Genet. Res., Camb. (1995), **66**, *pp.* 255–265 *With* 4 *text-figures* Copyright © 1995 Cambridge University Press

Using marker-maps in marker-assisted selection

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

[1] *Department of Applied Statistics, University of Reading, Reading RG6 2FN, UK*

[2] *Rosin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, UK*

(*Received 11 April 1995 and in revised form 19 June 1995*)

Summary

A method of using information on the location of markers to improve the efficiency of marker-assisted selection (MAS) in a population produced by a cross between two inbred lines is developed. The method is closer to mapping QTL than the selection index approaches to MAS described by previous authors. We use computer simulations to compare our method with phenotypic selection and two selection index approaches, simulations being performed on three genetic maps. The simulations show that whilst MAS can be considerably more efficient than phenotypic selection differences between the three MAS methods are slight. Which of the MAS methods is best depends on a number of factors: in particular the genetic map, the time scale under consideration and the population size are of importance.

1. Introduction

A number of papers have now been published examining the possibility of using marker-assisted selected (MAS) to improve the value of quantitative traits. Lande & Thompson (1990) developed a method which, rather than attempting to estimate the location of Quantitative Trait Loci (QTL), directly, uses multiple regression of phenotype on marker-type to estimate differences between marker groups, and then combines these so-called marker effects with phenotypic information using a selection index. The method works best when utilizing the linkage disequilibrium between markers and QTL created by crossing two inbred lines. Computer simulation studies (Gimelfarb & Lande, 1994*a*; Zhang & Smith, 1992, 1993) have confirmed that the use of markers can improve the efficiency of selection relative to selection based solely on phenotypes, particularly when the population size is large.

The only information about the relative positions of the markers used by Lande & Thompson (1990) is the linkage group to which a marker is assigned; in general a linkage map of markers exists, and so it seems sensible to consider whether the extra information about the estimated location of the markers embodied in this map can be used to improve the performance of MAS. In this paper we develop a method for map-based marker-assisted selection (MBMAS) and, using computer simulations, compare its performance with phenotypic selection and the approach to MAS detailed by Gimelfarb & Lande (1994*a*).

Our method is based on the ideas of *interval mapping*. First introduced by Lander & Botstein (1988) for a cross between two inbred lines, interval mapping provides estimates of the location and the 'effect-size' of QTL. A point on the genome is arbitrarily chosen and it is supposed that a QTL is present at this point. The expected genetic contribution of the marker interval in which this putative QTL is located to the trait under consideration can then be written as a function of the QTL effect for each possible combination of marker alleles at the marker loci flanking the QTL and, given the phenotypic values and marker-types of a number of individuals, maximum likelihood can be used to give an estimate of the effect of a QTL at that location. Also, a likelihood ratio test can be performed to test the hypothesis that a QTL exists at this location. By moving the putative QTL along the genome a likelihood map can be constructed from which it is possible, for each interval between markers, to test the hypothesis that a QTL exists in that interval and to estimate the QTL's position and effect.

* Corresponding author: John Whittaker, Department of Applied Statistics, University of Reading, PO Box 240, Whiteknights Road, Reading, RG6 2FN.

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press

It has been shown (see Haley & Knott, 1992; Martínez & Curnow, 1992) that proceeding as above but using least-squares rather than maximum likelihood to estimate the size of the QTL effect produces virtually identical results, despite involving the approximation of a mixture of normal distribution by a single normal distribution. Least-squares has computational advantages over likelihood methods, particularly when considering a number of QTL simultaneously, and is therefore the approach used here. More details follow in Section (i).

2. Method

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 alleles in the second line are labeled q_i and m_j in a corresponding fashion. We assume that all QTL lie within a marker interval, that is that there are no QTL between the last marker locus on a chromosome and the end of that chromosome.

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. From these we wish to construct an estimate $\hat{z}$ of the genetic value of the individual, z. We tackle this problem in two stages: first we estimate the size and position of the QTL, then we calculate $\hat{z}$. We describe the calculation of $\hat{z}$ first.

(i) *Estimation of an individual's genetic value*

The obvious estimate of z based on the phenotype y and marker-type $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ is the regression of z on y and $\mathbf{x}$, $E(z \mid y, \mathbf{x})$. We next derive an expression for $E(z \mid y, \mathbf{x})$.

Suppose that there are n QTL, and let the number of Q_i alleles at the ith QTL for a particular individual be described by g_i, so that if the locus is $Q_i Q_i$, $g_i = 1$, if $q_i q_i$, $g_i = -1$ and if the locus is a heterozygote $g_i = 0$, for $i = 1, 2, \ldots, n$. Then, assuming that the QTL combine additively between and within loci, we can write the genetic value of that individual as

$$z = \sum_{i=1}^{n} a_i g_i,$$

where a_i is the effect of the ith QTL. If we knew the location and effect of the QTL, we could calculate

$$\begin{aligned}\hat{z} &= E(z \mid y, \mathbf{x})\\ &= \sum_{g_1=-1}^{1} \sum_{g_2=-1}^{1} \cdots \sum_{g_n=-1}^{1} p(\mathbf{g} \mid y, \mathbf{x}) \sum_{i=1}^{n} a_i g_i\\ &= \sum_{\mathbf{g}} p(\mathbf{g} \mid y, \mathbf{x}) \sum_{i=1}^{n} a_i g_i,\end{aligned}$$

where $\mathbf{g} = (g_1, g_2, \ldots, g_n)$ and $p(\mathbf{g} \mid y, \mathbf{x})$ is the probability of getting a particular $\mathbf{g}$ given the phenotype and the marker-type. Writing $f(y \mid \mathbf{g}, \mathbf{x})$ for the probability density function of phenotype conditional on $\mathbf{g}$ and $\mathbf{x}$ and noting that given $\mathbf{g}$ this is independent of the marker-type,

$$\begin{aligned}p(\mathbf{g} \mid y, \mathbf{x}) &= \frac{\mathrm{f}(\mathrm{y} \mid \mathbf{g}, \mathbf{x}) p(\mathbf{g}, \mathbf{x})}{f(y, \mathbf{x})}\\ &= \frac{f(y \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{x}) p(\mathbf{x})}{f(y \mid \mathbf{x}) p(\mathbf{x})}\\ &= \frac{f(y \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{x})}{\Sigma_{\mathbf{g}} f(y \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{x})}.\end{aligned}$$

We shall assume that the phenotype conditional on genotype has a normal distribution with known variance so that

$$Y \mid \mathbf{g} \sim N\left(\sum_{i=1}^{n} a_i g_i, \sigma_e^2\right).$$

Furthermore, if we relabel the markers so that the ith QTL locus is flanked by the marker loci x_{i_l} and x_{i_r} (this notation introduces a degree of redundancy in that markers may have more than one label), we know that in the F_2 derived from two completely homozygous lines

$$p(\mathbf{g} \mid \mathbf{x}) = \prod_{i=1}^{n} p(g_i \mid x_{i_l}, x_{i_r}),$$

as we are assuming independence of recombination events; i.e., Haldane's (1919) mapping function. Given the position of the QTL relative to their flanking markers $p(\mathbf{g} \mid \mathbf{x})$ is easily calculated in the F_2. For subsequent generations, rather more work is required; a method of estimating $p(g_i \mid x_{i_l}, x_{i_r})$ from the marker frequencies is given in Appendix 1.

Replacing the actual QTL locations and effects by the estimates derived in Section (ii) below we have, in principle, an estimator for z. However, this formula for $\hat{z}$ involves summing over the 3^n possible QTL types, which makes this time-consuming when we need to calculate $\hat{z}$ for reasonable numbers of individuals. It is therefore useful to derive a computationally quick approximation for $\hat{z} = E(z \mid y, \mathbf{x})$. To do this, write $z_i = a_i g_i$ so that

$$\begin{aligned}E(z \mid y, \mathbf{x}) &= \sum_{i=1}^{n} E(z_i \mid y, \mathbf{x})\\ &= \sum_{i=1}^{n} \sum_{g_i=-1}^{1} a_i g_i p(g_i \mid y, \mathbf{x}).\end{aligned}$$

Working as above,

$$p(g_i \mid y, \mathbf{x}) = \frac{f(y \mid g_i, \mathbf{x}) p(g_i \mid \mathbf{x})}{f(y \mid \mathbf{x})}, \qquad (1)$$

and

$$f(y \mid \mathbf{x}) = \sum_{\mathbf{g}} f(y \mid \mathbf{g}) p(\mathbf{g} \mid \mathbf{x}),$$

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press

which is a mixture of Normal distributions. To approximate $E(z\,|\,y, \mathbf{x})$, we shall replace this mixture of distributions by a Normal distribution with the same mean and variance, so that we assume $Y|\mathbf{x} \sim N(\mu, \sigma^2)$ where

$$\mu = \sum_{i=1}^{n} a_i \sum_{g_i=-1}^{1} g_i p(g_i\,|\,x_{i_l}, x_{i_r})$$

and, if σ_e^2 is the environmental variance, and $v_i|\mathbf{x}$ is the variance due to the ith QTL given the marker-type,

$$v_i|\mathbf{x} = a_i\left[\sum_{g_i=-1}^{1} g_i^2 p(g_i\,|\,x_{i_l}, x_{i_r}) - \left(\sum_{g_i=-1}^{1} g_i p(g_i\,|\,x_{i_l}, x_{i_r})\right)^2\right],$$

where σ^2 is given by

$$\sigma^2 = \sigma_e^2 + \sum_{i=1}^{n} v_i\,|\,\mathbf{x}.$$

There is no covariance term here because of the conditioning on $\mathbf{x}$. Similarly we replace $f(y\,|\,g_i, \mathbf{x})$ in eqn (1), by assuming that $(Y|g_i, \mathbf{x}) \sim N(\mu_i, \sigma_i^2)$ where

$$\mu_i = \sum_{j=1, j\neq i}^{n} a_j \sum_{g_j=-1}^{1} g_j p(g_j\,|\,x_{j_l}, x_{j_r}) + a_i g_i$$

and

$$\sigma_i^2 = \sigma_e^2 + \sum_{j=1, j\neq i}^{n} v_j\,|\,\mathbf{x}.$$

The approximation to $E(z\,|\,y, \mathbf{x})$ so produced is computationally fast, and seems to agree with the exact method to an accuracy of more than 99% for five QTL and a heritability of 0·4. One would expect the accuracy of the approximation to improve as the number of QTL increases.

(ii) *Estimation of QTL effects*

To use $\hat{z}$ derived above, we need estimates of the locations and effects of the QTL present. It would be possible to use interval mapping methods to produce these, but this would be both computationally demanding and difficult to automate completely: see Martínez & Curnow (1992) for the problems of 'ghost' QTL, for instance. Also it is known that, while interval mapping can in general be relied on to place a QTL in the correct marker interval, fixing the position of the QTL within that interval can be unreliable, and mapping accurately multiple QTL in the same interval impossible.

Accordingly, we assume in estimating QTL effects that each interval contains at most one QTL, and that if an interval contains a QTL that QTL is located midway between the two flanking markers. In other words, we fit an interval effect that we hope will accurately reflect the QTL within that interval. In the rest of this paper we shall treat these interval effects as if they represented 'real' QTL. We can now fit a model with effects for any selected set of intervals by regression mapping. If the selected intervals are labeled 1, 2, ..., k, the phenotypic value of an individual whose marker-type at these intervals is represented by $\mathbf{x}$ can be written

$$y = \beta_0 + \sum_{i=1}^{k} a_i \sum_{g_i=-1}^{1} p(g_i\,|\,x_{i_l}, x_{i_r}) + \epsilon,$$

where β_0 is a mean term, a_i is the effect of the ith interval, x_{i_l} and x_{i_r}, the number of markers at the loci flanking that interval, g_i the genotype at the hypothesized QTL at the midpoint of that interval and ϵ an error term. Given the phenotypes and marker types of n individuals, denoted by y_j and $\mathbf{x}^j$ for $j = 1$, 2, ... n we can then solve for a_i using least-squares, i.e. by minimizing

$$\sum_{j=1}^{n}\left(y_j - \beta_0 - \sum_{i=1}^{k} a_i \sum_{g_i=-1}^{1} p(g_i|x_{i_l}^i, x_{i_r}^j)\right)^2$$

with respect to $\beta_0, \mathbf{a}$). If the error, ϵ, had a normal distribution this would give maximum likelihood estimates for $(\beta_0, \mathbf{a})$. However, even if the fitted model was correct so that all QTL segregating in the population were located at the midpoints of the selected intervals – which is extremely unlikely – and the underlying environmental error has a normal distribution the distribution of ϵ would be a mixture of normals because individuals with the same marker-types may have different genotypes.

The interval effects $\mathbf{a}$ can now be used to calculate $\hat{z} = E(z\,|\,y, \mathbf{x})$ as described above.

(iii) *Model selection*

For the above approach to be useful we need a procedure for deciding which set of intervals to include in our model. First we take each chromosome in turn and, using forward stepwise regression (Draper & Smith, 1981) find the model with the smallest residuals containing k of that chromosome's variables, for $k = 0, 1, 2, \ldots$, with no effects fitted on the other chromosomes. In selecting the best model for each chromosome – that is, selecting the optimum value for k – we are trying to determine how many variables to include in a regression. This is a difficult statistical problem, but a commonly used criterion is Mallows C_p (Draper & Smith, 1981).

For a particular model with p parameters – $(p-1)$ interval effects plus a mean term – Mallow's C_p is defined as

$$C_p = \frac{SSE}{\sigma^2} - n + 2p$$

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press

where SSE is the error sum of squares from the model under consideration, n is the number of observations – here the number of individuals – and $\hat{\sigma}^2$ is an estimate of the error variance σ^2 obtained by fitting the full model, that is fitting effects for all the intervals on a chromosome. The error variance here will include the variance due to QTL on other chromosomes. Note the form of C_p: the error sum of squares less a penalty for the number of parameters included in the model. We could choose as the best model the model minimizing C_p. Gilmour (1995) points out that this tends to result in overfitting and suggests the use of adjusted C_p, $\bar{C}_p$, given by

$$\bar{C}_p = C_p - \frac{2(k-p+1)}{n-k-3}$$

where k is the number of potential regressors; here k is the number of intervals on a single chromosome. Again this is the error sum of squares less a penalty for the number of parameters included in the model. We use $\bar{C}_p$ here, but it should be noted that even $\bar{C}_p$ presents a risk of overfitting: Gilmour (1995) gives a method for selecting models using hypothesis tests, but this has not been used as it would be difficult to automate.

The best models for each chromosome are pooled to give a full model in which effects are included for each interval selected by the chromosome by chromosome procedure. This full model can then be fitted to the data.

We have now developed a method for using the marker-map to estimate, in any generation, the expected genetic value of an individual, given the individual's phenotype and marker-type. These expected genetic values can be used for selection. We shall describe briefly an alternative approach introduced by Lande & Thompson (1990) and further developed by Gimelfarb & Lande (1994*a*) which uses information on the location of the markers only to divide markers into linkage groups, and then compare these two methods with selection based solely on phenotype using computer simulations.

(iv) *MAS by regression on markers*

In the analysis of Gimelfarb & Lande (1994*a*) selection is based on combining phenotype and the 'molecular score', S, in an index I such that

$$I = y + b_s S,$$

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

$$S = \sum_{i \in \mathscr{A}} \beta_i x_i.$$

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, i.e. by minimizing

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

where y_j and $\mathbf{x}^j$ are the phenotype and marker-type respectively of the jth individual. Note that this is equivalent to using a linear approximation

$$\hat{z}_L = y + \beta_0' + \sum_{i \in \mathscr{A}} \beta_i' x_i$$

to the regression function $E(z \mid y, \mathbf{x})$.

Gimelfarb and Lande (1994*a*) suggest using a two-stage process to select the markers to include in $\mathscr{A}$. In the first stage, each chromosome is taken in turn and the 'forward selection process' used to select a predetermined number of markers from those on that chromosome. Next, the markers so selected for each chromosome are pooled and the forward selection process used to select a number, again predetermined, of markers which make up the set $\mathscr{A}$. The simulations of Gimelfarb & Lande (1994*a*) show that an improvement in performance can be gained by repeating this selection procedure in each generation, so that the markers contributing to the marker score may change between generations.

A linear index incorporating phenotype and marker score is optimal when the marker score coefficient in the tth generation, b_S is given by

$$b_S(t) = \frac{(1-h^2(t))\,\sigma_y^2(t)}{h^2(t)\,\sigma_y^2(t) - \sigma_S^2(t)},$$

where σ_y^2(t) is the phenotypic variance, $\sigma_S^2(t)$ the variance accounted for by the markers and $h^2(t)$ the heritability of the trait, all in the tth generation. Substituting for

$$h^2(t) = \frac{\sigma_y^2(t) - \sigma_e^2(t)}{\sigma_y^2(t)},$$

where σ_e^2 is the environmental variance, this becomes

$$b_S(t) = \frac{\sigma_e^2}{\sigma_y^2(t) - \sigma_S^2(t) - \sigma_e^2}.$$

We have assumed that σ_e^2 is known in the simulations below.

It should be noted that this formula assumes that all the variance explained by the markers is genetic: the fact that the markers to be included in the index are selected because they explain a high proportion of the phenotypic variance will tend to result in this not being true. This leads to more weight being placed on the marker score than would be optimal; indeed, in the simulations described below we often found that b_S became negative. In this case we set b_S to 20, a value arbitrarily selected to place nearly all weight on the marker score.

Note that in using this method we must decide how many markers to include in the model: Gimelfarb & Lande (1994*a*) demonstrate that selecting either too many or too few markers reduces the efficiency of selection. There is no theory to guide us here; in simulations this decision may be made on the basis of trial and error but this is clearly not possible in practice. The obvious solution to this problem is to base a model selection procedure on Mallows C_p, in similar vein to Section (iii). This has been done.

3. Simulations

Four methods were compared using computer simulations. They were:

- selection based solely on phenotype;
- selection based on the multiple regression of phenotype on single markers, the marker score being combined with phenotypic information using an index, as described in Section (iv), with the number of markers included in the model fixed (at 6, as in the BASE parameter set of Gimelfarb & Lande, 1994*a*) (GLMAS);
- selection based on the multiple regression of phenotype on single markers, the marker score being combined with phenotypic information using an index, as described in Section (iv), with the number of markers selected using adjusted Mallows C_p (MGLMAS);
- selection using information on the location of markers as detailed in Section 2; that is, using the estimation procedure described in Section (ii) in conjunction with the model selection procedure of Section (iii) (MBMAS).

In each case the model fitting procedure was repeated every generation so that the intervals (for MBMAS) or markers (for GLMAS and MGLMAS) for which effects are fitted may change with time.

Simulations were done using two maps. 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 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. We stress that the simulated QTL were not constrained to the midpoints of the intervals. The effect of these QTL, a_i for $i = 1, 2, \ldots,$ 100 was generated assuming that the amount of additive genetic variance due to QTL may be approximated by a power series, as by Lande & Thompson (1990). That is, the variance due to the *i*th QTL is $ka^{(i-1)}$ for constants a and k. Here a is a shape parameter; we write a in terms of n_E, the effective number of loci (Lande, 1981),

$$n_E = \frac{(\Sigma_{i=0}^{99} a^i)^2}{\Sigma_{i=0}^{99} a^{2i}} \approx \frac{1+a}{1-a}.$$

The simulations were performed with $n_E = 10$, which is the value used by Gimelfarb & Lande (1994*a*). Given a, the constant k determines the genetic variance σ_G^2, for

$$\sigma_G^2 = \text{var}\left(\sum_{i=1}^{100} a_i g_i\right)$$
$$= \sum_{i=1}^{100} \text{var}(a_i g_i) + \sum_{i=1}^{100} \sum_{j<i} \text{cov}(a_i g_i, a_j g_j)$$
$$= 0{\cdot}5k^2 \sum_{i=1}^{100} a^{2(i-1)} + k^2 \sum_{i=1}^{100} \sum_{j<i} a^{(i+j-2)} \text{cov}(g_i, g_j)$$

We wish to choose σ_G^2 to give the desired heritability; given that the phenotypic variance was fixed at 1 in all simulations, we require $\overset{2}{\sigma_G} = h^2$. This is done by generating a large number of F_2 individuals with $k = 1$ and recording the genetic variance, V, for these individuals. The QTL can then be rescaled by setting $k = h^2/V$ and these rescaled effects used to generate a generation of F_2 individuals on which MAS can be performed. 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 (1994*a*); 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 (1994), simulations of two types were run. In the first, 'total coupling', the effects of QTL are 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 F_2 heritability.

Simulations were run from 20 generations with heritabilities 0·1 and 0·2, and 100, 200 and 400 individuals of each sex. The number of replicates was varied with the population size: 40 replicates for 100 and 200 individuals and 30 for 400 individuals. In every generation the top 20% of individuals of each sex are selected and paired at random: each pair is then assumed to produce five offspring of each sex.

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press

4. Results

Figures 1–4 contain the simulation results obtained for $n = 100$ and $n = 400$, where n is the number of individuals of each sex, with a heritability of 0·1. In all cases figures are given as percentages of the maximum genetic value obtainable, that is the genetic value of an individual possessing all favourable alleles. Standard errors for these results are acceptable: for example for

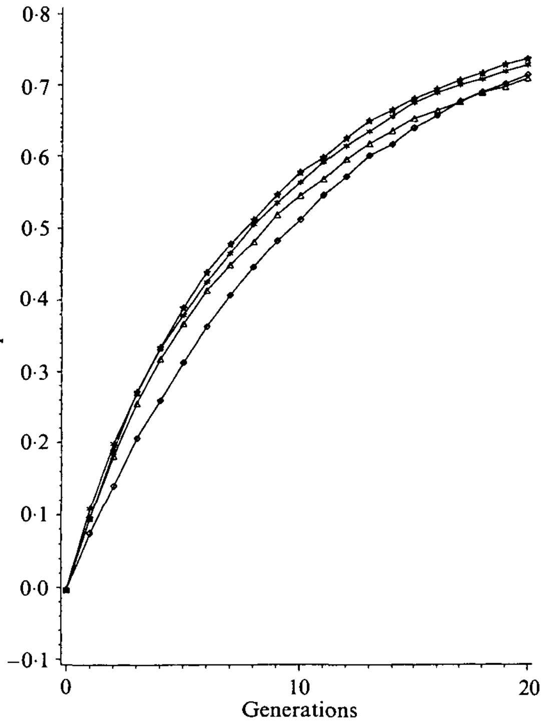

Fig. 1. Response to selection, as a proportion of the genetic maximum: map 1, $n = 100$, $h^2 = 0{\cdot}1$. ◇, Phenotypic selection; △, GLMAS; *, MBMAS; ☆, MGLMAS.

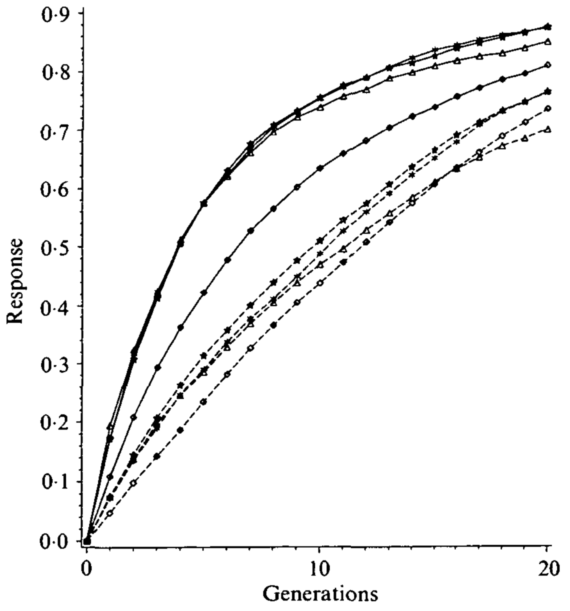

Fig. 2. Response to selection, as a proportion of the genetic maximum: map 2, $n = 100$, $h^2 = 0{\cdot}1$. ◇——◇, Phenotypic selection, coupling; △——△, GLMAS, coupling; *——*, MBMAS, coupling; ☆——☆, MGLMAS, coupling; ◇---◇, phenotypic selection, repulsion; △---△, GLMAS, repulsion; *---*, MBMAS, repulsion; ☆---☆, MGLMAS, repulsion.

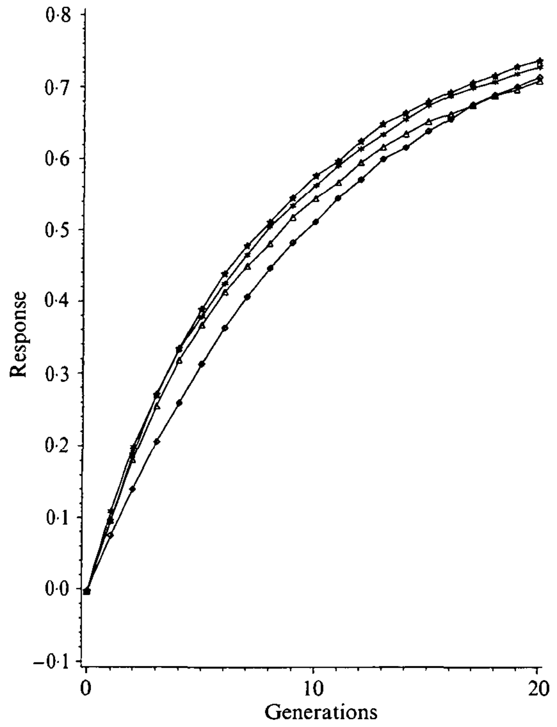

Fig. 3. Response to selection, as a proportion of the genetic maximum: map 1, $n = 400$, $h^2 = 0{\cdot}1$. ◇, Phenotypic selection; △, GLMAS; *, MBMAS; ☆, MGLMAS.

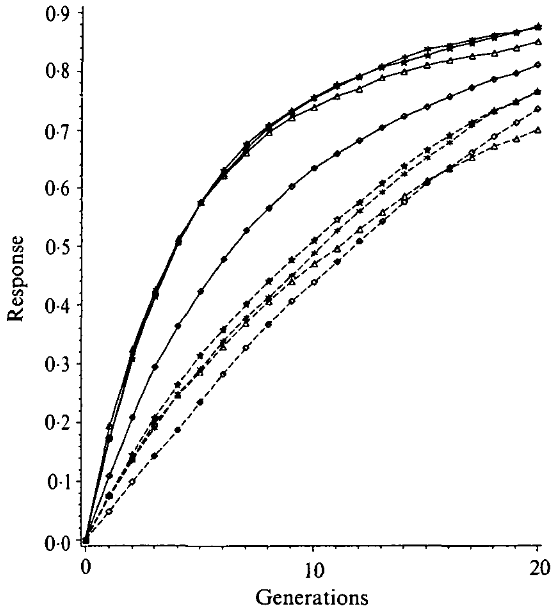

Fig. 4. Response to selection, as a proportion of the genetic maximum: map 2, $n = 400$, $h^2 = 0{\cdot}1$. ◇——◇, Phenotypic selection, coupling; △——△, GLMAS, coupling; *——*, MBMAS, coupling; ☆——☆, MGLMAS, coupling; ◇---◇, phenotypic selection, repulsion; △---△, GLMAS, repulsion; *---*, MBMAS, repulsion; ☆---☆, MGLMAS, repulsion.

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press

Table 1. *Mean number of effects fitted each generation* (*gen*): n = *400*, $h^2 = 0{\cdot}1$

	Map 1		Map 2, coupling		Map 2, repulsion	
gen	MBMAS	MGLMAS	MBMAS	MGLMAS	MBMAS	MGLMAS
0	18·4	22·2	19·3	21·2	19·8	22·4
1	18·2	22·6	20·0	21·2	21·7	25·3
2	18·4	23·7	21·0	23·2	20·8	24·1
5	14·4	20·4	17·6	21·5	19·0	24·2
10	8·6	17·8	13·3	18·0	18·3	22·7
15	6·0	15·8	10·0	15·8	14·2	20·0
20	5·4	15·7	8·0	15·2	9·1	17·0

Table 2. *QTL frequencies*: n = *400*, $h^2 = 0{\cdot}1$, *map 1*

		QTL effects				
	gen	0·1808	0·1479	−0·1210	−0·0990	−0·0810
Phenotypic selection	0	0·50	0·50	0·50	0·50	0·50
	1	0·57	0·56	0·43	0·52	0·47
	5	0·77	0·74	0·22	0·50	0·36
	10	0·92	0·89	0·08	0·43	0·23
	20	1·00	0·99	0·01	0·24	0·09
GLMAS	0	0·50	0·50	0·50	0·50	0·50
	1	0·59	0·64	0·40	0·55	0·46
	5	0·78	0·81	0·23	0·56	0·30
	10	0·91	0·93	0·11	0·52	0·22
	20	0·99	0·99	0·02	0·41	0·09
MGLMAS	0	0·50	0·50	0·50	0·50	0·50
	1	0·60	0·62	0·40	0·54	0·45
	5	0·82	0·86	0·21	0·50	0·27
	10	0·94	0·95	0·06	0·48	0·18
	20	0·99	0·99	0·01	0·32	0·05
MBMAS	0	0·50	0·50	0·50	0·50	0·50
	1	0·61	0·62	0·39	0·54	0·44
	5	0·85	0·84	0·18	0·53	0·28
	10	0·95	0·94	0·06	0·44	0·17
	20	1·00	1·00	0·00	0·28	0·06

map 1 with a heritability of 0·1 and 100 individuals of each sex standard errors in generation 10 are 0·0038, 0·0071, 0·0111 and 0·0071 for phenotypic selection, GLMAS, MGLMAS and MBMAS respectively. In general MGLMAS tends to have larger standard errors than the other methods.

Results for $n = 200$ lie between those for $n = 100$ and $n = 400$. Results for $h^2 = 0{\cdot}2$ show an improvement in the performance of phenotypic selection relative to the MAS methods. This is a general trend: as heritability increases so does the performance of phenotypic selection relative to the MAS.

It is immediately apparent that in all cases phenotypic selection is inferior to all the marker-assisted methods for at least ten generations and in most cases for 20 generations. Differences between the marker-assisted methods are generally small compared to the difference between marker-assisted methods and selection on the phenotype alone; however, the relative performance of the marker-assisted methods depends both on the time-horizon to be considered and on the genetic map. The advantage of marker-assisted methods over phenotypic selection increases with decreasing heritability and increasing population size, as found by Gimelfarb & Lande (1994*a*).

The maximum ratio of response to marker-assisted selection to response to phenotypic selection occurs in generation one or two and declines rapidly thereafter; in contrast the maximum difference between MAS and phenotypic selection occurs later, between generations 5 and 10, and declines more gradually. It is worth noting that, for all methods and maps, population size affects the rate of improvement in the first few generations much more markedly than it

Table 3. *Genetic and phenotypic variances*: $n = 400$, $h^2 = 0{\cdot}1$, *map 1*

		Genetic			Genetic due to markers		Phenotypic due to markers	
gen	pheno	GLMAS	MBMAS	MGLMAS	GLMAS	MGLMAS	GLMAS	MGLMAS
0	0·100	0·100	0·100	0·100	0·052	0·069	0·074	0·112
1	0·093	0·081	0·079	0·081	0·029	0·041	0·055	0·079
2	0·089	0·075	0·073	0·074	0·021	0·035	0·048	0·070
5	0·069	0·056	0·050	0·053	0·012	0·019	0·038	0·056
10	0·040	0·033	0·026	0·027	0·005	0·007	0·034	0·042
15	0·022	0·020	0·016	0·015	0·003	0·004	0·029	0·042
20	0·014	0·014	0·010	0·010	0·002	0·003	0·026	0·040

Genetic due to markers is the covariance of marker score with true genetic value; phenotypic due to markers the covariance of marker score with phenotypic value. Pheno is phenotypic selection.

does the total response to selection achieved after 20 generations.

Gimelfarb & Lande (1994*a*) noted that for map 2 MAS performs better in comparison with phenotypic selection when the QTL are in coupling rather than repulsion phase, because in coupling phase we essentially have 'good' and 'bad' chromosomes. Detecting and selecting for these good chromosomes is much easier than attempting to disentangle the individual alleles, as we must in repulsion phase. We see the same problem in map 1, where the sign of a QTL effect is chosen at random: the ratio of response to MAS to response to phenotypic selection for map 1 is generally slightly less than for map 2 in repulsion mode, probably because the markers are more widely spaced.

Table 1 contains the mean number of effects fitted by MGLMAS and MBMAS for each map with $n = 400$ and $h^2 = 0{\cdot}1$. As one would expect the number of effects fitted declines with time, and at each generation MGLMAS fits more effects than MBMAS. Surprisingly the mean number of effects fitted is relatively insensitive to changes in the other parameters, although there is a slight increase in mean number of effects with increasing n and h^2. This is to be expected: as n and h^2 increase so should the number of parameters it is possible to estimate accurately.

Differences between the marker-assisted methods are less significant, although the worst performer overall is clearly GLMAS. In all cases one, and usually both, of the Mallow-based approaches is superior to GLMAS in generations subsequent to the sixth; for map 1 the Mallow approaches are always superior whilst for map 2 in repulsion mode at least one of the Mallow-based approaches is superior after the third generation. GLMAS does best compared with the two methods fitting a variable number of effects using map 2 in coupling phase: indeed, for this situation GLMAS is the best performing method in the first few generations, particularly for $n = 100$. As both MGLMAS and MBMAS fit considerably more effects than GLMAS in early generations this suggestions the routines based on Mallow's C_p may be fitting too many parameters. MGLMAS and MBMAS perform almost identically, although MBMAS has a slight edge for map 1 with $n > 100$ and MGLMAS for map 2 in coupling mode.

Table 2 gives the frequencies of the Q_i allele at QTL 1, 3, 5, 7, 9 (where the QTL are ranked in order of magnitude of effect) for each selection method with map 1, $n = 400$ and $h^2 = 0{\cdot}1$. The most striking feature of these tables is how quickly all the loci except the seventh, which has effect $-0{\cdot}099$, become fixed. Examination of map 1 reveals that the marker bracket containing this locus (between the 66th and 67th markers) also contains QTL of effect 0·0600 and 0·0180, so this is not surprising. The more sophisticated methods (MGLMAS and MBMAS) are considerably more successful at changing the frequency of the seventh QTL.

5. Discussion

As we noted above, the optimum method in any particular case depends both on the map and the time horizon being considered. GLMAS fits few effects in comparison to MGLMAS and MBMAS and does best for map 2 in coupling mode, because we then have 'good' and 'bad' chromosomes and it is possible to select for these using relatively few markers, at least in early generations. As this is rather a special case we believe that the Mallow-based methods are to be preferred except when the objective is to maximize selection advance in the first or second post-F_2 generation.

The fact that MGLMAS and MBMAS fit considerably more effects than GLMAS in early generations suggest that an improvement in MGLMAS and MBMAS may result from modifying C_p so as to increase the penalty imposed for including an additional parameter. A harder, but potentially more fruitful approach, may be to attack the cause of this over-fitting: the fact that any variable selection procedure leads to over-estimation of the variance attributable to the selected variables. This leads to bias in the parameter estimates and, as mentioned in section 2·4, to an over-estimation in the genetic variance attributable to the markers. Some methods for reducing this bias exist (Miller, 1990) but application to this problem would be very computationally demanding.

That this bias is considerable is shown in Table 3, where the actual genetic and phenotypic variances attributable to the markers are displayed together with the genetic variance remaining in the population for map 1 with $n = 400$ and $h^2 = 0{\cdot}1$. Indeed, the phenotypic variance attributable to the markers is often greater than the genetic variance remaining in the population, as we mentioned in Section (iv). (Remember that the *actual* genetic variance attributable to the markers is in real populations unobservable, and that in computing the selection index we substitute the phenotypic variance attributable to the markers.)

Note that even if the actual genetic variance attributable to the markers was used in calculating the selection index the ratio of weight on marker score to weight on phenotype would be high enough to ensure that phenotype had little influence on selection, particularly in later generations. This is inevitable given the form of the selection index: with $I = y+b_S S$ we have

$$b_S(t) = \frac{\sigma_e^2}{\sigma_y^2(t)-\sigma_S^2(t)-\sigma_e^2}$$

and the genetic variance $\sigma_y^2(t)-\sigma_e^2$ tends to zeros as t increases. It follows that, no matter how small the variance due to the markers $\sigma_S^2(t)$, b_S must be large for large t.

It is at first surprising, the MBMAS does not seem to offer a considerable improvement over MGLMAS, as it makes use of extra information in the location of the markers and models the physical processes involved in selection more closely than does MGLMAS. There are two mathematical reasons why we might expect such an improvement. Firstly, because the known benefits of mapping QTL via interval rather than single marker methods should lead to the interval effects being better estimated than the marker effects (see, for example, Lander & Botstein, 1988). Secondly, one would hope that the $\hat{z}$ used in the interval method would be a better estimator of z than the linear function $\hat{z}_L$, because $\hat{z}$ acknowledges the non-linearity of $E(z \mid y, \mathbf{x})$. However, it is possible to show (Whittaker, Thompson & Visscher, personal communication) that when, as here, QTL combine additively within and between loci, $E(z \mid \mathbf{x})$ is a *linear* function of $\mathbf{x}$ in F_2 populations, i.e. that for m markers

$$E(z \mid \mathbf{x}) = k_0 + \sum_{i=1}^{m} k_i x_i$$

for a set of constants k_i. This result has considerable implications for marker-assisted selection and regression mapping, which we shall address elsewhere. Its importance here is in showing that $\hat{z}_L$ differs from $E(z \mid y, \mathbf{x})$ only in the treatment of the phenotype: as we have seen that in general little weight is placed on the phenotype we would expect our estimators $\hat{z}$ and $\hat{z}_L$ to be almost identical in the F_2. The simulation results presented here suggest that this is also true, at least approximately, in subsequent generations. This may also explain why MBMAS does best for map 1: one would expect this optimal treatment of the phenotype to be more important here than for map 2 as the markers are more widely spaced, and so less informative about the QTL present.

One might also draw from this the conclusion that more sophisticated methods of using information on the location of markers will not improve the efficiency of MAS greatly, and that there are only slight gains in performance to be gained by weighting the marker and phenotypic information more accurately. However, we see from Table 3 that after generation 10 over 80% of the genetic variance cannot be explained by the markers and this suggests that improved weighting of information will result in more weight being placed on the phenotype and improved response. More work is needed to resolve this question fully, but it does seem that improvements in efficiency should be more easily gained by trying to improve the estimates of QTL effects (for MBMAS) or marker effects (for MGLMAS). An obvious way of doing this is to develop a method of combining estimates across generations: this should have a similar effect to increasing the population size, and as we have seen population size has a major effect on the efficiency of MAS.

A number of simplifying assumptions have been made in this paper. Relaxation of any of these assumptions will reduce the advantage of MAS over phenotypic selection, but in general this reduction should not be too severe. For instance, the assumption that markers are evenly spaced should not be crucial, provided all inter-marker distances remain reasonably small. Also, one would expect that the assumption that QTL are additive could be relaxed without changing the results substantially, since additive effects can be estimated independently of non-additive effects (Whittaker, Thompson & Visscher, personal communication). This assertion is supported by Gimelfarb & Lande (1994*b*). The assumption that the lines are completely inbred and in complete linkage disequilibrium is probably of more importance.

6. Conclusions

The use of markers can offer a considerable improvement in response to selection over selection on the phenotype alone. However, incorporating information on relatives by using a family selection index gives higher genetic response rates than selection solely on individual phenotype, so we may have

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press

overstated the value of MAS in practice. It should also be stressed that the advantage of MAS declines as heritability increases: it is of most value for traits of low heritability. Using a marker map gives at best a slight improvement on the simple regression on markers approach. The conditions favouring each of the marker-assisted approaches are not altogether clear, but differences seem to be slight relative to the differences between phenotypic selection and any of the marker-assisted methods. We recommend the selection of markers using Mallows C_p (or some related criteria) as a more robust approach than the inclusion of a fixed number of markers.

J.C.W. was funded by the BBSRC. C.S.H. and R.T. acknowledge support from MAFF. We thank Peter Visscher for helpful discussions of this work.

References

Draper, N. R. & Smith, H. (1981). *Applied Regression Analysis*, 2nd edition. New York: Wiley.

Gilmour, S. G. (1995). The interpretation of Mallows' C_p statistic (*submitted*).

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

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

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.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Lande, R. (1981). The minimum number of genes contributing to quantitative variation between and within populations. *Genetics* **99**, 541–553.

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

Lander, E. S. & Botstein, D. (1988). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Martínez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical Applied Genetics* **85**, 480–488.

Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall.

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.

Appendix 1: Derivation of an approximation for $p(g_i \mid x_{i_l}, x_{i_r})$

The formula for $\hat{z}$ derived above requires $p(g_i \mid x_{i_l}, x_{i_r})$ to be known, which they are in the F_2, but in subsequent generations selection and recombination will change $p(g_i \mid x_{i_l}, x_{i_r})$ in a complicated manner. We shall attempt to estimate $p(g_i \mid x_{i_l}, x_{i_r})$ in later generations by reference to the marker-type frequencies, $p(M_{i_l} M_{i_r})$, $p(M_{i_l}, m_{i_r})$, $p(m_{i_l} M_{i_r})$ and $p(m_{i_l} m_{i_r})$. These are, of course, unknown, because we have no way of determining the phase of the double heterozygote ($x_{i_l} = 1, x_{i_r} = 1$). Ignore this problem for now and assume that the marker-type frequencies, $p(M_{i_l} M_{i_r})$, $p(M_{i_l} m_{i_r})$, $p(m_{i_l} M_{i_r})$ and $p(m_{i_l} m_{i_r})$ are known, so it makes sense to write

$$p_t(Q_i \mid M_{i_l} M_{i_r}) = \frac{p_t(M_{i_l} Q_i)\, p_t(M_{i_r} \mid M_{i_l} Q_i)}{p_t(M_{i_l} M_{i_r})}.$$

where the subscript t means that these are the probabilities in the tth generation. Assume that in the tth generation the linkage disequilibrium between the ith QTL locus and the left flanking marker is equal to the linkage disequilibrium between the ith QTL locus and the right flanking marker, and denote this by $L_i(t)$ so that

$$p_t(M_{i_l} Q_i) = p_t(M_{i_l})\, p_t(Q_i) + L_i(t), \qquad (1\,\text{A})$$

as usual, with corresponding expressions for the other pairs of alleles. It seems reasonable that

$$p_t(M_{i_r} \mid M_{i_l} Q_i) \approx p_t(M_{i_r} \mid Q_i) \qquad (2\,\text{A})$$

that is most of the information about the allele at a particular locus is provided by knowledge of the locus closest to it; this is true exactly in the F_2 but not for $t > 2$, as is shown in Appendix 2. Substituting this assumption into equation 1 A and then writing in terms of the linkage disequilibrium gives

$$p_t(Q_i \mid M_{i_l} M_{i_r}) \approx \frac{1}{p_t(M_{i_l} M_{i_r})} \times \left(p_t(M_{i_l})\, p_t(Q_i)\, p_t(M_{i_r}) + L_i(t)\, p_t(M_{i_l}) + L_i(t)\, p_t(M_{i_r}) + \frac{L_i(t)^2}{p_t(Q_i)} \right), \qquad (3\,\text{A})$$

with corresponding expressions for $p_t(Q_i \mid M_{i_l} m_{i_r})$, $p_t(Q_i \mid m_{i_l} M_{i_r})$ and $p_t(Q_i \mid m_{i_l} m_{i_r})$.

To use these expressions we need to known $p_t(Q_i)$ and $L_i(t)$. We shall now obtain an approximate expression for $L_i(t)$ in terms of the marker frequencies and the linkage disequilibrium between the markers, $D_i(t)$. Assuming $p_t(M_{i_r} \mid M_{i_l} Q_i) \approx p_t(M_{i_r} \mid Q_i)$ as above (eqn 2 A),

$$\begin{aligned} p_t(M_{i_l} M_{i_r}) &\approx p_t(M_{i_l} Q_i)\, p_t(M_{i_r} \mid Q_i) \\ &\quad + p_t(M_{i_l} q_i)\, p_t(M_{i_r} \mid q_i) \\ &= p_t(M_{i_r} Q_i) \frac{p_t(Q_i M_{i_r})}{p_t(Q_i)} \\ &\quad + p_t(M_{i_l} q_i) \frac{p_t(q_i M_{i_r})}{p_t(q_i)}. \end{aligned}$$

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press

Writing in terms of allele frequencies and linkage disequilibrium and simplifying we get

$$p_t(M_{i_l}M_{i_r}) = p_t(M_{i_l})\,p_t(M_{i_r}) + \frac{L_i(t)^2}{p_t(q_i)\,p_t(Q_i)}.$$

Producing the corresponding expression for $p_t(M_{i_l}m_{i_r})$, $p_t(m_{i_l}M_{i_r})$ and $p_t(m_{i_l}m_{i_r})$ shows our assumption that $p_t(M_{i_r}\,|\,M_{i_l}Q_i) = p_t(M_{i_r}\,|\,Q_i)$ is only true if $D_i(t)$, the linkage disequilibrium between the flanking marker loci, is

$$D_i(t) = \frac{L_i(t)^2}{p_t(q_i)\,p_t(Q_i)}.$$

Thus a reasonable estimator of $L_i(t)$ would be

$$\hat{L}_i(t) = \surd p_t(Q_i)\,p_t(q_i)\,\hat{D}_i(t),$$

where $\hat{D}_i(t)$ is the usual estimator of $D_i(t)$,

$$\hat{D}_i(t) = p_t(M_{i_l}M_{i_r})\,p_t(m_{i_l}m_{i_r}) - p_t(M_{i_l}m_{i_r})\,p_t(m_{i_l}M_{i_r}).$$

This still leaves the problem of estimating $p_t(Q_i)$; we do this using the conditional probabilities $p(Q_i\,|\,..)$ from the previous generation in the following manner. First consider the F_2. The probabilities $p_2\,(Q_i\,|\,..)$ and $p_2(Q_i)$ are known, as is the linkage disequilibrium,

$$L_i(2) = \frac{1}{4} - \frac{r_i}{2}$$

where r_i is the rate of recombination between the QTL and the flanking markers (see Appendix 2). Therefore we can work out $p_3(Q_i)$ by summing over the four possible marker types,

$$\begin{aligned} p_3(Q_i = p_2(Q_i\,|\,M_{i_l}M_{i_r})\,p_2(M_{i_l}M_{i_r}) + p_2(Q_i\,|\,M_{i_l}m_{i_r})\\ p_2(M_{i_l}m_{i_r})\\ + p_2(Q_i\,|\,m_{i_l}M_{i_r})\,p_2(m_{i_l}M_{i_r}) + p_2(Q_i\,|\,m_{i_l}m_{i_r})\\ p_2(m_{i_l}m_{i_r}) \end{aligned}$$

and using the equations for $p_2(Q_i\,|\,..)$ derived above (eqn 3 A). This value of $p_3(Q_i)$ can then be used to calculate $p_3(Q_i\,|\,..)$ using equations for $p_2(Q_i\,|\,..)$, and in turn $p_3(Q_i\,|\,..)$ and the marker frequencies such as $p_4(M_{i_l}M_{i_r})$ used to calculate $p_4(Q_i)$. Continuing the process allows us to track changes in gene frequency and linkage disequilibrium through time.

(i) Dealing with double heterozygotes

In the above we assume we can observe the marker frequencies such as $p(M_{i_l}M_{i_r})$. This is impossible for a double heterozygote $x_{i_l} = 1$, $x_{i_r} = 1$; we do not know whether its marker genotype is $M_{i_l}M_{i_r}/m_{i_l}m_{i_l}$ or $M_{i_l}m_{i_r}/m_{i_l}M_{i_r}$. We deal with this by first counting marker-types omitting the double heterozygotes to get $p'_t(M_{i_l}M_{i_r})$, $p'_t(M_{i_l}m_{i_r})$. Then we assign haplotypes to double heterozygotes using $p'_t()$ to generate $p_t()$, assuming that the marker combinations $M_{i_l}M_{i_r}$, $M_{i_l}m_{i_r}$, $m_{i_l}m_{i_r}$ and $m_{i_l}m_{i_r}$ occur in the same frequencies in the double heterozygote individuals as they do in the population with double heterozygote individuals removed. For example,

$$\begin{aligned} p_t(M_{i_l}m_{i_r}) = p'_t(M_{i_l}M_{i_r})\,(1 - p_t(x_{i_l} = 1, x_{i_r} = 1)\\ + p_t(x_{i_l} = 1, x_{i_r} = 1)\\ \times\left[\frac{p'_t(M_{i_l}m_{i_r})p'_t(m_{i_l}M_{i_r})}{p'_t(M_{i_l}m_{i_r})p'_t(m_{i_l}M_{i_r}) + p'_t(M_{i_l}M_{i_r})p'_t(m_{i_l}m_{i_e})}\right]. \end{aligned}$$

Appendix 2: Dependence of conditional allele frequencies on non-proximate marker-types

We shall show that in the F_j, $p(M_{i_l}\,|\,Q_i M_{i_r}) \neq p(M_{i_l}\,|\,Q_i)$ for $j > 2$.

Proof. Suppose that in the F_j we have the gametes

$$M_{i_l}Q_iM_{i_r}, M_{i_l}Q_im_{i_r}, M_{i_l}q_iM_{i_r}, M_{i_l}q_im_{i_r},$$
$$m_{i_l}Q_iM_{i_r}, m_{i_l}Q_im_{i_r}, m_{i_l}q_iM_{i_r}, m_{i_l}q_im_{i_r}$$

with frequencies x_1, x_2, …, x_8 respectively. Then,

$$p(M_{i_l}\,|\,Q_iM_{i_r}) = \frac{x_1}{x_1 + x_5}$$

whilst

$$p(M_{i_l}\,|\,Q_i) = \frac{x_1 + x_2}{x_1 + x_2 + x_5 + x_6},$$

so that $p(M_{i_l}\,|\,Q_iM_{i_r}) = p(M_{i_l}\,|\,Q_i)$ if and only if $x_2x_5 = x_1x_6$.

This completes the proof, for we have shown that $p(M_{i_l}\,|\,Q_iM_{i_r}) = p(M_{i_l}\,|\,Q_i)$ if and only if $p(M_{i_l}Q_im_{i_r})\,p(m_{i_l}Q_iM_{i_r}) = p(M_{i_l}Q_iM_{i_r})\,p(m_{i_l}Q_im_{i_r})$, which is in general false because of the effect of selection.

Note that in the F_2 we have

$$\begin{aligned} p(M_{i_l}Q_im_{i_r}) = p(m_{i_l}Q_iM_{i_r}) &= 0{\cdot}5r(1-r)\\ p(M_{i_l}Q_im_{i_r}) &= 0{\cdot}5(1-r)^2\\ p(M_{i_l}Q_im_{i_r}) &= 0{\cdot}5r^2 \end{aligned}$$

and similarly for q_i, where r is the chance of a recombination between the QTL and, say, the left-hand marker loci (remembering that we are supposing QTL are located in the centre of marker intervals). Thus,

$$\begin{aligned} p(M_{i_l}Q_im_{i_r})\,p(m_{i_l}Q_iM_{i_r}) &= 0{\cdot}25r^2(1-r)^2\\ &= p(M_{i_l}Q_iM_{i_r})\,p(m_{i_l}Q_im_{i_r}), \end{aligned}$$

which implies that $p(M_{i_l}\,|\,Q_iM_{i_r}) = p(M_{i_l}\,|\,Q_i)$ in the F_2. Furthermore, $p(M_{i_l}Q_i) = 0{\cdot}5(1-r)$ so, because $p(M_{i_l}Q_i) = p(M_{i_l})p(Q_i) + L_i$, we know that the linkage disequilibrium in the F_2 is $L_i = 0{\cdot}25 - 0{\cdot}5r$.

https://doi.org/10.1017/S0016672300034698 Published online by Cambridge University Press