Animal Science 1999, 68: 59-68

1357-7298/99/7223059$20·00

On the efficiency of marker-assisted introgression

P. M. Visscher[1] and C. S. Haley[2]

[1]*Institute of Ecology and Resource Management, University of Edinburgh, West Mains Road, Edinburgh EH9 3JG*
[2]*Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS*

Abstract

The efficiency of marker-assisted introgression programmes, expressed as genetic lag relative to a commercial population under continuous selection, was investigated using analytical methods. A genetic model was assumed for which the genetic variance in the introgression population was a function of the within-breed genetic variance and the initial breed difference. It was found that most of the genetic lag occurs in the latter stages of an introgression programme, when males and females which are heterozygous for the allele to be introgressed are mated to produce homozygous individuals. Reducing genetic lag through selection on genomic proportion by using genetic markers throughout the genome, i.e. *by selecting heterozygous individuals which resemble the recipient (commercial) population most, was effective if the initial breed difference was very large (e.g. 20 within-breed phenotypic standard deviations). In that case, selection solely on genetic markers could be practised to speed up genome recovery of the commercial line. If the initial breed difference is small, phenotypic or best linear unbiased prediction (BLUP) selection is superior in reducing genetic lag under the assumed genetic model.*

Keywords: *genetic markers, introgression, pigs, selection.*

Introduction

Marker-assisted introgression using backcrossing is an efficient way to incorporate a desired allele from a donor population into a commercial (elite) population (e.g., Hospital *et al.*, 1992). Markers are used to keep track of the allele which is introgressed and to select against the remainder of the donor genome, which is usually associated with inferior performance for loci other than that which is introgressed. After a number of generations of backcrossing, the final backcross population is intercrossed to create individuals which are homozygous for the desired allele. Relative to a continuously selected nucleus population, the population with the desired alleles is superior at that locus but inferior with respect to other loci. This is because (i) at each backcross generation fewer individuals can be selected for overall genetic merit because of a pre-selection for the allele to be introgressed, (ii) there may be a genetic lag due to using older individuals for breeding in the backcross population, and (iii) the initial difference between the donor and recipient population in overall economic merit may have been large. These points were clearly described by Gama *et al.* (1992), who investigated the introgression of a transgene into a nucleus pig population. Gama *et al.* (1992) used markers only to identify the transgene and did not use markers to select against the 'background genotype', i.e. the remainder of the donor genotype at each generation.

The genetic efficiency of the introgression process depends on the genetic value of the final commercial product relative to nucleus populations that are under continued selection and the extra costs associated with the introgression. Hospital *et al.* (1992) and Visscher *et al.* (1996) investigated the relative genetic lag in the case of inbred lines, while Gama *et al.* (1992) looked at the lag using outbred lines (pig populations) when selection was on phenotypes. The aim of this study is to expand on those studies by investigating the parameters that drive the genetic efficiency of gene introgression programmes, and to explore the reduction of genetic lag through selection on markers that identify the founder breed's genomes. New in our approach is (i) the formulation of a genetic model which associates the initial breed difference with polygenic genetic variation during a backcrossing programme, (ii) the

Table 1 *Introgression of allele Q from a homozygous QQ donor line (line A) into a homozygous qq recipient line (B).* t *is the generation of crossing*

t	Introgression population	Parental cross	Genotype at trait locus			Selected genotype
1	F_1	$B \times A$	Qq			Qq
2	BC_1	$B \times F_1$	Qq	qq		Qq
3	BC_2	$B \times BC_1$	Qq	qq		Qq
T	BC_{T-1}	$B \times BC_{T-2}$	Qq	qq		Qq
$T+1$	G_{T+1}	$BC_{T-1} \times BC_{T-1}$	Qq	qq	QQ	QQ
$T+2$	G_{T+2}	$G_{T+1} \times G_{T+1}$	QQ			QQ

derivation of simple expressions to calculate genetic lag of an introgression programme, and (iii) the investigation of novel introgression breeding schemes in which selection is based solely on genetic markers.

Material and methods

Assumptions and notation

Gene introgression programmes in plants and animals are usually initiated to improve the genetic mean for one particular trait, such as disease resistance or reproductive performance. We assume in this study that the aim is to maximize profit, in that we wish to maximize the efficiency of genetic value for overall economic merit. This is also in line with current practice in pig breeding populations, where genes for quantitative traits such as litter size are being introgressed (Rothschild *et al.*, 1996). To simplify matters, we assume that economic efficiency is proportional to genetic efficiency, although we discuss the economic aspects of introgression programmes in the **Discussion.**

We are interested in the genetic merit after $t = (T + 1)$ discrete generations, where t is the number of generations of crossing, so that $t = 1$ is the first crossbred generation (the F_1 individuals), $t = 2$ is the first backcross generation, and $t = T$ is the generation in which the backcross individuals are intercrossed to make the desired allele homozygous in generation $\{T + 1\}$. See Table 1 for a diagram summarizing the introgression programme.

The mean for overall genetic merit of the donor and recipient population at the start of the introgression programme are (2α) and D, respectively. Hence, the allele substitution effect of the desired gene is (α), whereas the difference in background genotype between the donor and recipient population is D.

The commercial (elite) population is assumed to have a genetic gain of δ units per generation, which is achieved by selecting the best males (m) and females (f), i.e. $\delta = \frac{1}{2}[\delta_m + \delta_f]$. At each (discrete) generation, males (or individuals of the sex with the largest reproductive rate) of the commercial populations are mated with females from the backcross population. Thus, if the first selection is at generation 1, then the total genetic gain of nucleus animals born in generation t is, $\frac{1}{2}(t - 1)(\delta_m + \delta_f) = (t - 1)\delta$.

Throughout this study, we assume that males from the commercial population are mated with backcross females during the backcrossing phase of the introgression programme. However, the alternative approach, i.e. using crossbred males and commercial females, may have practical advantages, and we will return to this approach in the **Discussion.**

At generation t, the selection response in the backcross population is δ_t. Note that we assume that response to selection in the backcross population is generation dependent but that response in the nucleus population is constant, i.e. it does not depend on t. A summary of the relative genetic value in nucleus and introgression population, apart from

Table 2 *Relative genetic lag of individuals in the nucleus and backcross population during the backcrossing phase.* t *is the generation of crossing. Selected males from the nucleus population are used both in the nucleus and backcross population. Selection of females in the backcross population is ignored*

	Progeny born in		Selected animals in nucleus	
t	Nucleus	Backcross	Females	Males
1	0	0	δ_f	δ_m
2	$\frac{1}{2}(\delta_m + \delta_f)$	$\frac{1}{2}\delta_m$	$\frac{1}{2}\delta_m + 3/2\,\delta_f$	$3/2\,\delta_m + \frac{1}{2}\delta_f$
3	$(\delta_m + \delta_f)$	$\delta_m + \frac{1}{4}\delta_f$	$\delta_m + 2\delta_f$	$2\delta_m + \delta_f$
4	$3/2(\delta_m + \delta_f)$	$3/2\delta_m + 5/8\delta_f$	$3/2\delta_m + 5/2\,\delta_f$	$5/2\delta_m + 3/2\delta_f$
...				
k	$\frac{1}{2}(k-1)(\delta_m + \delta_f)$	$\frac{1}{2}(k-1)\delta_m + [\frac{1}{2}(k-3) + \frac{1}{2}^{k-1}]\delta_f$	$\frac{1}{2}(k-1)\delta_m + \frac{1}{2}\,(k+1)\delta_f$	$\frac{1}{2}(k+1)\delta_m + \frac{1}{2}\,(k-1)\delta_f$
Lag at $t = k$		$[1 - \frac{1}{2}^{k-1}]\delta_f$		

Table 3 *Genetic merit of individuals in the nucleus and introgression population during the intercrossing phase*

	Progeny born in		Selected individuals in introgression population	
t	Nucleus	Backcross	Females	Males
T	$\frac{1}{2}(T-1)(\delta_m + \delta_f)$	$\frac{1}{2}(T-1)\delta_m + [\frac{1}{2}(T-3) + \frac{1}{2}^{T-1}]\delta_f$	δ_{fT}	δ_{mT}
$T+1$	$\frac{1}{2}T(\delta_m + \delta_f)$	$\frac{1}{2}(T-1)\delta_m + [\frac{1}{2}(T-3) + \frac{1}{2}^{T-1}]\delta_f + \delta_T$	$\delta_{f(T+1)}$	$\delta_{m(T+1)}$
$T+2$	$\frac{1}{2}(T+1)(\delta_m + \delta_f)$	$\frac{1}{2}(T-1)\delta_m + [\frac{1}{2}(T-3) + \frac{1}{2}^{T-1}]\delta_f + \delta_T + \delta_{(T+1)}$		
Lag at $t = T+2$		$(\delta_m + \delta_f) + [1 - \frac{1}{2}^{T-1}]\delta_f - \delta_T - \delta_{(T+1)}$		

$\delta_T = \frac{1}{2}[\delta_{fT} + \delta_{mT}]$
$\delta_{T+1} = \frac{1}{2}[\delta_{f(T+1)} + \delta_{m(T+1)}]$

the effect of genetic background (D) and the allele substitution effect of the allele to be introgressed (α), is given in Tables 2 and 3.

We refer to the population of crossbred individuals during the backcrossing phase as the 'backcross population' and to the populations in the later phases of the introgression programme as the 'intercross population'. When either of these populations is referred to, we use the term 'introgression population'.

Genetic value during introgression

In the above described simple scheme, it can be shown (following Gama *et al.*, 1992) that the genetic mean for genetic merit of progeny born at generation t relative to the genetic merit of the nucleus population, i.e. the difference between the introgression population and commercial population, is

$$\Delta_t = (\alpha) - (\tfrac{1}{2})^t D + \sum_{i=1}^{t-1} (\tfrac{1}{2})^i \delta_{t-i} - (1 - (\tfrac{1}{2})^{t-1})\delta_f \quad (1).$$

Equation (1) shows that the difference in genetic merit between the introgress and nucleus population is made up of three parts: the effect of the allele which is introgressed, the remainder of the donor genome, and the difference in selection response at each generation between the two populations. Since, without loss of generality, we have assumed that males from the nucleus are used in the introgression population, the latter part is the difference in response between females in the nucleus and females in the introgression population. If we could assume that the response to selection in the introgression population was equal in each generation, i.e. $\delta_i = \delta_j = \delta_{cb}$, (as was assumed for the nucleus population), then

$$\Delta_t = (\alpha) - (\tfrac{1}{2})^t D - (1 - \tfrac{1}{2}^{t-1})(\delta_f - \delta_{cb}) \quad (2).$$

From equation (2) it can be seen that if the initial line difference D is large relative to the response in the nucleus population, it largely determines the difference in genetic mean between the introgress and nucleus population in the early generations of the introgression programme, even after four or five generations of backcrossing.

If we further assume that t is large ($t \to \infty$), then

$$\Delta_\infty = (\alpha) - (\delta_f - \delta_{cb}) \quad (3).$$

If the alternative design is used in which backcross males are mated with commercial females, the genetic lag is, $\Delta_\infty = (\alpha) - (\delta_m - \delta_{cb})$.

After selection in generation T, males and females from the introgression population are mated *inter se*. Hence, males and females from backcross generation T who carry one copy of the desired allele are selected and mated. Their progeny in generation {T + 1} that are homozygous for the desired allele are selected to become the founder parents for the new commercial line. The total response for genetic merit of the individuals born at generation ($T + 2$), is easily determined using equation (1):

$$\Delta_{T+2} = (2\alpha) - (\tfrac{1}{2})^T D + \sum_{i=1}^{T-1} (\tfrac{1}{2})^i \delta_{T-i}$$

$$-(1 - (\tfrac{1}{2})^{T-1})\delta_f - (2\delta - \delta_T - \delta_{T+1}) \quad (4).$$

Relative to equation (1), there is an additional term, which describes the difference in response to selection between the nucleus population and the final generations of intercrossing. Hence, the total difference in response at generation $\{T + 2\}$ comprises the difference (i) of backcross progeny born at generation T, (ii) of selected parents at generation T, and (iii) of selected parents at generation $\{T + 1\}$. The selection intensities in the introgression population are different in each of these time periods because only one sex is selected during introgression and both sexes are selected during intercrossing. Furthermore, the proportion of individuals preselected for the desired alleles is different in generation T and $\{T + 1\}$, since half of the individuals are heterozygotes in generation T, and a quarter of individuals are homozygous in generation $\{T + 1\}$. Equation (4) clearly shows that, in addition to the remainder of the donor genome, most of the genetic lag occurs during the intercrossing phase.

Making the same simplifications as before (responses for the introgression population constant across generations, a large number of generations, same selection differential of males and females in the nucleus), gives a simple form for the relative genetic mean of the final new commercial population,

$$\Delta_\infty = (2\alpha) - 3(\delta - \delta_{cb}) \qquad (5).$$

Although this is somewhat unrealistic, since the response in the introgression population will not be equal during the three phases, it gives a rough indication of how large the effect of the introgressed allele should be before it is worthwhile to embark on an introgression programme. For example, in a pig nucleus breeding programme with a theoretical annual response to selection of about 2 to 3% of the mean performance (Smith, 1984), the substitution effect (α) of the favourable allele should be at least 3 to 4·5 % of mean performance, assuming no selection during the introgression phase ($\delta_{cb} = 0$). These figures are in agreement with those reported by Gama *et al.* (1992).

A further result from these equations is that we can compare the value of the introgression population under random selection *v.* continued selection at each generation. Using equation (4), and assuming that the response during backcrossing is constant per generation (δ_{cb}) gives,

$$\Delta_{T+2}(\text{selection}) - \Delta_{T+2}(\text{random selection}) = (1 - \tfrac{1}{2}^{T-1})\delta_{cb} + \delta_T + \delta_{T+1}$$

which reduces to $\{\delta_{cb} + \delta_T + \delta_{T+1}\}$ when T is large. Hence, compared with random selection, one round of selection is gained during backcrossing and two rounds of selection during the intercrossing phase.

Selection responses at each generation

The main difference in genetic mean between the nucleus and introgression population resulted from different responses to selection between these populations at each generation. The responses differ because of the following.

(i) In the backcross population, a pre-selection is made on those individuals that have received one copy of the desired allele from their crossbred parent. This pre-selection will, on average, exclude half of the individuals available for selection. At the generation of intercrossing, both sexes are selected rather than one sex only during the backcrossing phase. In the final generation of selection $(T + 1)$, preselection on markers linked to the alleles to be introgressed will exclude 3/4 of the population, since only 1/4 will be homozygous for the desired alleles. Hence, overall the intensity of selection will be lower in the introgression population.

(ii) Depending on the assumptions regarding the distribution and frequency of loci influencing the trait of interest within each of the founder populations, the heritabilities may differ between the commercial and introgression populations.

(iii) Individuals in the nucleus population are likely to be selected based on phenotypic observations on themselves and their relatives. However, in the introgression population it is possible to select on genomic proportion using genetic markers, i.e. select those individuals that are most similar to individuals from the recipient population. Hence, the accuracy of selection can be different between the two populations.

In the nucleus population, simple phenotypic selection is assumed,

$$\delta = \tfrac{1}{2}(\delta_m + \delta_f) = \tfrac{1}{2}(i_m h^2 \sigma_p + i_f h^2 \sigma_p) \qquad (6).$$

Genetic and environmental variances within the commercial population are defined as var_w and var_e, so that $h^2 = var_w/(var_w + var_e)$. We assume that var_w is constant, so that we do not have to take into account the reduction in genetic variance due to selection. A justification for this assumption may be that the commercial population already has reached an equilibrium genetic variance.

For the backcross population, we assume that the genetic variance can be partitioned into a variance

due to differences between the two founder breeds (var_b), and genetic variation within the breeds (var_w). For simplicity we assume that the within breed variance is the same for both breeds, although a weighted average is easily accommodated. The variance within the breeds is assumed to be constant, and as defined above. At each generation,

$$var_{a(t)} = var_{b(t)} + var_w \quad (7).$$

We assume that the variance due to variation between the breeds is a function of the original breed difference (D), following a genetic model with many linked loci with the breeds fixed for alternative alleles (Visscher and Haley, 1996). This model assumes that the genetic variance due to the difference between the two lines is proportional to the variance in genomic proportion (Hill, 1993), i.e.

$$var_{b(t)} = D^2 var_{gp(t)} \quad (8)$$

with $var_{gp(t)}$ the variance in genomic proportion at generation t of crossing (Hill, 1993). The heritability in the backcross population is defined as,

$$H_t^2 = var_{a(t)}/(var_{a(t)} + var_e) \quad (9).$$

This heritability will usually be larger than the heritability within the commercial population. For large t, or an initial breed difference of D = 0, the two heritabilities will be equal.

The proportion of genetic variance in the backcross population which is caused by between-line differences, is defined as,

$$c_t^2 = var_{b(t)}/var_{a(t)}.$$

In the extreme case of inbred lines (var_w = 0), this proportion becomes unity. If D = 0, it reduces to zero.

For the introgression population, we assume that selection is on genomic proportion (Visscher *et al.*, 1996; Visscher, 1996) based upon scoring m markers per chromosome. The reason for this is that we are interested in recovering the recipient genome as soon as possible, and wish to discard individuals not needed for breeding, i.e. those individuals which do not carry the desired allele and those with a large proportion of the donor genome. In this way, selection (on marker genotypes) can be done early in life, and phenotypes do not have to be collected for the individuals in the introgression population. The response to selection on markers is,

$$\delta_t = i_{cb}\, c_t\, r_t(m)\, \sigma_{a(t)} \quad (10).$$

with i_{cb} = the selection intensity of the sex in the backcross population which is used for introgression; $r_t(m)$ = the square root of the proportion of variance in genomic proportion at generation t which is explained by m markers per chromosome (Visscher, 1996); $\sigma_{a(t)}$ = the genetic standard deviation at backcross generation t.

Hence, we assume that the accuracy of selection (i.e. the correlation between the breeding value and the selection criterion) is, $\rho_t = c_t\, r_t(m)$. We have ignored the linkage drag around the allele to be introgressed. However, if there are many chromosomes, this effect can be ignored, because the genomic proportion over all chromosomes will be close to the average proportion from the non-carrier chromosomes (see Hospital *et al.*, 1992). In addition, in practice the allele which is introgressed may be a quantitative trait locus (QTL) allele, whose effect typically spans a region of 20 to 50 centi Morgan (cM), approximately coinciding with the size of the donor segment which remains after five generations of backcrossing (Stam and Zeven, 1981). Finally, we have assumed that the variance in genomic proportion in intercross generation T is the same as in backcross generation t ($t < T$). This is only an approximation, since the variance in genomic proportion for intercross populations is initially larger (i.e., an F_2 population relative to a first backcross population) but declines more rapidly due to recombination (Visscher and Haley, 1996). However, the variance in genomic proportion is small after a few generations of crossing (Hill, 1993), so that this assumption is unlikely to change the results significantly.

To make the comparison with phenotypic selection, we also calculate the gain in the introgression population under mass selection, using equation (6) but with the appropriate selection intensities and heritabilities for the introgression population.

Using equations (4), (6), and (10), the difference in genetic gain between the population with the introgressed allele(s) and the commercial population can be determined.

Results

Inbred lines

A special case was investigated for which the genetic variance within lines was zero, i.e. var_w = 0. The genetic gain in the 'nucleus' (this is more relevant to plant breeders than animal breeders) or elite population is obviously zero, and markers in the backcross population are used to recover the recipient genome as quickly as possible. This is the

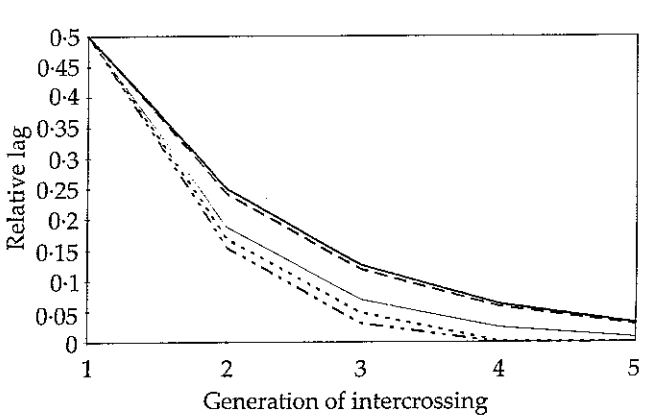

Figure 1 Relative genetic lag (proportion of donor genome remaining) in an introgression programme from an inbred line cross, for random selection, phenotypic selection, and marker selection: —— random; — —phenotypic (D=2); - - 1 marker per chromosome; ——·--5 markers per chromosome; ——— phenotypic (D=20).

scenario of Hospital *et al.* (1992) and Visscher *et al.* (1996).

The genetic lag in the backcross population when dealing with inbred lines is proportional to the line difference D when selection is on markers, so that results can be expressed as (Δ_{T+2}/D). This ratio, called the relative lag, which is equivalent to the proportion of the donor genome remaining, was plotted against the generation of intercrossing in Figure 1, for the case of random selection, selection on phenotypes in the introgression population, and for selection based on one or five markers per chromosome. Results are for 20 chromosomes, and a proportion selected of 25% for females (during backcrossing), and 2·5% for males (during intercrossing). For phenotypic selection, genetic lag was not proportional to the line difference D, because the heritability is non-linear in D. Therefore, results for phenotypic selection were presented for two extreme values of D, 2·0 and 20·0 respectively.

The conclusions are essentially the same as from Hospital *et al.* (1992) and Visscher *et al.* (1996): if the aim is to reduce lag to essentially zero, i.e. to recover the recipient genome almost entirely, phenotypic selection is about one generation of crossing faster than random selection for a large line difference (D = 20), and using markers speeds the process up by a further generation.

Outbred lines

For the case of outbred lines, i.e. the recipient line is under continued selection pressure, we express the lag in units of generations of genetic gain in the nucleus. This way, the results have an intuitive meaning and can be compared directly across species. Since in the case of outbred lines the results are not independent of (or proportional to) the initial line difference, we look at a range of values for D.

In Table 4, the lag is shown for two extreme values of D (2·0 and 20·0), and for each initial line difference, the lag is shown for the case of random selection in the introgression population, phenotypic selection, and selection on markers. For the latter case, five markers per chromosome were chosen, although results are very similar for two or more markers per chromosome (results not shown). The heritability for the trait under selection in the nucleus population was assumed to be 0·25, and proportions of 0·025 males and 0·25 of females are used to replace the population each generation.

For the case of D = 20, i.e. a very large initial line difference, phenotypic selection and marker selection perform very similar until generation of intercrossing 5. After that, markers do not explain much of the phenotypic variance, and phenotypic selection is more efficient. The lag with phenotypic

Table 4 *Genetic lag, in number of generations of gain in the nucleus population, for progeny born at generation* T+2, *when intercrossing to make the desired allele homozygous is performed at generation* T. *Initial breed difference is 2 or 20 phenotypic standard deviations,* h^2 *in nucleus is 0·25*

		T									
D	Selection	1	2	3	4	5	6	7	8	9	10
2	None	4·2	3·5	3·1	2·9	2·8	2·8	2·7	2·7	2·7	2·7
	Phenotypic	2·9	1·9	1·4	1·2	1·1	1·0	1·0	1·0	1·0	1·0
	Markers	4·2	3·0	2·7	2·6	2·5	2·6	2·6	2·6	2·6	2·7
20	None	24·2	13·4	8·1	5·4	4·0	3·4	3·0	2·9	2·8	2·7
	Phenotypic	22·9	9·5	4·3	2·2	1·4	1·1	1·0	1·0	1·0	1·0
	Markers	24·2	9·2	3·9	2·0	1·5	1·5	1·7	2·0	2·2	2·3

selection keeps reducing until an asymptote is reached, whereas the lag with marker selection reaches a minimum and then increases again until it also reaches an asymptote. The phenotypic selection asymptotes to a lag of 1·0 generations, while the marker selection case asymptotes to 2·7 generations of genetic gain at generation 20 (results for later generations are not presented). The lag is minimum at $T = 5$ (1·5 generations), and then increases again because there is very little between line variance left, and the nucleus is under continuous selection. As expected, the random selection case also exhibits an asymptotic lag of 2·7 generations of gain.

For D = 2, results are very different. Genetic markers explain very little from the beginning and marker selection is not much better than random selection. Both exhibit an asymptotic lag of 2·7 generations of genetic gain.

Since phenotypic and marker selection perform differently depending on when the intercross phase commences for the case of D = 20, the optimum point of switching over from marker selection to phenotypic selection was investigated for D = 2·0 and D = 20·0. The lag at generation $T = 4$ and $T = 6$ of intercrossing was compared while varying the generation of switching selection from $t = 1$ (phenotypic selection throughout) to $t = T + 1$ (marker selection throughout). Results are shown in Table 5. So, for example, at generation of switching selection $t = 2$, selection in the F_1 generation was on markers (no response assumed), and at the first backcross generation ($t = 2$) and beyond, selection was on phenotypes. The lag under random selection is shown for comparison. For $T = 4$, this lag is 5·39 (D = 20) and 2·89 (D = 2) generations of genetic gain. For $T = 6$, the values are 3·38 and 2·75, respectively.

Under our assumed model and parameters, the optimum generation for changing selection strategies is very flat (Table 5). For a large initial line difference (D = 20), selection solely on markers is just as good as phenotypic selection when considering the lag for $T = 4$, while for $T = 6$, phenotypic selection should be practised in at least the last generation of crossing. For a smaller line difference (D = 2), phenotypic selection should be practised early on, although not much gain is lost if the switch from marker to phenotypic selection is made before generation 4 of crossing. In practice, taking costs into account, the lag at generation of intercrossing of 4 ($T = 4$) may well be the most relevant. In that case, phenotypic information does not appear necessary under our model, at least for a large initial line difference. Hence, in that case, if marker costs were sufficiently low, selection solely on markers could be practised.

Table 5 *Genetic lag (in number of generations of genetic gain in the nucleus) at generation of intercrossing of 4 (LAG4) and 6 (LAG6) when the selection criterion changes from selection on markers to phenotypic selection at generation of crossing of t. The lag is the difference in polygenic breeding value between animals born in the nucleus and introgression population at generation 6 and 8, respectively. Selection proportions were 0·025 for males and 0·25 for females*

	D† = 2		D† = 20	
t	LAG4	LAG6	LAG4	LAG6
1	1·21	1·04	2·22	1·14
2	1·27	1·05	2·27	1·16
3	1·34	1·07	2·23	1·14
4	1·51	1·11	2·16	1·13
5	2·55	1·20	1·96	1·11
6		1·40		1·13
7		2·57		1·52
random selection	2·89	2·75	5·39	3·38

† Initial breed difference (in phenotypic standard deviations in the nucleus population).

Discussion

It was shown that the polygenic lag in a population used for introgressing an allele into a nucleus population was caused partly during the backcrossing phase and partly during the intercrossing phase. The latter part is more important because during intercrossing a larger proportion of individuals have to be pre-selected for their marker genotype for the allele to be introgressed and superior genes from the nucleus population are not used during this phase. For a large initial difference between the founder population, the remainder of the donor genome can have a substantial contribution to genetic lag in the early generations of the introgression programme. It would be possible to take the population to fixation of the favourable allele more slowly, for example by selecting both heterozygous (Qq) and homozygous (QQ) females at generation $T + 1$, and homozygous QQ males. The population could then be fixed at generation $T + 3$, by selecting QQ males and females as parents in generation $T + 2$. Although this approach would give more opportunity to select for other traits in generation $T + 1$ and $T + 2$, in particular in females, the duration of the introgression programme would be extended by at least one generation.

The size of a donor chromosome segment around the allele which is introgressed can be substantial in the absence of selection. For example, Stam and Zeven (1981) show that the average length is about

32 cM in the sixth backcross generation, for a chromosome of 100 cM length. Hence, if the initial line difference D is large, and of the order of the number of chromosomes, the effect of this linkage drag on the genetic lag can be substantial. This effect was ignored in this study. However, in practice we might be interested in introgressing a donor region of roughly that size, because it corresponds to the confidence region of a typical QTL estimated from an experimental population. Hence, if we are interested in introgressing a QTL allele, it may be justified to ignore the effect of linkage drag on genetic lag.

Selection on markers throughout the genome to recover the recipient genotype as fast as possible is a good selection policy if the line difference is large, so that there is a high correlation between the genomic composition and breeding value of an individual. This assumes a genetic model in which the line difference is explained by many genes spread throughout the genome, with the line with highest value containing mainly plus alleles and the low line mainly minus alleles (Visscher and Haley, 1996). It is difficult to assess how realistic or unrealistic this genetic model is, because experiments to explain genetic differences within and between breeds are being carried out at present. Based upon our assumptions, selection on markers is just as good or better than selection on phenotypes. In practice, selection will usually be based on best linear unbiased prediction (BLUP) breeding values, in which family information is used. In that case, phenotypic selection may be superior to marker selection. However, selection on markers may be preferred because it is easy to perform, as soon as the individuals are born. Hence, those individuals that carry one copy of the desired allele can be identified at birth, and non-carriers can be sold or discarded immediately. Marker selection during backcrossing will be of particular advantage if the quantitative trait of interest (total economic merit, for example) is difficult to measure. An alternative approach is to combine marker and BLUP selection in a different way if limiting the genotyping costs is important. For example, the best individuals based on BLUP values could be ranked, and top ranked individuals would be genotyped sequentially until the desired number of heterozygotes has been selected. Or, if the heterozygotes have been kept after testing at birth, only the best animals based upon BLUP values could be genotyped for markers throughout the genome, and selection would be on a combination of BLUP values and marker information, with more information on BLUP values in later generations. However, the assumptions for a BLUP evaluation may break down when dealing with an introgression breeding programme, since there will be large changes in gene frequencies over time, heterogeneity of variance due to large gene effects, and non-normality of data.

Since the majority of genetic lag occurs in the latter phases of the introgression programme, efforts to reduce the lag should be focused on these phases. Since the selection intensity in females is particularly reduced in generations T and $\{T + 1\}$, it may be worthwhile to use artificial means to increase female fertility. Multiple ovulation and embryo transfer may be a cost-effective way of reducing lag in the latter stages of introgression programmes, even for fecund species such as pigs. An additional advantage of using larger family sizes is that the information of phenotypes and markers can be used to map the allele to be introgressed more precisely, if its exact location was not known at the start of the introgression programme. Hence, larger family sizes can assist in making sure that the desired allele is made homozygous in the final generation.

We have assumed throughout that during the backcrossing phase, the best males from the nucleus are used both in the nucleus and backcross populations. In some cases it may be more practical to use the second best females from the nucleus, and mate them with selected males from the backcross population. In that case, the difference in genetic mean during backcrossing between the two selection strategies (genetic mean of backcross individuals for selection of males in nucleus minus genetic mean of backcross individuals for selection of females in nucleus) may be written as,

$$\Delta_{t(m)} - \Delta_{t(f)} = \sum_{i=1}^{t-1} (\frac{1}{2})^i [\delta_{f(t-i)} - \delta_{m(t-i)}] + (1 - (\frac{1}{2})^{t-1})[\delta_m - \delta'_f] \qquad (11)$$

with $\Delta_{t(m)}$ ($\Delta_{t(f)}$) the genetic gain at generation t of backcrossing if males (females) from the nucleus are used in the backcross population, and δ'_f the genetic superiority of the second best females in the nucleus, which are mated with backcross males. If the selection responses during backcrossing are relatively constant, and t becomes large, then equation (11) reduces to,

$$\Delta_{\infty(m)} - \Delta_{\infty(f)} = [\delta_{f(cb)} - \delta_{m(cb)}] + [\delta_m - \delta'_f] \qquad (12)$$

with subscript cb denoting individuals from the backcross population. Although the first term in

equation (12) will usually be negative (higher selection pressure on males is possible in the backcross population), the second term will be much more positive because selected males in the nucleus are very superior in genetic merit relative to the next best females. Hence, usually it is more efficient to use males from the nucleus population. In practice, pig breeding companies may keep females from the commercial line for an extra litter if they are used both for normal within-line selection and for introgression matings, rather than using the next best females. In that case, the difference between the alternative schemes, i.e. either using backcross males or females, will be small, and practical considerations will decide which breeding programme is adopted. If the aim early on in the introgression programme is to reduce the contribution of the donor genome and in the latter stages to reduce lag, one suggestion is to use females from the nucleus population early on and males at the final stages. Hence, a hybrid programme in which the selected sex in the introgression population changes from males to females could be an efficient introgression breeding programme.

For all the previous results, the focus of attention has been the marginal genetic gain. In practice we are dealing with both marginal gains and an extra cost of an introgression programme. To discuss the cost aspect, we assume that the main extra cost of the introgression programme is in genotyping the individuals. Assume that the cost of tissue sampling and DNA preparation is c, and that the cost per marker genotyping is c_m. If necessary, a fixed cost associated with each individual which is kept can easily be accommodated. Assume that the total number of marker genotypes per genotyped individuals is one (for the introgressed locus) plus $\frac{1}{2}m$ times the number of chromosomes (n_c). The factor of $\frac{1}{2}$ is because only those individuals carrying one copy of the desired allele will be genotyped for the remaining markers. If intercrossing takes place at generation $t = T$, then backcross progeny born at $\{T-2\}$ generations are genotyped (starting at the first backcross generation). In addition, individuals born at generation T and $\{T + 1\}$ will be genotyped. For each of the backcross generations, $\frac{1}{2}N$ individuals are genotyped for the allele to be introgressed, and $\frac{1}{4}N$ individuals for all chromosomes, since only one sex is eligible for selection. For individuals born in generation T, all are genotyped for the allele to be introgressed, and $\frac{1}{2}$N for the remainder of the genome. Finally, all N individuals born at $\{T + 1\}$ are genotyped, but only $\frac{1}{4}N$, i.e. those homozygous for the allele to be introgressed, are genotyped for the rest of the genome.

In total, DNA will be extracted from $N(\frac{1}{2}T + 1)$ individuals, and the genomic proportion will be determined for $N(T + 1)/4$ individuals. Hence, the total marginal marker cost of the introgression programme is:

$$C_{T+1} = N[(c + c_m)(\tfrac{1}{2}T + 1) + mn_c c_m(T + 1)/4] \quad (13).$$

The total cost of an introgression programme heavily depends on the cost of genotyping. Genotyping costs are evolving rapidly, and it is very difficult to predict what kind of genetic markers will be used in, say, 5 years time, and what the cost per genotyped individual will be. Some scientists predict that it will be feasible to sequence the complete genome of individuals for a reasonable cost in the not so distant future (Lander, 1996).

Assuming present day costs of \$10 for DNA extraction and preparation, and \$3 per marker genotype per individual, and five markers per chromosome for 18 autosomes, the total cost for an introgression programme in which intercrossing is carried out at generation $T = 5$, is $N(10 + 3)(5/2 + 1) + N \times 18 \times 5 \times 3(6/4) = N\$450{\cdot}5$. Hence, for a pig nucleus of $N = 400$, e.g. 40 litters per generation, the total marker cost of the programme at today's prices would be about \$180,200, spread out over six generations. The minimum marker cost, assuming phenotypic selection rather than marker selection, would be $N\$45{\cdot}5$ = \$18,200. Given the cost for genotyping, these estimates are likely to be underestimates, because not all markers will be completely informative (i.e. fixed for alternative alleles in the two founder breeds), and usually more than a single marker will be used to introgress a QTL, to reduce the risk of loosing the QTL allele due to imprecision in the estimate of its location (Visscher *et al.*, 1996). However, the total genotyping cost for selection against the donor genome will be smaller than given in equation (13), because once individuals are homozygous for part of the recipient genome, such regions (or whole chromosomes) need not be typed any longer in their progeny. Hence, in the latter stages of the introgression programme, it is expected that only few animals need to be genotyped for markers on all chromosomes.

In practice, the cost-benefit ratios of gene introgression programmes do not solely depend on extra genetic gain relative to extra costs. Risk management, e.g. introgressing a disease resistance gene, or marketing, e.g. selling a product which nobody else has available, may well be more important in determining the success of introgression programmes.

Acknowledgements

This work was supported by the BBSRC, MAFF, and the Marker Assisted Selection Consortium of the UK pig breeding industry. We thank John Gibson and Hein van der Steen for many useful comments and discussions, and the referee for constructive comments.

References

Gama, L. T., Smith, C. and Gibson, J. P. 1992. Transgene effects, introgression strategies and testing schemes in pigs. *Animal Production* **54:** 427-440.

Hill, W. G. 1993. Variation in genetic composition in backcrossing programs. *Journal of Heredity* **84:** 212-213.

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

Lander, E. S. 1996. The new genomics: global views of biology. *Science* **274:** 536-539.

Rothschild, M. F., Jacobson, C., Vaske, D. A., Tuggle, C. K., Wang, L., Short, T. H., Eckardt, G. R., Sasaki, S., Vincent, A., McLaren, D. G., Southwood, O., Van der Steen, H., Mileham, A. and Plastow, G. 1996. The estrogen receptor locus is associated with a major gene influencing litter size in pigs. *Proceedings of the National Academy of Sciences of the United States of America* **93:** 201-205.

Smith, C. 1984. Rates of genetic change in farm livestock. *Research and Development in Agriculture* **1:** 79-85.

Stam, P. and Zeven, A. C. 1981. The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica* **30:** 227-238.

Visscher, P. M. 1996. Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. *Journal of Heredity* **87:** 136-138.

Visscher, P. M. and Haley, C. S. 1996. Detection of putative quantitative trait loci in line crosses under infinitesimal genetic models. *Theoretical and Applied Genetics* **93:** 691-702.

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

(*Received* 21 *July* 1997—*Accepted* 4 *August* 1998)