

## Using Markers in Gene Introgression Breeding Programs

Frédéric Hospital, Claude Chevalet and Philippe Mulsant

Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, Centre de Toulouse,  
F-31326 Castanet, France

Manuscript received February 21, 1992  
Accepted for publication August 20, 1992

### ABSTRACT

We investigate the use of markers to hasten the recovery of the recipient genome during an introgression breeding program. The effects of time and intensity of selection, population size, number and position of selected markers are studied for chromosomes either carrying or not carrying the introgressed gene. We show that marker assisted selection may lead to a gain in time of about two generations, an efficiency below previous theoretical predictions. Markers are most useful when their map position is known. In the early generations, it is shown that increasing the number of markers over three per non-carrier chromosome is not efficient, that the segment surrounding the introgressed gene is better controlled by rather distant markers unless high selection intensity can be applied, and that selection on this segment first can reduce the selection intensity available for selection on non-carrier chromosomes. These results are used to propose an optimal strategy for selection on the whole genome, making the most of available material and conditions (*e.g.*, population size and fertility, genetic map).

AS genomic molecular markers become available in certain species, questions are being raised about their use in breeding programs. In the case of selection for a quantitative trait, marker assisted selection programs can be undertaken (LANDE and THOMPSON 1990), but it is hard to evaluate their potential efficiency in the absence of sufficient data on the relation between genotype and phenotype. However, reliable predictions can be made where only one gene of interest from a "donor" is introgressed into the genome of a selected or cultivated "recipient" by recurrent backcrosses to the recipient genotype. Since this problem involves only the genomic composition ("donor" or "recipient" genotypes) of the species, breeds, strains, lines or populations in question, theoretical predictions about it can be quite realistic.

In each generation of the breeding program, the offspring that carry the introgressed gene are chosen, then among these, those carrying the lowest proportion of donor genes at other loci are selected. Assuming that the introgressed gene is correctly identified and selected at each generation, we can study selection aimed only at reducing the proportion of the donor genome among the carriers of the introgressed gene. In the breeding program, two aspects of this problem can be considered separately: first, reduction of the length of the donor-type segment carried along with the introgressed gene to an acceptable size, and second, recovery of the composition of the recipient genome as quickly as possible.

In the absence of selection, the evolution of the proportion of recipient genome around the intro-

gressed gene has been given a complete solution by HANSON (1959) and STAM and ZEVEN (1981), while for the other chromosomes the solution is well known (Equation 1 below). With selection, the problem is more difficult. Some data have been provided by experimentation (YOUNG and TANKSLEY 1989), and recently HILLEL *et al.* (1990) have addressed the problem from a theoretical point of view in order to evaluate the potential use of DNA fingerprints provided by molecular probes revealing variable number tandem repeats (VNTR).

The aim of this paper is to extend these analyses by answering the following questions: (1) What is the expected short-term efficiency of selection using markers, as a function of the density and the distribution of the markers? (2) How can several markers surrounding the introgressed gene best be used? (3) Is it possible to combine selection on the whole genome and selection near the introgressed gene?

### MODEL AND METHODS

The criterion used to evaluate the response to selection is the expected proportion of recipient genome at some backcross generation ( $g$ ), measured after selection and denoted  $G(g)$ . Up to six backcross generations are considered, the original  $F_1$  population being generation zero. We assume that for all markers, alleles from the recipient population and from the donor line can be distinguished and are codominant. Tabulated values may be compared to the value obtained without selection on a non-carrier chromosome:

$$G(g) = 1 - (1/2)^{g+1}. \quad (1)$$

The  $G$  value may be defined for the single chromosome that carries the introgressed gene, or for the other chromosomes, or for the whole genome. The genome considered in most cases consisted of 20 chromosome pairs of 100 cM each, one of them possibly bearing the introgressed gene at its center.

**Population and selection:** Analytical predictions were made assuming an infinite population size. The effect of finite population size was investigated by simulation. In this case, the size  $N$  of the population refers to the number of zygotes that carry the donor allele at the locus of the introgressed gene.

Selection for reproductive individuals that carry as many recipient alleles as possible at  $M$  marker loci is based on the index

$$I = \sum_i a_i V_i \quad (i = 1, \dots, M) \quad (2)$$

where  $a_i$  is the weight of the  $i$ th marker locus and  $V_i$  is equal to zero or one for the donor or recipient allele respectively. To control the region bearing the introgressed gene, markers in the vicinity of the gene can be used to perform a prior selection aimed at reducing the length of the donor chromosomal segment carried along with the gene. This prior selection is obtained by assigning a large enough weight in the index to these markers.

The *proportion selected*,  $s$ , is the ratio of the number of individuals used for reproduction to the preceding number  $N$  of zygotes (*i.e.*, an expected half of the total number of zygotes). In plant breeding it is possible to reproduce populations with only one or a few individuals, and  $s$  could be as small as  $1/N$ . In animal breeding,  $s$  must be large enough so that on average a reproducing individual gives birth to at least  $(2/s)$  offspring. Note that the proportion selected  $s$  has low values when selection is strong, contrary to the *selection intensity*  $i$ .

**Analytical formulas** are detailed in the appendices. They refer to three approaches:

*Response at independent loci:* Here, we consider a set of  $L$  independent loci, and estimate  $G$  by the expected proportion of these loci that carry alleles of the recipient genome. Although this approximation does not take account of the genomic structure, it allows one to derive simple predictions for the chromosomes that do not carry the introgressed gene. Standard methods of quantitative genetics are used to derive the change with time of the exact mean and variance of  $G$  without selection, and to derive an approximation to the expected values of these quantities at the  $n$ th generation when selection is applied only during the  $g$ th generation ( $1 \leq g \leq n$ ).

*Recipient genome content on the chromosome carrying the introgressed gene:* Where feasible,  $G$  was calculated

by analysis, assuming a continuous set of loci along the chromosome, and the Haldane mapping function between distance  $d$  (in Morgans) and recombination rate  $r$ :

$$d = - (1/2) \ln(1 - 2r). \quad (3)$$

These analyses follow the approach introduced in the theory of junctions by FISHER (1949) and developed by HANSON (1959) and STAM and ZEVEN (1981), whose calculations without selection are extended to the case with selection. The scheme supposes a marker locus in the vicinity of the introgressed gene. Recombinant individuals that carry the donor allele of the introgressed gene and the recipient allele at the marker locus are selected. Analysis is restricted to the calculation of mean values, and to the case of a single selection step as in APPENDIX A.

*Joint response to selection:* The preceding two approaches are used to deal with combined selection for the mean recipient genome content on the chromosome carrying the introgressed gene and for the overall recipient genome content on the other chromosomes. As before, a single step of selection is considered.

**Computer simulations:** When no exact solution was available by standard analysis, computer simulations were carried out. The program describes individuals at the genomic level. The genome is represented by many loci evenly distributed on the genetic map. The density used in most calculations is one locus every 5 cM. Higher densities were tested but were found to yield the same results (however, denser maps were used when selected markers very close to the introgressed gene were considered). Crossing overs are generated according to a Poisson distribution assuming no interference. At each locus two codominant alleles are characteristic of either the recipient or the donor line, so that the estimated  $G$  value is the proportion of loci carrying recipient alleles. Some loci are considered to be observable markers, and submitted to selection based on index (Equation 2). The number and positions of the markers, as well as the position of the introgressed gene, are either specifically assigned or else randomly chosen by the program. Starting from  $F_1$  individuals, the program simulates six backcross generations with the recurrent parent. Most simulations were carried out with  $N = 200$ , and 1000 runs per case. Trials made with a larger population size (1000) gave similar results.

## RESULTS

**Selection on the chromosomes that do not carry the introgressed gene:** For a single selection step, the calculations of APPENDIX A (Equation A-1) for independent loci show that selection is more effective the later it is applied. For continuous selection, the esti-

TABLE 1

Proportion (*G*) of recipient genome under continuous selection

Proportion selected	Generation	Selection			No selection
		Markers	Non-markers	Total	
0.02	1	0.856	0.840	0.841	0.750
	2	0.980	0.959	0.961	0.875
	3	1.000	0.986	0.987	0.938
	4	1.000	0.993	0.994	0.969
	5	1.000	0.996	0.997	0.984
	6	1.000	0.998	0.998	0.992
0.05	1	0.841	0.827	0.829	0.750
	2	0.970	0.950	0.952	0.875
	3	1.000	0.985	0.987	0.938
	4	1.000	0.993	0.993	0.969
	5	1.000	0.996	0.997	0.984
	6	1.000	0.998	0.998	0.992
0.1	1	0.828	0.816	0.817	0.750
	2	0.959	0.942	0.943	0.875
	3	1.000	0.984	0.985	0.938
	4	1.000	0.992	0.993	0.969
	5	1.000	0.996	0.996	0.984
	6	1.000	0.998	0.998	0.992
0.3	1	0.802	0.794	0.794	0.750
	2	0.935	0.923	0.924	0.875
	3	0.986	0.974	0.975	0.938
	4	1.000	0.991	0.992	0.969
	5	1.000	0.996	0.996	0.984
	6	1.000	0.998	0.998	0.992
0.5	1	0.786	0.780	0.781	0.750
	2	0.918	0.909	0.910	0.875
	3	0.974	0.965	0.966	0.938
	4	0.995	0.987	0.988	0.969
	5	1.000	0.995	0.995	0.984
	6	1.000	0.997	0.998	0.992

Twenty non-carrier chromosomes with two selected marker loci on each. Selection: simulation results, mean *G* values after selection averaged over either marker loci only (Markers), or non-marker loci only (Non-markers), or whole genome (Total). No selection: Equation 1, expected *G* values without selection for comparison.

mated measure *G* was determined by simulations. Each chromosome was 100 cM long and selection was applied at two marker loci per chromosome, located at 20 cM from the distal telomeric loci. These locations had been previously found to be most efficient (results not shown). Table 1 shows the results obtained under continuous selection for six generations, for five values of *s* (2, 5, 10, 20 and 50%). Results concerning the marker loci (third column) show that fixation occurs very rapidly at high selection intensity, markers becoming useless after fixation. The fourth column shows that response at non-selected loci is lower. This is illustrated in Figure 1, showing the mean distribution of *G* at each locus along a chromosome.

The effect of increasing the number of selected marker loci is shown in Figure 2, where mean *G* values are given for six generations under continuous selection, with up to ten marker loci per 100 cM. The loci

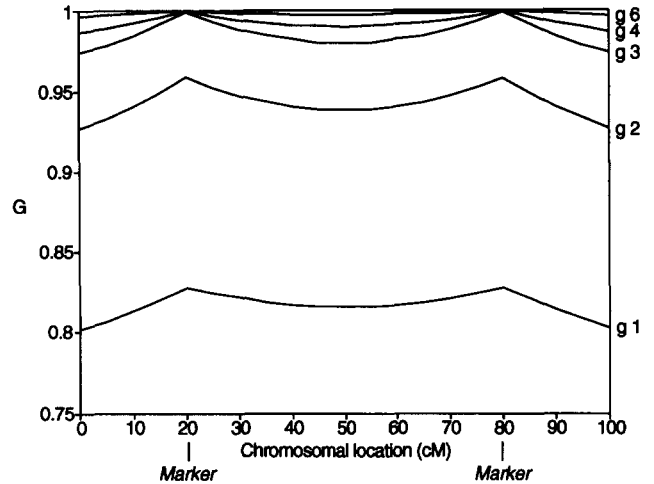


FIGURE 1.—Local response to selection along a non-carrier chromosome. Data of Table 1 plotted along a virtual chromosome. Each point corresponds to the average response over the 20 non-carrier chromosomes at the same locations. Generations 1 to 6 of continuous selection from bottom to top, with *N* = 200 and *s* = 0.1.

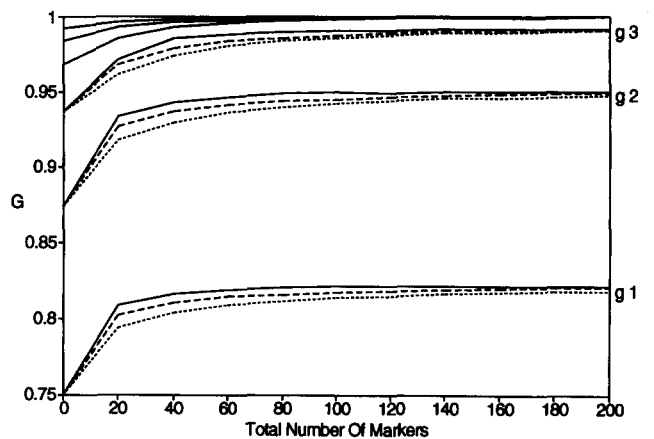


FIGURE 2.—Effect of number and localization of selected marker loci. Twenty non-carrier chromosomes with *N* = 200 and *s* = 0.1, generation 1 to 6 from bottom to top. Solid lines: markers evenly spaced on each chromosome; dashed lines: equal number of markers assigned to each chromosome, with position random; dotted lines: marker positions completely random on the whole genome.

were either evenly spaced along a known genetic map, or randomly located on each chromosome, or randomly located on the whole genome. We find that increasing the number of selected markers above three per 100 cM has poor efficiency in early generations. In later generations higher densities are relatively more efficient, but all *G* values are already close to 1. In practice then, two markers per 100 cM should be sufficient to get the highest possible response in the early generations if the markers are localized. Using markers with unknown chromosomal location is less useful. It was found that these results were not affected by increasing the number of chromosomes up to 30 pairs, or the population size up to 1000 (data not shown).

Selecting various proportions of individuals during

TABLE 2

Effect of different proportion selected in the first two generations on mean  $G$  values after the second generation

Proportion selected in first generation	Proportion selected in second generation				
	1.00	0.30	0.10	0.05	0.02
1.00	0.875	0.897	0.910	0.914	0.920
0.30	0.909	0.924	0.932	0.936	0.941
0.10	0.925	0.936	0.945	0.947	0.951
0.05	0.933	0.942	0.949	0.951	0.956
0.02	0.941	0.949	0.955	0.957	0.961

Simulation results with twenty non-carrier chromosomes.

the first two generations only is considered in Table 2. Results are in qualitative accordance with the tendency suggested by the analytical approach for independent loci, which showed that a single selection step is more efficient the later it is applied.

**Selection involving only the chromosome carrying the introgressed gene:** In this section the genetic system involves a single chromosome, and the gene to be introgressed is located at its center. The chromosome length investigated is 100 cM as in the previous analyses. One marker locus on each side of the introgressed gene is considered, at a distance  $d$  from it (selection schemes with more than two markers on the chromosome are considered in the DISCUSSION). Response to selection on the carrier chromosome is lower than on non-carriers. This is due to the existence of a segment of donor genome carried along with the introgressed gene. Here, the use of selection is mostly devoted to the reduction of this segment, so that markers position ( $d$ ) becomes a critical parameter.

The results of calculations (APPENDIX B, and application in APPENDIX C) on the interaction between the proportion selected  $s$  and the recombination rate  $r$  that corresponds to distance  $d$  during the first generation are shown in Figure 3. Each curve refers to a given value of  $s$ , and shows the expected value of  $G$  on the carrier chromosome as a function of  $r$ . For small proportions  $s$  of selected individuals, there are two local optima at  $r_1$  and  $r_2$  given by

$$r_1^2 + 2r_1(1 - r_1) = s$$

and

$$r_2^2 = s.$$

For  $r$  equal to or slightly larger than  $r_1$ , selection allows all single recombinant genotypes to be retained, and results in a mean  $G$  value nearly equal to 0.75. Using more distal markers ( $r_1 < r < r_2$ ), an increased average response is obtained which is maximal for  $r = r_2$ , when only double recombinants are selected. For larger values of  $r$ , the response decreases following Equation C-1 (APPENDIX C). Double recombinants for markers at  $r = r_2$  have a larger  $G$  value than single recombinants for more proximal markers at  $r = r_1$ , so

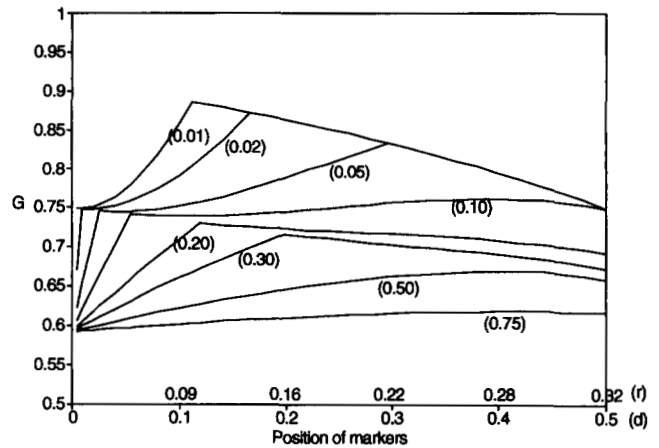


FIGURE 3.—Effect of proportion selected and marker positions on response at carrier chromosome. Numerical solution of equations in APPENDICES B and C. Each curve shows expected  $G$  values after one generation of selection, for a given value of proportion selected (between parentheses). One pair of selected markers surrounding the introgressed gene.  $r$ , recombination fraction between the introgressed gene and each marker;  $d$ , corresponding distance in Morgans. See text for explanations.

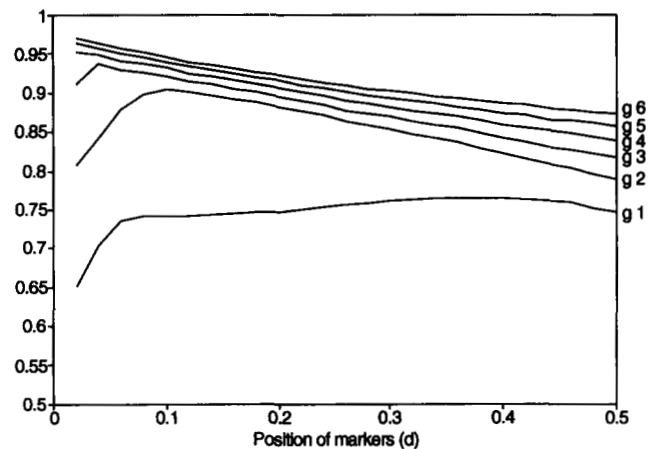


FIGURE 4.—Same as Figure 3 for proportion selected  $s = 0.10$  and 6 generations of continuous selection (from bottom to top). Simulation results.

that using two markers at  $r = r_2$  would be an optimal solution in the first generation. But in this case, the markers reach fixation after the first generation and become useless for selection in the subsequent generations, whereas choosing markers at  $r = r_1$  would allow the same process to take place in the second generation (selection of single recombinants for the second marker), eventually resulting in a higher  $G$ . Such schemes with continuous selection were studied by means of simulations (see below). For larger proportions  $s$  of individuals selected the point  $r = r_2 = s^{1/2}$  no longer exists. Moreover, with large values of  $s$  the response becomes a decreasing function of  $r$  as soon as  $r > r_1$ , so that the optimum moves to  $r_1$ .

The effect of gene-marker distance under continuous selection is shown in Figure 4 for a constant proportion ( $s = 0.10$ ) selected in the first six genera-

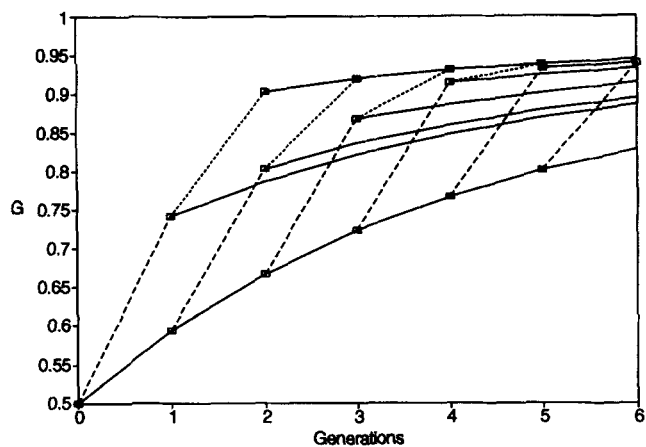


FIGURE 5.—Effect of the time when selection is applied on response at carrier chromosome ( $s = 0.1$ ,  $d = 10$  cM). Analytical and simulation ( $N = 200$ ) results. Solid lines: evolution without selection; dashed lines: first generation of selection; dotted lines: second generation of selection. Squares indicate possible shift between selection and no selection.

tions. As recombination events accumulate over time, the apparent recombination rate between a marker and the introgressed gene increases, so that optimal values for  $r$  are smaller when later generations are considered. Figure 4 clearly shows that the choice of the markers depends on the number of generations during which selection is to be performed. Markers close to the introgressed gene are useful in later generations only (for this intermediate value of  $s$ ) whereas more distal markers should be used in the short term.

The effect of the time when a single selection step is applied is illustrated in Figure 5. These results were obtained by numerical solution of the equations in APPENDIX B, and fit the simulation results well. In most cases with two markers a single selection step is more efficient if applied later as shown by Figure 5. This tendency is in accordance with that found previously for selection on non-carrier chromosomes. Note however that exceptions to this rule can be found when markers are far from the introgressed gene (results not shown).

Applying selection more than once was investigated by simulation. Results of two successive selection steps are included in Figure 5. The case of continuous selection is not shown, but appeared to be equivalent to selecting only during the first two generations. Hence, in this example ( $N = 200$ ,  $d = 10$  cM,  $s = 0.10$ ), it turns out that fixation of both markers occurs within two generations of selection regardless of when selection is performed (cases with two non-successive selection steps give the same outcome).

**Selection on the whole genome:** Table 3 presents results of simulations similar to those presented in Table 1, but with one of the 20 pairs of chromosomes bearing the introgressed gene at its center. The mean

TABLE 3  
Continuous selection on both carrier and non-carrier chromosomes

Generation	Carrier chromosome		Non-carrier chromosome	
	Selection	No selection	Selection	No selection
1	0.636	0.607	0.817	0.750
2	0.747	0.684	0.943	0.875
3	0.855	0.741	0.983	0.938
4	0.891	0.784	0.993	0.969
5	0.903	0.817	0.996	0.984
6	0.913	0.842	0.998	0.992

Same as Table 1 with one carrier and nineteen non-carrier chromosomes. Simulation results, proportion selected  $s = 0.10$ .

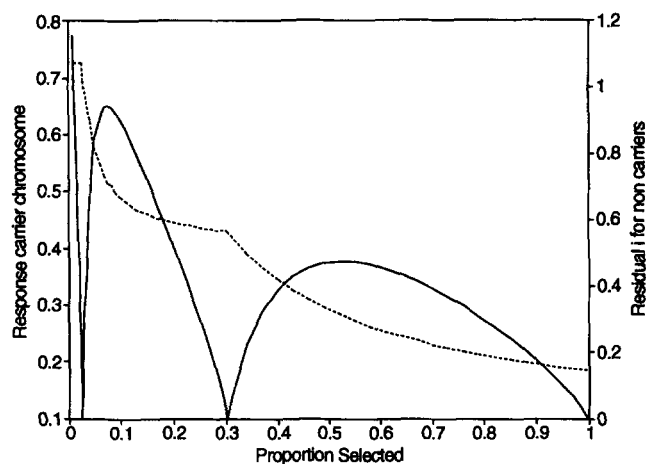


FIGURE 6.—Selection on the whole genome.  $G$  values after selection on carrier chromosome (dotted line) and residual intensity  $i$  for selection on non carriers (solid line) for different proportions selected. Analytical results for one generation of selection with  $d = 20$  cM.

response over the whole genome is only slightly lower than with no introgressed gene, but the results highlight the large difference between the chromosome carrying the gene and the other chromosomes. When using an index with equal weights, selection on the markers has poor efficiency for the carrier chromosome.

If markers are available near the introgressed gene as well as on other chromosomes it is possible to combine selection on all these markers. In the first generation an analytical analysis is possible (APPENDIX C) under the approximation whereby the whole genome is described by independent loci. Figure 6 shows the joint pattern of expected  $G$  values after selection on the carrier chromosome, and of the remaining intensity  $i$  of selection that can be exerted on the other chromosomes. Assuming that the first priority of selection is to reduce the length of the donor segment surrounding the introgressed gene, individuals selected in this first step of selection may then be screened according to their  $G$  value for the rest of the genome, provided they are more numerous than needed to reproduce the population. Conversely, if

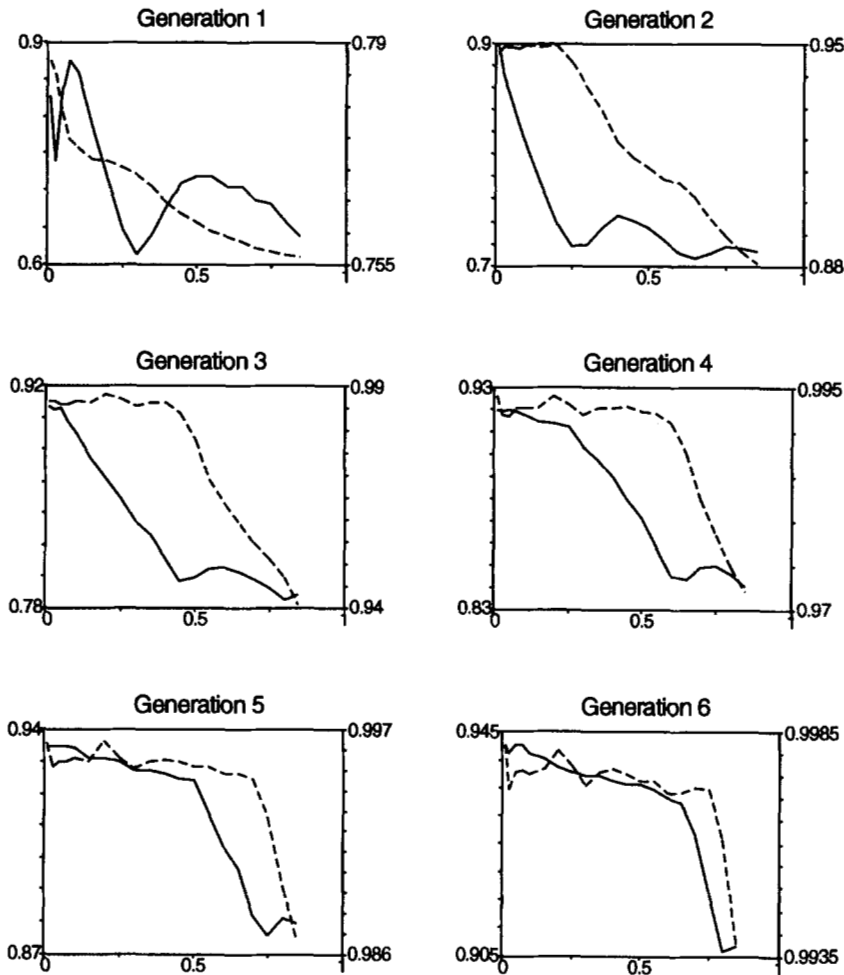


FIGURE 7.—Same as Figure 6 for six generations of continuous selection. Simulation results. Abscissa: proportion selected. Left ordinates and dashed lines:  $G$  values after selection on carrier chromosome; right ordinates and solid lines:  $G$  values after selection on non-carriers. Note the different scales for ordinates.

recombinant individuals for the carrier chromosome are not numerous enough, the additional (non-recombinant) individuals needed may be chosen according to their overall recipient genome content. Finally, if the number of recombinants for the carrier chromosome is equal to the number of individuals needed to reproduce the population, other chromosomes are chosen at random as far as their recipient genome content is concerned. Hence, the combinations ( $s$ ,  $r_1$ ) and ( $s$ ,  $r_2$ ) of parameters that were found optimal in the preceding section when only the carrier chromosome was considered, are the worst for the non-carrier chromosomes when the whole genome is considered. The results of simulations for several generations of continuous combined selection are shown in Figure 7 for both carrier and non-carrier chromosomes. The points with reduced response move to larger values of  $s$  as later generations are considered, so that the situation in later generations may seem less dramatic than the one described above for the first generation. Note however that the values of  $G$  for non-carrier chromosomes must be compared to the corresponding values with no selection (0.75, 0.875, ...).

## DISCUSSION

**Efficiency of markers for overall recipient genome recovery:** From a quantitative and practical point of view the predicted gain in time expected from marker selection is about two backcross generations. If the proportion selected is less than ten percent, the recipient genome content after selection in the third generation is equal to or larger than that expected without selection at the fifth generation. Although significant, these predictions are less optimistic than those given by HILLEL *et al.* (1990). According to our results (Table 1), two generations of continuous selection are needed to get a response equal to their predictions for a single selection step (Table 3 with 40 segments, in HILLEL *et al.* 1990). Also, their calculations for continuous selection (Table 5, HILLEL *et al.* 1990) predict quasi-fixation within two generations, a result at variance with ours (Table 2). Note that the difference would be even greater if we had considered markers of unknown chromosomal location as they did. In fact the results of these authors may be compared with ours only for the response at the selected marker loci. Their analysis considers segments of chromosomes but deals with them as if they

were independent loci, so that their results are comparable only to our restricted approach developed in APPENDIX A. The difference stems from the effects of recombination between the selected marker loci and the surrounding loci (Figure 1). A practical and important consequence of these discrepancies is that we find that stopping selection after the second backcross generation is not a correct strategy. As shown in Table 1, substantial gains from selection are still expected at the third generation for any selection intensity, and all the more as selection is less intense. This trend is in accordance with the qualitative analytical result of APPENDIX A. Therefore, if marker assisted introgression is to be applied, it should be performed during at least three generations. Moreover, if only low selection intensity ( $s > 0.10$ ) can be applied, selection should proceed longer in order to achieve a savings of about two generations over a six generation interval.

Using markers to monitor gene introgression was found to be most useful if linkage maps are available. A proper choice of evenly spaced markers allows a minimal number of them to be used. If only markers of unknown chromosomal location are used, as would be the case with VNTRs, the response is reduced (Figure 2). The increased efficiency due to mapped markers may not seem worth the work needed to establish the genetic map. However, one should note that the additional gain due to mapped markers may represent up to  $\frac{1}{3}$  of the advance due to selection with non-mapped markers. Furthermore, with non-mapped markers one takes the risk that some chromosomes are under no control. This risk is eliminated if every chromosome can be marked by randomly located syntenic loci (provided for example by probes from chromosome specific libraries). In this latter case the response is only slightly lower than with evenly spread markers.

A density of two to three markers per 100 cM seems to be an optimal choice, since increasing this density is of small benefit. This result may seem somewhat surprising, and contrary to current opinion. It can be explained by considering the distribution of donor genes in the genome. In early generations, few recombination events have occurred, so that donor genes should be represented by a few long segments on each chromosome. Since one selected marker by segment is enough to control it, only a few markers per chromosome are needed. As recombination events accumulate over time, the number of donor segments spread over the genome increases, and their length decreases, so that more and more markers are needed to control all of them. A higher density of markers might then be useful in later generations to eliminate these donor segments more rapidly. Indeed, this seems to be the case since in later generations the use

of many markers appeared to be *relatively* more efficient than using only a few (results not shown). But practical gains are very low since all  $G$  values are then over 0.99. Hence, increasing the number of markers does not seem efficient under the special conditions of gene introgression. Investigating this efficiency in a more general framework of marker assisted selection would require further analyses, not restricted to average genome content, but also devoted to the distribution of segment lengths (STAM and ZEVEN (1981).

**Use of markers surrounding the introgressed gene:** Our analysis has focused on the simple case where a single pair of markers controlling the introgressed gene is used in the selection scheme. Knowing the intensity of selection that can be applied, it provides a basis for choosing such a pair when several markers are available on the carrier chromosome. The interesting result is the potential use of rather distal markers during the first generations, which may provide a significant recovery of the recipient genome, even if the length of the donor segment carrying the introgressed gene is not reduced to the minimum in this process. Indeed, if proximal markers are also available, it is not harmful to wait until later generations to select for them, since the efficiency of selection generally increases with time. Furthermore, selecting only for very proximal markers is clearly ineffective if the population size is not very large.

Hence, we suggest using a more complex strategy of selection to make the most of *all* the markers available on the carrier chromosome. This strategy is suggested by the results presented in this paper concerning the classical case of a single pair of markers controlling the introgressed gene. It consists of giving preference to the selection of proximal recombinants as soon as they arise, but not ignoring more distal recombinations. In practice, this would be easy to do when typing the individuals. Some theoretical results that give an idea of the efficiency of such a strategy are presented in Table 4. They were obtained by using a selection index such that the weight of a proximal single recombinant is larger than the sum of weights of all more distal double recombinants. Table 4 compares the response to selection expected with this multi-marker strategy (last column), to that expected with four single pair strategies, for small to large population sizes. As expected the optimal distance for each single pair of markers depends on the population size, but for each population size the response obtained with the multi-marker strategy is higher than this optimum. After six generations the response obtained with the multi-marker strategy for small  $N$  is even slightly higher than for any other strategy for large  $N$ , so that it may be more interesting to type a few individuals for several markers spread along the chromosome, than to type many individuals



TABLE 4

## Use of several markers surrounding the introgressed gene

Population size	Generation	Selection on single pair				Multi-marker selection
		1 cM	2 cM	5 cM	10 cM	
20	3	0.793	0.834	0.901	0.905	0.918
	6	0.912	0.947	0.953	0.940	0.977
50	3	0.820	0.873	0.916	0.914	0.928
	6	0.942	0.964	0.955	0.939	0.982
100	3	0.861	0.898	0.933	0.917	0.937
	6	0.970	0.968	0.959	0.938	0.981
200	3	0.865	0.912	0.927	0.918	0.942
	6	0.975	0.968	0.955	0.943	0.981

Four pairs of markers are located at 1, 2, 5 and 10 cM on each side of the introgressed gene.  $G$  values after selection performed on an index combining the four pairs of markers (last column), as explained in the text, can be compared to  $G$  values after selection performed on only one of the four pairs. Simulation results, proportion selected  $s = 0.10$ .

for a few very proximal markers. Moreover, reducing the number of individuals may accelerate the experiment and reduce its cost, especially for animal species.

Whereas a significant improvement in recipient genome recovery is expected with the use of markers linked to the introgressed gene, as compared to the expectations under no selection (STAM and ZEVEN 1981), it is worth stressing that when the proportion selected exceeds ten percent it does not seem possible to get a process of introgression on the carrier chromosome that is faster than the random process on non-carrier chromosomes. This is probably due to the short time interval considered. The first aim of selection on the carrier chromosome is to sort out recombinants near the introgressed gene. Once this is achieved the carrier chromosome, except for the segment surrounding the gene, behaves like the non-carrier chromosomes and can be submitted to selection. However, most of the markers selected during the initial step will have reached fixation, and then introgression will proceed as if without selection, unless other polymorphic markers can still be found on the chromosome.

**Selection on the whole genome:** Two subsets can be distinguished in the total foreign genome present in a gamete from a crossed parent. The first is the continuous segment surrounding the introgressed gene on the carrier chromosome, and the second is the other segments spread on non-carrier chromosomes or on telomeric parts of the carrier chromosome. The efficiency of selection is not the same for these two subsets. Since, by definition, any individual candidate for selection is carrying the introgressed gene, the foreign genome around this gene is strongly retained, and selection is less efficient. Moreover, the results show that it does not seem possible to simultaneously optimize selection for both the carrier and

non-carrier chromosomes. Hence, a decision must be taken as to the efficiency of selection for each part of the genome.

Selection could be performed using an index on the markers of all chromosomes, according to the principle described in the previous section. For example, in order to attain equal efficiencies of selection for the two subsets of foreign genome, the relative weight in the index of the markers devoted to the control of the foreign segment surrounding the introgressed gene should be equal to the proportion of the total foreign genome due to this segment. This value will depend on the genome size and the generation considered. Some examples are shown in Table 5. Such an index could be further refined to take into account qualitative considerations concerning the distance between recipient and donor genomes (for example, if the donor is a wild species, priority might be given to eliminating detrimental genes on non-carrier chromosomes), or the presence of favourable genes near the introgressed one that should be kept with it on the carrier chromosome.

## CONCLUSION

Our results show that markers are useful in gene introgression programs. The selection strategy to be applied should take into account the available material and conditions, and should determine when and how strongly to select, and which markers to use. A common approach is to select as early as possible, and as strongly as possible on markers very close to the introgressed gene, generally without paying much attention to non-carrier chromosomes. We have shown that this is not always the best strategy. Three of our results might be kept in mind when seeking an optimal selection strategy in gene introgression programs:

First, for the carrier chromosome, the response to selection is a function of both the distance between the markers and the introgressed gene, and the proportion of selected individuals.

Second, once a first step of selection has been performed on the carrier chromosome, it is still possible to select on non-carrier chromosomes, although the intensity of selection available for the latter depends on that applied to the former.

Third, the later the selection is performed, the more efficient it is. We showed this point both analytically (for a single generation of selection) and by simulation (for several generations of continuous selection).

The first point shows that the strategy commonly advocated is valid only if it is possible to reproduce the population with very few individuals, otherwise too few of the selected individuals will be recombinants for markers close to the introgressed gene.



**TABLE 5**  
**Distribution of foreign genome without selection on a gamete from a crossed parent**

Generation	Carrier chromosome		Non-carrier chromosome Total (c)	Proportion of foreign segment in total foreign genome		
	Segment (a)	Total (b)		n = 10 Chr (x)	n = 20 Chr (x)	n = 30 Chr (x)
1	0.7842	0.8114	0.5000	0.1476	0.0761	0.0512
2	0.6275	0.6662	0.2500	0.2152	0.1159	0.0793
3	0.5126	0.5538	0.1250	0.3053	0.1750	0.1227
4	0.4271	0.4660	0.0625	0.4153	0.2583	0.1874
5	0.3622	0.3968	0.0313	0.5342	0.3657	0.2780
6	0.3111	0.3410	0.0156	0.6460	0.4877	0.3918

(a) length of foreign segment surrounding the introgressed gene relative to carrier chromosome length; (b) proportion of foreign genome on carrier chromosome; (c) proportion of foreign genome on non-carrier chromosomes; (x) proportion of (a) relative to the overall foreign genome of the gamete.  $x = a/[b + (n - 1)c]$  where  $n$  is the number of chromosomes.

Hence, for intermediate selection intensity, a better strategy is either to select in the short term for markers further from the gene, or to wait for a few generations until recombinants for markers close to the gene can accumulate in the population. This choice will depend on the genetic map available for the species selected. Of course, if many markers are available, selection could be performed both in the short term for markers far from the gene, and later for markers closer to it. Whereas the general belief is that one need pay attention only to markers close to the introgressed gene, our results emphasize the importance of distant markers in most cases.

Thanks are due to FRANÇOIS RODOLPHE, ALAIN CHARCOSSET, ISABELLE OLIVIERI, DANIEL WALLACH and NIGEL GRIMSLEY for careful reading of the manuscript.

LITERATURE CITED

FISHER, R. A., 1949 *The Theory of Inbreeding*. Oliver & Boyd, Edinburgh.  
 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, A. CAHANER and U. LAVI, 1990 DNA fingerprint applied to gene introgression breeding programs. *Genetics* **124**: 783-789.  
 LANDE, R., and R. THOMPSON, 1990 Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**: 743-756.  
 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.  
 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 A

**Response to selection at  $L$  independent loci:** We consider a set of  $L$  independent marker loci where

alleles from the donor and recipient lines can be distinguished, and assume that selection is performed in a population of infinite size. The trait selected for is the frequency  $G$  of recipient alleles at the marker loci. If selection is applied at generation ( $g$ ), the expected response to selection depends on the proportion of selected individuals, and on the variance of  $G$  among zygotes of the  $g$ th generation.

In a  $g$ th generation backcrossed individual, let  $X(g)$  be the number of recipient alleles received from its crossed parent. The proportion of recipient alleles is then

$$G(g) = [L + X(g)]/(2L) = 1/2 + (1/2)(X(g)/L).$$

In the next generation, assuming independence between loci, the number  $X(g + 1)$  in an offspring can be written as

$$X(g + 1) = X(g) + Y$$

where  $Y$  is a random binomial variable with parameters  $[L - X(g)]$  and  $(1/2)$ . Then conditional moments of  $X(g + 1)$  are

$$E[X(g + 1)|X(g)] = X(g) + (1/2)[L - X(g)]$$

$$\text{Var}[X(g + 1)|X(g)] = (1/4)[L - X(g)]$$

From this we get

$$E[X(g + 1)] = (1/2)[L + E[X(g)]]$$

$\text{Var}[X(g + 1)] = (1/4)(L - E[X(g)]) + (1/4)\text{Var}[X(g)]$  using initial conditions in  $F_1$  that  $X(0) = 0$ ,  $\text{Var}[X(0)] = 0$  we then have

$$E[X(g)] = L[1 - (1/2)^g]$$

$$\text{Var}[X(g)] = L[(1/2)^g - (1/4)^g].$$

Without selection we thus have

$$E[G(g)] = 1 - (1/2)^{g+1}$$

$$\text{Var}[G(g)] = [(1/2)^g - (1/4)^g]/(4L).$$

Now we consider that selection is applied for the first time during the  $g$ th generation, with a proportion ( $s$ ) of selected individuals. If  $L$  is large enough, the distribution of  $G$  is nearly Gaussian, and the expected advance due to selection can be approximated by

$$\text{Int}(s) \cdot (\text{Var}[G(g)])^{1/2}$$

where  $\text{Int}(s)$  is the expectation of the truncated reduced normal distribution with tail frequency ( $s$ ). After selection the expected frequency of recipient marker alleles is then approximately equal to

$$E^*[G(g)] = 1 - (1/2)^{g+1} + \text{Int}(s) \cdot \left(\frac{1}{2}\right) \cdot \left(\frac{(1/2)^g - (1/4)^g}{L}\right)^{(1/2)}$$

If no selection is applied during the next generations, the expected proportion follows the usual recurrence relation

$$E[G(k + 1)] = (1/2)E[G(k)] + (1/2) \text{ for } k \geq g.$$

The overall result after  $n$  generations, involving only selection during the  $g$ th generation is thus approximately equal to:

$$E^{(g)}[G(n)] = 1 - \frac{1}{2} \left[ (1/2)^n - \text{Int}(s) \cdot (1/2)^{n-g} \left(\frac{(1/2)^g - (1/4)^g}{L}\right)^{(1/2)} \right] \quad (\text{A1})$$

This expression is maximal if ( $g$ ) is maximal, that is, the highest response is obtained if selection is performed during the last generation.

APPENDIX B

**Response around the introgressed gene:** A similar analysis can be done on the chromosome carrying the introgressed gene, although the analysis is complicated by taking account of recombinations. Following HANSON (1959) and STAM and ZEVEN (1981), we assume that the number of crossing overs along a chromosomal segment of length ( $u$ ) Morgans follows a Poisson law with mean ( $u$ ), and we compute the frequency of the recipient genome at any locus ( $x$ ) on the chromosome. (Strictly speaking, we consider probability densities.) We consider one half of the chromosome, from the locus of the introgressed gene (locus 0) to one telomere (locus at distance  $l$ , the length of the half chromosome), and on this half chromosome a marker locus at distance  $d$  from the introgressed gene. Let  $x(u, g)$  be the frequency of recipient alleles at locus  $u$ , generation  $g$ .

**Densities of recipient alleles in offspring:** Without taking account of the marker locus, the proportion of recipient alleles at locus  $u$  in the next generation can

be written as

$$x(u, g + 1) = r(u) \cdot 1 + (1 - r(u)) \cdot x(u, g) \quad (\text{B1})$$

for  $0 < u < l$ , where  $r(u)$  is the recombination rate corresponding to a distance  $u$  on the chromosome map by Haldane's mapping function

$$r(u) = (1/2) \cdot [1 - \exp(-2u)].$$

Equation B1 gives the change with time of the mean recipient genome content on the chromosome carrying the introgressed gene. Integrating over the half-chromosome length gives

$$G(g) = \frac{1}{2} + \frac{1}{2l} \int_0^l x(u, g) du.$$

Since  $x(u, 0) = 0$  in  $F_1$  one gets

$$G(g) = 1 - \frac{1}{2l} \int_0^l [1 - r(u)]^g du.$$

If selection is applied on the marker, the recurrence relation for  $x(u, g)$  must take account of the allele carried at the marker locus:

(a) Consider first the case where a recombination occurred between the gene and the marker: the offspring carries a recipient allele at the marker locus. This case occurs with probability  $r(d)$ . For a locus ( $u$ ),  $0 < u < d$ , the recombination may have taken place before, or after locus ( $u$ ). Splitting the recombination fraction  $r(d)$  into the two terms corresponding to both conditions gives

$$r(d) = [1 - r(u)] \cdot r(d - u) + r(u) \cdot [1 - r(d - u)].$$

Then

$$x(u, g + 1) = \frac{[1 - r(u)] \cdot r(d - u)}{r(d)} \cdot x(u, g) + \frac{r(u) \cdot [1 - r(d - u)]}{r(d)} \cdot 1 \quad (\text{B2})$$

for  $0 < u < d$ . From the marker to the telomere, we get a relation similar to that given in Equation B1:

$$x(u, g + 1) = r(u - d) \cdot x(u, g) + [1 - r(u - d)] \cdot 1 \quad (\text{B3})$$

for  $d < u < l$ .

(b) The case of no recombination (the offspring carries a donor allele at the marker locus), occurs with probability  $[1 - r(d)]$ . In this case, the probability of recombination between the introgressed gene and a locus ( $u$ ),  $0 < u < d$ , is equal to the probability that recombinations occur in both segments  $]0, u[$  and  $]u, d[$ . The total probability of no recombination between 0 and  $d$  is then split into two terms:  $1 - r(d) = [1 - r(u)] \cdot [1 - r(d - u)] + r(u) \cdot r(d - u)$ , so that

$$x(u, g + 1) = \frac{[1 - r(u)] \cdot [1 - r(d - u)]}{1 - r(d)} \cdot x(u, g) + \frac{r(u) \cdot r(d - u)}{1 - r(d)} \cdot 1 \tag{B4}$$

for  $0 < u < d$ . From the marker to the telomere we then have the same relationship as in Equation B1 without selection:

$$x(u, g + 1) = r(u - d) \cdot 1 + [1 - r(u - d)] \cdot x(u, g) \tag{B5}$$

for  $d < u < 1$ .

**Selection for recombinants at the marker locus:** Now we consider selection for recombinant individuals at the marker locus during the  $g$ th generation, assuming as in APPENDIX A that no selection is practiced during other generations, and that the  $n$ th generation ( $n \geq g$ ) is observed. The response to selection is measured by the integrated mean recipient genome content  $G(n)$ , and depends on the time of selection ( $g$ ), on the distance ( $d$ ), and on the available selection intensity ( $i$ ) which is a function of the proportion  $s$  of selected individuals.

From the beginning of the process to the time of recombination, densities change according to Equations B4 and B5; at the time of recombination, densities are given by Equations B2 and B3; then from the time of recombination to the time of selection, densities change following Equation B1. Since selection is applied at generation ( $g$ ), the expected proportion of non recombinant individuals is

$$[1 - r(d)]^g$$

and the complementary proportion  $R(g)$  of recombinants is the sum over  $\alpha$  of the probabilities

$$r(d) \cdot [1 - r(d)]^{\alpha-1}$$

that the recombination event took place between generations  $(\alpha - 1)$  and  $(\alpha)$ ,  $\alpha = 1, 2, \dots, g$ . Combining these probabilities with the evolutionary rules (B4) and (B5) before recombination, (B2) and (B3) at recombination, and then (B1), yields two densities

$$x^R(u, g), \text{ and } x^{NR}(u, g)$$

for recombinant and non recombinant individuals in generation  $g$ .

The relative proportion of recombinant to non recombinant individuals is increased in the selected population. Assuming for simplicity that the introgressed gene is in the middle of the chromosome, with two half chromosomes of length ( $l$ ) and markers at distance ( $d$ ) on each side, selecting a proportion ( $s$ ) of individuals yields the following densities  $x^*(u, g)$  in the selected population ( $R(g)$  is the expected proportion of recombinants on one half chromosome)

if:

$$s < R(g)^2$$

then only double recombinants are selected for, so that

$$x^*(u, g) = x^R(u, g)$$

if:

$$R(g)^2 < s < R(g)^2 + 2 \cdot R(g) \cdot [1 - R(g)],$$

then a mixture of double and single recombinants are kept and so

$$x^*(u, g) = \frac{R(g)^2}{s} x^R(u, g) + \frac{s - R(g)^2}{s} \left[ \frac{1}{2} x^R(u, g) + \frac{1}{2} x^{NR}(u, g) \right] = \frac{s + R(g)^2}{2s} x^R(u, g) + \frac{s - R(g)^2}{2s} x^{NR}(u, g).$$

if:

$$s > R(g)^2 + 2 \cdot R(g) \cdot [1 - R(g)],$$

all three genotypes at the marker may be kept. Then

$$x^*(u, g) = \frac{R(g)}{s} x^R(u, g) + \frac{s - R(g)}{s} x^{NR}(u, g).$$

Thereafter, from generations ( $g$ ) to ( $n$ ), densities change according to Equation B1, starting with their values  $x^*$  obtained after selection during the  $g$ th generation.

Results were obtained numerically by calculating the preceding densities at a large number of distinct loci (100 and 1000 per Morgan), and summing them along the chromosome length to get mean recipient genome content  $G$ .

### APPENDIX C

**Combined selection in the first generation:** Selection aimed at reducing the length of the donor segment surrounding the introgressed gene may be combined with global selection for the recipient genome on other chromosomes, provided that the selection intensity is sufficiently great.

We assume that selection is performed during the first backcross generation. As in APPENDIX B, we assume a pair of markers surrounding the gene at a distance ( $d$ ) corresponding to a recombination fraction  $r = r(d)$ , on each half chromosome of length ( $l$ ). Recombinant individuals that carry the recipient allele are kept. The expected recipient genome content  $G$  is derived, both in recombinant and non recombinant offspring, from Equations B2, B3 and B4, B5, respectively. Similarly, the expected length  $\mu$  of the donor genome segment around the gene can be calculated in both genotypes. One gets the following expressions. (Note that the mean results of STAM and ZEVEN (1981) can be derived by combining these expressions.)

In double recombinants (proportion  $r^2$ ):

$$G^{RR} = \frac{1}{2} + \frac{1}{2} \frac{1 - \exp[-2(l-d)]}{2l} \tag{C1a}$$

$$\mu^{RR} = 2 \frac{[1 - \exp(-d)]^2}{1 - \exp(-2d)}. \tag{C1b}$$

In single recombinants (proportion  $2r(1-r)$ ):

$$G^{NR} = \frac{1}{2} - \frac{1 - \exp(-2d)}{4.l.[1 + \exp(-2d)]} \tag{C2a}$$

$$\mu^{NR} = \frac{[1 - \exp(-d)]^2}{1 - \exp(-2d)} + \frac{1 - \exp(-2d)}{1 + \exp(-2d)} + 2 \frac{\exp(-d) - \exp(-l)}{1 + \exp(-2d)}. \tag{C2b}$$

In non-recombinants (proportion  $(1-r)^2$ ):

$$G^{NN} = \frac{1}{2} - \frac{1 - \exp(-2d)}{2.l.[1 + \exp(-2d)]} \tag{C3a}$$

$$\mu^{NN} = 2 \frac{1 - \exp(-2d)}{1 + \exp(-2d)} + 4 \frac{\exp(-d) - \exp(-l)}{1 + \exp(-2d)}. \tag{C3b}$$

In a first step, selection sorts out double recombinant individuals, then single recombinants. Then a second step of selection for global recipient genome content may be considered (see RESULTS for details). Under the approximation of APPENDIX A, the response de-

pends on the selection intensity which is still available.

As in APPENDIX A, let  $\text{Int}(\cdot)$  be the intensity function for normal variates. Then the effective intensities obtained when the global frequency of selected individuals is equal to ( $s$ ) are:  
for

$$s < r^2: \text{ then } i = \text{Int}\left(\frac{s}{r^2}\right)$$

for

$$r^2 < s < r^2 + 2r(1-r),$$

then:

$$i = \frac{s - r^2}{s} \text{Int}\left(\frac{s - r^2}{2r(1-r)}\right)$$

for

$$s > r^2 + 2r(1-r),$$

then:

$$i = \frac{s - r^2 - 2r(1-r)}{s} \text{Int}\left(\frac{s - [r^2 + 2r(1-r)]}{(1-r)^2}\right).$$

Given this set of equations in  $G$ ,  $\mu$  and  $i$ , one can compute the joint expected response to selection. Shown here for the first generation, they can be readily extended to selection during the  $g$ th generation (provided no selection was done before), by replacing the ( $r$ ) value by the  $R(g)$  recombinant fraction after  $g$  generations.