Copyright © 1996 by the Genetics Society of America

Marker-Assisted Introgression in Backcross Breeding Programs

Peter M. Visscher, Chris S. Haley and Robin Thompson

Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland

Manuscript received December 12, 1995
Accepted for publication August 28, 1996

ABSTRACT

The efficiency of marker-assisted introgression in backcross populations derived from inbred lines was investigated by simulation. Background genotypes were simulated assuming that a genetic model of many genes of small effects in coupling phase explains the observed breed difference and variance in backcross populations. Markers were efficient in introgression backcross programs for simultaneously introgressing an allele and selecting for the desired genomic background. Using a marker spacing of 10–20 cM gave an advantage of one to two backcross generations selection relative to random or phenotypic selection. When the position of the gene to be introgressed is uncertain, for example because its position was estimated from a trait gene mapping experiment, a chromosome segment should be introgressed that is likely to include the allele of interest. Even for relatively precisely mapped quantitative trait loci, flanking markers or marker haplotypes should cover ~10–20 cM around the estimated position of the gene, to ensure that the allele frequency does not decline in later backcross generations.

CROSSES between inbred lines and crosses between outbred lines have been used to explain the genetic nature of breed or line differences in species of agricultural interest and to detect quantitative trait loci (QTL) in backcross (BC) and F_2 populations (*e.g.*, PATERSON *et al.* 1988; STUBER *et al.* 1992; ANDERSSON *et al.* 1994). Besides understanding the basis of genetic variation, such experiments may detect QTL that are of commercial interest. For example, ROTHSCHILD *et al.* (1994) found evidence for a QTL affecting litter size in a cross between commercial Large White pigs and Chinese Meishan pigs. The favorable QTL allele, which had an effect of approximately one piglet and originated from the Meishan, would be of economic benefit in commercial pig populations. In several recent studies it has been shown that genetic markers can be used to introgress genes from one line into another (SMITH *et al.* 1987; HILLEL *et al.* 1990, 1993; GROEN and TIMMERMANS 1992; HOSPITAL *et al.* 1992; GROEN and SMITH 1995). Genetic markers could be used in two ways in introgression programs: (1) using markers for the gene that is to be introgressed and (2) using markers to select for (or against) a particular background genotype. In all recent studies it was assumed that the genotype at the gene to be introgressed was known exactly, hence there was no need to use markers to aid its introgression and attention was focused on the background genotype. In fact the genotype of genes to be introgressed will not usually be known as they will often be QTL, which can only be genotyped imprecisely, if at all, using phenotypic information, or major loci (*e.g.*, coat color), which more often than not have dominance/recessive relationships and hence heterozygotes cannot be distinguished from one of the homozygotes.

Having located a desired gene and genetic markers associated with it in an "inferior" breed, the aim of the introgression phase is to fix the favorable alleles in the commercial population with as little as is possible of the remainder of genome from the inferior breed. The usually proposed route (*e.g.*, SMITH *et al.* 1987; HILLEL *et al.* 1990; GROEN and TIMMERMANS 1992; HOSPITAL *et al.* 1992) is a number of generations of backcrossing a population that carries the allele to be introgressed (from the donor population, *i.e.*, the inferior breed) to a recipient population (*i.e.*, the commercial breed), followed by an *inter se* cross to make the desired allele homozygous.

HILLEL *et al.* (1990, 1993) proposed that DNA fingerprints could be used for introgression of alleles in backcross populations by selecting for or against a certain genomic background. Their theory is based on a number of "chromosome segments" that effectively are unlinked loci. However, there are some problems with their theory and results. In general, the class of minisatellite markers (fingerprints) are not very suitable for use in introgression programs (the markers are dominant and the fingerprint loci will often not be mapped, but are known to be nonrandomly distributed across the genome). HILLEL *et al.* (1990, 1993) implicitly assume that the proportion of the genome from one line (or breed) in a (back)cross is the same as the proportion of unlinked markers from the same line. This is not correct, because it ignores recombination around the marked loci, as was also noted by HOSPITAL *et al.* (1992) and GROEN and SMITH (1995), hence the results from HILLEL *et al.* (1990, 1993) are unrealistic. HILLEL

Corresponding author: Peter M. Visscher, Institute of Ecology and Resource Management, University of Edinburgh, West Mains Rd., Edinburgh EH9 3JG, Scotland. E-mail: peter.visscher@ed.ac.uk

et al. (1990) derive an equation for the variance of the proportion (in percentage) of segments originating from the desired genome line, var $[G(A)]$. For backcross generation t, the result of applying their equation is, var $[G(A)] = 625/(Nt)$ (Table 2 of HILLEL *et al.*), with N the number of segments (marker loci). We believe this result is incorrect, and that the correct expression is (see also APPENDIX A and HOSPITAL *et al.* 1992)

$$\text{var}\,[G(A)] = (2500/N)\tfrac{1}{2}^{t}(1 - \tfrac{1}{2}^{t}).$$

This result is for the case of no selection. Furthermore, HILLEL *et al.* (1990) do not take account of a reduction in the variance of var $[G(A)]$ when animals (or plants) are selected on the number of markers from the desired line; this reduction in variance can be substantial.

GROEN and TIMMERMANS (1992) presented a simulation study on the use of genetic markers to increase the efficiency of introgression. When comparing phenotypic selection and selection using markers in a backcross program, they conclude that not much benefit is to be expected from using markers. However, as pointed out by the authors, this conclusion depends on the parameters used (in particular on the effectiveness of phenotypic selection). Using the parameters of GROEN and TIMMERMANS (1992), selection using markers has a small advantage over phenotypic selection for at least the first three backcross generations.

The thorough study of HOSPITAL *et al.* (1992) deals in detail with introgression in backcross breeding programs. The authors do not consider selection on a quantitative trait. For the chromosome with the allele to be introgressed, the authors only consider two well-placed markers. One of the main conclusions of HOSPITAL *et al.* (1992) was that retrieving the recipient's genome was approximately two generations faster if markers were used.

GROEN and SMITH (1995) followed on from the study of GROEN and TIMMERMANS (1992) and studied the efficiency of phenotypic selection and selection on markers in backcross and intercross programs. Selection was in two stages: first, animals were selected carrying the allele to be introgressed, and among those animals, those with the best phenotype or those with the largest number of markers from the recipient line were selected. GROEN and SMITH (1995) concluded that selection using markers is always inferior to selection for phenotypes. However, the assumption with regards to the distribution of gene effects for the quantitative trait under consideration in the donor and recipient line is important. GROEN and SMITH (1995) assumed QTL of equal size, with a frequency of 0.7 in the recipient line, and 0.6 in the donor line. In this case most of the genetic variation is *within* lines, so that selection on markers giving between line information is not expected to contribute much.

The efficiency of the marker-assisted introgression program depends on (1) the frequency of the introgressed allele in the final population and (2) the genetic progress for other traits. As most studies have assumed that the allele to be introgressed can be identified without error by a single marker, the frequency of the allele during the backcross phase remains at 50%. In practice, a single marker or a marker bracket associated with the QTL or major gene is likely to be used and the recombination fraction between the gene and marker(s) may be larger than zero, so that the frequency of the allele may be substantially <50% after a few generations of backcrossing.

The aims of this paper are as follows: (1) to investigate by simulation the relative genetic gain in a backcross program by using only markers, only phenotypes, or an index of markers and phenotypes, (2) to investigate by simulation the frequency of the gene to be introgressed when its position is not known exactly, and (3) to introduce an infinitesimal model to explain breed differences. All three areas are novel, since in previous publications no explicit genetic model was investigated for the background genotype, and the gene to be introgressed was assumed to be known.

MODELS AND METHODS

We assume a (back)cross between inbred lines, with 100% informative markers. In all our calculations we use HALDANE's (1919) mapping function without interference.

Gene to be introgressed *vs.* background genotype: Assume that the donor line is fixed for a gene of interest that we wish to introgress into a recipient population. This gene may be a major gene or a QTL, and has a significant effect on a component of economic performance. We do not consider any other genetic variation for the trait with which the gene to be introgressed is associated. Apart from the gene to be introgressed, we assume there is a quantitative trait that is associated with another component of economic performance. We term the genotype of an individual for this trait the background genotype and assume that the genotype of the donor line is inferior for this trait.

Infinitesimal model of linked loci for background genotype: To model genotypes in different backcross generations, we choose a model in which a large number of linked loci, all of equal and small effect and in coupling phase in the inbred lines, explains the breed difference and the variance in the first backcross or an F_2 population (APPENDIX A). The reason for choosing this model is because of its similarity with the "standard" infinitesimal model (*e.g.*, BULMER 1971) implicitly assumed for most livestock selection programs, its relative simplicity, and similarity to models used to predict the variance of the proportion of the genome that originates from either inbred line (STAM and ZEVEN 1981; HILL 1993), and because of lack of information on "more realistic" genetic models. Table 1 shows the

TABLE 1

Breed differences and heritability in a first backcross population assuming a large number of linked loci in coupling phase

Interval (cM)	D^a	h^{2b}
50	0.25	0.3
	0.50	1.1
	1.00	4.4
	2.00	15.5
100	0.25	0.2
	0.50	0.9
	1.00	3.4
	2.00	12.4
200	0.25	0.2
	0.50	0.6
	1.00	2.3
	2.00	8.6

Analytical results for a single interval, assuming a residual standard deviation of unity.

[a] Breed difference in residual standard deviations.

[b] Heritability (×100) in the first backcross population.

relationship between the breed difference and heritability for a single chromosome or chromosome segment in the first backcross when the underlying genetic model is our infinitesimal coupling model.

Selection criteria: When introgressing the gene of interest, selection is in two stages: first, all individuals that have the desired marker genotype are selected (which indicate that the individuals carry one copy of the gene to be introgressed), second, we select among these individuals for the desired background genotype.

If our interest is solely in recovering the recipient genome, and we use only marker information, a selection index was calculated as, $I_m = \Sigma b_i M_i$ $(i = 1, m)$. M_i is the value of marker i, which is either 0 (marker originates from donor line) or 1 (marker originates from recipient line). The optimum weights for b_i, at least in the case of no selection, are calculated in a manner analogous to standard selection index theory (HAZEL 1943), *i.e.*, $\mathbf{b} = \mathbf{P}^{-1}\mathbf{G}$, where $\mathbf{b}$ is a vector of index weights, $\mathbf{P}$ the covariance matrix of individual marker observations, and $\mathbf{G}$ a vector with covariances between marker observations and the proportion of the genome that comes from the recipient population. Note that with our genetic model, the breeding value of an individual is proportional to the genetic composition of its genome, *i.e.*, the proportion of the genome that is from the recipient population. In the absence of selection, index weights for any marker spacing and any backcross population were derived elsewhere (VISSCHER 1996). It was shown therein that for equally spaced markers, including markers at the chromosome ends, the relative weights for the end markers were ½ and 1.0 for all other markers (VISSCHER 1996). Hence, when using markers only to recover the recipient genome, the selection criterion in this study was created by adding up the number of markers that an individual has from the recipient line (and counting a half if the marker is at the end of a chromosome). Note that the marker index score is not calculated from regression of phenotypes on individual markers, as was done in marker-assisted selection (MAS) simulation studies (ZHANG and SMITH 1992; GIMELFARB and LANDE 1994).

The selection criteria for combining phenotypes and marker information is a straightforward extension from the previous selection index. Now we select on

$$I = b_m I_m + b_p P,$$

with b_m and b_p the weights given to the marker index and the phenotype (P), respectively. These weights are found by maximizing the correlation between the index and the breeding value of an individual, assuming the infinitesimal coupling model. It then follows that,

$$b_m \propto (1 - h^2)$$

$$b_p \propto h^2(1 - p),$$

with p the proportion of genetic variance explained by the marker index I_m, *i.e.*, $p = \text{cov}^2(A, I_m)/(\text{var}(I_m)\,\text{var}(A))$, and A is the additive genetic value. Under no selection, this proportion is identical to the proportion of the variance in genetic composition that is explained by the markers (VISSCHER 1996). The relative index weights are the same as those of LANDE and THOMPSON (1990).

When applying selection, we used the selection index weights that were derived for the case of random selection.

Simulation: One or 20 chromosomes, each of 100 cM in length, were simulated using $n = 101$ loci per chromosome. Simulated loci were 1 cM apart, and all loci have an effect on the background genotype. For the background genotype, alleles had effects of $-\alpha/2$ or $+\alpha/2$, depending whether they originated from the donor or recipient line, respectively. Per chromosome, the maximum genotypic value was $2n(\alpha/2) = n\alpha = 100\%$ of the recipient genotypic value, so that the F_1 was 0%, and without any selection the average genotypic value in backcross generation $t(G_t)$ was $(1 - {}^1/_2{}^t)100\%$. Crossovers were generated assuming HALDANE's (1919) mapping function without interference. For the simulations with introgression, the allele to be introgressed was at 25 cM from one end of chromosome *1*.

Given an arbitrary value for the population environmental variance (var_E), and given the population heritability in the first backcross generation (h^2), the population genetic variance for the first backcross generation before selection was calculated as $\text{var}_A(BC_1) = \text{var}_E\, h^2/(1 - h^2)$. The breed difference, D, was calculated from $\text{var}_A(BC_1)$ assuming an infinitesimal model (Equation A2 in APPENDIX A), and α was calculated as $D/(2n)$. For each sample of individuals, the heritability and genetic variance may differ from the population values because of sampling.

Some of the loci were designated to be marker loci. A marker allele has a small effect (α) on the background genotype, but its main use was to mark from which breed a chromosome segment was derived.

The population structure was the same in all simulations: each of 10 males was mated to 10 females each producing eight progeny (four males and four females), giving a population size (N) of 800. Options that were varied in different simulation runs were as follows.

Introgression: In some scenarios, no introgression was performed. For these cases, selection was only on the background genotype.

Selection for background genotype: Selection was either random, on phenotypes, on a marker score, or on an index of phenotypes and markers. In the case of the index, index weights were derived assuming no selection.

Number of markers per chromosome: Numbers were 1, 2, 3, 6, or 11. Markers were equidistantly placed, with always a marker at either end of the chromosome. For a single marker its position was in the middle of the chromosome.

Heritability in the first backcross generation: h^2 was 10% or 40%.

Selection intensity: Either males (proportion selected = 2.5%) or females (proportion selected = 25%) were selected across matings and backcrossed to the recurrent population.

Sampling of QTL position: In some simulations, the position of the gene to be introgressed was sampled from a normal distribution with mean 25 cM (the true position) and standard deviation σ cM. This simple method was employed to mimic uncertainty in the QTL position when the position (and effect) are estimated from a QTL mapping experiment. Markers closest to the estimated position of the QTL were used for introgression. If only a single marker was used for introgression and the estimated position was in the middle of a marker interval, the marker closest to the center of the chromosome was used for introgression.

Combinations of the above scenarios gave many simulation results, and therefore only the most informative results are presented. For each scenario, 100 replicates were simulated.

RESULTS

Selecting for background genotype only: Table 2 shows simulation results from background selection using a varying number of markers per chromosome, for 1 or 20 chromosomes. For a single chromosome, the conclusions from VISSCHER (1996) are confirmed, *i.e.*, increasing the number of markers from, say, three to 11 gives a substantial increase in the rate of genome recovery. However, as Table 2 clearly shows, the gain in real terms is small with a realistic number of chromosomes. Table 2 can be compared directly with results from HOSPITAL *et al.* (1992) who used a similar model, and results agree well. Their results are in terms of the proportion of the genome which is from the recurrent line, say P_t, which is related to G_t in this study as, $P_t = G_{t+1}$. (Their P_t was measured among the selected individuals in generation t, whereas our G_{t+1} was measured as the mean of all individuals in generation $t + 1$ before selection.) For example, for three evenly spaced markers per chromosome in backcross generation 2, we found an average genotypic value of 83.9 (Table 2) for 20 chromosomes, whereas HOSPITAL *et al.* (1992) found a value of 84.1 for a similar proportion selected (2.0% instead of our 2.5%) and using two well placed markers.

Selection on markers was superior to selection on phenotypes, at least in the first few generations, which is a consequence of the genetic model used in the simulation. When calculating index weights for phenotypes and markers in the absence of selection, it can be shown that the marker index score gets all the weight in the index for the parameters used in this study. This also explains why the results for selection on markers and an index are similar. Therefore, only results for marker and phenotypic selection are presented subsequently. When selection is on markers, the markers become fixed after a few rounds of selection (results not shown), and the slightly increasing average genotypic value in later generation is from repeatedly backcrossing to the superior breed. For example, with three, six and 11 markers, all markers were fixed by backcross generation 3, 4, and 5, respectively, when the proportion selected was 2.5%.

Simultaneous introgression and selection for background genotype: Average frequencies of the introgressed gene and average genetic values for the quantitative trait are presented in Table 3. The number of markers used for introgression is m_i. The position of the gene was assumed to be known (at 25 cM), and markers were either at the gene to be introgressed ($m_i = 0$) or linked to it ($m_i = 1$ or 2). Markers were at positions 0, 20, 40, 60, 80, and 100 cM, so that in the case of a single marker used for introgression it was 5 cM distant from the gene, and in the case of two flanking markers the distances were 5 and 15 cM. Table 3 shows that even with random selection we may lose the gene because of chance recombination. For example, for $m_i = 1$, the frequency was 41.8% in the fifth generation.

With selection on the background genotype, the decrease in the frequency of the gene to be introgressed below 50% is larger. Essentially we are selecting in favor of recombinations between the gene and the nearest marker locus, and even when a single marker is only 5 cM away, the frequency of the gene may drop to ~30% in generation 5 (Table 3). When flanking markers are used for introgression, the frequency of the gene stays

TABLE 2

Average genetic value in backcross populations undergoing background selection only

Selection	No. markers	BC generation							
		1 chromosome				20 chromosomes			
		2	3	4	5	2	3	4	5
Phenotypes	0	87.8	97.0	99.3	99.8	77.9	90.5	95.9	98.2
Markers	1	90.3	94.3	96.9	98.7	82.9	94.6	97.4	98.7
	2	94.3	97.2	98.8	99.2	82.5	94.5	98.1	99.0
	3	98.1	99.3	99.6	99.9	83.9	96.2	99.1	99.6
	6	99.6	99.8	99.8	99.9	84.6	96.8	99.7	99.9
	11	99.1	100.0	100.0	100.0	84.5	96.9	99.9	100.0
Index	1	93.9	98.3	99.3	99.8	82.9	94.7	97.7	98.9
	2	96.7	99.0	99.7	99.9	82.6	94.6	98.1	99.2
	3	98.5	99.4	99.8	99.9	84.0	96.2	99.2	99.6
	6	99.7	99.9	100.0	100.0	84.6	96.8	99.7	99.9
	11	99.9	100.0	100.0	100.0	84.6	96.9	99.8	100.0

Simulation results. The genetic value is relative to the value of the superior breed. $N = 800$, and 10 out of 400 males are selected each generation. $h^2 = 10\%$ in the first BC generation.

at 50%. However, a larger proportion of the genome of the donor breed is selected, and response to selection for the quantitative trait is low (average genetic value is ~80% in generation 5).

Location of gene to be introgressed is uncertain: In Table 4, frequencies of the QTL are shown for situations where the position of the QTL is unknown. The position was sampled from a normal distribution with standard deviation of either 3 or 6 cM. These standard deviations are realistic for QTL mapping experiments presently carried out in plant and animal populations (KNOTT and HALEY 1992). As expected, the gene frequency reduces dramatically for a large sampling variance combined with background selection on markers. For example, for both sampling standard deviations, the frequency decreased to ~23% in generation 5. Results for $\sigma = 3$ and $\sigma = 6$ were very similar when introgression was on a single marker. This results mainly from the density of the marker map and the position of the QTL. In our case, the position of the QTL was

TABLE 3

Frequency of introgressed gene and genetic value for different backcross generations

t^a	m_i^b	Selection on background genotype					
		Random		Phenotype		Markers	
		f(QTL)[c]	G[d]	f(QTL)	G	f(QTL)	G
2	0	50	60.8	50	66.0	50	85.5
	1	47.5	61.0	46.2	67.1	38.3	86.3
	2	49.8	57.5	49.7	60.7	49.5	77.0
3	0	50	68.4	50	76.9	50	94.8
	1	44.9	69.3	41.1	78.1	37.9	90.2
	2	49.4	63.2	48.5	68.9	49.0	80.9
4	0	50	74.4	50	84.0	50	95.2
	1	43.0	75.1	37.1	84.2	36.6	90.8
	2	49.0	67.3	48.4	74.2	48.4	81.6
5	0	50	78.6	50	88.0	50	95.7
	1	41.8	79.2	32.0	88.2	35.1	91.5
	2	48.7	70.3	48.5	77.8	47.7	82.2

Simulation results for a single chromosome. A marker map of six markers per chromosome was used. $h^2 = 10\%$, $N = 800$, and 10 males are selected each generation. The gene to be introgressed is a 25 cM from one end of the chromosome.

[a] Backcross generation.

[b] Number of markers used for gene introgression. $m_i = 0$ indicates that the marker used for introgression is at the gene itself.

[c] Allele frequency, *i.e.*, the proportion of individuals carrying one copy of the desired allele.

[d] Average genetic value (in %), relative to the value of the recipient population.

TABLE 4

Frequency of introgressed gene when the position of the QTL is estimated

		Selection on background genotype					
		Random		Phenotypes		Markers	
Generation	m_i^a	$\sigma = 3^b$	$\sigma = 6$	$\sigma = 3$	$\sigma = 6$	$\sigma = 3$	$\sigma = 6$
2	1	48.2	47.9	45.9	45.2	36.5	35.1
	2	49.9	49.2	50.0	48.1	48.5	43.7
	4	49.7	49.9	49.9	49.8	49.8	49.8
3	1	45.8	44.7	41.8	39.8	27.3	25.0
	2	49.2	47.5	49.9	46.4	46.5	39.1
	4	49.7	49.8	49.6	49.6	49.7	49.7
4	1	44.0	42.1	38.2	35.0	23.4	23.5
	2	48.8	47.7	49.4	43.9	46.3	39.6
	4	49.7	49.8	49.2	49.7	49.9	49.2
5	1	41.5	39.8	35.1	30.8	22.9	23.1
	2	48.3	46.4	49.0	41.6	46.5	39.1
	4	49.6	49.2	49.6	49.9	49.9	49.5

Simulation results for a single chromosome. h^2 in the first generation was 10%. The position of the gene to be introgressed was sampled from a normal distribution with mean 25 cM and variable standard deviation. The marker map consisted of 11 equidistal markers.

[a] Number of markers used for introgression.

[b] Standard deviation (in cM) of the estimated QTL position about the true position.

at 25 cM, and the four nearest markers were at positions 20, 30, 10, and 40 cM. Therefore, with both sampling standard deviations, the most likely markers to be selected for the introgression are those at position 20 and 30. Selecting on the four nearest markers is equivalent to selection for marker haplotypes.

If a single marker is used for introgression and there are two markers at equal distance from the (estimated) position of the QTL, the marker closest to the center of the chromosome should be chosen for introgression. The reason for this is that with selection on the background genotype the probability of selecting individuals with a recombination between the QTL and the marker is larger when the marker is closest to a chromosome end. For example, if the position of the gene to introgress is at 25 cM, and there are markers at 20 and 30 cM and background genotype selection is based on markers, introgressing using the marker at 30 cM gave a gene frequency of 42% in the second BC generation, while introgressing on the marker at 20 cM gave a frequency of 30% (results not shown in tables).

In APPENDIX B we present a method to calculate probabilities of maintaining the QTL in backcross populations, assuming that the true position of the QTL is normally distributed around the estimated position. In Table 6 minimum marker distances are presented that correspond to high probabilities that a QTL is inside a marker bracket and the desired QTL allele from the donor breed is associated with the marker alleles from the donor breed in the first backcross generation. For example, a chromosome segment flanked by markers that are 20.4 cM apart should be selected for a probability larger than 0.95 that a QTL is inside this bracket, assuming that the true QTL position is normally distributed around the estimated position with a standard deviation (σ) of 5 cM. The optimum marker spacing is the result of finding the balance between the effects of uncertainty of the true QTL position and double recombination within the marker bracket (see APPENDIX B for more details). When the position of the QTL is estimated poorly, it may be impossible to find a bracket that gives a high probability of maintaining the

TABLE 5

Genetic value in backcross populations depending on the proportion of animals selected and the heritability in the first generation

		Selection on background genotype			
		Phenotypes		Markers	
Generation	p[a]	0.1[b]	0.4[b]	0.1[b]	0.4[b]
2	2.5	66.0	74.5	85.5	85.9
3	2.5	76.9	85.1	94.8	94.7
4	2.5	84.0	90.1	95.2	95.0
5	2.5	88.0	92.9	95.7	95.3
2	25	62.3	64.9	69.6	69.5
3	25	71.9	76.2	83.0	82.6
4	25	78.4	83.1	88.9	88.8
5	25	83.0	87.1	92.2	92.1

Simulation results for a single chromosome. The genetic value is relative to the value of the superior breed. A marker map of six evenly spaced markers was used, and an additional marker at 25 cM was used for introgression. $N = 800$.

[a] Proportion selected (in %).

[b] Heritability values.

TABLE 6

Minimum marker distances for given probabilities that a QTL is inside a marker bracket and the desired QTL allele associated with markers from the donor population

		σ^b				
		1	2	3	4	5
$P^a_{2(1)} \geq$	0.95	4.0	8.0	12.0	16.2	20.4
	0.975	4.6	9.2	14.0	19.0	25.2
	0.99	5.2	10.8	17.8	[c]	[c]

[a] The probability of a QTL being located within a marker bracket and the desired QTL allele from the donor breed associated with the marker alleles from the donor breed in the first backcross generation. The estimated position of the QTL is assumed to be midway between the flanking markers, at location 50 cM on a 100 cM chromosome, and its true position is sampled from a normal distribution.

[b] Standard deviation (in cM) of the estimated QTL position about the true position.

[c] $P_{2(1)} < 0.99$ for all marker spacings.

QTL in the backcross population. For example, for $\sigma =$ 5 cM, there is no marker spacing that gives a probability ≥ 0.99 of maintaining the QTL (Table 6); a narrow marker interval increases the chance of the true QTL being outside the interval, while a wide interval increases the chance of double recombinants.

Influence of heritability and selection intensity: Average genetic values for the background genotype for a proportion selected of either 2.5 or 25%, and a heritability in the first backcross population of either 10 or 40% are in Table 5. For these simulations, it was assumed that the position of the gene to be introgressed was known, and that a marker at the gene itself was available. Hence, the frequency of the introgressed gene remained constant close to 50%. Increasing the proportion selected from 2.5 to 25% decreased the response to selection, in particular for selection on markers. For a large heritability, selection on markers is still superior because of the underlying genetic model. Obviously, the response to selection on markers does not vary with heritability, because the proportion of the genetic variance explained by the markers remains constant.

DISCUSSION

Implications for breeding programs: Introgression of desirable alleles using markers may have several advantages over introgression using phenotypic information only. For the allele to be introgressed, marked chromosome segments ensure that the correct segment of donor genome is incorporated into the recipient line. For nonadditive acting alleles, using markers may be the only way to ensure a successful introgression program. For the background genotype, using markers gives a direct estimate of the proportion of the donor genome that is still present in each backcross generation. This may be preferred over phenotypic selection, in particular for traits with low heritabilities that are difficult to measure (*e.g.*, age and sex-limited traits).

Still, deciding on an introgression program may be a risky undertaking for a breeding company, because usually the "donor" breed will be inferior with respect to the main traits of economic importance. For example, in plant breeding the donor line may be carrying disease resistance genes, but will be inferior with respect to yield. Similarly in pig breeding, a donor breed may be superior for litter size, but inferior with respect to growth traits. To keep the risk to a minimum, the breeder has to be sure that there is a gene (QTL) worthwhile to be introgressed, and during the introgression phase the gene should not be lost. To keep the gene it was found that selecting on marker haplotypes covering the likely position of the QTL would maintain the frequency of the QTL at 50% (Table 4). With the assumed genetic model, selecting on marker scores gives faster rates of genome recovery than selecting on phenotypes. In practice, breeders may use family and pedigree information for selection purposes, and not all markers will be informative, so that selection using phenotypic records may become more competitive.

Optimum spacing of markers for introgression: If the gene to be introgressed was detected in a QTL mapping experiment, and we have an estimate of the standard errors of the effect and position of the QTL, an optimum marker spacing for introgression was derived (APPENDIX B). For practical situations, a marker distance of 10–20 cM seems appropriate. If the location of the QTL is estimated very poorly, it may be better to select on a marker haplotype based on more than two markers. In that case the width of the chromosome segment should take care of the uncertainty of the QTL position, while the markers should be spaced to give a low probability of there being undetected double recombinants.

Infinitesimal model of linked effects: To investigate the efficiency of simultaneously introgressing a gene and selection on some quantitative trait, a genetic model had to be assumed. Due to the lack of knowledge about the distribution of genes affecting quantitative traits, both within and between breeds, we chose an infinitesimal model, *i.e.*, many genes with the same effect with the breeds fixed for alternative alleles. This model is similar to the "coupling phase" model of GIMELFARB and LANDE (1994), but with a different assumption regarding the QTL effects (GIMELFARB and LANDE assumed that gene effects follow a geometric series, whereas we assume all genes have the same effect), and the number of QTL. ROBERTSON (1977) proposed an infinite locus model for a finite chromosome assuming linkage equilibrium between all pairs of loci across the population, and DEKKERS and DENTINE (1991) used a similar model to investigate the variance explained by genetic markers in outbred populations. In practice,

the nature of genetic variation within and between breeds will depend on which two breeds are under consideration. In some cases breeds may differ from a common origin through genetic drift (*e.g.*, wild-type plant or animal populations), whereas in other cases breeds may differ through artificial selection. Frequencies of genes affecting quantitative traits and determining the breed difference D will probably be different for these cases. Our genetic model may be unrealistic, but has some relatively simple properties (APPENDIX A). Elsewhere, we show that our model has implications for QTL mapping experiments (VISSCHER and HALEY 1996).

Population structure: Although the heritability and selection intensity were varied to investigate whether they influenced the efficiency of selection on the background genotype, the population structure, in terms of the number of male and female parents and population size, was not altered. Changing the population structure is unlikely to change the conclusions dramatically, assuming the genetic model of many linked loci. Even selecting a single individual, as is possible in plant breeding, could be seen as an effect of selection intensity. Only in very large populations would it be possible to pick out the ideal genotype, *i.e.*, the individual with the desired introgressed gene, and the recipient genome recovered in a single generation. With many chromosomes assorting independently, the required population size may well be too large (or costly) to achieve this. For example, consider the simple example of c pairs of chromosomes of equal length in the first backcross generation. Assuming Haldane's mapping function, the probability of having an individual with the complete genotype of the recipient population is $[^1/_2(1 - r_m)]^c$, with r_m the recombination rate between the chromosome ends. Hence, for a genome of 10 chromosomes of length 1 Morgan, the probability of recovering the recipient genome is $[^1/_2(1 - 0.4323)]^{10} = 3.4 \times 10^{-6}$. This gives a required population size to expect a single individual with the desired genotype in the order of 300,000.

Future work: Topics that have not been addressed in this study are the introgression of multiple alleles simultaneously, and introgression of alleles when genetic markers are not fully informative. These areas need further research, because they are of great importance to practical plant and animal breeders.

We thank NAOMI WRAY, JOHN WHITTAKER, BILL HILL for helpful comments. FRED HOSPITAL and another referee are thanked for their careful reading of previous manuscripts, and their many useful suggestions. P.M.V. was funded by the Marker Assisted Selection Consortium of the U.K. pig industry (Cotswold Pig Development Company Ltd., J.S.R. Farms Ltd., Pig Improvement Company/National Pig Development Company, Newsham Hybrid Pigs Ltd., and the Meat and Livestock Commission) and by the Ministry of Agriculture, Fisheries and Food (MAFF), the Department of Trade and Industry, and the Biotechnology and Biological Sciences Research Council (BBSRC). C.S.H. and R.T. acknowledge support from MAFF, BBSRC and the European Commission.

LITERATURE CITED

ANDERSSON, L., C. S. HALEY, H. ELLEGREN, S. A. KNOTT, M. JOHANSSON *et al.*, 1994 Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science **263:** 1771–1774.

BULMER, M. G., 1971 The effect of selection on genetic variability. Am. Nat. **108:** 45–58.

DEKKERS, J. C. M., and M. R. DENTINE, 1991 Quantitative genetic variance associated with chromosomal markers in segregating populations. Theoret. Appl. Genet. **81:** 212–220.

GIMELFARB, A., and R. LANDE, 1994 Simulation of marker assisted selection in hybrid populations. Genet. Res. **63:** 39–47.

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.

GROEN, A. F., and M. M. J. TIMMERMANS, 1992 The use of genetic markers to increase the efficiency of introgression—a simulation study. Proceedings of the XIX World's Poultry Congress **1:** 523–527.

HALDANE, J. B. S., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. **8:** 299–309.

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

HAZEL, L. N., 1943 The genetic basis for constructing selection indexes. Genetics **28:** 476–490.

HILL, W. G., 1993 Variation in genetic composition in backcrossing programs. J. Hered. **84:** 212–213.

HILLEL, J., T. SCHAAP, A. HABERFIELD, A. J. JEFFREYS, Y. PLOTZKY *et al.*, 1990 DNA fingerprints applied to gene introgression in breeding programs. Genetics **124:** 783–789.

HILLEL, J., M. VERRINDER GIBBINS, R. J. ETCHES, 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.

KNOTT, S. A., and C. S. HALEY 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. **60:** 139–151.

LANDE, R., and R. THOMPSON, 1990 Efficiency of marker assisted selection in the improvement of quantitative traits. Genetics **124:** 743–756.

MATHER, K., and J. L. JINKS, 1971 *Biometrical Genetics.* Chapman and Hall Ltd., London.

PATERSON, A. H., E. S. LANDER, D. HEWITT, S. PETERSON, S. E. LINCOLN *et al.*, 1988 Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature **335:** 721–726.

ROBERTSON, A., 1977 Artificial selection with a large number of linked loci, pp. 307–322 in *Proceedings of the International Conference on Quantitative Genetics.* Iowa State University Press, Ames, IA.

ROTHSCHILD, M. F., C. JACOBSON, D. A. VASKE, C. K. TUGGLE, T. H. SHORT *et al.*, 1994 A major gene for litter size in pigs. Proceedings of the 5th World Congress on Genetics Applied to Livestock Production **21:** 225–228.

SMITH, C., T. H .E. MEUWISSEN and J. P. GIBSON, 1987 On the use of transgenes in livestock improvement. Anim. Breed. Abstr. **55:** 1–10.

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.

STUBER, C. W., S. E. LINCOLN, D. W. WOLFF, T. HELENTJARIS and E. S. LANDER, 1992 Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics **132:** 823–839.

VISSCHER, P. M., 1996 Proportion of the variation in genetic compo-

sition in backcrossing programs explained by genetic markers. J. Hered. **87:** 136–138.

VISSCHER, P. M., and C. S. HALEY, 1996 Detection of putative quantitative trait loci in line crosses under infinitesimal genetic models. Theoret. Appl. Genet. (in press).

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

ZHANG, W., and C. SMITH, 1992 The use of marker assisted selection with linkage disequilibrium. Theoret. Appl. Genet. **83:** 813–820.

Communicating editor: B. S. WEIR

APPENDIX A

Model for many linked QTL: The aim of this appendix is to present a genetic model of many linked loci that explains breed differences and genetic variances in backcross populations derived from inbred lines.

Suppose we have n equally spaced loci on a chromosome (or interval) of length L. If alternative alleles are fixed in the two breeds for all loci, and all loci have equal additive effects (allele substitution effect α), then

$$D = \text{trait difference for parental breeds} = 2n\alpha,$$

$$\text{or } \alpha = D/(2n).$$

For backcross (BC) generation t, the total additive genetic variance is

$$\begin{aligned}\text{var}_A\,(BC_t) &= \sum_{i=1}^{n}\sum_{j=1}^{n}[(\tfrac{1}{2}^t)(1 - r_{ij})^t\alpha^2] - n^2(\tfrac{1}{4}^t)\alpha^2 \\ &= \alpha^2 n^2[(\tfrac{1}{2}^t)\sum_{t=1}^{n}\sum_{j=1}^{n}(1 - r_{ij})^t(1/n)(1/n) - (\tfrac{1}{4}^t)] \\ &= \left(\frac{D^2}{4}\right)[(\tfrac{1}{2}^t)\sum_{i=1}^{n}\sum_{j=1}^{n}(1 - r_{ij})^t(1/n)(1/n) - (\tfrac{1}{4}^t)], \quad \text{(A1)}\end{aligned}$$

with r_{ij} the recombination fraction between loci i and j. The term inside the square brackets is similar to the equations of STAM and ZEVEN (1981) and HILL (1993) for the variance of the proportion of the genome from the recurrent population. Essentially the genetic variance under the assumed model is just a scaled version of the variance of the proportion of the genome from one of the inbred lines. For large n, and assuming HALDANE's (1919) mapping function without interference, (A1) can be approximated by (see also HILL 1993)

$$\text{var}_A\,(BC_t) = \frac{D^2}{4}\left[1/(2L^2)(\tfrac{1}{4}^t)\sum_{i=1}^{t}\binom{t}{t}1/(i^2)(2iL - 1 + e^{-2iL})\right], \quad \text{(A2)}$$

with L the length of the chromosome (block) in Morgans. For the first BC generation, (A2) may be written as

$$\text{var}_A\,(BC_t) = (\tfrac{1}{4})(D^2/4)(1 - r_m/L)/L,$$

with r_m the recombination fraction between the chromosome ends. Extension to multiple chromosomes is straightforward (HILL 1993). With the "standard" infinitesimal model (BULMER 1971) the genotypic value at any locus is of order (O) $1/n^{1/2}$, and the variance at each locus is $O(1/n)$. With a very large number of linked loci on a chromosome (block), the variance in backcross populations is of $O(1/n^2)$, so values at individual loci should be of order $(1/n)$. However, we cannot both have finite genic and total variance for infinite n (the genic variance goes to zero for large n with finite genetic variance).

APPENDIX B

Optimum marker distance for introgression: Given that the position of a QTL is estimated with error, which available markers should we choose in introgression programs? If we take the markers too close to the estimated position of the QTL, we may lose the desired allele because the QTL is actually outside the bracket we are selecting. If we take the markers too far, we may lose the desired QTL allele because of double recombinants.

Assumptions: Given an estimate of a position of a QTL of μ Morgan on a chromosome of length L Morgan, we assume that the true position of the QTL is normally distributed about μ with a standard error of prediction of σ. Hence we do not know the true position of the QTL, but on average the estimate of the position is correct. (Note that the error variance is that of the true position given the estimated position, and not the other way around.)

We are interested in the probability of having the true QTL in backcross generation t, given one or two markers around the estimated position that originate from the same breed. Hence, if we have a cross between two inbred lines with QTL_1 and M_1 from the donor line and QTL_2 and M_2 from the recipient line. We wish to calculate

$$\text{Prob}\,(QTL_1|M_1) \quad \text{and} \quad \text{Prob}\,(QTL_1|M_1M_1).$$

Haldane's mapping function without interference is assumed throughout.

Single marker: Suppose we only have a single marker on a chromosome of length L. The position of the marker and the estimated position of the QTL are y and μ, respectively. Let $f_t(x)$ be the probability of no recombination between the true QTL position and the marker at generation t, and $g(x)$ the normal density function with mean μ and standard deviation σ. Then

$$P_t = \text{Prob}_t\,(QTL_1|M_1) = \int_0^L f_t(x)g(x)\,dx, \quad \text{(B1)}$$

with $f_t(x) = (1 - r(|y - x|))^t = (\tfrac{1}{2})^t\,(1 + e^{-2|y-x|})^t$ and

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{[-1/2(x-\mu)^2]/\sigma^2}.$$

Integrating (B1) gives

$$P_t = \int_0^y f_t(x)g(x)\,dx + \int_y^L f_t(x)g(x)\,dx = \int_0^y (\tfrac{1}{2})^t \times (1 + e^{-2(y-x)})^t \frac{1}{\sqrt{2\pi\sigma^2}} e^{[-(x-\mu)^2]/2\sigma^2} dx + \int_y^L (\tfrac{1}{2})^t (1 + e^{-2(x-y)})^t \frac{1}{\sqrt{2\pi\sigma^2}} e^{[-(x-\mu)^2]/2\sigma^2} dx$$

$$= (\tfrac{1}{2})^t \sum_{i=0}^{t} \binom{t}{i} \Big[e^{2i(\mu-y+i\sigma^2)}[\Phi(d_1/\sigma + 2i\sigma) - \Phi((\mu - y)/\sigma + 2i\sigma)] + e^{2i(y-\mu+i\sigma^2)}[\Phi(d_2/\sigma + 2i\sigma) - 2\Phi((y - \mu)/\sigma + 2i\sigma)] \Big], \quad \text{(B2)}$$

where d_1 is the distance between the "start" of the chromosome and the estimated QTL position (μ Morgans); d_2 is the distance between the "end" of the chromosome and the estimated QTL position ($L - \mu$ Morgans); and $\Phi(x)$ is the cumulative normal density function. If the marker is at the estimated QTL position, *i.e.*, $y = \mu$, then Equation B2 simplifies to

$$P_t = (\tfrac{1}{2})^t \sum_{i=0}^{t} \binom{t}{i} e^{2(i\sigma)^2} \Big[\Phi(d_1/\sigma + 2i\sigma) + \Phi(d_2/\sigma + 2i\sigma) - 2\Phi(2i\sigma) \Big]. \quad \text{(B3)}$$

If the estimate of the QTL is near the end of a chromosome, the optimum position of the marker, which gives the largest value of P_t can be determined using Equation B2. However, only for large values of σ does the best position of a single marker differ from the estimated position. For example, if the estimate of the QTL position is at 0 cM on a chromosome of length 100 cM, *i.e.*, $\mu = 0$, the best marker position for $\sigma = 1, 2, 3, 4, 5$ is 0.7, 1.3, 2.0, 2.7, and 3.2 cM, respectively. These values were calculated using (B2) assuming a truncated normal density function at the estimated position of the QTL. Hence there is a possibility that the QTL is not on the chromosome. If the estimated position of the QTL is elsewhere on the chromosome, then the optimum position of the marker is at the estimated position.

Marker bracket: If we select on marker bracket M_1M_1, there are three possibilities with respect to the true QTL position, given that the estimated QTL position is within the marker bracket and given that marker bracket M_1M_1 is selected: (1) The true QTL is outside the bracket and is "to the left" of the first marker. (2) The true QTL is inside the bracket. (3) The true QTL is outside the bracket and is "to the right" of the second marker. The corresponding probabilities are termed $P_{1(t)}$, $P_{2(t)}$, and $P_{3(t)}$, respectively. In total, the probability that the true QTL is selected given selection on marker bracket M_1M_1 is $P_t^* = P_{1(t)} + P_{2(t)} + P_{3(t)}$. In practice, $P_{2(t)}$ may be the most important probability, because the chromosome segments outside the marker brackets may be under selection.

$P_{1(t)}$ and $P_{3(t)}$ are calculated using Equation B2,

$$P_{k(t)} = (\tfrac{1}{2})^t \sum_{i=0}^{t} \binom{t}{i} e^{2(id_3 + (i\sigma)^2)} [\Phi(d_4/\sigma + 2i\sigma) - \Phi(d_3/\sigma + 2i\sigma)], \quad k = 1,3, \quad \text{(B4)}$$

where d_3 is the absolute value of the distance between the marker and the estimated QTL position (in Morgans) and d_4 is the distance from the estimated QTL position to the "start" ($P_{1(t)}$) or "end" ($P_{3(t)}$) of the chromosome.

For $P_{2(t)}$, the positions of the markers (relative to the start of the chromosome) are y and z, respectively. The function $h_t(x)$ is the probability of no double recombination between the position of the true QTL (x), and the flanking markers. Then,

$$P_{2(t)} = \int_y^z h_t(x)g(x)\,dx,$$

with

$$h_t(x) = (1 - r_1(x - y))^t(1 - r_2(z - x))^t/(1 - r_m)^t = (\tfrac{1}{4})^t (1 + e^{-2(x-y)})^t(1 + e^{-2(z-x)})^t/(1 - r_m)^t,$$

where r_1 is the recombination fraction between the true QTL position and the first marker position; r_2 is the recombination fraction between the true QTL position and the second marker position; and r_m is the recombination fraction between the two markers (distance = $z - y$).

After some tedious algebra, it can be shown that

$$P_{2(t)} = \frac{(\tfrac{1}{4})^t}{(1 - r_m)^t} \sum_{i=0}^{t} \sum_{j=0}^{t} \binom{t}{i}\binom{t}{j} c_{ij}[\Phi(d_5/\sigma + 2(j - i)\sigma) + \Phi(d_6/\sigma - 2(j - i)\sigma) - 1], \quad \text{(B5)}$$

where $c_{ij} = \exp\{2[j - i)\sigma]^2 - 2(id_5 + jd_6)\}$; d_5 is the distance between the first marker and the estimated QTL position, *i.e.*, $d_5 = \mu - y$; and d_6 is the distance between the second marker and the estimated QTL position, *i.e.*, $d_6 = z - \mu$.