Genet. Res., Camb. (1997), **69**, *pp.* 137–144. *With 4 figures.* Printed in the United Kingdom © 1997 Cambridge University Press

Optimal weighting of information in marker-assisted selection

J. C. WHITTAKER[1]*, C. S. HALEY[2] AND R. THOMPSON[2,3]

[1] *Department of Applied Statistics, University of Reading, Whiteknights Road, PO Box 240, Reading RG6 2FN, UK*
[2] *Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK*
[3] *Rothamsted Experimental Station, Harpenden, Herts AL5 2JQ, UK*

(*Received 27 August 1996 and in revised form 20 January 1997*)

Summary

In crosses between inbred lines linear regression can be used to estimate marker effects; these marker effects then allow marker-assisted selection (MAS) for quantitative traits. Weighting of marker and phenotypic information in MAS requires estimation of genetic variance associated with the markers: the usual estimators are biased, resulting in too much weight being placed on marker information relative to phenotypic information. In this paper we develop a cross-validation method to remove this bias, and show by simulation that response to selection using this method is almost as high as that achieved using optimal weighting of marker and phenotypic information.

1. Introduction

Lande & Thompson (1990) suggested that the linkage disequilibrium between genetic markers and quantitative trait loci (QTL) created when two inbred lines are crossed could be used to facilitate marker-assisted selection (MAS). They used multiple regression of phenotype on marker-type to select markers associated with the trait through linked QTL, estimated the effect on the trait associated with these selected markers, and then combined these marker effects with phenotypic information using a selection index. Computer simulations (Gimelfarb & Lande, 1994a; Whittaker *et al.*, 1995) have shown that the method is more effective than selection on phenotype alone when population sizes are large and heritability low. Zhang & Smith (1992, 1993) obtained similar results when comparing selection on the BLUP estimate of an individual's genetic value with selection based on an index combining marker effects and the BLUP estimate.

However, problems arise in calculating the relative weighting of marker and phenotypic information, because the same data are used to select the markers that affect the trait and to estimate relative weights given to marker and phenotypic information. This leads to overestimation of the magnitude of the marker effects and overestimation of the variance explained by the markers, so that too much weight is put on the marker score relative to the phenotypic information, and selection response is reduced. We describe these problems more fully in Section 2, and suggest a solution based on cross-validation (Efron & Tibshirani, 1993). Simulation results indicate that this solution works well and gives a useful improvement in selection response in comparison with existing methods.

2. Methods

We shall consider a cross between two inbred lines, each assumed homozygous (for different alleles) at all loci. We label the alleles at the ith QTL in the first line Q_i, and the alleles at the jth marker locus M_j. The corresponding alleles in the second line are labelled q_i and m_j.

For each individual in the population we know the phenotype y and the number of M_i alleles at the ith marker locus, x_i, so that the marker genotype of an individual is described by $\mathbf{x} = (x_1, x_2, \dots, x_n)$. From these we wish to construct an estimate $\hat{z}$ of the genetic value of the individual, z. Ideally we would use the regression of z on y and $\mathbf{x}$, $E(z \mid y, \mathbf{x})$, but this is difficult to evaluate (Whittaker *et al.*, 1995). Lande & Thompson (1990) suggested using a linear approximation, so that

$$\hat{z} = b_0 y + b_1 s$$

* Corresponding author. Telephone: +44 (0)1734 318023. e-mail: j.c.whittaker@reading.ac.uk.

https://doi.org/10.1017/S0016672397002711 Published online by Cambridge University Press

where, for any individual, the marker score s is given by

$$s = \beta_0 + \sum_{i \in \mathscr{A}} \beta_i x_i.$$

Here β_i is the additive effect associated with the ith marker and $\mathscr{A}$ is the set of markers for which effects have been fitted. The effects β_i are fitted by multiple linear regression, that is by minimizing

$$\sum_{j=1}^{n} (y_j - \beta_0 - \sum_{i \in \mathscr{A}} \beta_i x_i^j)^2,$$

where y_j and $\mathbf{x}^j$ are the phenotype and marker-type respectively of the jth individual. Whittaker *et al.* (1996) showed that in F_2 populations

$$E(z \mid \mathbf{x}) = \beta_0 + \sum_{i \in \mathscr{A}} \beta_i x_i,$$

and in the Appendix we show that this also holds for generations subsequent to the F_2 provided there is no selection, so the main approximation here is in the linear combination of marker and phenotypic information. An algorithm to select the set of important markers $\mathscr{A}$ is necessary: we use the one based on Mallow's C_p described by Whittaker *et al.* (1995).

Expressions for the relative weights of marker and phenotypic information **b** are in principle easily calculated (Lande & Thompson, 1990), for by standard theory (Falconer, 1989)

$$\mathbf{b} = \mathbf{P}^{-1}\mathbf{G}$$

where

$$\mathbf{P} = \begin{bmatrix} \operatorname{var}(y) & \operatorname{cov}(y,s) \\ \operatorname{cov}(y,s) & \operatorname{var}(s) \end{bmatrix}$$

and

$$\begin{aligned}\mathbf{G} &= (\operatorname{cov}(z,y), \quad \operatorname{cov}(z,s))^T \\ &= (\operatorname{var}(z), \quad \operatorname{cov}(z,s))^T.\end{aligned}$$

However, for this to be of use we need estimates of $\operatorname{var}(z)$ and $\operatorname{cov}(z,s)$. Previous work (Lande & Thompson, 1990; Gimelfarb & Lande, 1994*a*) has tended to estimate $\operatorname{cov}(z,s)$ by $\operatorname{cov}(y,s)$, arguing that all the phenotypic variance explained by the markers is by definition genetic. However, as noted above, a variable selection technique has been used to choose the subset of markers to include in the marker score s. Markers are selected because they explain a high proportion of the phenotypic variance so using the same data to select markers and to estimate marker effects clearly leads to $\operatorname{cov}(y,s)$ overestimating $\operatorname{cov}(z, s)$ and hence to overestimation of the weight to be placed on marker score in the selection index. Indeed, this leads to virtually all weight being placed on the marker score (Whittaker *et al.* 1995; Gimelfarb & Lande, 1994*a*), so that the index is effectively equivalent to selection on the markers alone.

Zhang & Smith (1992, 1993) avoid this problem by generating two independent sets of F_2 data from the same population, applying their marker selection procedure to one to give $\mathscr{A}$ and then obtaining unbiased marker effects for this $\mathscr{A}$ from the other set of data. The same set of markers $\mathscr{A}$ is then used in all subsequent generations, so that bias in estimates due to the marker selection procedure is eliminated. However, this is clearly not an efficient use of data: much of the information from the first of these F_2 data sets is wasted. Also, Gimelfarb & Lande (1994a) have shown that MAS is more efficient if the marker selection procedure is repeated every generation. We shall repeat the marker selection procedure every generation whilst estimating $\operatorname{cov}(z,s)$ using cross-validation (Shao, 1993).

Cross-validation estimates are constructed as follows. We split the data set into two parts, denoted by the set $\mathbf{S}$ and its complement $\mathbf{S}^c$ respectively. The data in $\mathbf{S}$ are used to select and estimate marker effects; these marker effects are then used to calculate marker scores s for the individuals in $\mathbf{S}^c$, and $\operatorname{cov}(z,s)$ estimated by $\operatorname{cov}(y,s)$ calculated over all individuals in $\mathbf{S}^c$. There is no bias here, because no data point contributing to $\operatorname{cov}(y,s)$ was used in the marker selection procedure. Averaging over a number of sets $\mathbf{S}$ gives the cross-validation estimate of $\operatorname{cov}(z,s)$.

This gives a range of possible cross-validation estimates, varying in the choices of $\mathbf{S}$ used. We minimize the computational cost of the estimation procedure by splitting the data into two parts of equal size, and using each half as $\mathbf{S}$ in turn. This gives the following procedure. The data are divided into two halves, to one of which the variable selection procedure is applied to select a set of markers and produce estimates of marker effects. For every individual in the other half of the data these marker effects are used to calculate a marker score $s^{(1)}$ and, using this section of data only, $\operatorname{cov}(y^{(1)}, s^{(1)})$ calculated. The process is then repeated, swapping the roles of the sections of data, to get a score $s^{(2)}$ for the other section of data and hence $\operatorname{cov}(y^{(2)}, s^{(2)})$. *All* the data are then used to select $\mathscr{A}$ and produce the marker effects used in calculating the marker scores s actually used in selection. We could use

$$0{\cdot}5(\operatorname{cov}(y^{(1)}, s^{(1)}) + \operatorname{cov}(y^{(2)}, s^{(2)}))$$

to estimate $\operatorname{cov}(z,s)$, but this will result in an underestimate because we would expect to be able to select markers explaining more of the variation in z using all the data than using any part of it. We make a rough correction using $n, n^{(1)}$ and $n^{(2)}$, the number of markers used in calculating $s, s^{(1)}$ and $s^{(2)}$ respectively, to get the estimate of $\operatorname{cov}(z,s)$:

$$0{\cdot}5n\left(\frac{\operatorname{cov}(y^{(1)}, s^{(1)})}{n^{(1)}} + \frac{\operatorname{cov}(y^{(2)}, s^{(2)})}{n^{(2)}}\right).$$

This correction is *ad hoc* but seems to result in satisfactory performance.

https://doi.org/10.1017/S0016672397002711 Published online by Cambridge University Press

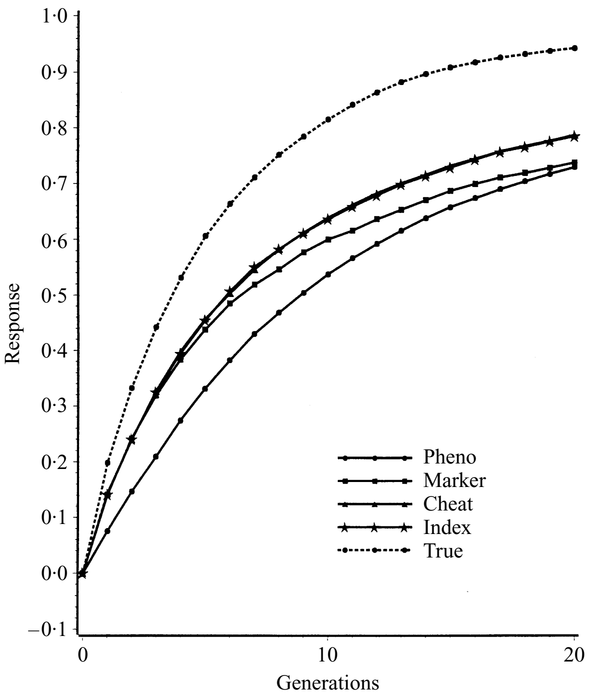

Fig. 1. Selection response for map 1 (20 chromosomes, 100 QTL, positive and negative alleles allocated at random between the two lines as described in Section 3), $n = 400$, $h^2 = 0{\cdot}1$.

Two methods for the estimation of var(z) were considered. The first assumed that the environmental variance σ_e^2 was known, for instance from a very large F_1 generation, with var(z) estimated by

$$\text{var}(y) - \sigma_e^2.$$

The second method is based on the use of family structure. The simulations in Section 3 are based on full sib families. For such a family structure let f be the number of sibs in a family and $\bar{y}_f$ be the mean phenotypic value of such a group of sibs. Then,

$$\text{var}(y-\bar{y}_f) = \text{var}(y) - \text{cov}(y, \bar{y}_f)$$

because

$$\text{cov}(y, \bar{y}_f) = \text{var}(\bar{y}_f)$$

and

$$\text{cov}(y, \bar{y}_f) = \frac{1}{f}[\text{var}(y) + (f-1)\,\text{cov}(y, y_s)]$$

where cov(y, y_s) is the covariance between the phenotypic values of sibs. If we write ρ for the correlation between the genetic values of sibs we have

$$\text{cov}(y, y_s) = \text{cov}(z, z_s) = \rho\,\text{var}(z)$$

and this implies that

$$\text{var}(z) = \frac{1}{\rho}\left[\text{var}(y) - \frac{f}{f-1}\text{var}(y-\bar{y}_f)\right].$$

We can calculate the phenotypic variance var(y) and the variance of family deviations var$(y-\bar{y}_f)$ but ρ is unknown. In the absence of selection, $\rho = 0{\cdot}5$, but this is not true in selected populations. Ignoring changes in QTL frequencies (i.e. assuming an infinitesimal genetic model), selection reduces the between-family variance whilst leaving the within-family variance unchanged. Here we will have changes in gene frequency but we still expect that in selected populations $\rho < 0{\cdot}5$. However, estimation of ρ has proved difficult so we have ignored the effect of selection and assumed $\rho = 0{\cdot}5$ in all generations. There is no family information in the F_2 so we have selected on markers alone in the F_2 generation.

A priori we expect to put positive weight on both marker score and phenotypic value. We therefore allow only positive values of b_0 and b_1. If, for instance, $b_0 < 0$ we set $b_0 = 0$ and $b_1 = 1$ so that selection is exclusively on marker score. If both $b_0 < 0$ and $b_1 < 0$ we set $b_0 = 1$ and $b_1 = 0$ so that selection is exclusively on phenotype.

3. Simulations

Four methods were compared using computer simulations. They were:

Selection based solely on an individual's phenotypic value (PHENO).

Selection based solely on the marker score s (MARKER).

Selection based on the combination of s and y in an index as described in Section 2, with the optimal, but in practice unknown, weights used in the selection index. That is, the true values of cov(z, s) and var(z) are used in calculating the weights **b**, with the marker scores s estimated as above. Note that by true values of cov(z, s) and var(z) we mean the actual values in the finite simulated population rather than the expected infinite population values (CHEAT).

Selection based on the combination of s and y in an index as described in Section 2, with cov(z, s) calculated using the cross-validation method described there and var(z) estimated using family structure (INDEX).

For each method the model-fitting procedure was repeated every generation so that markers for which effects are fitted may change with time. Note that CHEAT should give an upper bound to the methods discussed here. To improve on CHEAT we must either incorporate additional information, for example by using information on relatives to augment the phenotype information, or improve the marker selection and estimation procedure. In order to get a rough idea of the increase in selection response that is possible by improving the marker selection and estimation procedure, we considered selection on

https://doi.org/10.1017/S0016672397002711 Published online by Cambridge University Press

marker score where the individual marker effects are obtained by regressing genetic value on marker-type rather than regressing phenotypic value on marker-type. This should produce marker effects which are almost, but not exactly, optimal. Selection response obtained using this method (TRUE) will therefore give an upper bound to the performance of MAS. Note that this upper bound will be less than that obtained by direct selection on genetic value.

Simulations were done using the two maps described by Whittaker *et al.* (1996). For both maps QTL are assumed to combine additively both between and within loci. The first had 20 chromosomes, each of length 1 morgan; 5 marker loci were spaced evenly along each chromosome, with a marker located at each end of every chromosome. Using the Haldane (1919) mapping function this gives the probability of recombination between two adjacent markers on the same chromosome to be 0·1967. Locations for 100 QTL were chosen from a uniform distribution, with the effect of these QTL, a_i, for $i = 1, 2, \ldots, 100$ generated assuming that the amount of additive genetic variance due to QTL may be approximated by a power series, as in Lande & Thompson (1990). Positive and negative alleles were allocated at random between the two lines. We refer to this map as map 1.

In addition we used the map from Gimelfarb & Lande (1994a); this has 110 markers evenly spread over 10 chromosomes of length 1 morgan, with 25 QTL placed randomly on the map. Using the Haldane mapping function this gives the probability of recombination between two adjacent markers on the same chromosome to be 0·0906. We refer to this map as map 2. As in Gimelfarb & Lande (1994a), simulations of two types were run. In the first, 'total coupling', the effects of QTL were in the same direction; that is one of the initial lines has all the alleles with positive effect on that chromosome and the other line all the negative effects. In 'total repulsion', QTL effects alternate in sign along the genome. Neither total repulsion nor total coupling is likely to occur in reality: they represent extreme cases and are included here to show the range of possible behaviour.

Note that results are grouped according to the heritability in an F_2 population. The fact that QTL on the same chromosome are positively correlated in coupling phase but negatively correlated in repulsion phase means that, with environmental variance fixed, QTL effects must be larger in repulsion than in coupling to give the same heritability in the F_2.

Simulations were run for 20 generations with heritabilities 0·1 and 0·2, and 200 and 400 individuals of each sex. The number of replicates was varied with the population size: 60 for 200 individuals and 40 for 400 individuals. In every generation the top 20% of individuals of each sex was selected and paired at random: each pair is then assumed to produce five offspring of each sex.

4. Results and discussion

The results obtained are shown in Figs. 1–4 for 400 individuals of each sex and heritabilities of 0·1 and 0·2. Results for 200 individuals of each sex are similar in that the ordering of the methods is unchanged, though the differences between the methods are reduced;

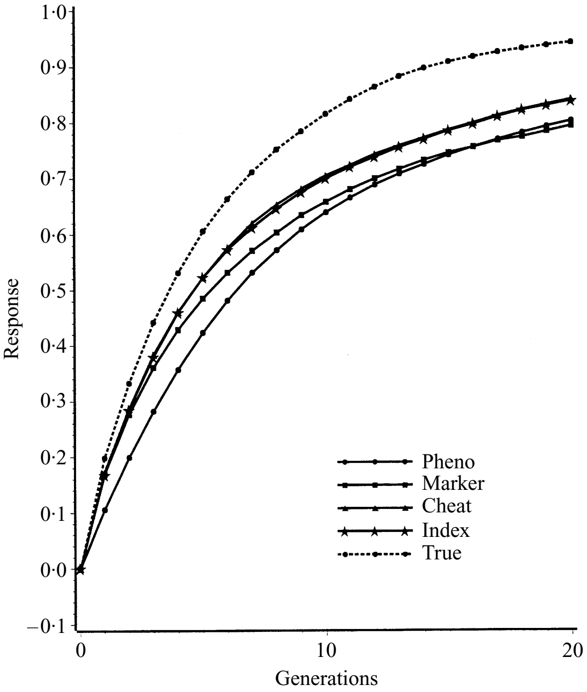

Fig. 2. Selection response for map 1 (see caption for Fig. 1), $n = 400$, $h^2 = 0{\cdot}2$.

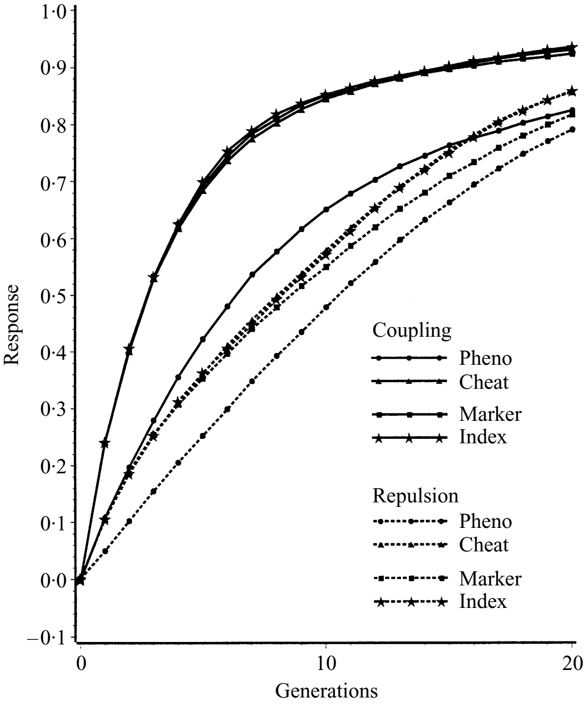

Fig. 3. Selection response for map 2 (10 chromosomes, 25 QTL, positive and negative alleles allocated in coupling or repulsion as described in Section 3), $n = 400$, $h^2 = 0{\cdot}1$.

https://doi.org/10.1017/S0016672397002711 Published online by Cambridge University Press

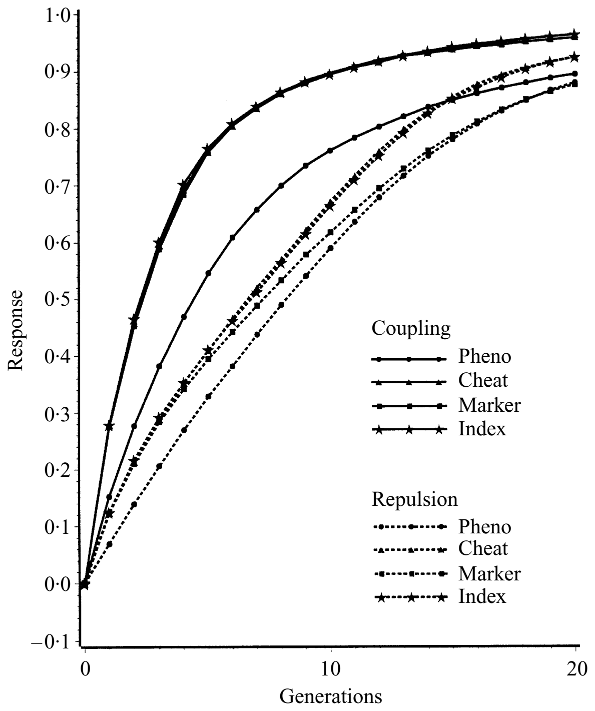

Fig. 4. Selection response for map 2 (see caption for Fig. 3), $n = 400$, $h^2 = 0{\cdot}2$.

Table 1. *Standard errors of accumulated selection response for map 2, repulsion* (*see caption for Fig. 3*), $n = 400$, $h^2 = 0{\cdot}1$

gen	PHENO	MARKER	CHEAT	INDEX
0	0·0034	0·0034	0·0034	0·0034
1	0·0083	0·0105	0·0092	0·0105
2	0·0121	0·0153	0·0155	0·0142
3	0·0156	0·0213	0·0153	0·0176
4	0·0163	0·0220	0·0184	0·0199
6	0·0193	0·0265	0·0263	0·0241
8	0·0223	0·0321	0·0345	0·0319
10	0·0266	0·0387	0·0403	0·0316
15	0·0258	0·0593	0·0346	0·0465
20	0·0232	0·0617	0·0365	0·0349

these results are therefore not shown here. Standard errors of accumulated selection response for map 2 in repulsion, 400 individuals of each sex, $h^2 = 0{\cdot}1$ are given in Table 1. Standard errors for other maps and other values of n and h^2 are similar. Results for TRUE have been given only for map 1 for the sake of clarity: results for map 2 are similar. In all cases results are given as percentages of the maximum genetic value obtainable, that is the genetic value of an individual possessing all favourable alleles.

As usual, the marker-assisted methods are increasingly favoured, in comparison with selection on phenotype, by increasing population size and decreasing heritability. The advantage of INDEX over PHENO and MARKER is clear for all parameter values with map 1 and map 2 in repulsion mode, with CHEAT in most cases performing marginally better than INDEX. For map 2 in coupling mode, for all parameter values, CHEAT, MARKER and INDEX are virtually indistinguishable, all performing much better than PHENO. This is reasonable: we know MARKER works very well for map 1 (Whittaker *et al.*, 1995; Gimelfarb & Lande, 1994*a*) because it can identify good and bad chromosomes instead of having to separate linked good and bad QTL, so we might expect that phenotypic information is of little benefit in selection using this map. In all cases TRUE does dramatically better than all other methods; clearly there is considerable room for improvement in marker selection and estimation. Comparison with the results in Whittaker *et al.* (1995) is interesting. MGLMAS in Whittaker *et al.* (1995) used the approach described in Lande & Thompson (1990) to combine marker and phenotypic information and, as explained above; this causes the mean selection index coefficient for marker score to be very high relative to that for phenotypic value. Thus we might expect the performance of MGLMAS to be close to that of MARKER. In fact, selection response using MGLMAS is about midway between that obtained using MARKER and that obtained using INDEX. Presumably this is because the mean selection index coefficients are distorted by a few very large values, with MGLMAS assigning sensible weights to phenotype a reasonable proportion of the time.

It is a little surprising to find that INDEX does so well relative to CHEAT, particularly given that the weights given to phenotypic and marker information in INDEX can be quite different from the optimal weights (Table 2). It seems that selection is reasonably robust to errors in the selection index weights. This is presumably because marker score and phenotype are correlated so that the position is similar to selection indices involving information on relatives, where Sales & Hill (1976) have shown that selection is reasonably robust to errors in the selection index weights.

We know (e.g. Whittaker *et al.* 1995) that the relative efficiency of MARKER and PHENO changes with time, PHENO doing relatively better in later generations. This is partly a consequence of MARKER being more successful in fixing QTL than PHENO in early generations, so that the heritability in populations selected using MARKER is lower than that in populations selected using PHENO, and partly due to recombination eroding marker QTL correlation. We would expect weight to be transferred from marker score to phenotype with time, and Table 2 confirms that this does happen.

The other notable feature of Table 2 is the number of times one of the components of **b** would be negative if not for the constraint discussed in Section 2. Since this is a much rarer occurrence for the optimal weights, it would seem that the major cause is error in the variance estimation procedure: as the genetic variance is very low in later generations, small errors

https://doi.org/10.1017/S0016672397002711 Published online by Cambridge University Press

Table 2. *Estimated and optimal selection index coefficients, together with the number of runs for which each was negative for map 2, repulsion,* $n = 400$, $h^2 = 0{\cdot}1$

	Estimated					Optimal				
Gen	b_1	b_2	$b_1 < 0$	$b_2 < 0$	$b_1, b_2 < 0$	b_1	b_2	$b_1 < 0$	$b_2 < 0$	$b_1, b_2 < 0$
1	0·0762	0·9238	21	1	1	0·0677	0·9323	0	0	0
2	0·0720	0·9280	18	0	0	0·0831	0·9169	0	0	0
3	0·1643	0·8357	13	2	1	0·0897	0·9103	0	0	0
4	0·1878	0·8122	9	2	1	0·1090	0·8910	0	0	0
6	0·3011	0·6989	11	6	1	0·1389	0·8611	0	0	0
8	0·2149	0·7851	11	5	0	0·1525	0·8475	0	0	0
10	0·2737	0·7263	10	6	0	0·1485	0·8515	0	0	0
15	0·3653	0·6347	7	8	2	0·1724	0·8276	0	0	0
20	0·3614	0·6386	13	11	1	0·2422	0·7578	0	3	0

Table 3. *Actual and estimated genetic variances and and covariances for map 2, repulsion,* $n = 400$, $h^2 = 0{\cdot}1$, *with root mean square errors (r.m.s.e.); ρ is the correlation between the genetic values of sibs*

Gen	Actual var (z)	Est. var (z)	r.m.s.e.	ρ	Actual cov (z, x)	Est. cov (s, z)	r.m.s.e.	cov (y, z)
1	0·0847	0·0461	0·0536	0·3572	0·0572	0·0545	0·0225	0·1233
2	0·0765	0·0578	0·0432	0·3796	0·0489	0·0520	0·0239	0·1177
3	0·0664	0·0546	0·0402	0·3837	0·0373	0·0429	0·0239	0·1083
4	0·0632	0·0302	0·0498	0·4110	0·0304	0·0263	0·0234	0·0909
6	0·0554	0·0406	0·0331	0·4268	0·0239	0·0235	0·0177	0·0889
8	0·0548	0·0507	0·0396	0·4439	0·0241	0·0274	0·0228	0·0922
10	0·0539	0·0406	0·0336	0·4349	0·0212	0·0201	0·0224	0·0849
15	0·0378	0·0289	0·0368	0·4504	0·0146	0·0148	0·0148	0·0745
20	0·0181	0·0166	0·0305	0·4678	0·0065	0·0088	0·0182	0·0643

in estimation can cause the variance estimates to become negative and this leads to one of the components of **b** becoming negative. Similar constraints on the components of **b** were used in MGLMAS (Whittaker *et al*., 1995).

Table 3 gives variance estimates and actual values for one parameter set, together with the root mean square error (r.m.s.e.) of the estimates. The r.m.s.e. of an estimate $\hat{x}$ of x over n replicates is

$$\sqrt{\left[\frac{1}{n}\sum_{i=1}^{n}(\hat{x}-x)^2\right]}$$

and is a more appropriate measure of variability than standard error here because the variances we are trying to estimate are themselves random variables. It seems that our cross-validation estimate of cov (z, s) is approximately unbiased but very variable. It is a much better estimate of cov (z, s) than cov (y, s): cov (y, s) gives estimates that are heavily biased upwards, particularly in later generations when we may have cov $(y, s) >$ var (z). The estimates of var (z) are biased downwards because we have assumed in calculating them that the correlation between the genetic values of sibs is 0·5, which is greater than the true values. Again, these estimates are rather variable. Estimating genetic variance by var $(y) - \sigma_e^2$ gave better variance estimates but had no effect on selection response. We would expect that downward bias in our estimates of var (z) would result in b_1 being reduced from its optimal value, whereas we can see from Table 2 that in fact it is increased. This is because the constraint that b_i be non-negative pushes the estimated selection index coefficients towards putting equal weight on phenotype and marker-type. Runs with **b** unconstrained resulted in lower selection response, presumably because errors in estimation caused negative weight to be put on phenotype or marker score when the selection index coefficients should have been positive.

Fortunately, these estimation problems do not affect the selection response greatly in the situations discussed here. This may not be true for other maps or parameter values, so we shall discuss briefly ways of improving the variance estimates. Firstly, consider the estimation of the covariance of marker score and genetic value. This is essentially equivalent to estimating the prediction error in regression models, a problem on which there is a large literature (Breiman, 1992). The most promising approaches have been data re-use methods such as the boot-strap or cross-validation: we have used a very simple form of cross-validation, but it seems likely that more sophisticated methods such as the balanced incomplete cross-validation of Shao (1993), the 0·632 bootstrap (Efron & Tibshirani, 1993) or Breiman's (1992) little bootstrap would give better variance estimates, though with much higher computational cost. It is rather harder to see how the estimates of genetic variance could be improved.

5. Conclusions

If correctly weighted, both marker and phenotypic information are of value in MAS, with marker information important in early generations and phenotypic information of greater importance later on. Estimation of weights is difficult; the methods discussed here are far from ideal, but are good enough to produce better response than selection on either markers or phenotypic information alone in most of the situations considered. Indeed, it seems that the performance of MAS is robust with respect to the relative weighting of marker and phenotypic information, and this means that methods discussed here give almost as great a selection response as the optimal weighting. It would therefore seem that any further improvements in selection response under MAS must be gained by improving the components of the selection index. One way to do this is to incorporate information on relatives, for instances by using BLUP methodology (Zhang & Smith, 1992, 1993) or selection indices. We believe the cross-validation approach developed here would be of value in correctly weighting phenotypic and marker information in more complex models, but have used linear fixed effect models here for the sake of simplicity and computational ease. Alternatively, selection response can be increased by improving the procedure used to select and estimate marker effects, for instance by using information on previous generations in this procedure. We hope to discuss these points elsewhere.

Appendix. Linearity of $E(z\,|\,\mathbf{x})$

Consider a QTL flanked by markers, with recombination fractions r_L and r_R between the QTL and the left and right flanking markers respectively and recombination fraction θ between the markers. Let the number of Q alleles at the QTL be $g+1$ and the number of M_L or M_R alleles at the left- and right-hand flanking markers be x_L+1 and x_R+1 respectively. Then writing $P_t(ABC)$ for the frequency of ABC gametes in the F$_t$th generation, where A is M_L or m_L, B is Q or q and C is M_R or m_R, we have in an infinite population with no selection and random mating

$$\begin{aligned}P_t(ABC) = {} & (1-\theta-r_L r_R)\,P_{t-1}(ABC)\\ & +r_L(1-r_R)\,P_{t-1}(BC)\,P_{t-1}(A)\\ & +r_R(1-r_L)\,P_{t-1}(AB)\,P_{t-1}(C)\\ & +r_R r_L\,P_{t-1}(AC)\,P_{t-1}(B) \qquad \text{(A 1)}\end{aligned}$$

for $t > 2$. Now suppose that

$$\begin{aligned}P_{t-1}(MQm) &= P_{t-1}(mqM)\\ P_{t-1}(mQm) &= P_{t-1}(MqM)\\ P_{t-1}(mQM) &= P_{t-1}(Mqm)\\ P_{t-1}(MQM) &= P_{t-1}(mqm). \qquad \text{(A 2)}\end{aligned}$$

It is easy to check that (A 1) implies that corresponding relations hold at generation t also. All F$_1$ individuals have one MQM and one mqm gamete, so that

$$\begin{aligned}P_2(MQm) &= P_2(mqM) = 1-\theta-r_R r_L\\ P_2(mQm) &= P_2(MqM) = r_L(1-r_R)\\ P_2(mQM) &= P_2(Mqm) = r_R(1-r_L)\\ P_2(MQM) &= P_2(mqm) = r_R r_L\end{aligned}$$

and, by induction, (A 2) holds for all $t \geqslant 2$. It follows immediately that

$$\begin{aligned}P_t(Q\,|\,Mm) &= P_t(q\,|\,mM)\\ P_t(Q\,|\,mm) &= P_t(q\,|\,MM)\\ P_t(Q\,|\,mM) &= P_t(q\,|\,Mm)\\ P_t(Q\,|\,MM) &= P_t(q\,|\,mm). \qquad \text{(A 3)}\end{aligned}$$

Let

$$\begin{aligned}\lambda &= P(Q\,|\,Mm)-P(q\,|\,Mm)-P(Q\,|\,mm)+P(q\,|\,mm)\\ \rho &= P(Q\,|\,mM)-P(q\,|\,mM)-P(Q\,|\,mm)+P(q\,|\,mm)\end{aligned}$$

where the t subscripts have been suppressed for convenience. It is easy to check using (A 3) that

$$E(g\,|\,x_L\,x_R) = \lambda x_L+\rho x_R,$$

and extending to multiple QTL and flanking markers as in Whittaker *et al.* (1996) we see that

$$E(z\,|\,\mathbf{x}) = \beta_0+\sum_{i\in\mathscr{A}}\beta_i x_i,$$

as required.

C.S.H. acknowledges support from MAFF, and J.C.W. from BBSRC. We thank Robert Curnow and Peter Visscher for helpful discussions of this work.

References

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: *X*-fixed prediction error. *Journal of the American Statistical Association* **87**, 738–754.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Falconer, D. S. (1989). *Introduction to Quantitative Genetics*, 3rd edn. New York: Longman.

Gimelfarb, A. & Lande, R. (1994*a*). Simulation of marker-assisted selection in hybrid populations. *Genetical Research, Cambridge* **63**, 39–47.

Gimelfarb, A. & Lande, R. (1994*b*). Simulation of marker assisted selection for non-additive traits. *Genetical Research, Cambridge* **64**, 127–136.

https://doi.org/10.1017/S0016672397002711 Published online by Cambridge University Press

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distance between loci of linked factors. *Journal of Genetics* **8**, 299–309.

Lande, R. & Thompson, R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**, 743–756.

Sales, J. & Hill, W. G. (1976). Effect of sampling errors on efficiency of selection indices. 1. Use of information from relatives for single trait improvement. *Animal Production* **22**, 1–17.

Shao, J. (1993). Linear model selection via cross-validation. *Journal of the American Statistical Association* **88**, 486–494.

Whittaker, J. C., Curnow, R. N., Haley, C. S. & Thompson, R. (1995). Using marker-maps in marker-assisted selection. *Genetical Research*, *Cambridge* **66**, 255–265.

Whittaker, J. C., Thompson, R. & Visscher, P. (1996). On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77**, 23–32.

Zhang, W. & Smith, C. (1992). Computer simulation of marker-assisted selection utilizing linkage disequilibrium. *Theoretical and Applied Genetics* **83**, 813–820.

Zhang, W. & Smith, C. (1993). Simulation of marker-assisted selection utilizing linkage disequilibrium: the effects of several additional factors. *Theoretical and Applied Genetics* **86**, 492–496.

https://doi.org/10.1017/S0016672397002711 Published online by Cambridge University Press