IMPROVING THE EFFICIENCY OF BACKCROSSING PROGRAMS USING DNA MARKERS

I. R. Franklin

CSIRO Animal Production, Locked Bag 1, Delivery Centre, Blacktown, NSW 2148

SUMMARY

Repeated backcrossing, either to upgrade to a new breed or strain, or to transfer desirable genes from one breed to another, is time consuming and expensive. This study describes some of the theory of progress towards homozygosity, with particular attention to the proportion of the genome associated with genetic markers. In general, selection for markers is much more effective than selecting for the phenotype of the recurrent parent, and optimal strategies for the use of genetic markers are discussed.

Keywords: Backcrossing, introgression, genetic markers

INTRODUCTION

In plants, perhaps the most important application of recurrent backcrossing is introgression – the transfer of one or more genes (eg. for disease resistance) from one variety into another. In animals, single genes of economic importance are rare, but recurrent backcrossing is commonly used to upgrade from one strain to another. In some cases, an exotic breed may be available only as semen, and recurrent backcrossing may be the only means to regenerate the properties of the desired exotic strain. Any procedure that accelerates the rate of progress is highly desirable. This paper considers some of the theoretical aspects of recurrent backcrossing and, in particular, the use of DNA markers.

BREED REPLACEMENT

Consider first the most straightforward of backcrossing programs – breed or strain replacement. Four or five generations are often more than adequate to recreate almost all of the economic value of the desired breed, especially if the two breeds do not differ greatly in economic efficiency. Nevertheless, there is often an obsession with breed purity, and breeders may go to considerable expense to upgrade by embryo transfer when a simple backcrossing program using AI may be both sufficient and more cost-effective. However, even four generations in sheep or cattle can take many years.

The algebra of such programs is straightforward. Denote the donor and recurrent strains by P1 and P2. In the first generation, the offspring contain 50% of each genome, in the second generation the offspring contain 75% of the P2 genome, and in the n^{th} generation, the fraction that is P2 is $1 - (\_)$". However, even if a particular generation is, say, 75% exotic on average, the variation about this mean value may be considerable (Franklin 1977; Hill 1993). The variance about the mean is

$$V(p) = \frac{1}{8L^2}\frac{1}{4^n}\prod_{i=1}^{n} C_i^n \frac{1}{i^2}\left(2iL - v + \prod_{j}^{v} e^{-2il_j}\right) \qquad (1)$$

where $L = \Sigma l_j$ is the total genome length in Morgans and $\{l_j\}, j=1,v$ are the lengths of each of the v chromosomes. Table 1 shows an example, derived from Hill, to illustrate the variation expected. In this example, there are 30 chromosomes with sizes 1.5, 1.0 and 0.5 Morgans, in equal numbers, so that the total genome size (L) is 30.

Table 1. Mean and standard deviation of the contribution of the recurrent parent

Generation	F_1	B_1	B_2	B_3	B_4	B_5
Mean	0.50	0.75	0.875	0.9375	0.9687	0.9844
SD	0.0	0.0351	0.0286	0.0204	0.0140	0.0094

Consider, for example, generation B2. The mean proportion of the genome derived from P2 is 0.875, but individuals will vary, with 95% probability, from 0.82 to 0.93 P2. How do we exploit this variability? There are two possibilities, which are not mutually exclusive. The first is to select for a polygenic trait that exhibits a large difference between the two breeds. The second is to use genetic markers to select individuals with a higher proportion of the desired genome.

The first point to note is that the first generation contains no variation in the proportion of the genome derived from each parental line. Hence, any form of selection for genomic proportion is ineffective until the B1 or subsequent generations. In these later generations, the genome is divided into chromosome segments that are alternatively heterogenic (P1/P2) or homogenic (P2/P2) with respect to parental origin. In the B1 generation, the segments are, on average, of equal length, and in later generations heterogenic segments become smaller as recombination breaks them down.

Mass selection for a polygenic trait. First, consider selection for a quantitative trait for which the original strains differ substantially. The variance in any generation can be partitioned into three components: an environmental variance, V_E , a genotypic variance arising from genes segregating within the each of the parental lines, $V_{G(w)}$, and a variance due to the variation in genomic proportion, $V_{G(B)}$. The total variance, V_P, is $V_{G(B)} + V_{G(w)} + V_E$. Now suppose that the trait is scaled additive, and that the difference between the two parents is D. Then $V_{G(B)} = D^2 V(p)$, where $V(p)$ is defined by (1). When we select individuals based on their phenotype, the selection differential applied to the proportion of the genome that is homogenic is

$$S_p = b_{pP} S_P \ ,$$

where b_{pP} is the regression of the genomic proportion (p) on the phenotype P. This regression is

$$D\frac{V(p)}{V_P} \quad - \quad \frac{1}{D}\left(\frac{V_{G(B)}}{V_P}\right) \quad - \quad \frac{1}{D}h_B^2 \ ,$$

defining h_B^2 as the proportion of the total variance due to segregation among parental chromosome segments. Selection in subsequent generations will be less effective because of the decline in $V_{G(B)}$. Selection for a trait associated with the difference between the lines has little effect on increasing the proportion of the genome derived from the recurrent parent, unless (a) the heritability of the trait is high, and (b) the two lines differ markedly for the selected trait.

Selection for markers. Consider first a single marker on a chromosome of length l, and that we select individuals that are homozygous for alleles derived on from the recurrent parent (P_2). The proportion of the chromosome that is homogenic for P_2 is a function of the chromosome size and the position of the marker on the chromosome. In B_1 individuals, the expected proportion of a chromosome homogenic for P_2 among individuals homozygous at the marker locus is

$$\tfrac{1}{2}l + \tfrac{1}{4}(2 - e^{2x} - e^{2(l-x)}) \qquad (3)$$

where x is the marker position (Stam and Zeven 1981). Figure 1 shows the relative proportion of the chromosome derived from P_2 as a function of chromosome length (l = 0.5,1.0,2.0) and the relative map position of the marker.

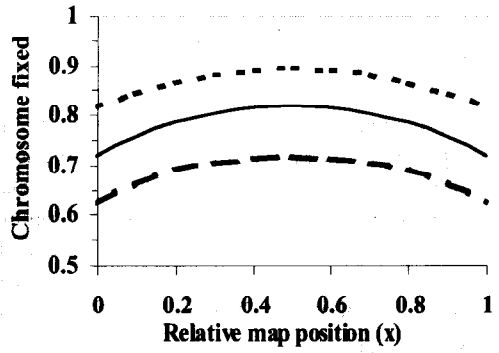

Figure 1. The proportion of chromosome homogenic as a function of a marker's relative position on the chromosome. Legend: --- l=0.5 , — l=1.0 , -- l=2.0.

We can make two observations. First, the centrally located markers carry with them a greater proportion of the chromosome. However, the difference is not great – a marker located between $0.3l$ and $0.7l$ is almost as efficient. Secondly, while markers on small chromosomes carry with them relatively more of the chromosome, in absolute terms markers on large chromosomes are more effective. For example, a centrally located marker on a chromosome of unit length has an expected homogenic length of 0.816 Morgans. For a chromosome of length 2, the expected homogenic length is 0.6227x2 = 1.245 Morgans.

Obviously, we can increase the fraction homogenic by using two or more markers. For example, suppose that we select for two markers on a chromosome of unit length at positions 0.33 and 0.66. The length of chromosome homogenic is 0.904 compared to 0.816 for a single marker. If there are three markers (at 0.1, 0.5, 0.9) the expected proportion homogenic is now 0.971. There is, however, a cost. If we select for homozygosity at a single marker, we have to discard, in the B1 generation, 50% of all individuals. Selecting for two markers, we discard 62% of individuals. The third case requires that we cull 70%. Hence, it is wasteful to use more than a single marker per chromosome; considerable selection intensity is lost for little gain in the proportion of the genome fixed.

How do we implement a selection program using marker data? First, we choose one informative, centrally located marker per chromosome (or one per arm for large metacentric chromosomes). The next task is to select individuals with as many loci as possible fixed for alleles derived from the P2 strain, given constraints on the selection differential. The probability that k of n loci are fixed is

$$\Pr(k) = C_k^n q^k (1-q)^{n-k} \qquad (4)$$

where q = 0.5 for B_1, 0.75 for B_2 *etc.* Selection on markers is continued each generation, and with good pedigree data, there is no need to test for specific markers if they have been fixed in the parents in previous rounds.

An example. For simplicity, assume an organism with 30 chromosomes, each of unit length, and suppose that we are able to select the top 10% of the population. Consider, first, selection based on a phenotype that differs in the two parents. Let V_E =3, $V_{G(w)}$ = 1, and D = 10. In the B_1 generation, $V(p)$ = 0.001183, and therefore $V_{G(B)}$ = 0.1183, h_B^2 = 0.0287. Then, $S_P = i\sigma_P$ = 3.56, and from (2) we find S_p = 0.0102. In other words, the proportion of the genome that is P_2 among the selected parents has increased only marginally – from 0.75 to 0.76. Alternatively, using marker selection, we choose, using (4), individuals with 19 or more markers homozygous. The mean proportion of the genome selected is obtained by summing, over all chromosomes, the expected proportion homogenic as a function of the marker genotype. In this case, the expected value of p is 0.8008. This example illustrates that selection on marker genotype is much more effective than selecting on a quantitative trait; here, markers produce approximately five times the response.

INTROGRESSION

The principles discussed above apply equally when the purpose of the backcrossing program is to introduce a specific allele from one strain to another. In this case we select, in addition, for heterozygosity at markers in the vicinity of the locus or loci involved. Clearly, the optimum strategy is to choose markers that flank, as closely as possible, the loci of interest. The key point is that the background genotype on the chromosome containing the introgressed locus remains substantially heterozygous - approximately 50% after five generations if the markers are at 0.45 and 0.55 on a chromosome of unit length. Some aspects of marker assisted introgression are discussed in Visscher *et al* (1996), Hospital and Charcosset (1997), and Visscher and Haley (1999), among others.

CONCLUSIONS

The efficiency of a backcrossing program may be enhanced if it is accompanied by selection, either for some phenotype characteristic of the recurrent parent, or for genetic markers. Marker selection appears, in general, to be much superior. At present, the major drawbacks to using markers are the difficulty in choosing enough markers with alleles not shared by the two parental strains, and the costs of typing these markers. However, both are probably temporary problems. As DNA marker technology turns toward single nucleotide polymorphisms (SNPs) and DNA chip technology to score them, we will be able to design haplotypes that differentiate two strains with a high degree of reliability, and to score a large number of these polymorphisms simultaneously.

Finally, I have undervalued the role of selection for a quantitative trait by considering, as the sole goal, the reconstruction of the genome of the recurrent parent. In practice, as for example when trying to reduce the fibre diameter of a flock by strain replacement, the response due to variation within strains may be considerable, and selection is worthwhile if the parents differ little for the trait.

REFERENCES

Franklin, I.R. (1977) *Theor. Pop. Biol.* **11**:60
Hill, W.G. (1993) *J. Heredity* **84**:212
Hospital, F. and Charcosset, A. (1997) *Genetics* **147**: 1469
Stam, P. and Zeven, A.C. (1981) *Euphytica* **30**:227
Visscher, P.M., Haley, C.S. and Thompson, R. (1996) *Genetics* **144**:1923
Visscher, P.M. and Haley, C.S. (1999) *Animal Science* **68**: 59