Animal Genetics,
1997, **28**, 181–187

Marker-assisted introgression using non-unique marker alleles I: selection on the presence of linked marker alleles

A M van Heelsum, P M Visscher, C S Haley

Summary

This paper investigates marker-assisted introgression of a major gene into an outbred line, where identification of the introgressed gene is incomplete because marker alleles are not unique to the base populations (the same marker allele can occur in both donor and recipient population). Those markers are used to identify the introgressed allele as well as the background genotype. The effect of using those markers, as if they were completely informative on the retention of the introgressed allele, was examined over five generations of backcrossing by using a single marker or a marker bracket for different starting frequencies of the marker alleles. Results were calculated by using both a deterministic approach, where selection is only for the desired allele, and by a stochastic approach, where selection is also on background genotype. When marker allele frequencies in donor and recipient population diverged from 1 and 0 (using a diallelic marker), the ability to retain the desired allele rapidly declined. Marker brackets performed notably better than single markers. If selection on background marker genotype was applied, the desired allele could be lost even more quickly than expected at random because the chance that the allele, which is common in the donor line, is present on the locus identifying the introgressed allele and is surrounded by alleles common in the recipient line on the background marker loci, will descend from the donor line (double recombination has taken place), is a lot smaller than the chance that this allele will stem from the recipient line (in which the allele occurs in low frequency). Marker brackets again performed better. Preselection against marker homozygotes (producing uninformative gametes) gave a slightly better retention of the introgressed allele.

Keywords: animal breeding, marker-assisted introgression, outbred populations, quantitative trait loci

A M van Heelsum
C S Haley
Roslin Institute
(Edinburgh), Roslin,
Midlothian EH25 9PS,
UK
P M Visscher
Institute of Ecology and
Resource Management,
University of
Edinburgh, Edinburgh
EH9 3JG, UK

Correspondence: A M van Heelsum.

Accepted 17 February 1997

Introduction

Different breeds of plants or livestock can have different traits that make them of interest for commercial use. In traditional crossbreeding schemes to create a synthetic, breeds would be crossed and the offspring would be selected on the beneficial characteristics of all parental breeds. If one of the breeds has only one or a few interesting traits, ideally we would like to transfer only the gene(s) controlling this trait to a commercial population, which outperforms the first breed for other traits and leaves the rest of the genome, the 'background genotype', unchanged.

Introgression of the desired gene(s) can be performed by using backcrossing, where the donor breed (supplying the genes to be introgressed) is crossed with the (commercial) recipient breed, the crossbred generation is crossed with the recipient again, and so on, until most of the genome of the backcrossed animal descends from the recipient population, except for the introgressed genes. Intercrossing can then take place to make the desired alleles homozygous.

Introgression can only be carried out effectively if it is possible to identify the descent of the alleles at the loci of the introgressed genes as well as for the background genotype. Recent progress in mapping markers, major genes and QTLs (quantitative trait loci) makes this increasingly feasible.

A number of earlier studies have looked into the problems and benefits of introgression by using markers to aid recovery of the background genotype and identification of the desired alleles (MAI or Marker Assisted Introgression). Hillel *et al.* (1993) were very optimistic about the benefits of applying MAI to a poultry breeding scheme. They concluded that if (intensive) selection on background genotype was carried out by using DNA fingerprinting, only two generations of backcrossing were needed to recover almost all of the recipient genotype (assuming complete identification of the introgressed allele). Hillel *et al.* (1993) treated the fingerprint markers as completely independent markers and ignored their dominant character, non-random distribution over the genome and any recombination between chromosome segments and markers.

Hospital *et al.* (1992) also assumed direct, complete identification of the introgressed gene, but used a more realistic approach that accounted for recombination events. Minimizing the donor segment around the introgressed gene was studied separately from the recovery of recipient genome on the other chromosomes. Hospital *et al.* (1992) were also less optimistic about the number of backcross generations needed to successfully introgress a gene and recover the background genotype of the recipient; this depended on the quality and quantity of the markers used.

Visscher *et al.* (1996) used markers to identify the introgressed gene. They studied a gene at a known position as well as at a sampled position, mimicking a realistic situation where the exact position of the introgressed gene is not known. They assumed marker alleles to be unique to the alternative base populations and therefore fully informative in the first cross (F_1). The predictive value of the markers will only become limited over generations because of recombination whereas, in reality, markers might often start as not unique to the base populations.

Groen & Smith (1995) looked at a situation where less informative markers were used: the marker allele M1, which had a frequency of 100% in the recipient population, also occurred in the donor population with a frequency of 20%. This means that marker allele M2 is still unique to the donor population; therefore the introgressed QTL-allele can still be fully identified in the F_1, and the introgression results will not be impeded.

In most of the alternatives shown by Groen & Smith (1995), marker alleles used to identify introgressed allele as well as background genotype were fixed in the alternative base populations. However, the difference between the frequencies of the 'good' background QTL-alleles in donor and recipient populations was very small (0·7 and 0·6, respectively). This implies that the markers will pick up very little genetic variance and that the within-line variances are large. Therefore, their conclusion that phenotypic selection outperforms selection on genomic similarity is, given the parameters, not surprising.

Gama *et al.* (1992) compared within-line phenotypic selection (within the recipient line) with introgression, to determine the genetic lag, caused by the introduction of an inferior animal (or breed) into a highly selected commercial nucleus, in order to introgress a gene. The effect of the introgressed gene has to be greater than the genetic lag to make the scheme profitable. Another reason for the lag is that selection on production traits, other than those influenced by the introgressed gene, cannot be as intensive as in the purebreeding nucleus. The lag becomes greater if the within-line selection is based on Best Linear Unbiased Prediction (BLUP), including information on relatives, because the response in the purebreeding (recipient) population would be higher.

All the studies mentioned above assumed that the introgressed allele would be easily identifiable by direct observation or by a marker allele unique to the donor population. This might be a good starting point for introgression studies but the consequences, if this assumption is invalid, should be examined. In this study, markers were used to identify the introgressed gene (with known position), of which the marker alleles were not necessarily unique to the alternative base populations. This introduced an extra uncertainty to the introgression process. It was assumed that the allele frequencies in the base populations were known, although the donor population itself might no longer be available (and even if the starting frequencies in the donor population were unknown a fair indication of their frequencies could be obtained by comparing the available crossbred and recipient populations). This study aimed to investigate the severity of problems that the extra uncertainty can cause, by loss of the introgressed allele, for different strategies of using the non-unique markers.

Models and methods

Introgression using one closely linked marker; a two-locus model

In order to investigate the efficiency of a marker that is closely linked to the gene of interest, one can start with a very simple situation: a chromosome segment of 1 centimorgan (cM) in length, the gene of interest at one end and the marker locus at the other end. The trait gene was assumed to have a major influence on the desirable trait, which cannot be measured phenotypically at the moment of selection (e.g. disease resistance or fertility traits) and cannot be identified by using molecular analysis.

The marker and major gene were both diallelic. The donor population was fixed for the trait-allele T1 and the recipient population was fixed for trait-allele T2. The frequency of marker allele A1 was p in the donor population and q in the recipient population (therefore the frequency of A2 was 1-p in the donor and 1-q in the recipient population). From those allele frequencies, the frequencies of every possible genotype in the F_1 and in the recipient population could be calculated. The gamete frequencies could then be calculated, given the recombination fraction r and, from this, the genotype frequencies could be calculated for the next generation; this was carried out for the following five generations of backcrossing. Assuming Haldane's mapping function, the recombination rate between marker and trait locus was 0·0099.

Selection started in the first backcross generation (BC_1). All the selected genotypes were required to have either one or two copies of the marker allele with the highest frequency in the donor population on the marker locus. It was assumed that p is always greater than or equal to q, so A1 was the marker allele that was most frequently linked with the desired-trait allele T1. Therefore, selection was on A1, and all A1A1 and A1A2 genotypes, either heterozygous or homozygous for the major gene (T1T2 or T2T2), were selected. Only the two genotypes homozygous for the alternative marker allele A2 (either containing no or one copy of the desirable major gene allele) were not selected.

The proportion of animals with the desired trait allele was given by calculating the sum of the genotype frequencies of genotypes having a copy of T1. Ideally this will stay at 50% over the subsequent backcross generations, but when p and q are not equal to 1 and 0, respectively, the marker will not be fully informative, so selection will be less efficient and the percentage will be lower than 50%. In later generations, recombination will make the predictive value of the marker even lower.

Starting frequencies p/q were chosen as 1·0/0·0 (fully informative marker), 0·99/0·01, 0·95/0·05, 0·9/0·1 and 0·5/0·5 (completely uninformative marker); so, apart from in the first case, both marker alleles can come from both populations. To investigate the effect of only one allele occurring in both populations, the percentage of animals having a copy of T1, in backcross generation five, was calculated for a starting frequency p of 1, with q varying between 0 (fully informative) and 1 (completely uninformative).

Selection on heterozygosity. Selecting individuals with at least one copy of marker allele A1 meant that marker homozygotes as well as heterozygotes were selected. One of the copies of the marker allele always comes from the recipient parent, so half of the gametes of the homozygous (A1A1) animals will be 'false positive'. To avoid the production of false positives, an extra selection criterion can be added: heterozygosity at the marker locus.

This strategy was investigated for the single marker case with a p and q of 0·9 and 0·1, respectively, again over five generations of backcrossing. Only animals heterozygous for the marker were selected from BC_1 onwards. If F_1 animals were typed they could be selected for heterozygosity on the marker locus as well, so retention of the introgressed trait allele for heterozygosity selection, started in the F_1, was also investigated.

Introgression using one or two markers: a stochastic approach

To study more complicated situations, a stochastic simulation was implemented. For every animal a genome was simulated, consisting of one chromosome with a number of loci 1 cM apart. One of the loci contained a major gene, at a known position. The major-gene alleles were assumed to be fixed in the base populations, giving allele T1 a frequency of 100% in the donor population and allele T2 a frequency of 100% in the recipient population. Selection was on markers, because the major gene was assumed to be not readily identifiable. The marker alleles again were not fully informative: the frequencies could vary, where the frequency of allele 1 of all markers (e.g. A1, B1, etc.) was p in the donor population and q in the recipient population. Three different combinations of starting frequencies p and q were investigated: 1·0/0·0, 0·9/0·1 and 0·5/0·5. In all simulations recombination between loci followed Haldane's mapping function.

The simulated population had a size of 800: 10 sires were selected to be used on 10 dams each; every dam had eight offspring (four males and four females), thus resembling a nucleus pig-breeding population. The F_1 was formed by crossing the donor and the recipient population, after which the crossbred animals were backcrossed to the recipient population for five generations, creating BC_1–BC_5. In the five backcross generations, selection took place among males, which were mated to females from the recipient population.

Two-locus model. To test the programme, the first situation simulated was a two-locus chromosomal segment for each animal, mimicking the calculations from the deterministic approach mentioned earlier on, so results could be compared. Again the segment was assumed to be 1-cM long with, on one end, the diallelic marker and, on the other end, the diallelic major gene. Results were averaged over 100 replicates.

Multilocus model. The next situation considered was a multilocus model, where every individual was assumed to have a genome length of 100 cM. Markers were equally spaced over the genome in 10-cM intervals with markers on either end of the (single) chromosome; therefore, there were a total of 11 markers, equally spaced over the chromosome (position 1, position 11, etc. to position 101). The major gene was fixed on position 30 if only one marker was used for identification of the major gene-allele or on position 35 if a marker bracket was used. For identification of the major-gene allele the closest available marker(s) were used, being the marker on position 31 in the one marker case and the markers on position 31 and 41 if a marker bracket was used.

Selection took place in two steps. The first step was aimed at preserving the desired trait allele, so only animals with at least one copy of the thought-to-be desirable marker allele were selected (A1 if a single marker was used, A1 and B1 if a marker bracket was used). The second step was aimed at recovering the background genotype of the recipient as quickly as possible by selection on markers, and was either at random or on a marker score. This score (defined by Visscher 1996) is essentially calculated as the proportion of marker loci thought to be descending from the donor population, those with the most '1'-alleles over all marker loci (including the loci nearest the major gene locus) being selected. Again, results are the mean value of 100 replicates.

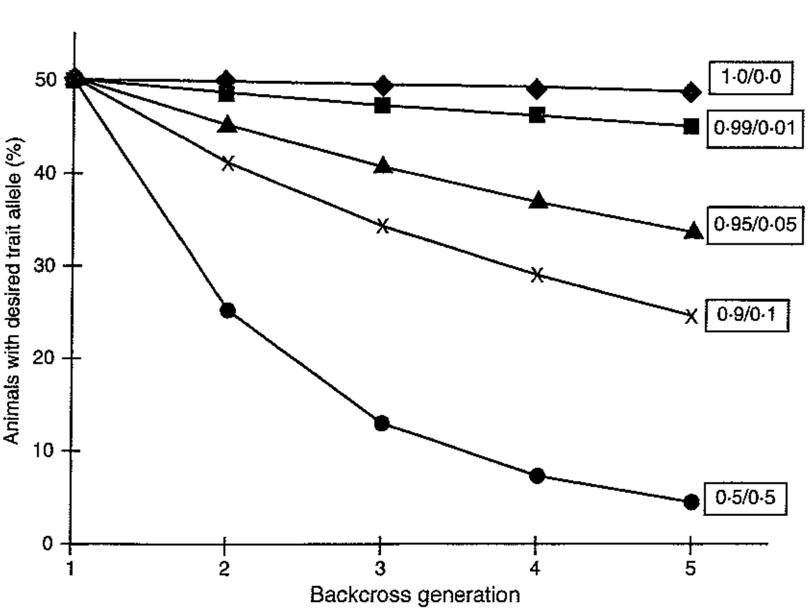

Fig. 1. Percentage of animals with the desired trait allele over five generations of backcrossing. A single, diallelic marker was used to identify the major gene allele for five different, opposite marker allele frequencies in the donor and recipient population.

Results

Deterministic approach

The percentages of animals that still have the desired (introgressed) trait allele over five generations of backcrossing, for the two-locus deterministic model and five different starting frequencies (1·0/0·0, 0·99/0·01, 0·95/0·05, 0·9/0·1 and 0·5/0·5), are given in Fig. 1. If the starting frequencies are 1·0 and 0·0 all A1 alleles originally stem from the donor population and all A2 alleles stem from the recipient population. The marker is then fully informative in the F_1, but, with recombination, the percentage of animals with the desirable trait allele decreases slightly over generations (e.g. 48·0% in BC_5).

Where starting frequencies are 0·5 and 0·5 the marker is completely uninformative in the F_1, so the percentage of animals with desired trait allele should halve every generation. This is true until BC_3, but in BC_4 the percentage is 6·6; slightly higher than the expected 6·25%, and in BC_5 the discrepancy is even greater (3·7% compared with the expected 3·1%). Because we keep selecting on A1-marker alleles, linkage disequilibrium will eventually be induced between marker and major gene, even if absent in the first backcross generation. Selection starts only in BC_1, so disequilibrium has not yet been induced and the frequency will halve, as expected, in BC_2. The selected BC_2 parents will produce more A1-gametes that originate from their BC_1-parent than A1-gametes originating from their recipient parent and, because T1 can come only from the BC_1-parent, the fraction of T1 within A1-gametes is higher than the fraction of T1 within A2-gametes. This effect will become visible in the selected group in BC_3 (not in BC_3 itself because the fraction of T1 gametes overall from BC_2 is just one of eight, as first expected) and in the gametes the BC_3-animals produce. This means that the fraction of animals with a desired trait allele in BC_4 will not halve but will be slightly higher than half, and so on in later generations. This effect depends only on the recombination rate between marker and major gene: if the recombination rate is 0·5, marker and major gene segregate completely independently of each other so no linkage can be induced but, if the recombination rate is close to 0, the effect will be maximal and very close to the results shown in Fig. 1 (for a recombination rate of 0·0099).

The curve for starting frequencies 0·9/0·1 in Fig. 1 is almost midway between those for starting frequencies 1·0/0·0 and 0·5/0·5, indicating that the relationship between starting frequencies and loss of the introgressed allele is not linear.

Figure 1 indicates that starting frequencies have to be extreme to avoid loss of the desired trait allele. Even if the A1-marker allele is fixed in the donor population, the occurrence of this allele in the recipient population as well can seriously frustrate the introgression process. This is shown in Fig. 2, where the percentage of animals with the desired trait allele in BC_5 is plotted against the frequency of A1 in the recipient population (q), when A1 is fixed in the donor population. The percentage of animals with the desired allele drops very rapidly for increasing values of q.

Table 1 shows the results of adding selection for heterozygosity. The preservation of the desired trait allele is somewhat better than it is without discarding homozygous animals: if the extra selection is started in BC_1, 28·4% of the animals still possess the introgressed allele in BC_5; if started in the F_1 this value is 31·2% whereas, without the extra selection, only 23·9% of the animals still had the desired allele in BC_5 (Fig. 1).

Table 1. Percentage of animals with the desired trait allele over five generations of backcrossing calculated using a deterministic model*

	Extra selection started in:	
Backcross generation	F_1	BC_1
1	50·0	50·0
2	44·1	40·7
3	39·3	35·9
4	35·0	31·9
5	31·2	28·4

*One diallelic marker was used, which was 1 cM from the major gene, with initial frequencies of marker allele A1 of 0·9 and 0·1 in donor and recipient population respectively; there was added selection against animals homozygous for A1, starting in the F1 (first cross) or in the BC1 (first backcross generation).

Stochastic approach

The average values of 100 simulated replicates for the two-locus model (results not shown) gave results that were very close to those calculated using the deterministic approach (Fig. 1); no significant difference between the methods was found. Differences between the simulated multilocus model and the deterministic model were slightly higher, but almost always convincingly non-significant. However, comparison of the simulated multilocus model and the simulated two-locus model, showed some differences that were near a 5% significance level, even over 100 replicates. There does not appear to be a consistent trend in those differences. It should be noted that the standard deviations of the percentages produced by the stochastic multi-locus model can be rather large: from 2–17% for p/q values of 1·0/0·0, from 4–23% for p/q values of 0·9/0·1 and from 5–12% for p/q values of 0·5/0·5 (BC_2–BC_5).

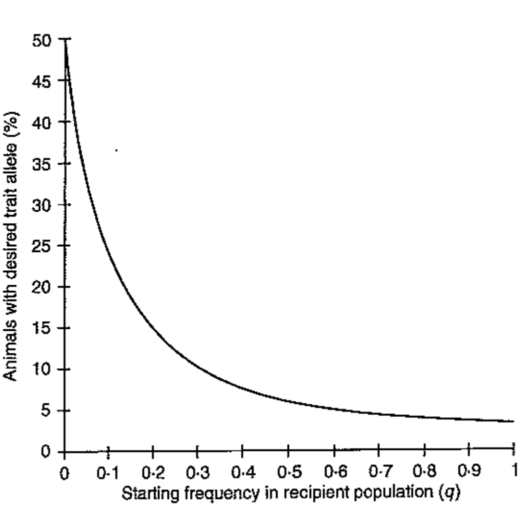

Fig. 2. Percentage of animals with a desired trait allele in the fifth backcross generation (BC_5). A single diallelic marker was used to identify the trait allele at a distance of 1-cM from the major gene, where p (frequency of the first marker allele in the donor population) equals 1·0 and q (frequency of the first marker allele in the recipient population) ranges from 0·0–1·0.

The limited difference in introgressed gene retention between the two-locus and multilocus model (without selection on background genotype) seems to justify the conclusion that as long as there is no selection on background genome, the background genome will not influence the outcome of the introgression process. So, if the preservation of an introgressed trait allele using non-unique markers is under scrutiny, it will be sufficient to consider only the markers used to identify the major gene.

If selection also takes place on background genotype (using markers), the situation is quite different. Even with starting frequencies of 1 and 0 (completely informative markers), and with one marker used to identify the trait allele, only 42·3% (SE 1·7) of the animals in BC_5 still have a copy of the desired trait allele. When a marker bracket is used this percentage is 49·7% (SE 0·2). Without background selection, and using a marker bracket, 50·0% (SE 0·2) still have the desired trait allele in BC_5; this percentage is 48·5% (SE 0·5) if a single marker is used. Starting frequencies of 0·5/0·5 gain a major gene preservation, which is not significantly different from uninformative markers in the deterministic approach, for all four selection strategies (results not shown).

In Fig. 3 the differences between the four different selection strategies are plotted for starting frequencies of 0·9 and 0·1. Again, the marker bracket performs much better than a single marker. Selection on background genotype has a detrimental effect on the retention of the introgressed trait allele; in the single marker case values are even lower than for no selection on the introgressed allele at all.

Discussion

The main conclusion of this work is that using non-unique marker alleles as if they were unique can lead to disaster. In particular, if selection is carried out on background genotype, as well as on presence of the introgressed trait allele, the chance that the introgressed allele will be lost is unacceptably high (Fig. 3). Even if heterozygosity selection is added, it will be hard for any commercial company to justify the expense of an introgression programme if, after five backcross generations, the chance of resulting animals still having the introgressed allele is only 31% (Table 1).

The sudden loss of the introgressed trait allele when selection on background genotype is carried out is not surprising when we realize which genotype is most preferred using both selection criteria. Ideally, we want a genome that stems exclusively from the recipient population (identified by 'two'-marker alleles), except for the trait allele, which has to come from the donor population (identified by 'one'-marker alleles). This can be achieved only by double recombination on both sides of the introgressed region (consisting of the major gene and one or two markers). Double recombination is relatively rare; it is much more likely that the whole chromosome section will stem from a recipient gamete with only 'two'-alleles except on the introgressed gene marking sites (e.g. the chance of finding such a gamete produced by F_1 in the single marker case is $(q)^{m-1} * (1\text{-}q)$, where m is the number of markers on the whole genome). Therefore, even if the frequency of the A1-allele is very low in the recipient population, it will act as a very effective way of selecting against the desirable trait allele.

In this study, selection for background genotype was on markers, but it is likely that selection on phenotype would give the same results. It is much more likely that the stretch of genome coding for a favourable phenotype from the recipient has an 'unlikely' allele for the marker identifying the introgressed allele, than that double recombination has occurred around the introgressed gene and the marker allele really comes from the donor.

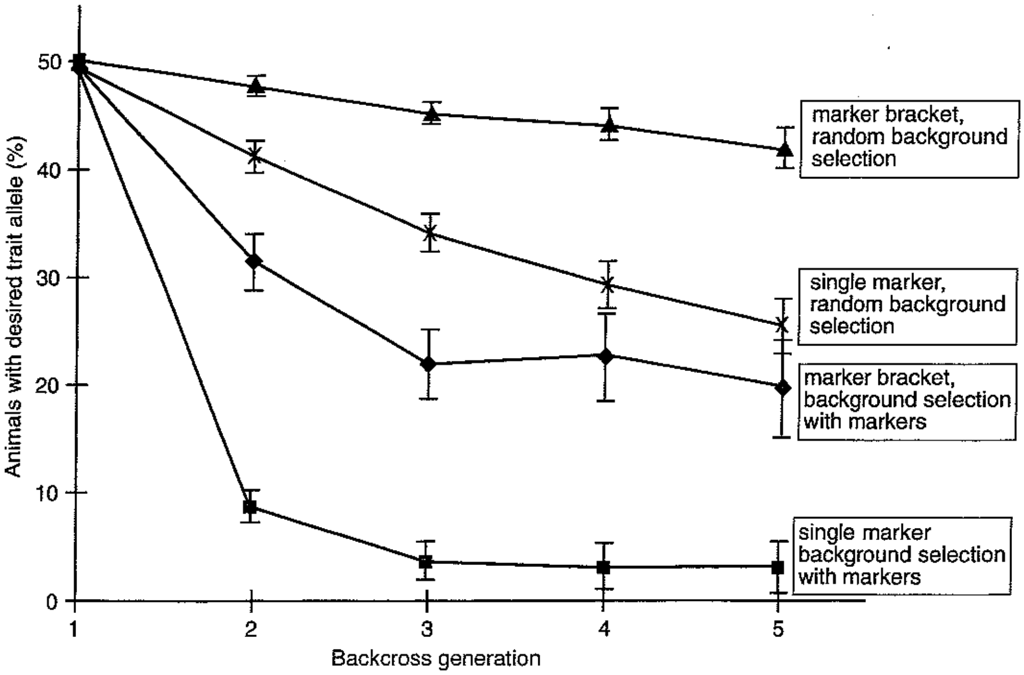

Fig. 3. Percentage of animals with the desired trait allele over five generations of backcrossing. Either a single marker or a marker bracket was used to identify the trait allele, either with or without selection on background genotype (initial frequency of first marker allele in donor and recipient population 0·9 and 0·1, respectively). The data points represent the average value of 100 replicates, vertical bars represent 95% confidence intervals.

Keeping the percentage of animals with the introgressed allele at 50% over more generations of backcrossing will not be realistic unless the trait gene itself can be identified and direct selection can take place, rather than indirect selection on linked markers. With indirect selection there will always be some decrease, over generations, owing to recombination. For the sole purpose of not losing the introgressed allele it would be best to keep the number of backcross generations to a minimum, and start intercrossing as quickly as possible to fix the allele in the population. However, this implies that the numbers of recombinants around the introgressed gene have not had the chance to build up, so it is likely that there still is quite a substantial stretch of donor genome around the gene (Hospital *et al.* 1992); therefore, as far as minimizing the donor contribution in the background genotype is concerned, this is not the best way. Presumably the right moment to change from backcrossing to intercrossing depends on the reliability of the identification of the trait gene and the background genotype (Hillel *et al.* 1990), the magnitude of the difference between donor and recipient population (Gama *et al.* 1992) and the rate of selection in both populations. If we expect the donor population to have more valuable genes than just the gene for introgression (as in Groen & Smith 1995), intercrossing might start sooner, together with selection on phenotype.

The identification of the major-gene alleles was more reliable if a marker bracket was used, rather than a single marker, and if allele frequencies were as different as possible in the alternative populations. If markers are shown not to give a fairly good preservation of the trait allele the obvious thing to do would be to look for better markers (more closely linked or even the major gene itself) or to have more extreme starting frequencies. This might not always be possible, or it might be undesirable to wait any longer to start the introgression process, so it might still be useful to search for strategies that can deal with non-unique markers.

In all scenarios, of which results are shown in Figs 1 & 2, all animals are selected that fulfil the selection criterion of having at least one copy of the thought-to-be desirable marker allele, without making any distinction between genotypes within this group. In case of one diallelic marker only three marker genotypes exist and adding the heterozygosity criterion means that all three genotypes are treated separately. However, if a marker bracket is used, no distinction is made between genotypes homozygous for one marker but heterozygous for the other, where the probability of those genotypes to contain the desirable trait allele can be quite different (owing to selection in previous generations, or different starting frequencies of the markers). Eventually, the aim is to end up with as many animals as possible still possessing the introgressed allele so, in the intercrossing phase following the backcrossing phase, the allele can be fixed in the newly created synthetic population. Using the probability of the allele to be present would be a logical selection criterion in the backcrossing phase. It might make the selection more effective and a better preservation of the desired allele could be achieved. This will be subject of further study.

Acknowledgements

We wish to thank Bill Hill for many useful comments. This research was supported by the Marker Assisted Selection Consortium of the UK pig industry (Cotswold Pig Development Company Ltd, J.S.R. Farms Ltd, Pig Improvement Company/National Pig Development Company, Newsham Hybrid Pigs Ltd, Rattlerow Farms Ltd and the Meat and Livestock Commission) and by the Ministry of Agriculture, Fisheries and Food (MAFF), the Department of Trade and Industry (DTI) and the Biotechnology and Biological Sciences Research Council (BBSRC).

References

Gama L.T., Smith C. & Gibson J.P. (1992) Transgene effects, introgression strategies and testing schemes in pigs. *Animal Production* **54**, 427–40.

Groen A.F. & Smith C. (1995) A stochastic simulation study on the efficiency of marker-assisted introgression in livestock. *Journal of Animal Breeding and Genetics* **112**, 161–70.

Hillel J., Schaap T., Haberfield A. *et al.* (1990) DNA fingerprints applied to gene introgression in breeding programs. *Genetics* **124**, 783–9.

Hillel J., Verrinder Gibbins A.M., Etches R.J. & Shaver D.McQ. (1993) Strategies for the rapid introgression of a specific gene modification into a commercial poultry flock from a single carrier. *Poultry Science* **72**, 1197–211.

Hospital F., Chevalet D. & Mulsant P. (1992) Using markers in gene introgression breeding programs. *Genetics* **132**, 1199–210.

Visscher P.M. (1996) Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. *Journal of Heredity* **87**(2), 136–7.

Visscher P.M., Haley C.S. & Thompson R. (1996) Marker assisted introgression in backcross breeding programs. *Genetics* **144**, 1921–30.