Made in United States of America
Reprinted from THE JOURNAL OF HEREDITY
Vol. 87, No. 2, March/April 1996

Proportion of the Variation in Genetic Composition in Backcrossing Programs Explained by Genetic Markers

P. M. Visscher

For a randomly mating backcrossing population, the proportion of the variance of genetic composition explained by genetic markers was derived for any generation. It was shown that nearly all the variance can be explained by placing three or more markers per chromosome. It was found that a single marker in the middle of a chromosome explains more variance than two markers at the ends.

Recently, Hill (1993), using results from Franklin (1977), derived the variation in genetic composition in backcrossing populations, that is, the variance in the proportion of the genome that is from the recurrent (or nonrecurrent) population. Breed societies may be interested in knowing how much variation there is in the true proportion of the desired breed in so-called upgrading programs. For example, among animals in a second backcross population (with a mean proportion of desired breed of 87.5%), some genomes may contain 85% and some 90% of the recurrent breed. In this study we extend Hill's results and show what proportion of that variance can be explained by genetic markers. Using markers gives a direct estimate of the mean proportion of the genome that is from the recurrent breed, and the variance explained by the markers gives an estimate of the accuracy of estimation.

We follow Hill's notation as closely as possible. Let t denote the generation of backcrossing, with $t = 1$ being the first backcross generation. $N_{i(l)}$ denotes the proportion of alleles at locus i which is from the nonrecurrent breed in generation t. Hence, $N_{i(l)} = 0$ if neither allele originates from the non-recurrent breed, and equals 1/2 if one allele originates from the nonrecurrent breed. Let N_t be the average of the $N_{i(l)}$ summed over all loci. For a single locus,

$$E(N_{i(l)}) = (\tfrac{1}{2})^{t+1}$$

$$\text{var}(N_{i(l)}) = (\tfrac{1}{4})(\tfrac{1}{2})^t(1 - (\tfrac{1}{2})^t)$$

Hence, for $t = 1$, the mean and variance of the proportion of alleles at a single locus that are from the nonrecurrent breed are 0.25 and 0.0625. For any two loci,

$$\text{cov}(N_{i(l)}, N_{j(l)}) = (\tfrac{1}{4})(\tfrac{1}{2})^t\,[\,(1 - r_{ij})^t - \tfrac{1}{2}^t],$$

where r_{ij} is the recombination fraction between loci i and j (see also Hill 1993).

For the whole genome, $E(N_t) = (\tfrac{1}{2})^{t+1}$, and var($N_t$) was derived by Hill (1993). Assuming Haldane's (1919) mapping function without interference, and an infinite number of loci per chromosome, Hill (1993) showed that the variance in the proportion of the genome from the nonrecurrent population is

$$\text{var}(N_t) = (\tfrac{1}{4})\left[1/(2L^2)(\tfrac{1}{4}^t) \times \sum_{i=1}^{t}\binom{t}{i}(1/i^2) \times \left(2iL - \nu + \sum_{j=1}^{\nu} e^{-2il_j}\right)\right], \quad (1)$$

where L is the total map length in Morgans, $\Sigma l_j = L$, and ν is the number of chromosomes. Without loss of generality, we consider only a single chromosome in subsequent derivations. For the first backcross generation, and for a single chromosome, Equation 1 reduces to

$$\text{var}(N_t) = (\tfrac{1}{4})^2(1 - r_m/L)/\,L, \quad (2)$$

where r_m is the recombination rate between the chromosome ends. Now consider that some of the loci are marker loci, and assume that all marker loci are fully informative. Let $X_{i(l)}$ ($i = 1, \ldots, m$) be an observed value of $N_{i(l)}$, and let X_t be an estimate of N_t, the proportion of the genome from the nonrecurrent breed, based solely on markers. The problem now is how to combine the estimates provided by each marker (since each marker $X_{i(l)}$ is either 0 or ½) into an overall estimate which has the largest correlation with N_t. This problem is analogous to predicting a breeding value of an individual when different sources of phenotypic information are available, and is solved using selection index theory (Hazel 1943). Let **X** be an m × 1 vector of observed $X_{i(l)}$ for an individual, **V** the m × m covariance matrix of **X**, and **y** the m × 1 vector with covariances between $X_{i(l)}$ and N_t. The m × 1 vector **b** contains the weights for each marker. Hence, $X_t = \mathbf{b'X}$, and the optimum index weights are $\mathbf{b} = \mathbf{V}^{-1}\mathbf{y}$.

The proportion of the variance of N_t which is explained by the markers is

$$R^2 = (\mathbf{b'Vb})/\text{var}(N_t) = \mathbf{y'V}^{-1}\mathbf{y}/\text{var}(N_t)$$

To find the values of **b** for a particular marker spacing, we need to know the elements of **V** and **y**. Elements of **V** are straightforward:

$$\mathbf{V}_{ij} = (\tfrac{1}{4})(\tfrac{1}{2})^t\,[(1 - r_{ij})^t - (\tfrac{1}{2})^t] \quad (3)$$

The covariance between marker i, which is at distance d_i from the start of the chromosome (and distance $L - d_i$ from the end of the chromosome), and N_t was derived using Hill's results, assuming Haldane's mapping function without interference, that is, $r_{13} = r_{12} + r_{23} - 2r_{12}r_{23}$ (Haldane 1919), thus

$$\text{cov}(X_{i(l)}, N_t) = y_i = (\tfrac{1}{4})(\tfrac{1}{4})^t(\tfrac{1}{2})\sum_{j=1}^{t}\binom{t}{j}(1/(2j)) \times [2 - e^{-2jd_i} - e^{-2j(L-d_i)}]. \quad (4)$$

We illustrate the preceding equations with a simple example for a single marker positioned in the middle of the chromosome at backcross generation 1. Then,

$$\text{var}(X_{1(1)}) = 1/16,$$

$$\text{cov}(X_{1(1)}, N_1) = (1/16L)(1 - (1 - 2r_m)^{1/2}),$$

and $\text{var}(N_1)$ was shown in Equation (2). Hence,

$$\mathbf{b}_1 = [1 - (1 - 2r_m)^{1/2}]/L \quad \text{and}$$
$$R^2 = [1 - (1 - 2r_m)^{1/2}]^2/(L - r_m)$$

For a chromosome of length $L = 1.0$ (and $r_m = 0.432$), $\mathbf{b}_1 = 0.632$ and $R^2 = 0.704$. Using Equations 1, 3, and 4, the correlation between the estimate of the proportion of the genome that is from the nonrecurrent breed (X_i) and the true proportion of variation in the genome (N_i) can be calculated for any number of markers, for any marker spacing, and for any backcross generation.

When markers are equally spaced along the chromosome, and assuming two markers at the chromosome ends, the relative weights for the end markers are 1/2, and all other markers have a weight of 1 (see Appendix). Optimum marker locations can be found using a search algorithm. Assuming that the optimum marker spacing is symmetric relative to the center of the chromosome (so that the ith and $(m - i + 1)$ marker are at equal distance from the chromosome ends), the search algorithm has to consider placing only $m/2$ (even number of markers) or $m/2 - 1$ (odd number of markers) markers at optimum positions.

Results

We present results for only a single chromosome of length 100 cM. In Table 1, the proportion explained by 1 to 11 evenly spaced markers is shown for backcross generations 1 to 3. For any generation, no selection was assumed in previous generations. More than five fully informative evenly spaced markers per chromosome, corresponding to a marker distance of 25 cM, seems sufficient because the amount of variance explained is greater than 92% for the first three backcross generations. A single marker in the middle of the chromosome explains more of the genetic composition than two markers at the ends, 70% versus 58%, respectively, for generation 1 (Table 1).

The optimum position of m markers along a chromosome was predicted by maximizing R^2 for the locations d_i for all markers. This was done numerically by calculating R^2 for different values of d_i ($i = 1, m/2$), the positions of the first $m/2$ markers. The marker positions were varied at 0.1 cM intervals. For a chromosome of length 100 cM the optimum marker positions are presented in Table 1. Given the end markers, the optimum position of the remaining markers is found by placing them equidistantly (results not shown). Therefore, only the position of the first marker is presented in Table 1. The calculations for the case of 11 markers was done assuming this property, because then the search is only over the first (and last) marker. In general, the optimum spacing of markers improves the R^2 substantially relative to the case of markers at the ends of the chromosome. This is mainly because the markers that were placed at the chromosome ends in the equal spacing scenario do not provide that much information (see also the Appendix). Three to four well-placed markers explain most of the variation in genetic composition in the first three backcross generations.

Table 1. The proportion of the variance in genetic composition (R^2) explained by m markers in a random mating backcross population for equal spacing (with markers at the chromosome ends) and optimum spacing of markers

		Equal spacing	Optimum spacing	
m	BC	$R^2 \times 100$	$R^2 \times 100$	Position of first marker (cM)
1	1	70.4	70.4	50.0
	2	63.4	63.4	50.0
	3	56.5	56.5	50.0
2	1	58.0	88.5	27.5
	2	49.6	84.2	28.0
	3	41.2	79.3	28.6
3	1	86.7	94.2	18.3
	2	81.6	91.7	18.8
	3	75.7	88.6	19.2
4	1	93.9	96.6	13.6
	2	91.0	95.0	13.9
	3	87.6	93.0	14.2
5	1	96.5	97.8	10.8
	2	94.8	96.7	11.0
	3	92.6	95.3	11.2
11	1	99.4	99.5	4.7
	2	99.1	99.3	4.8
	3	98.7	99.0	4.9

Discussion

The effect of the number of evenly spaced markers on the variance in genetic composition they explain was shown in Table 1. Perhaps surprisingly, a single marker in the middle of the chromosome is more informative than two markers at the chromosome ends.

The ideal position of the markers (Table 1) seems approximately the same for any BC generation. Two well-spaced markers provide more information than three equally spaced markers, and in general, $m - 1$ optimally spaced markers seem better than m equally spaced ones. Hospital et al. (1992) found the ideal position for two markers using simulation, and reported that the best position was 20 cM from the end for a chromosome of 100 cM, which is slightly closer from the chromosome ends than our more exact values of 27.5 cM. Our conclusion, that few markers (two to four) provide adequate coverage in a backcrossing program is in agreement with the conclusion of Hospital et al. (1992), who used two well-placed markers in a backcross introgression program to select against the donor genome.

Some major assumptions in this work were that recombination takes place without interference, and that recombination rates do not vary along the chromosome. In practice, these assumptions will be violated. However, these assumptions are unlikely to change the main conclusions of this work, that is, that individual markers can "mark" relatively large chromosome segments. The conclusions of the present study are not affected by a different recombination rate between the sexes because the contribution of the recurrent breed was always fixed (0% of the nonrecurrent breed at each generation). Hence, the conclusions are valid for practical backcross programs in which individuals of the recurrent breed who are mated to individuals from the crossbred population are usually of the same sex.

The results were presented assuming that we are interested in the genetic composition of the whole chromosome (or genome). In practice, the interest may be in particular regions of the genome that are likely to contain genes of interest. However, placing many more markers in gene-dense areas would not dramatically increase the proportion of genetic composition explained for that region, and a complete coverage of the genome does not require many markers. Furthermore, there may still be genes of large effect in regions that are generally thought to contain few genes.

In practice, not all the markers will be fully informative, and if the aim of marker genotyping is to pick up all the variation in backcross generations (for example, in a cross between outbred lines), more than three or four markers per 100 cM may be used to end up with a fully informative

marker map which covers the chromosome adequately.

The results from this study confirm those from simulation studies (Hospital et al. 1992), which showed that markers can be used efficiently in backcross introgression breeding programs where the aim is to introgress a small part of the donor breed and simultaneously recover the recipient genome as quickly as possible. Using 2 to 11 markers per chromosome to select against the donor genome sped up the genome recovery by approximately two generations relative to random selection of individuals with the introgressed gene (Hospital et al. 1992).

The results shown here have implications for quantitative trait loci (QTL) mapping studies. In such experiments two breeds are usually crossed that are very different for a quantitative trait of interest, and marker and phenotypic data from a backcross or F_2 population are analyzed to find evidence for QTL. However, the standard null hypothesis usually is that there are no QTL segregating in a particular region, whereas we know that there must be genes somewhere in the genome which can explain the (large) breed difference. Alternatively, we may wish to test a genetic model of many linked loci which are fixed for alternative alleles in two breeds. Such a model would predict the relative weights of regression coefficients for individual markers if we would perform a multiple regression of phenotypes on all markers on a chromosome. The relative weights (regression coefficients) follow from the results presented in this study. These weights can be tested against the observed regression coefficients, and such a test was found to work well in simulation studies in that it could discriminate between genetic models based on a single QTL and models based on many linked QTLs (Visscher and Haley, in press).

Appendix

Relative Index Weights for Equally Spaced Markers

For the first backcross generation, ignoring subscripts for $t = 1$, let $X = \mathbf{b'X}$ be an index of individual marker scores, with $\mathbf{b}$ an m × 1 vector of weights for marker scores X_i ($X_i = 0$ or 1/2), and $\mathbf{X}$ an m × 1 vector with observed values X_i. Markers are assumed to be evenly spaced along a chromosome with length L (Morgans). The index weights $\mathbf{b}$ are calculated so as to maximize the correlation between X and N. Index weights are calculated as $\mathbf{b} = (\mathbf{var(X)})^{-1}\mathbf{y} = \mathbf{V}^{-1}\mathbf{y}$, with $\mathbf{y}$ a vector of covariances with $\mathbf{y}_i = \mathrm{cov}(X_i,N)$.

To show that the weights of b_1 and b_m are 1/2 relative to the weights for the other markers, we first show that $\mathbf{V}^{-1}$ is tridiagonal for the first backcross generation. Because we assume evenly spaced marker loci, all elements of $\mathbf{V}$ are functions of r, the recombination rate between adjacent marker loci. The matrix $\mathbf{V}$ has a special form in that for a particular value of $|i - j|$ all elements are identical. For example, for $|i - j| = 0$, $V_{ij} = 1/16$, for $|i - j| = 1$, $V_{ij} = (1/8)(1/2 - r)$, and for $|i - j| = 2$, $V_{ij} = (1/8)(1/2 - 2r(1 - r))$. In general, $V_{ij} = (1/16)[(e^{-2d})^{|i-j|}]$, with $d = L/(m - 1)$. The matrix $\mathbf{V}$ is an example of an autoregressive matrix, which means that element $\mathbf{V}_{ij}$ is proportional to the product of elements $\mathbf{V}_{ik}$ ($k = 1, j - 1$). This is true also if markers are not spaced evenly along the chromosome.

Because of the special form of $\mathbf{V}$ (an autoregressive matrix), its inverse is always a tridiagonal matrix with elements

$$V^{11} = V^{mm} = 16/(1 - e^{-4d}),$$

$$V^{ii} = 16(1 + e^{-4d})/(1 - e^{-4d}) \quad (\text{for } i > 1 \text{ and } i < m),$$

$$V^{ij} = -16e^{-2d}/(1 - e^{-4d}) \quad (\text{for } |i - j| = 1), \text{ and}$$

$$V^{ij} = 0 \quad (\text{for } |i - j| > 1).$$

The covariance between X_k and Z is

$$y_k = [\tfrac{1}{2}(1 - e^{-2d(k-1)}) + \tfrac{1}{2}(1 - e^{-2d(m-k)}]/(16L)$$

Multiplying $\mathbf{V}^{-1}$ with $\mathbf{y}$ gives, to a constant of proportionality,

$$b_1 = b_m \propto \tfrac{1}{2}(1 - 2e^{-2d} + e^{-4d}), \quad \text{and}$$

$$b_k \propto (1 - 2e^{-2d} + e^{-4d}), \quad \text{and}$$

$$b_k/b_1 = b_k/b_m = 2.$$

For $t > 1$, the inverse of $\mathbf{V}$ is not tridiagonal, but off-diagonals for $|i - j| > 1$ are relatively small, and the relative weights for b_1 and b_m are very close to 1/2. This was found empirically by calculating the regression coefficients for various combinations of t and m.

From the Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, Scotland. This work was funded by the Marker Assisted Selection Consortium of the U.K. pig industry (Cotswold Pig Development Company Ltd., J.S.R. Farms Ltd., National Pig Development Company, Newsham Hybrid Pigs Ltd., Pig Improvement Company, and the Meat and Livestock Commission) and by MAFF, DTI, and the BBSRC. I thank Robin Thompson, Chris Haley, Bill Hill, Sara Knott, and Naomi Wray for many helpful comments and discussions.

References

Franklin IR, 1977. The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. Theor Pop Biol 11:60–80.

Haldane JBS, 1919. The combination of linkage values and the calculation of distances between the loci of linked factors. J Genet 8:299–309.

Hazel, 1943. The genetic basis of constructing selection indices. Genetics 28:476–490.

Hill WG, 1993. Variation in genetic composition in backcrossing programs. J Hered 84:212–213.

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

Visscher PM, Haley CS, in press. Detection of quantitative trait loci in line crosses under infinitesimal genetic models. Theor Appl Genet.

Received February 22, 1995
Accepted September 6, 1995

Corresponding Editor: Leif Andersson