	72	0.927	146	0.913	293
	2	0.997	19	0.996	44	0.994	95	0.993	207	0.991	445
	3	0.999	20	0.999	47	0.998	105	0.998	231	0.997	509
	4	1.000	21	0.999	49	0.999	111	0.999	249	0.999	555
	5	1.000	21	1.000	50	1.000	115	0.999	262	0.999	590
20	1	0.942	17	0.914	35	0.887	72	0.861	146	0.835	293
	2	0.991	21	0.987	50	0.983	116	0.978	262	0.974	591
	3	0.997	24	0.995	58	0.994	139	0.992	331	0.991	781
	4	0.998	25	0.998	64	0.997	157	0.996	382	0.995	927
	5	0.999	26	0.999	67	0.998	169	0.998	419	0.997	1036
40	1	0.890	17	0.840	35	0.793	72	0.748	146	0.706	293
	2	0.975	25	0.963	63	0.951	154	0.939	373	0.927	901
	3	0.990	30	0.985	82	0.980	220	0.975	584	0.970	1541
	4	0.995	34	0.992	98	0.989	277	0.987	776	0.984	2169
	5	0.997	37	0.995	111	0.993	325	0.992	945	0.990	2745
60	1	0.845	17	0.776	35	0.714	72	0.656	146	0.603	293
	2	0.956	28	0.934	74	0.913	192	0.893	491	0.873	1252
	3	0.980	37	0.971	109	0.961	320	0.952	927	0.942	2684
	4	0.989	44	0.984	141	0.979	446	0.973	1405	0.968	4417
	5	0.993	49	0.990	169	0.987	567	0.983	1895	0.980	6323

Foreground selection only at $t = 1$. q, number of QTLs; S, length of the confidence interval at $\alpha_{CI} = 0.01$ risk level; m, number of markers; $P_{Q|M}$, efficiency of foreground selection; $N_{0.01}$, minimum population size. $L = 150$ cM, $x_0 = 75$ cM, $[x_{inf}, x_{sup}] = [x_0 = S/2, x_0 + S/2]$.

several QTLs linked on the same chromosome would be multi-normal. In fact, it is likely that the distributions for different QTLs on the same chromosome would not be independent, but the theory in this domain remains unexplored. Hence, this problem was not considered formally. However, with same numbers m of markers per QTL, only slight positive modifications are expected compared with the previous case since (i) for a given QTL (l), considering additional markers on the chromosome can only increase $P_{Q|M(l)}$ and (ii) $P_{Q|M}$ values reported in Tables 5 and 3 are already high under the hypothesis of independence.

Combined foreground and background selection on the carrier chromosome: To perform background selection on a chromosome carrying a QTL, we consider two additional markers (denoted y_1 and y_2) surrounding the markers x_i devoted to foreground selection. At each generation all individuals carrying the donor type allele at all markers x_i are selected in the first step. Then, among these individuals one or several individuals carrying the most alleles of recipient type at markers y_1 and y_2 are selected in the second step. In the present framework of background selection (see METHODS), we request that at least one recombination event takes place on both sides of the QTL (between y_1 and x_1, and between x_m and y_2). As already noticed by YOUNG and TANKSLEY (1989), the probability that this goal is fulfilled is much lower in a single generation than in two generations, when single recombination events can take place one at a time at each generation. Our calculations provide an analytic demonstration of this (see APPENDIX). The minimal population size for a single generation project is then unrealistically large in most cases (results not shown). Therefore we only consider here a two-generations project.

Numerical applications of the analytic approach described above are presented in Table 6 for a single QTL in BC_2. The parameters (L, m, S) considered are the same as in Tables 5 and 3. In addition, we consider different lengths S^* of the segment [y_1, y_2] contained between markers devoted to background selection. The percentage of recipient type genome (*genomic similarity*) on the carrier chromosome is expected to increase when S^* decreases thus, the positions of markers y determine the efficiency of background selection. We wish now to investigate the efficiency of foreground selection, given that background selection is performed. Considering background selection on y_1 and y_2 modifies the conditional probability of having the requested allele at the QTL. Then, it also modifies the optimal positions of markers x_i devoted to foreground selection. Hence, new optimal positions were derived in the case of background selection for each set of parameters. In some cases (see below), true optimal positions of markers x_1 and x_m can be very close to y_1 and y_2, respectively, requiring very large population sizes (data not shown). To avoid this, a lower bound of 1 cM was artificially set for the distance between these markers in the

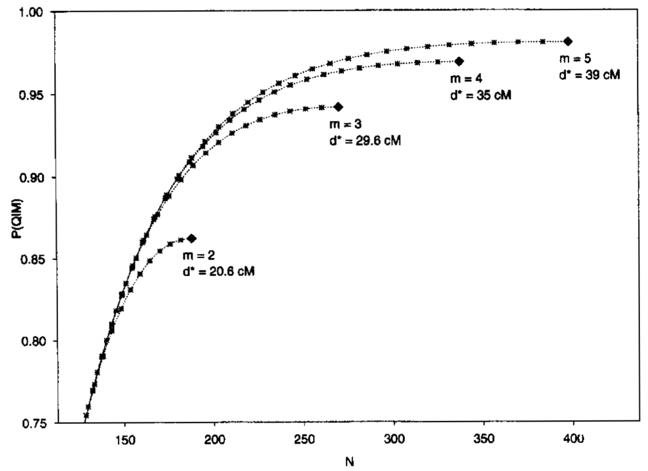

FIGURE 1.—Effect of marker spacing on the relation between foreground selection efficiency and minimum population size in the third backcross generation. Abscissa: minimum population size $N_\alpha[t]$. Ordinate: conditional probability $P_{Q|M}[t]$. For each number m of markers, each dotted line represents the couples of values ($N_\alpha[t]$, $P_{Q|M}[t]$) corresponding to the variations of markers positions (see explanations in text). The points identified by the diamonds correspond to optimal marker spacing ($d = d^*$). Stars show variation of d from d^* to 0 by step -1 cM. Results for $q = 4$ QTLs, $\alpha = 0.01$, $S_{0.01} = 40$ cM, $L = 150$ cM, $x_0 = 75$ cM and $t = 3$.

numerical optimizations. This limited population sizes, while only slightly reducing $P^*_{Q|M}$ when compared with the optimum.

Though all marker positions were optimized to produce the results in Table 6, not all positions are given. Since these positions are symmetrical with respect to x_0, they are partially described in Table 6 by parameter $d = x_m - x_1$ (except for $m = 1$). Strictly speaking, this only gives exact positions for up to three markers. Yet, parameter d determines the length ($(S^* - d)/2$) of the segments on which at least one recombination event is requested and is the most relevant in the calculation of minimal population sizes. As expected, when S^* is large compared with S, optimal positions of markers x are close to the ones obtained for foreground selection only (Table 2), though the corresponding conditional probabilities are lower, even when $S^* = L$. Optimal marker spacing should fulfil two conditions: (1) provide a good control of the chromosomal region inside $[x_1, x_m]$ and (2) a good control of the region outside $[x_1, x_m]$. When background selection is performed, the outside region is bounded by markers y_1 and y_2, so that condition (2) is best fulfilled when the distance between markers x and y is close to zero. The important consequence is that the optimal distance between markers x_1 and x_m *increases* when S^* decreases (until it is bounded). For short confidence intervals, both conditions can be fulfilled with only two markers, giving a distance d close to S^* and a high $P^*_{Q|M}$. For longer confidence intervals, only condition (1) is fulfilled with two markers, giving a greater distance between markers x and y and a lower $P^*_{Q|M}$. It is then useful to work with more than two markers per QTL, so that inner markers deal with condition (1), while outer markers deal with condition (2).

The minimum numbers of individuals that should be genotyped in first ($n[1]$) and second ($n[2]$) generations to obtain with 1% risk at least one individual with requested genotype at the end of the program (generation 2) are given in Table 6 for all sets of parameters. These numbers were derived numerically from (A22), allowing population sizes to be possibly different at each generation. When more than one couple ($n[1]$, $n[2]$) of solutions were found, the couple with (1) minimal sum then (2) minimal difference was retained, providing the best repartition of experimental means over the two generations. For large distances S^* between background markers, population sizes are not much greater than the ones obtained with foreground selection only for a single QTL, as expected. In this case, using constant or variable population sizes approximately leads to the same total number of individuals to be genotyped over the two generations. For lower values of S^*, minimal population sizes are greatly increased, compared with foreground selection for a single QTL. In this case, using variable population sizes at each generation allows the total number of individuals to be substantially reduced. For example, for a two-generations project with two markers and $S = 10$ cM, constant numbers of individuals that should be genotyped at each generation are 995, 202, 101, 49 and 34 for $S^* = 10$, 20, 30, 50 and 70 cM, respectively. For a three-generations project (results not shown), the difference between total numbers of individuals with constant or variable population sizes is reduced. Note that the risk considered here is not the risk of a failure of the experiment, as was the case in Tables 5 and 3, so that minimum population sizes given in Table 6 are not mandatory.

The design of an experiment should both insure a high probability $P^*_{Q|M}$ of having the donor type allele at the QTL and reduce the length of the donor-type segment of genome retained around the QTL, which depends on S^* and d. For a given confidence interval S, it is seen from Table 6 that increasing the distance S^* between markers y leads to an increase in probability $P^*_{Q|M}$ and a reduction in population size, while increasing the length of the donor segment. Increasing the number of markers increases both $P^*_{Q|M}$ and population size. It has to be noticed that the effect of a variation in the number of markers on $P^*_{Q|M}$ is more important than the effect of a variation in S^*. Hence, the design of a background selection experiment could be optimized by (1) choosing enough markers to insure a high probability $P^*_{Q|M}$ then (2) adjusting the distance S^* with respect to the sought efficiency and the available experimental means. In addition, the positions of markers x could be modified with respect to their opti-

TABLE 6

Effect of background selection on foreground selection efficiency, marker spacing and minimum population size in a two-generation backcross program

		$m = 1$			$m = 2$				$m = 3$				$m = 4$				$m = 5$			
S	S^*	$P^*_{Q\|M}$	$n[1]$	$n[2]$	$P^*_{Q\|M}$	d	$n[1]$	$n[2]$	$P^*_{Q\|M}$	d	$n[1]$	$n[2]$	$P^*_{Q\|M}$	d	$n[1]$	$n[2]$	$P^*_{Q\|M}$	d	$n[1]$	$n[2]$
10	10	0.687	118	200	0.976	8	624	1077	0.978	8	628	1076	0.978	8	629	1076	0.978	8	625	1080
	20	0.838	62	100	0.995	10	129	221	0.999	12	166	280	0.999	14	222	383	1.000	14	227	379
	30	0.889	43	67	0.995	10	67	111	0.999	12	78	124	0.999	12	77	125	1.000	14	88	143
	50	0.928	27	42	0.996	8	33	54	0.999	10	35	58	1.000	12	39	61	1.000	14	42	65
	70	0.944	21	31	0.996	8	24	38	0.999	10	25	40	1.000	12	28	41	1.000	12	28	41
20	20	0.683	62	100	0.972	18	679	1177	0.983	18	684	1187	0.982	18	687	1187	0.985	18	692	1187
	30	0.781	43	67	0.982	18	118	196	0.995	24	245	417	0.998	26	375	638	0.999	28	752	1311
	40	0.830	32	52	0.985	16	61	97	0.996	22	82	139	0.998	24	97	157	0.999	26	110	185
	60	0.878	23	36	0.988	14	33	52	0.996	20	41	62	0.998	22	43	67	0.999	24	46	72
	80	0.901	20	27	0.989	14	26	37	0.996	18	28	41	0.998	22	30	46	0.999	24	32	48
40	40	0.673	32	52	0.941	30	153	259	0.976	38	818	1406	0.983	38	822	1427	0.985	38	831	1432
	50	0.729	27	42	0.952	28	72	117	0.983	40	171	287	0.992	48	903	1547	0.995	48	898	1573
	60	0.766	23	36	0.958	26	49	75	0.985	38	80	130	0.993	44	115	189	0.996	50	191	319
	80	0.811	20	27	0.965	24	30	48	0.987	34	41	62	0.993	40	48	76	0.996	46	60	93
	100	0.836	17	23	0.968	24	25	36	0.988	34	31	45	0.994	40	34	53	0.996	44	36	60
60	60	0.662	23	36	0.910	38	74	128	0.963	56	475	807	0.978	58	979	1683	0.983	58	993	1705
	70	0.700	21	31	0.922	36	52	81	0.968	52	107	177	0.983	64	344	592	0.989	68	1074	1856
	80	0.728	20	27	0.930	34	39	61	0.971	50	64	108	0.985	60	105	175	0.991	68	187	311
	100	0.766	17	23	0.933	32	29	42	0.975	48	42	62	0.987	56	52	79	0.992	64	66	104
	120	0.789	15	20	0.944	32	24	34	0.977	46	29	47	0.988	54	36	55	0.992	60	41	64

Combined foreground and background selection with one QTL at $t = 2$. m, number of foreground selection markers x; S, length of the confidence interval at $\alpha_{CI} = 0.01$ risk level; $S^* = y_2 - y_1$, distance between background selection markers; $P^*_{Q|M}$, efficiency of foreground selection under combined foreground and background selection; $d = x_m - x_1$, marker spacing; $n[1]$, $n[2]$, minimum population size in first and second generation, respectively. $L = 150$ cM, $x_0 = 75$ cM, $[x_{inf}, x_{sup}] = [x_0 - S/2, x_0 + S/2]$.

mal values, as was proposed previously. The conclusions drawn from Figure 1 apply in the case of background selection. Moreover, the shape of the optimum depends on S^*, and flattens when S^* decreases, so that using several markers with altered positions is even more interesting when the distance S^* between markers y is short.

Remark on minimal population size: It should be noted that the calculations above were done in a framework in which a single individual can be selected at each generation. This is relevant to most plant breeding programs that aim at homozygous inbred line development, when the size of the progeny of selected individuals is high enough. In most animal species, but also some plant species, more individuals with the desired genotype are needed for a successful introgression program, when inbreeding has to be limited or when progeny size is small. For instance in livestock species, it may be practical to use males from the recipient population rather than males from the donor population, to reduce genetic lag for other traits. In this situation, females of the crossbred population are selected on their marker genotype, and because of their small reproductive capacity, many have to be selected. Although the number of individuals that have to be selected at a given generation does not affect conclusions on optimal marker positions, it limits selection intensity for both foreground and background selection.

For foreground selection only the probability of obtaining at least k individuals with the donor allele at all the markers over t generations is

$$P_N[t] = \left(\sum_{i=k}^{N} C_N^i P_M^i (1 - P_M)^{N-i}\right)^t, \qquad (14)$$

and the corresponding minimum number of individuals is obtained by solving

$$P_N[t] = 1 - \alpha, \qquad (15)$$

which can be done numerically. Using (14) and (15) instead of (2) and (4), respectively, would lead to larger populations sizes than the ones given in Tables 2–5. Hence, the maximal number of unliked QTLs that can be monitored simultaneously may have to be revised downward.

For combined foreground and background selection, the extension of (A.16) – (A.19) is more complicated and was not considered here. In any case, this would also lead to much larger populations sizes than the ones in Table 6, so that applying strong background selection on the carrier chromosome in the case where many more than one individual have to be selected at each generation would be hardly possible.

CONTROL OF GENETIC BACKGROUND

The aim of background selection is to accelerate the return to the recipient parent genotype, compared with what would be obtained at random. Markers devoted to background selection could be located on the chromosomes carrying the introgressed QTLs (carrier chromosomes) or on the other chromosomes (noncarrier chromosomes). Without background selection, the percentage of donor type genome on noncarrier chromosomes is classically reduced by one-half on an average at each generation (50% in the F_1, 25% in BC_1, 12.5% in BC_2, *etc.*). The reduction is slower on carrier chromosomes, because foreground selection induces a linkage drag of donor type genome around the introgressed QTL, or better said around the markers devoted to the control of the QTL. Hence, whereas markers devoted to background selection on noncarrier chromosomes should be placed so as to control the whole chromosome, markers on the carrier chromosomes should first be used to reduce the length of the donor type segment of genome dragged along with the QTL. The case of known gene introgression ($S = 0$) was investigated by HOSPITAL *et al.* (1992). When the introgressed gene is a QTL ($S > 0$), placement of the markers should in addition take account of the uncertainty on the localization of the gene, but the conclusions of HOSPITAL *et al.* (1992) still apply to the case of QTL introgression. We want here to provide a brief overview of the efficiency of background selection in the case of the introgression of several QTLs.

In the previous section we focused on a single carrier chromosome, and we restricted it to the case where it is possible to obtain at least one individual with the requested genotype at *all* markers x and y. This restriction provides the best efficiency for background selection on the carrier chromosome, but implies that fairly large numbers of individuals must be manipulated, even for a single QTL. The analytic approach could be formally extended to any number of QTLs but would certainly lead to unrealistic population sizes. Also, this approach does not deal with background selection on noncarrier chromosomes. Yet, even when it is unlikely that one individual with the perfect genotype can be obtained, it is always possible to contemplate performing background selection among the individuals carrying the donor allele at all markers x (provided that the population size recommended in the foreground selection section is used). In this situation, one would like to know what efficiency of background selection can be expected, on both carrier and noncarrier chromosomes. This was studied through computer simulations.

We consider a backcross breeding program aimed at introgressing four QTLs in a diploid species with 10 chromosome pairs, each of 150 cM long. Each QTL is located on a different chromosome, with its expected position at the center of the chromosome ($x_l = x_0 = 75$ cM) and confidence interval of length $S(1\%) = 40$ cM. We consider up to four generations of selection (foreground and background). Background selection is performed using two markers (y) on each carrier chromosome, at positions 45 and 105 cM ($S^* = 60$ cM), and three markers per noncarrier chromosome at positions 25, 75 and 125 cM. Foreground selection is performed at each generation using three markers (x) per QTL, at positions 56, 75 and 94 cM (optimal positions in Table 6).

Simultaneous design: First, we consider a backcross breeding program where at each generation foreground selection is applied to the four QTLs simultaneously. Given the positions of markers x, the minimal population size (for foreground selection only) is $N_{0.01} = 377$. We then consider genotyping 400 individuals at each generation.

The results of simulations when background markers on carrier and noncarrier chromosomes were assigned equal weights (see METHODS) are presented in Table 7. Note that the probability of having the donor type allele at the QTLs is computed only on the chromosomes originating from the non-recurrent parent (so that a probability of 1 corresponds to heterozygosity at all the QTLs) as was the case in the APPENDIX and in Tables 2–6, while genomic similarity is computed over both chromosomes for each pair. The similarity on carrier chromosomes was measured on the segments $[0, x_1[$ and $]x_m, L]$ for each chromosome, and $Pr(Q)$ was estimated from the genotype at discrete loci on the segments $]y_1, y_2[$. Note that in Table 7, $Pr(Q)$ after foreground selection can be compared to $P_{Q|M}$, whereas $Pr(Q)$ after background selection cannot be compared to $P^*_{Q|M}$ since the individual retained after the background selection step in the simulations may not have the recipient type allele at all markers y.

In the conditions of Table 7 with four QTLs, the number N' of individuals carrying the donor type allele at markers x for all QTLs is small (6.4 on an average). Yet, the results in Table 7 underline that background selection of a single individual among this limited number of individuals is still efficient and allows a gain of approximately one or two generations to reach given proportion of recipient genome, when compared with the proportions obtained with no background selection. A clear acceleration of the return to the recipient parent is then obtained with limited additional costs (genotyping only N' individuals for the background selection markers). In the first generation, average similarity after background selection is slightly higher than similarity before foreground selection. The gain due to background selection is lower in the following generations, and average similarity after background selection then remains below similarity before foreground selection in the same generation due to the linkage drag around the QTLs.

TABLE 7

Combined foreground and background selection with four QTLs

Gen	Rep	n	$Pr(Q)$	Genomic similarity		
				Carrier	Noncarrier	Average
1	999	400	0.062	0.750 (0.750)	0.750 (0.750)	0.750 (0.750)
		6.4	0.974	0.601 (0.601)	0.750 (0.750)	0.701 (0.701)
		1	0.981	0.614 (0.601)	0.826 (0.750)	0.756 (0.701)
2	998	400	0.061	0.807 (0.800)	0.913 (0.875)	0.878 (0.850)
		6.4	0.956	0.687 (0.677)	0.912 (0.875)	0.838 (0.809)
		1	0.956	0.705 (0.677)	0.943 (0.875)	0.864 (0.809)
3	998	400	0.060	0.852 (0.839)	0.972 (0.938)	0.932 (0.905)
		6.4	0.933	0.758 (0.736)	0.972 (0.938)	0.901 (0.871)
		1	0.929	0.781 (0.736)	0.981 (0.938)	0.915 (0.871)
4	994	400	0.058	0.891 (0.868)	0.990 (0.969)	0.957 (0.936)
		6.4	0.909	0.817 (0.779)	0.991 (0.969)	0.933 (0.906)
		1	0.909	0.836 (0.779)	0.992 (0.969)	0.940 (0.906)

Simulation results over a total of 1000 replicates with equal marker weights. Three rows for each backcross generation (Gen) give the results before selection, after foreground selection or after background selection. For each generation, results are averaged over the number of replicates (Rep) in which at least one individual with the donor allele at all markers x was obtained ($N' \geq 1$). n, number of individuals; $Pr(Q)$, probability of having the donor type allele at all QTLs; Genomic similarity, recipient type genome content on carrier chromosomes, noncarrier chromosomes or on the average. The numbers in parentheses give the results with foreground selection only (no background selection). Ten chromosome pairs, each 150 cM long. S (1%) = 40 cM, x_0 = 75 cM and S^* = 60 cM.

Some elements for the optimization of background selection on carrier *vs.* noncarrier chromosomes were given by HOSPITAL *et al.* (1992), but as stated by these authors, such an optimization depends on particular conditions. Hence, rather than a complete overview, we just investigated four strategies for background selection: background selection on both carrier and noncarrier chromosomes with either (i) priority to carrier chromosomes, (ii) equal weights (conditions of Table 7), (iii) priority to noncarrier chromosomes, or (iv) background selection on noncarrier chromosomes only. As expected, strategy (i) does the best on carrier chromosomes, while strategies (iii) and (iv) are better for noncarrier chromosomes. But on an average over the entire genome, the best efficiency was obtained with strategy (ii) at generations 1-3 (Table 7), while strategy (i) gave the best results in generation 4 only. This is consistent with the results of HOSPITAL *et al.* (1992): since favorable recombination events are rarer on carrier chromosomes than on noncarrier chromosomes, due to linkage drag, giving priority to markers on carrier chromosomes is mostly interesting for programs that are run over several generations. Also, it is clear from Table 7 that background selection on noncarrier chromosomes is no longer efficient in generation 4, and that it is not worth while increasing this efficiency since similarity is already above 0.99 on noncarrier chromosomes before selection. Average similarity can then only be increased through selection on carrier chromosomes. Note that the average similarity over the entire genome is highly dependent on the respective numbers and lengths of carrier and noncarrier chromosomes, and also on the segments on which it is calculated for carrier chromosomes. Hence, this conclusion may not hold for all situations. The important point to be noticed is that strategies (iii) and (iv) give identical results for noncarrier chromosomes, while (iii) still does better than (iv) on carrier chromosomes. Hence, even when priority is given to background selection on noncarrier chromosomes, it is still worth considering markers on carrier chromosomes, as the latter can help in discriminating between individuals with identical scores on noncarrier chromosomes.

The effect of the proportion of individuals selected on background selection efficiency was investigated by HOSPITAL *et al.* (1992, Tables 1 and 2 for noncarrier chromosomes). Note that the proportion selected is the ratio $1/N'$, not $1/N$. In the conditions of our Table 7, this ratio equals to 0.16 but, the efficiency of background selection on noncarrier chromosomes in Table 7 is slightly higher than in HOSPITAL *et al.* (Table 1) for a proportion selected of 0.10, because the total genome size of noncarrier chromosomes considered by these authors was greater (20 chromosomes of 100 cM). Conversely, the efficiency on carrier chromosomes in Table 7 is lower than in HOSPITAL *et al.* (Table 3), because these authors only considered one single carrier chromosome.

Background selection efficiency could be increased by increasing N', and hence the total number N of individuals genotyped at each generation. But the results of HOSPITAL *et al.* (Table 1) show that the increase in selection efficiency on noncarrier chromosomes expected from a reduction of proportion selected below

0.10 is limited. Moreover, expected values of N'/N (P_M) are low for most situations investigated in this study, so that increasing N' necessitates the genotyping of an additional large number of individuals. For example, simulations conducted in the same conditions as in Table 7, but with 800 individuals genotyped at each generation instead of 400 give a genomic similarity of 0.723, 0.959 and 0.881 for carrier chromosomes, noncarrier chromosomes and average, respectively, at generation 2, and 0.866, 0.993 and 0.951, respectively, at generation 4. Thus, in situations where minimal population size for foreground selection (Table 3) is high, a large increase in the cost of the experiment only leads to moderate increase in the efficiency of background selection and is then hardly justified. But, increasing the number of genotyped individuals may be justified for background selection on carrier chromosomes when only a limited number (one or two) of QTLs is considered.

Pyramidal design: Since the minimum number of individuals that need to be genotyped increases exponentially with the number of QTLs (Equations 9 and 10), one could consider a more complex experimental design where (a) QTLs are first monitored one by one, to benefit from a higher background selection intensity, and then (b) favorable alleles at different QTLs are accumulated in the same genotype. With four QTLs located on different chromosomes as in the previous case, consider the four strains L_1, L_2, L_3 and L_4 derived by backcrossing from the original parents and selected for the donor allele at each of the four QTLs, respectively. The phase (b) of the design could consist in (b1) crossing a selected individual of L_1 to a selected individual of L_2 ($L_1 \times L_2$ cross) and a selected individual of L_3 to a selected individual of L_4 ($L_3 \times L_4$ cross), selecting individuals with desired genotype at the two QTLs in both crosses, then (b2) crossing selected individuals of ($L_1 \times L_2$) to selected individuals of ($L_3 \times L_4$). We wish now to compare the efficiency of this pyramidal design with the efficiency of the former simultaneous design in Table 7. In the pyramidal design, it is possible to perform background selection in phases (a), (b1) and (b2). The phase (a) is in fact a single-QTL introgression program, and the efficiency of selection can be predicted easily with the model used for Table 7. Predicting the efficiency of selection in phases (b1) and (b2) would require specific calculations that were not considered here. Rather, we just want to mimic what could be achieved in the pyramidal design, through simulations of a single-QTL introgression program, with the same expected value of N'. It is important to notice that the efficiency predicted by a single-QTL program overestimates the efficiency of a true pyramidal design, in particular because the genetic material in phases (b1) and (b2) of the pyramidal design may be homozygous for the donor type allele at some loci.

In the conditions of the simultaneous design in Table 7, the probability of having the donor type allele at all markers x in generation 4 is $1 - \alpha = 0.99$, for a total of 1600 individuals genotyped. The same probability can be reached in the pyramidal design by genotyping 20 individuals per cross per generation in phase (a), 50 individuals per cross in phase (b1), and 320 individuals in phase (b2), giving a total of only 580 individuals. With these population sizes, the expected number N' of individuals available for background selection is 7.1, 6.3, and 5 in phases (a), (b1), and (b2), respectively. These values of N' are roughly equivalent to the ones in Table 7, so that the efficiency of background selection is expected to be approximately the same as in Table 7 on either carrier or noncarrier chromosomes. This was confirmed by simulations (results not shown). The pyramidal design should then provide approximately the same efficiency as the simultaneous design with almost one third of the individuals, but with greater experimental complexity.

The efficiencies of the two designs could also be compared at equal costs (*i.e.*, same total number of individuals genotyped). It is unlikely that increasing population size in phase (b2) is valuable, and we only considered increasing population size in phases (a) and (b1). To keep to approximately the same total number of individuals genotyped as in Table 7 (1600), we then considered a pyramidal design involving the genotyping of 94 individuals per cross per generation in phase (a), 265 individuals per cross in phase (b1), and 320 individuals in phase (b2). With these population sizes, the expected number N' of individuals available per cross for background selection in phases (a) and (b1) is 33.3. The efficiency of background selection for such a design in phases (a) and (b1) was then (over) estimated from the simulation of three generations of a single-QTL introgression program with population size $N = 94$. Since background selection is expected to be efficient on the carrier chromosome in a single-QTL program, we used a selection index in which priority was given to markers on the carrier chromosome. The corresponding results are shown in Table 8.

The results in Table 8 show that background selection efficiency for noncarrier chromosomes in the pyramidal design over four generations is comparable to the one of the simultaneous design (Table 7), while efficiency for the carrier chromosome is greater (0.927 similarity instead of 0.836). This is not only due to the difference in the weights attributed to both chromosomes types, since even when priority is given to the carrier chromosomes, efficiency in the simultaneous design is lower (results not shown). The average similarity over the entire genome given in the ninth column in Table 8 is calculated for a true single-QTL program (one carrier and nine noncarrier chromosomes). Efficiency of a pyramidal design was (over) estimated from the same data by computing average similarity if there

TABLE 8

Combined foreground and background selection with one single QTL

				Genomic similarity			
Gen	Rep	*n*	*Pr*(*Q*)	Carrier	Noncarrier	Average (1c)	Average (4c)
1	1000	94	0.501	0.750 (0.750)	0.750 (0.750)	0.750 (0.750)	0.750 (0.750)
		33.2	0.994	0.602 (0.601)	0.749 (0.750)	0.738 (0.739)	0.700 (0.701)
		1	0.993	0.772 (0.601)	0.796 (0.750)	0.794 (0.739)	0.788 (0.701)
2	1000	94	0.496	0.886 (0.800)	0.898 (0.875)	0.897 (0.869)	0.894 (0.850)
		33.3	0.987	0.815 (0.677)	0.899 (0.875)	0.892 (0.860)	0.871 (0.809)
		1	0.988	0.901 (0.677)	0.934 (0.875)	0.931 (0.860)	0.923 (0.809)
3	1000	94	0.495	0.950 (0.839)	0.967 (0.938)	0.965 (0.930)	0.961 (0.905)
		33.4	0.982	0.917 (0.736)	0.967 (0.938)	0.963 (0.923)	0.950 (0.871)
		1	0.982	0.919 (0.736)	0.985 (0.938)	0.980 (0.923)	0.963 (0.871)
4	1000	94	0.488	0.960 (0.868)	0.993 (0.969)	0.990 (0.961)	0.982 (0.936)
		33.1	0.976	0.931 (0.779)	0.993 (0.969)	0.988 (0.955)	0.972 (0.906)
		1	0.976	0.932 (0.779)	0.993 (0.969)	0.988 (0.955)	0.973 (0.906)

Same as Table 7 with a single QTL and priority to markers on the carrier chromosome for background selection. (1c), average similarity computed with one carrier and nine noncarrier chromosomes (true situation); (4c), average similarity computed from same data as if there were four carrier and six noncarrier chromosomes.

were four carrier and six noncarrier chromosomes (11th column in Table 8). This indicates that at equal costs over four generations, the pyramidal design is expected to be more efficient.

If the donor parent may carry deleterious genes close to the QTLs, as would be the case in introgression from a wild species into a commercial breed, then an as perfect as possible return to the recipient parent is mandatory, even on the carrier chromosomes. Using a pyramidal design over three or four generations with large population sizes would then be preferable. Conversely, when the genetic distance between the donor and the recipient parent is small, the level of similarity that should be required is lower: one could then use a pyramidal design with minimal population size, to reduce the cost of the experiment. Note however that the pyramidal design necessitates that in phases (b1) and (b2) the genotypes of backcrossed progenies for at least the markers controlling the QTLs are available before crossing, what is not mandatory with the simultaneous design if all individuals can be crossed to the recipient parent at reasonable costs.

Conclusions: We provided a general framework for the optimization of the use of molecular markers in backcross breeding programs aimed at introgressing one to several QTLs, which can be used to derive specific applications. Also, some general conclusions can be drawn from the particular cases investigated in this article.

In general, it is worth using at least three markers to control each QTL in the foreground selection step. The positions of these markers should be optimized, if possible, with respect to the probability of presence of the donor allele at the QTLs, but also with respect to the minimum population size needed to obtain individuals with the requested genotype at all markers, since the probability of obtaining such individuals can be very sensitive to marker spacing. With optimally positioned markers, it is possible to manipulate up to four unlinked QTLs simultaneously with population sizes of a few hundred individuals (in the case when only one or a few individuals can be used for reproduction). A greater number of QTLs can be manipulated if the QTLs are linked, if very large population sizes can be considered, and/or if the precision of the gene location is high.

Also, it clearly appears that it is worth considering selection for recipient parent marker alleles (background selection) on both carrier and noncarrier chromosomes, even if this selection is performed among a restricted number of individuals resulting from the foreground selection step. Background selection then allows a gain of about one or two generations. The position of background selection markers on the carrier chromosomes must be taken into account in the optimization of foreground selection marker positions. Performing background selection on the carrier chromosome does not reduce much the probability of presence of the donor allele at the QTLs, but can necessitate very large population sizes. Compared with a program monitoring several QTLs simultaneously, a pyramidal program treating QTLs one by one in a first step, when feasible, would provide the same efficiency of background selection with smaller population size, or higher efficiency with the same population size.

We thank Peter Visscher for showing us his submitted paper and for helpful comments on earlier versions of this manuscript, including the discussion of the animal breeding point of view. We also thank Isabelle Goldringer and Steve Openshaw for careful reading and helpful comments on earlier versions of this manuscript. The Mathematica software package was supported by AIP Institut National de la Recherche Agronomique No. 93/4924 via the MMM group.

LITERATURE CITED

Darvarsi, A., A. Weinreb, V. Minke, J. I. Weller and M. Soller, 1993 Detecting marker QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

Groen, A. F., and C. Smith, 1995 A stochastic simulation study on the efficiency of marker-assisted introgression in livestock. J. Anim. Breed. Genet. **112:** 161–170.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci by using molecular markers. Heredity **69:** 315–324.

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.

Hillel, J., A. M. Verrinder Gibbins, R. J. Etches, and D. McQ. Shaver, 1993 Strategies for the rapid introgression of a specific gene modification into a commercial poultry flock from a single carrier. Poultry Sci. **72:** 1197–1211.

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

Hospital, F., C. Dillmann, and A. E. Melchinger, 1996 A general algorithm to compute multilocus genotype frequencies under various mating systems. Comput. Appl. Biosci. **12:** 455–462.

Knapp, S. J., W. C. Bridges, and D. Birkes, 1990 Mapping quantitative trait loci using molecular marker linkage maps. Theor. Appl. Genet. **79:** 583–592.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Mangin, B., B. Goffinet, and A. Rebai, 1994 Constructing confidence intervals for QTL location. Genetics **138:** 1301–1308.

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

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)* Ed. INRA, Paris.

Stuber, C., and P. Sisco, 1992 Marker-facilitated transfert of QTL alleles between elite inbred lines and responses in hybrids. Proc. 46th Annual Corn and Sorghum Research Conference, American Seed Trade Assoc., **46:** 104–113.

Tanksley, S. D., 1983 Molecular markers in plant breeding. Plant. Mol. Biol. Rep. **1:** 3–8.

Visscher, P. M., 1996 Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. J. Hered. **87:** 136–138.

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 Publishing Company, Redwood City, CA.

Young, N. D., and S. D. Tanksley, 1989 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: B. S. Weir

APPENDIX

We consider a single QTL located on a chromosome of total length L. In a "donor" × "recipient" backcross breeding program, we want to calculate the conditional probabilities associated with marker-QTL haplotypes, and the minimum population size for background selection (see METHODS for more details).

Foreground selection: At each generation, selection is for the donor type allele at m markers located at positions $x_1, \ldots, x_m$ on the chromosome $(0 \leq x_1 \leq \cdots \leq x_m \leq L)$. The conditional probability $P_{Q|M}[t]$ of having in generation t the donor type allele at the QTL, given that we have the donor type allele at all markers x_i was derived by Visscher *et al.* (1996) for one or two markers. Extending their approach to any number m of markers, we can write the following:

$$P_{Q|M}[t] = P_1[t] + \sum_{k=1}^{m-1} P_{2,k}[t] + P_3[t] \quad \text{(A.1)}$$

with

$$P_1[t] = \int_0^{x_1} (1 - r[x, x_1])^t g[x]\, dx \quad \text{(A.2)}$$

$$P_{2,k}[t] =$$

$$\int_{x_k}^{x_{k+1}} \frac{(1 - r[x_k, x])^t (1 - r[x, x_{k+1}])^t}{(1 - r[x_k, x_{k+1}])^t} g[x]\, dx \quad \text{(A.3)}$$

$$P_3[t] = \int_{x_m}^{L} (1 - r[x_m, x])^t g[x]\, dx, \quad \text{(A.4)}$$

where $g[x]$ is the probability that the true given the estimated position of the QTL is x, and where $r[x_k, x_{k+1}]$ is the recombination rate between loci x_k and x_{k+1}.

Assuming that $g[x]$ follows a normal distribution, with mean x_0 and variance σ^2, and that recombination rates follow Haldane mapping function with no interference, we have (Visscher *et al.* 1996)

$$\begin{aligned} P_1[t] = {} & \left(\frac{1}{2}\right)^t \sum_{i=0}^{t} \left(C_t^i \frac{1}{\sigma\sqrt{2\pi}} \right. \\ & \times \int_0^{x_1} \exp\left[-\frac{(x - (x_0 + 2i\sigma^2))^2}{2\sigma^2} \right. \\ & \left.\left. + 2i(i\sigma^2 + x_0 - x_1) \right] dx \right) \\ = {} & \left(\frac{1}{2}\right)^{t+1} \sum_{i=0}^{t} \left(C_t^i \exp[2i(i\sigma^2 + x_0 - x_1)] \right. \\ & \left. \times \left\{ \operatorname{erf}\left[\frac{2i\sigma^2 + x_0}{\sqrt{2}\sigma}\right] - \operatorname{erf}\left[\frac{2i\sigma^2 + x_0 - x_1}{\sqrt{2}\sigma}\right] \right\} \right) \end{aligned} \quad \text{(A.5)}$$

$$\begin{aligned} P_{2,k}[t] = {} & \\ & \frac{(^1/_2)^t}{(1 + \exp[-2(x_{k+1} - x_k)])^t} \sum_{i=0}^{t} \sum_{j=0}^{t} \left(C_t^i C_t^j \frac{1}{\sigma\sqrt{2\pi}} \right. \\ & \times \int_{x_k}^{x_{k+1}} \exp\left[-\frac{(x - (x_0 - 2(i - j)\sigma^2))^2}{2\sigma^2} \right. \\ & \left.\left. + 2(i - j)^2\sigma^2 - 2(i - j)x_0 + 2(ix_k - jx_{k+1}) \right] dx \right) \end{aligned}$$

$$= \frac{(^1/_2)^{t+1}}{(1 + \exp[-2(x_{k+1} - x_k)])^t} \times \sum_{i=0}^{t} \sum_{j=0}^{t} \Bigg(C_t^i C_t^j \exp[2(i-j)^2\sigma^2 - 2 \times (i-j)x_0 + 2(ix_k - jx_{k+1})] \times \left\{ \mathrm{erf}\left[\frac{2(i-j)\sigma^2 - x_0 + x_{k+1}}{\sqrt{2}\sigma}\right] - \mathrm{erf}\left[\frac{2(i-j)\sigma^2 - x_0 + x_k}{\sqrt{2}\sigma}\right]\right\}\Bigg) \qquad (A.6)$$

and

$$P_3[t] = \left(\frac{1}{2}\right)^t \sum_{i=0}^{t} \Bigg(C_t^i \frac{1}{\sigma\sqrt{2\pi}} \times \int_{x_m}^{L} \exp\left[-\frac{(x - (x_0 - 2i\sigma^2))^2}{2\sigma^2} + 2i(i\sigma^2 - x_0 + x_m)\right] dx\Bigg)$$

$$= \left(\frac{1}{2}\right)^{t+1} \sum_{i=0}^{t} \Bigg(C_t^i \exp[2i(i\sigma^2 - x_0 + x_m)] \times \left\{ \mathrm{erf}\left[\frac{2i\sigma^2 - x_0 + L}{\sqrt{2}\sigma}\right] - \mathrm{erf}\left[\frac{2i\sigma^2 - x_0 + x_m}{\sqrt{2}\sigma}\right]\right\}\Bigg), \qquad (A.7)$$

where erf is the error function defined by

$$\mathrm{erf}[x] = \frac{2}{\sqrt{\pi}} \int_0^x \exp[-y^2]\,dy. \qquad (A.8)$$

Background selection: *Markers-QTL haplotype:* At each generation, selection is now for the donor type allele at markers $x_1, \ldots, x_m$ and for the recipient type allele at markers y_1 and y_2 ($0 \leq y_1 < x_1 \leq \cdots \leq x_m < y_2 \leq L$). We want to calculate the conditional probability $P^*_{Q|M}$ of having the donor type allele at the QTL, given that we have the donor type allele at markers x and the recipient type allele at markers y.

We need first to calculate the probabilities associated with the requested genotypes. Haplotype frequencies can be extended from the results given for three loci by VISSCHER and THOMPSON (1995) or using the algorithm derived by HOSPITAL *et al.* (1996). Denoting by + the allele derived from the donor parent, and by − the allele derived from the recurrent parent, the probability $P^*_M[t]$ of having the requested allele at all markers (x_i's and y_j's) in generation t is then

$$P^*_M[t] = \Pr(y_1^- x_1^+ \cdots x_m^+ y_2^-) = (1 - (1 - r[y_1, x_1])^t)\,(P_M)\,(1 - (1 - r[x_m, y_2])^t), \qquad (A.9)$$

where P_M is the probability defined in METHODS, Equation 1.

Background selection on markers y_1 and y_2 is aimed at reducing the proportion of donor type genome on both sides of the QTL. Hence, when computing the probability of having the requested allele at the QTL, we are only interested in the cases where the true position of the QTL lies within the segment $]y_1, y_2[$. Let $P^*_{Q|M}[t]$ be the probability of having the donor allele at the QTL in generation t, given that we have the requested allele at all the markers x and y. We can write

$$P^*_{Q|M}[t] = P^*_1[t] + \sum_{k=1}^{m-1} P^*_{2,k}[t] + P^*_3[t] \qquad (A.10)$$

with

$$P^*_1[t] = \int_{y_1}^{x_1} \frac{(1 - (1 - r[y_1, x])^t)(1 - r[x, x_1])^t}{1 - (1 - r[y_1, x_1])^t} g[x]\,dx \qquad (A.11)$$

$$= \frac{(^1/_2)^{t+1}}{1 - (\frac{1}{2})^t(1 + \exp[-2(x_1 - y_1)])^t} \times \sum_{i=0}^{t} \Bigg(C_t^i \exp[2i(i\sigma^2 + x_0 - x_1)] \times \left\{ \mathrm{erf}\left[\frac{2i\sigma^2 + x_0 - y_1}{\sqrt{2}\sigma}\right] - \mathrm{erf}\left[\frac{2i\sigma^2 + x_0 - x_1}{\sqrt{2}\sigma}\right]\right\}\Bigg)$$

$$- \frac{(^1/_2)^{2t+1}}{1 - (^1/_2)^t(1 + \exp[-2(x_1 - y_1)])^t} \times \sum_{i=0}^{t} \sum_{j=0}^{t} \Bigg(C_t^i C_t^j \exp[2(i-j)^2\sigma^2 - 2(i-j) \times x_0 + 2(iy_1 - jx_1)] \times \left\{ \mathrm{erf}\left[\frac{2(i-j)\sigma^2 - x_0 + x_1}{\sqrt{2}\sigma}\right] - \mathrm{erf}\left[\frac{2(i-j)\sigma^2 - x_0 + y_1}{\sqrt{2}\sigma}\right]\right\}\Bigg) \qquad (A.12)$$

$$P^*_{2,k}[t] = P_{2,k}[t] \qquad (A.13)$$

$$P^*_3[t] = \int_{x_m}^{y_2} \frac{(1 - r[x_m, x])^t(1 - (1 - r[x, y_2])^t)}{1 - (1 - r[x_m, y_2])^t} g[x]\,dx \qquad (A.14)$$

$$= \frac{(^1/_2)^{t+1}}{1 - (^1/_2)^t(1 + \exp[-2(y_2 - x_m)])^t} \times \sum_{i=0}^{t} \Bigg(C_t^i \exp[2i(i\sigma^2 - x_0 + x_m)]$$

$$\times \left\{\mathrm{erf}\left[\frac{2i\sigma^2 - x_0 + y_2}{\sqrt{2}\sigma}\right] - \mathrm{erf}\left[\frac{2i\sigma^2 - x_0 + x_m}{\sqrt{2}\sigma}\right]\right\}\Bigg)$$

$$- \frac{(^1/_2)^{2t+1}}{1 - (^1/_2)^t(1 + \exp[-2(y_2 - x_m)])^t}$$

$$\times \sum_{i=0}^{t}\sum_{j=0}^{t}\Bigg(C_t^iC_t^j \exp[2(i-j)^2\sigma^2 - 2(i-j)$$

$$\times x_0 + 2(ix_m - jy_2)]$$

$$\times \left\{\mathrm{erf}\left[\frac{2(i-j)\sigma^2 - x_0 + y_2}{\sqrt{2}\sigma}\right]\right.$$

$$\left. - \mathrm{erf}\left[\frac{2(i-j)\sigma^2 - x_0 + x_m}{\sqrt{2}\sigma}\right]\right\}\Bigg). \qquad \text{(A.15)}$$

Minimal population size: In a given experiment, the probability of obtaining at least one individual with the requested genotype at all markers x_i's and y_j's can be calculated in the following framework: at each generation t, among a total of $n[t]$ backcross progenies, a single individual is selected if it has the donor type allele at all markers $x_1, \ldots, x_m$ (as in the foreground selection case) and in addition to this, one of the following conditions is met in the given order. (1) The individual has the recipient type allele at both markers y_1 and y_2. (2) The individual has the recipient type allele at y_1 or (exclusive) y_2. (3) The individual has the recipient type allele at none of markers y_1 and y_2. When a single individual is selected, this scheme is equivalent to selecting on the sum of recipient type alleles carried by the individuals.

Let $z_{11}[t]$, $z_{10}[t]$, $z_{01}[t]$ and $z_{00}[t]$ be the probabilities that the single individual selected in generation t has the recipient type allele at both markers y_1 and y_2, y_1 only, y_2 only or none, respectively. To simplify the notations we note $r_1 = r[y_1, x_1]$ and $r_2 = r[x_m, y_2]$. The probabilities z can be derived by recursion as follows:

$$z_{11}[t] = \{1 - (1 - P_M)^{n[t]}\}z_{11}[t-1]$$
$$+ \{1 - (1 - r_2P_M)^{n[t]}\}z_{10}[t-1]$$
$$+ \{1 - (1 - r_1P_M)^{n[t]}\}z_{01}[t-1]$$
$$+ \{1 - (1 - r_1r_2P_M)^{n[t]}\}z_{00}[t-1] \qquad \text{(A.16)}$$

$$z_{10}[t] = \{(1 - r_2P_M)^{n[t]} - (1 - P_M)^{n[t]}\}z_{10}[t-1]$$
$$+ \frac{r_1(1-r_2)}{r_1(1-r_2) + r_2(1-r_1)} \times \{(1 - r_1r_2P_M)^{n[t]}$$
$$- (1 - (r_1 + r_2 - r_1r_2)P_M)^{n[t]}\}z_{00}[t-1] \qquad \text{(A.17)}$$

$$z_{01}[t] = \{(1 - r_1P_M)^{n[t]} - (1 - P_M)^{n[t]}\}z_{01}[t-1]$$
$$+ \frac{r_2(1-r_1)}{r_1(1-r_2) + r_2(1-r_1)} \times \{(1 - r_1r_2P_M)^{n[t]}$$
$$- (1 - (r_1 + r_2 - r_1r_2)P_M)^{n[t]}\}z_{00}[t-1] \qquad \text{(A.18)}$$

$$z_{00}[t] = \{(1 - (r_1 + r_2 - r_1r_2)P_M)^{n[t]}$$
$$- (1 - P_M)^{n[t]}\}z_{00}[t-1], \qquad \text{(A.19)}$$

where P_M is the probability defined in METHODS, Equation 1.

The initial state in the F_1 ($t = 0$) being

$$z_{11}[0] = z_{10}[0] = z_{01}[0] = 0 \qquad \text{(A.20)}$$

$$z_{00}[0] = 1. \qquad \text{(A.21)}$$

The minimum numbers of individuals ($n[1], \ldots, n[t]$) that should be genotyped at each generation so that at least one individual with requested genotype at all markers x and y is obtained in generation t with risk α^* can be derived by solving

$$z_{11}[t] = 1 - \alpha^*. \qquad \text{(A.22)}$$

Though algebraic solutions to this equation cannot be found, sets of numerical solutions can be obtained easily.