Theor Appl Genet (1994) 89:305–312

A. J. Wright · R. P. Mowers

Multiple regression for molecular-marker, quantitative trait data from large F_2 populations

Received: 22 October 1993 / Accepted: 11 February 1994

Abstract Molecular marker-quantitative trait associations are important for breeders to recognize and understand to allow application in selection. This work was done to provide simple, intuitive explanations of trait-marker regression for large samples from an F_2 and to examine the properties of the regression estimators. Beginning with a(−1, 0, 1) coding of marker classes and expected frequencies in the F_2, expected values, variances, and covariances of marker variables were calculated. Simple linear regression and regression of trait values on two markers were computed. The sum of coefficient estimates for the flanking-marker regression is asymptotically unbiased for an included additive effect with complete interference, and is only slightly biased with no interference and moderately close (15 cM) marker spacing. The variance of the sum of regression coefficients is much more stable for small recombination distances than variances of individual coefficients. Multiple regression of trait variables on coded marker variables can be interpreted as the product of the inverse of the marker correlation matrix **R** and the vector **a** of simple linear regression estimators for each marker. For no interference, elements of the correlation matrix **R** can be written as products of correlations between adjacent markers. The inverse of **R** is displayed and used to illustrate the solution vector. Only markers immediately flanking trait loci are expected to have non-zero values and, with at least two marker loci between each trait locus, the solution vector is expected to be the sum of solutions for each trait locus. Results of this work should allow breeders to test for intervals in which trait loci are located and to better interpret results of the trait-marker regression.

Key words Trait-marker regression · Selection

Communicated by A. R. Hallauer

A. J. Wright · R. P. Mowers (✉)
ICI Seeds, Research Department, 2369 330th Street, Slater, Iowa 50244, USA

Introduction

Molecular markers are beginning to show promise in helping plant breeders develop inbreds for improved hybrid performance. Often the breeder begins with an F_1 of two inbreds and either backcrosses to a recurrent inbred or selfs to get an F_2 population and then makes selections among individuals by evaluating performance in hybrid cross combination. Molecular markers can be used in backcrossing to locate genetic factors and estimate the size of their effect (Lander and Botstein 1989). Population improvement or inbred development by selection might be enhanced using molecular marker information in a selection index (Lande and Thompson 1990; Zehr et al. 1992; Dudley 1993). It is therefore necessary to develop methods of relating phenotypic trait data to molecular marker data for both of these marker-assisted breeding methods.

The initial method used to associate phenotypic trait data with marker genotypic classes was by contrasting homozygous class means using *t*-tests, as proposed by Soller et al. (1976) and illustrated by Stuber et al. (1987). Lander and Botstein (1989) and Knott and Haley (1992) identified recognizable shortcomings in the single-locus analyses: a downward bias in estimated effect, loss of power in significance tests, high probability of false positives when tests are conducted at many marker loci, and inability to estimate location of genetic factors.

Maximum likelihood estimation or quasi-maximum likelihood methods provide a somewhat more sophisticated analysis technique to solve some of the disadvantages of the *t*-tests. Weller (1986) applied a combination of the method of moments and maximum likelihood to estimate means of marker classes and the location of a genetic factor in an F_2 of a cross between inbred lines. Lander and Botstein (1989) and Knapp et al. (1990) improved upon this by proposing analyses for pairs of markers that flank a genetic factor. Luo and Kearsey (1989, 1991) presented a method similar to Weller's for F_2 and backcross or doubled haploid populations.

More recently, Knott and Haley (1992) simulated data for an F_2 population to show that the use of adjacent (flanking) marker pairs improves the power for detection of genetic factors, gives more accurate estimates for the effect and position, and makes the method less sensitive to violations of assumptions such as non-normality.

Several authors have proposed using multiple regression of phenotypic trait measurements on molecular marker variables to detect and estimate effects of genetic factors. Knapp et al. (1990) introduced linear models useful for estimating means of genotypes for backcross and F_2 populations, and they illustrated coding the independent variables for marker classes, equating expected values of associated regression coefficients to parameters, and interpreting the regression coefficients to estimate effect and location of a genetic factor. Martinez and Curnow (1992) used regression on marker allele frequencies for a backcross, minimizing the residual sum of squares by iterating hypothetical recombination fraction distances (t) from zero to the distance between markers. They stated that the residual sum of squares, RSS(t), behaves in a way approximately inversely proportional to the LOD score of the Lander and Botstein (1989) maximum likelihood method of interval mapping. They also illustrated how using flanking marker methods with 2 true effects could also result in a "ghost" effect, and suggested regression with three or more markers to alleviate this situation. Haley and Knott (1992) illustrated a similar regression method for F_2 data and showed that results are nearly identical with those of maximum likelihood. Moreno-Gonzalez (1992a, 1992b) further developed regression models for backcross and F_2 generations from the cross of two inbred lines, and used simulation of a backcross population to show that stepwise multiple regression allows one to detect relatively small additive and dominance effects for independently segregating genetic factors. The wealth of statistical models presented by Moreno-Gonzalez illustrates the flexibility and general applicability of regression for molecular marker analyses.

Maximum likelihood and multiple regression methods have many similarities, but regression has advantages in its computational simplicity. Martinez and Curnow (1992) stated that if the error term in a regression equation were a normally distributed random variable rather than one from a mixture of normal distributions, the regression method would be maximum likelihood. They cited the ability to study performance of procedures algebraically as an advantage for the regression method. Haley and Knott (1992) derived the relationship in test statistics between the likelihood ratio test and the F test of regression and recommended regression as the method of choice, being less complex, computationally faster, and more general than maximum likelihood.

Our goal in this work is to simplify and extend the theory of the multiple regression approach to lend insight and explore properties of regression estimators. We concentrate in this paper on estimating additive effects from F_2 populations. The specific objectives of our work are:

1. To provide a simple and intuitive interpretation of regression of trait variables on molecular marker data for a large F_2 population.

2. To get unbiased (or nealy unbiased) estimates of the additive effect of a single genetic factor with flanking-marker regression estimators.

3. To interpret the signs of regression coefficients as a guide to placement of genetic factors within intervals.

4. To give theoretical variances of regression coefficients so that properties of the estimators are better known and so that theoretical avenues rather than just simulation approaches may be pursued.

Theory

The proposed method involves computing the multiple regression of the trait phenotype (Y) onto the marker genotype, taking the frequency of one parent's alleles (0, 1, 2, or equivalently coding as $-1, 0, 1$) at each marker as the independent variate. The particular case for which these results apply is for a large sample of plants from an F_2 population derived from selfing the cross of two inbreds. Results extend easily to backcross populations.

We begin by establishing some background in the relationships of variables for two markers, and we examine simple linear regression of a measured trait on a single marker variate. These results lead to a very intuitive explanation of the matrix solution of a multiple regression of trait phenotype on marker genotype variables. Next, we present a method to estimate the effect of a trait locus with flanking markers by summing the two regression coefficients and, finally, give algebraic solutions and variances for multiple regression.

The main properties of multiple regression can be illustrated by the simplest appropriate genetic model that has two markers (1 and 2) flanking a single trait locus. For this two-marker model there are four types of gametes produced when selfing the F_1: M_1M_2, m_1m_2, M_1m_2, m_1M_2 with gamete frequencies $(1-c)/2$, $(1-c)/2$, $c/2$, and $c/2$ respectively. Here M_1 and m_1 are the two marker alleles at locus 1, M_2 and m_2 are the marker alleles at locus 2, and c is the recombination fraction between the two loci. These gametes, when randomly combined for the F_2 population, result in the frequencies in Table 1. From this table we can compute expectations and covariances for large samples from an F_2 population. The expected values of X_1 and X_2 are zero, each has variance 0.5, their covariance is $(0.5-c)$, and their correlation is $(1-2c)$.

Table 1 Genotypes, frequencies, and contrast coefficients for large-sample F_2 populations. This table illustrates a linear contrast coding for X_1 and X_2 with coefficients $-1, 0$, and 1.

Genotype	Frequency	X_1	X_2	Y	Value for linear Y at locus 2
$M_1M_1M_2m_2$	$(1-c)^2/4$	1	1	y_1	α
$M_1M_1M_2m_2$	$2c(1-c)/4$	1	0	y_2	0
$M_1M_1m_2m_2$	$c^2/4$	1	−1	y_3	$-\alpha$
$M_1m_1M_2M_2$	$2c(1-c)/4$	0	1	y_4	α
$M_1m_1M_2m_2$	$[2(1-c)^2+2c^2]/4$	0	0	y_5	0
$M_1m_1m_2m_2$	$2c(1-c)/4$	0	−1	y_6	$-\alpha$
$m_1m_1M_2M_2$	$c^2/4$	−1	1	y_7	α
$m_1m_1M_2m_2$	$2c(1-c)/4$	−1	0	y_8	0
$m_1m_1m_2m_2$	$(1-c)^2/4$	−1	−1	y_9	$-\alpha$

Simple linear regression on a single marker

The simple linear regression of a measured trait, Y, on one of the loci, say X_1, gives the estimated additive effect for that marker locus. If a true single genetic factor controlling trait Y is located a recombination distance c_1 from locus 1, and the factor controlling Y has alleles Q and q (thought of as in the same arrangements as M_2 and m_2 in Table 1), the simple linear regression of Y on X_1 approaches, as n gets large,

$$b = [(1-c_1)^2(y_1 - y_9) + 2c_1(1-c_1)(y_2 - y_8) + c_1^2(y_3 - y_7)]/2.$$

This formula gives insight into estimation of additive effects with simple linear regression at each locus. If Y is exactly linear with true additive effect α, and $y_1 \ldots y_9$ are measured as deviations from the mean, then

$$y_2 = y_8 = 0 \quad \text{and} \quad y_1 = y_7 = -y_3 = -y_9 = \alpha.$$

For this single factor with linear effect:

1. $V(Y) = \alpha^2/2$
2. $\mathrm{Cov}(X_1, Y) = 0.5\,\alpha(1-2c_1)$, and
3. $b = \alpha(1-2c_1)$.

This shows that the true additive effect of a trait is underestimated by a factor of $(1-2c_1)$ when simple linear regression is used to estimate α. For example, if $c_1 = 0.1$, $(1-2c_1) = 0.8$, and we estimate the effect as 80% of its true value. Flanking marker methods using maximum likelihood have been developed to alleviate this bias problem. Saturating the genome with markers to reduce map distances also diminishes this source of bias. Regression with two loci flanking the gene of interest can also reduce this bias, as shown below.

Multiple regression for large samples from an F_2

The multiple regression of a vector of n observations $\mathbf{y}$ (assumed to have mean zero) on molecular marker information $\mathbf{X}$, as coded in Table 1, has model equation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where $\mathbf{e}$ is assumed distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$, and with p marker loci, $\mathbf{X}$ has dimension $n \times p$, and $\boldsymbol{\beta}$ has dimension $p \times 1$. A solution vector of estimated regression coefficients is:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The marker map information is contained in the matrix $\mathbf{X}'\mathbf{X}$. We see from Table 1 that $\mathbf{x}_i'\mathbf{x}_j/n$ approaches the covariance of any two markers i and j as the sample size n gets large. Thus, $\mathbf{X}'\mathbf{X}/n$ approaches a variance-covariance matrix with elements $(0.5-c_{ij})$, where c_{ij} is the recombination fraction between loci i and j. We define $c_{ii} = 0$, and for two different loci, we assume $c_{ij} \leq 0.5$. In $\mathbf{X}'\mathbf{X}/n$, each diagonal element approaches the variance of a marker (0.5), and off-diagonal elements go to the covariances $(0.5-c_{ij})$. The matrix $2\,\mathbf{X}'\mathbf{X}/n$ approaches the correlation matrix $\mathbf{R}$ among the marker variables.

$$\mathbf{R} = (r_{ij}) = (1-2c_{ij}),$$

where r_{ij} is the correlation between variables X_i and X_j. Therefore, information from $\mathbf{X}'\mathbf{X}/n$ may be used to map marker loci.

In the same way that $2\mathbf{X}'\mathbf{X}/n$ approaches a correlation matrix among marker variables, $2\mathbf{X}'\mathbf{y}/n$ approaches a vector $\mathbf{a}$ of correlations between markers and true genetic factors. Strictly speaking, for a single true additive effect, $2\mathbf{X}'\mathbf{y}/n$ approaches the vector of marker-trait correlations multiplied by the true additive effect α if different from 1. Using Table 1 and single-gene trait genotypes QQ, Qq, and qq in place of M_2M_2, M_2m_2, m_2m_2, we see that, as n gets large, $\mathbf{X}'\mathbf{y}/n$ approaches the vector $\alpha(0.5-c_i), i = 1 \ldots p$, where c_i is the recombination distance between locus i and the single true additive genetic factor (with effect α). In general, then, the matrix $2\mathbf{X}'\mathbf{y}/n$ approaches the vector $\mathbf{a} = \alpha\,(1-2c_i)$, and, as we saw earlier, $\mathbf{a}$ is the vector of expected values, as n gets large, for simple linear regression estimates for each marker locus.

Multiplying these component parts, we have that $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$ approaches $\mathbf{R}^{-1}\mathbf{a}$ as n gets large. The solution vector for multiple regression is the product of the inverse of the matrix of estimated molecular marker correlations and the vector of simple linear regression estimators. The single-locus regressions, which are homozygous class mean comparisons, are adjusted by the relationships with other marker loci by $(\mathbf{X}'\mathbf{X})^{-1}$. Intuitively, the marker correlation structure (map information) is used to adjust the simple linear regressions in the solution vector for multiple regression.

Although this solution appears simple, in practice there are difficulties in application. The regression coefficients may have high variances due to high correla-

tions between markers. The multicolinearity among closely-spaced markers also causes difficulties in computing and interpreting the partial regression coefficients. However, the large-sample properties of the F_2 allow a simple, intuitive interpretation of multiple regression on marker variables.

Large-sample F_2 regression with two markers

Suppose there is a single additive genetic factor at Q, located between markers 1 and 2. We compute the regression of trait values on the flanking marker variables as:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \rightarrow \mathbf{R}^{-1}\mathbf{a},$$

$$\mathbf{R} = \begin{bmatrix} 1 & (1-2c) \\ (1-2c) & 1 \end{bmatrix},$$

where c is the recombination fraction between the marker loci. The inverse of $\mathbf{R}$ is

$$\mathbf{R}^{-1} = 0.25(c-c^2)^{-1}\begin{bmatrix} 1 & (2c-1) \\ (2c-1) & 1 \end{bmatrix}.$$

We also have that

$$\mathbf{a} = \alpha\begin{bmatrix} 1-2c_1 \\ 1-2c_2 \end{bmatrix},$$

where c_1 and c_2 are the recombination fractions, with Q, of loci 1 and 2, respectively. Finally,

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \rightarrow 0.5\alpha(1-c)^{-1}\begin{bmatrix} 1-2c_2+(c_2-c_1)/c \\ 1-2c_1+(c_1-c_2)/c \end{bmatrix}$$

as n gets large.

An estimator of the true additive effect α can be obtained as the sum of the partial regression coefficients from the flanking-marker regression because (b_1+b_2) approaches $\alpha(1-c)^{-1}(1-c_1-c_2)$ as n gets large. For complete interference and Q between loci 1 and 2 ($c_1+c_2=c$), b_1+b_2 approaches α. Thus, in the case of no double recombination between loci 1 and 2, (b_1+b_2) is asymptotically unbiased for α. For the case of no interference ($c=c_1+c_2-2c_1c_2$) and c less than 0.15, there is only slight bias. For example, α is estimated as 0.98 of its true value for $c_1=c_2=0.10$ when using the sum of the two partial regression coefficients, and 0.94 of its value for $c_1=c_2=0.15$.

The sum of the regression coefficients for flanking-marker regression has a second important property, namely that the theoretical variance of (b_1+b_2) approaches a value not dependent on c^{-1}, but on $(1-c)^{-1}$, and therefore is bounded no matter how small c becomes. The theoretical variance of each of the regression coefficients is obtained from $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$, with expectation $2\mathbf{R}^{-1}(\sigma^2/n)$ for the large F_2 case. The variance of each individual coefficient approaches $0.5(c-c^2)^{-1}\cdot(\sigma^2/n)$, and the variance of the sum approaches $2(1-c)^{-1}(\sigma^2/n)$. Table 2 illustrates the effect of the size of c on these variances. Johnston (1984, p 241) uses this same example to illustrate the variability in estimates of individual partial regression coefficients in the presence of high multicolinearity in contrast to the fairly precise estimation of their sum.

It can also be noted that $b_1/(b_1+b_2) \rightarrow c_2/c$ gives an estimate of the position of Q within the marked chromosome segment. However, for complementary reasons to those given above, this estimate is likely to have a high variance and its properties will not be explored further.

The signs of the coefficients in a two-marker regression can be used to indicate direction from the flanked interval to a linked single additive genetic factor. Suppose the true genetic factor is located at Q and that we have complete interference (no double crossovers). The theoretical solution vectors for the three possible arrangements are:

Case 1. True order is M_1-Q-M_2, that is, $c_1+c_2=c$.

$$\mathbf{R}^{-1}\mathbf{a} = \alpha\begin{bmatrix} c_2/c \\ c_1/c \end{bmatrix}.$$

This is the case in which (b_1+b_2) has expected value α.

Case 2. True order is $Q-M_1-M_2$, that is, $c_1+c=c_2$.

$$\mathbf{R}^{-1}\mathbf{a} = \alpha\begin{bmatrix} (1-c_2)/(1-c) \\ -c_1/(1-c) \end{bmatrix}.$$

Case 3. True order is M_1-M_2-Q, that is, $c_1=c+c_2$.

$$\mathbf{R}^{-1}\mathbf{a} = \alpha\begin{bmatrix} -c_2/(1-c) \\ (1-c_1)/(1-c) \end{bmatrix}.$$

If α is positive, then for Case 1 both coefficients are positive and for Cases 2 and 3, the positive coefficient is on the side closest to locus Q and the negative coefficient is on the side opposite Q.

Table 2 Theoretical variances and covariance, in units of σ^2/n, for regression coefficients of flanking-marker regression for values of recombination fraction (c) between markers

c	$\mathrm{Var}(b_1)=\mathrm{Var}(b_2)$	$\mathrm{Cov}(b_1,b_2)$	$\mathrm{Var}(b_1+b_2)$[a]
0.500	2.000	0.0	4.000
0.250	2.667	−1.333	2.667
0.100	5.556	−4.444	2.222
0.010	50.505	−49.495	2.020
0.001	500.501	−499.499	2.002

[a] $\mathrm{Var}(b_1+b_2)=\mathrm{Var}(b_1)+\mathrm{Var}(b_2)+2\,\mathrm{Cov}(b_1,b_2)$

Large-sample F_2 regression for blocks of markers and trait loci

The above model for two markers will now be extended to include a block of several linked marker loci and trait loci, and the restriction of no double crossovers will be lifted. The general solution for the no-interference, large F_2 regression is the expected value, as sample size gets large, of:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

where $\mathbf{X}'\mathbf{X}$ is the matrix of sums of squares and products of marker allele frequencies, and $\mathbf{X}'\mathbf{y}$ the vector of marker/trait sums of products. To examine the effect of double recombination, consider three loci (1, 2, 3) linked in this sequence, with recombination rates c_{12} and c_{23}. These are the smallest units of chromosome region to be considered, and these probabilities of recombination are defined net of any multiple crossovers. Effective recombination of 1 and 3 occurs when there is only one net crossover in the region between 1 and 3. Then

$$\begin{aligned} c_{13} &= [1 - c_{12}]c_{23} + c_{12}[1 - c_{23}] \\ &= c_{12} + c_{23} - 2c_{12}c_{23}, \text{ so that} \\ r_{13} &= [1 - 2c_{13}] = 1 - 2c_{12} - 2c_{23} + 4c_{12}c_{23} \\ &= [1 - 2c_{12}][1 - 2c_{23}] = r_{12}r_{23}, \end{aligned}$$

where r is the correlation between the subscripted loci.

It may be noted that the influence of double crossovers is to introduce product terms in c. Hence, this model will be referred to as multiplicative, in contrast to the model for complete interference. It can be shown that in general the correlation of two loci is the product of the stepwise correlations between the loci that separate them. This allows the matrix of marker correlations $\mathbf{R}$ to be written, a typical non-diagonal term of which is

$$r_{ij} = \prod_{i=1}^{j-1} r_{i,i+1}.$$

Appendix 1 contains the matrix $\mathbf{R}$ for the multiplicative model.

Solution of the equations $\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{a}$ starts with finding the inverse of $\mathbf{R}$. The inverse has terms of three distinct types, the first and last differing from the remainder along the leading diagonal, and is given in Appendix 1. There are off-diagonal terms only in positions immediately adjacent to the diagonal, all other entries being zero. More formally, the three non-zero terms in each row correspond to entries $(i-1, i)$, (i, i), and $(i, i+1)$ and are:

$-r_{i-1,i}/(1 - r_{i-1,i}^2)$,

$(1 - r_{i-1,i}^2 r_{i,i+1}^2)(1 - r_{i-1,i}^2)^{-1}(1 - r_{i,i+1}^2)^{-1}$, and

$-r_{i,i+1}/(1 - r_{i,i+1}^2)$, respectively.

The upper left and lower right entries of $\mathbf{R}^{-1}$ are $(1 - r_{12}^2)^{-1}$ and $(1 - r_{p-1,p}^2)^{-1}$, respectively. The correlation matrix inverse is the same form as that for an autocorrelation series (Johnston 1984, p 311) if all markers are equally spaced.

Finally, the vector $\mathbf{a}$ is needed. It will first be assumed that there is only one trait locus, Q, and the elements of $\mathbf{a}$ will depend on its position. To confirm that the solutions are not affected by the different type of diagonal terms at the extremes, two situations can be examined. Assume first that the trait locus lies between markers 1 and 2 at recombination distances c_1 and c_2 from each, and with correlations r_1 and r_2 between each trait locus and Q. From the multiplicative relation of correlations

$r_1 r_2 = r_{12}$ and

$$\mathbf{a}' = \alpha(r_1, r_2, r_2 r_{23}, r_2 r_{23} r_{34}, \ldots).$$

It can then be confirmed that the product $\mathbf{R}^{-1}\mathbf{a}$ leads to a vector

$$\boldsymbol{\beta} = \alpha/(1 - r_1^2 r_2^2) \begin{bmatrix} r_1(1 - r_2^2) \\ r_2(1 - r_1^2) \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

all terms except β_1 and β_2 being zero. Secondly, assume that Q is between loci 3 and 4. Then, with similar definitions for r_3 and r_4,

$$\mathbf{a}' = \alpha(\mathbf{r}_{12}\mathbf{r}_{23}\mathbf{r}_3, \mathbf{r}_{23}\mathbf{r}_3, \mathbf{r}_3, \mathbf{r}_4, \mathbf{r}_4\mathbf{r}_{45}, \ldots)$$

and the third and fourth elements of $\boldsymbol{\beta}$ are found to have a similar form to those above. In general, all nonflanking markers have zero coefficients, and those for flanking markers i and j take the form

$$\begin{aligned} \beta_i &= \alpha r_i(1 - r_j^2)/(1 - r_i^2 r_j^2) \\ &= 4\alpha c_j[1 - c_j][1 - 2c_i]/(1 - [1 - 2c_i]^2[1 - 2c_j]^2). \end{aligned}$$

The extension to more than one trait locus is straightforward provided there is no epistasis. The observed marker/trait covariance vector can then be considered as the sum of contributions from each trait locus, and the solution is similarly the sum of separate $\boldsymbol{\beta}$ vectors for each. So for two trait loci:

$$\boldsymbol{\beta}^* = \mathbf{R}^{-1}(\mathbf{a}_1 + \mathbf{a}_2) = \boldsymbol{\beta}_1 + \boldsymbol{\beta}_2.$$

Thus, provided that trait loci are separated by at least two marker loci, their effects can be separately es-

timated, in spite of the fact that they themselves may be linked. When no marker intervenes, their effects are amalgamated as an apparent single locus.

An estimate of a single-locus trait effect α is obtained as $b_i + b_j$, which has expected value

$$\beta_i + \beta_j = \alpha(r_i + r_j)/(1 + r_i r_j)$$

$$= \alpha(1 - c_i - c_j)/(1 - c_i - c_j + 2c_i c_j),$$

for Q in the interval defined by loci i and j. The resulting downward bias in $\hat{\alpha}$ for a locus flanked by markers i and j is given by

$$2c_i c_j/(1 - c_i - c_j + 2c_i c_j) = 2c_i c_j/(1 - c_{ij}).$$

The bias is small, being only 2.4% when $c_i = c_j = 0.1$ and 12% when they are both 0.2.

The asymptotic variance of $\hat{\alpha}$ for a locus flanked by markers i and j is:

$$\mathrm{Var}(\hat{\alpha}) = \frac{2\sigma^2}{n(1 + r_{ij})}[(1 + r_{ij} r_{i-1,i}^2)/(1 - r_{i-1,i}^2)$$

$$+ (1 + r_{ij} r_{j,j+1}^2)/(1 - r_{j,j+1}^2)].$$

In terms of recombination frequencies, this variance is

$$\mathrm{Var}(\hat{\alpha}) = (\sigma^2/n)\left[\frac{1}{2c_{i-1,i}(1 - c_{i-1,i})} + \frac{1}{2c_{j,j+1}(1 - c_{j,j+1})} - \frac{2(1 - 2c_{ij})}{(1 - c_{ij})}\right].$$

This quantity has some unexpected properties, being relatively invariant to changes in c_i and c_j, but strongly affected by the distance of the nearest distal markers ($c_{i-1,i}$ and $c_{j,j+1}$). For $c_{ij} = 0.2$, the coefficient of (σ^2/n) drops from 50 to 10 as these distal recombination distances increase from 0.02 to 0.1. When there is no linkage to distal markers, then $\mathrm{Var}(\hat{\alpha}) = 4(1 + r_{ij})^{-1} \cdot \sigma^2/n = 2(1 - c_{ij})^{-1}\sigma^2/n$. From the point of view of estimating α, it is therefore important to eliminate close distal markers once flanking markers have been identified.

Discussion

The inclusion of parameters for location and gene effects in flanking marker regression models typically leads to non-linear forms that either have to be solved by iteration or transformed into linear forms that are not appropriately constrained (Knapp et al. 1990). One suggested solution has been to preassign values for the location parameter and search for the model with the lowest residual sum of squares (Haley and Knott 1992; Martinez and Curnow 1992). While these models explicitly deal with one trait locus, Moreno-Gonzalez (1992a, b) assumed that trait loci lay midway between flankers so as to provide a linear model that could be generalized to multiple loci. However, this approach requires the markers to be preassigned into pairs to potential flankers.

The important distinction of the present regression model is that no parameters accounting for gene effect or location are explicitly included, and it can be considered a special property of the linear arrangement of genes along the chromosome that the usual regression coefficients can be simply interpreted in these terms. Only markers immediately flanking trait loci are expected to have non-zero values, offering a simple means of recognizing and testing for active chromosome segments. In principle, the approximate position of the trait locus within this segment can also be estimated when double crossovers can be ignored. With complete interference, the sum of the coefficients is an unbiased estimator of the effect of the trait locus they enclose, and there is only a small bias when there is no interference and reasonably dense marking. Provided each trait locus is flanked by its own pair of markers, the magnitude and sign of its effect can be estimated independently of those of all others even when these are linked in a complex system.

The simplicity of the regression model allows various conventional methods of testing for the presence of trait loci, although not all properties have been examined in this specific application. The t-tests for individual coefficients or for adjacent pairs have different properties with respect to the occurrence of false positives and negatives and need to be critically evaluated to find a system that offers the optimum balance.

The theory has been formulated in terms of individuals in an F_2 family. However, the essential property on which the method depends is the form of the matrix R and is directly applicable to any population for which the additive correlation between linked loci is a simple function of $(1–2c)$. Apart from F_2 per se, this applies to any generation in which separate plant populations are derived from each F_2 individual, such as seed bulks or testcrosses. It also applies to backcrosses and doubled haploids, but not to populations of individuals in later generations of selfing. In particular, the correlation among recombinant inbred lines is known to be $(1–2c)/(1 + 2c)$. Even in these cases the regression is likely to be useful, although the specific properties of zero distal coefficients and unbiased estimates are no longer expected.

The estimated effect in regression of trait values on coded marker variables is the least squares linear effect of an allele substitution at the trait locus Q, which is equal to one-half the difference between homozygotes. It was shown by Mather and Jinks (1982) that a dominance effect at a locus in F_2 is independent of the additive effect, and it can similarly be shown that the

dominance effect at a trait locus is independent of the linear allele dose at a marker linked to it. The same is true of epistatic effects, so whereas there are correlations between effects of similar type [all squares and products of $(1-2c)$ terms], those between effects of different types are zero. Hence, although these effects are a source of error, the estimation of additive effects by regression remains unbiased. There should be no difficulty in principle in extending the regression model to allow estimation of dominance effects if required. However, in much of breeding practice only additive and additive epistatic effects can be selected for, and in this context dominance effects are of limited interest. This is clearly true in species in which a homozygous line variety is the objective, and in corn (*Zea mays* L.) and other species potential hybrid parents are compared using testcrosses in which no dominance deviation is expressed.

While the method offers a simple means of labelling and selecting trait alleles, a further attraction is its direct relationship to quantitative selection theory, which is almost entirely formulated in terms of least squares linear models (Falconer 1989). In fact, the regression equation itself is an optimum index for marker-based selection and leads very simply to marker-trait indices of the type discussed by Lande and Thompson (1990).

Summary

Multiple regression is shown to be a method of establishing marker-trait associations to estimate additive effects either for F_2 per se or in crosses to a tester. In this regression, each marker is represented by its own variate, coded to correspond to allele frequency. The method is less concerned with precise trait-gene location than with estimation of additive effects. Regression has a simple, intuitive interpretation, being the product of the inverse of the marker correlation matrix with the vector of simple linear regression estimators. The large-sample solutions are given for regression on two flanking markers and for the general no-interference case, along with variances of these asymptotic estimators. This work should help breeders apply molecular marker technology to their practice.

Appendix

The correlation matrix R and its inverse for the (no-interference) or multiplicative model

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{12} & 1 & r_{23} & \cdots & r_{2p} \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ r_{1p} & r_{2p} & \cdot & \cdot & 1 \end{bmatrix}$$

and under multiplicative model,

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & r_{12}r_{23} & r_{12}r_{23}r_{34} & \cdots & r_{12}r_{23}\cdots r_{p-1,p} \\ r_{12} & 1 & r_{23} & r_{23}r_{34} & \cdots & r_{23}\cdots r_{p-1,p} \\ r_{12}r_{23} & r_{23} & 1 & r_{34} & \cdots & r_{34}\cdots r_{p-1,p} \\ r_{12}r_{23}r_{34} & r_{23}r_{34} & r_{34} & 1 & \cdots & r_{45}\cdots r_{p-1,p} \\ \cdot & \cdot & \cdot & \cdot & \text{etc.} & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & r_{p-1,p} \\ r_{12}r_{23}\cdots r_{p-1,p} & r_{23}\cdots r_{p-1,p} & r_{34}\cdots r_{p-1,p} & r_{45}\cdots r_{p-1,p} & \cdots & 1 \end{bmatrix}$$

This correlation matrix has inverse:

$$\mathbf{R}^{-1} = \begin{bmatrix} (1-r_{12}^2)^{-1} & -r_{12}(1-r_{12}^2)^{-1} & 0 & 0 & 0 \\ -r_{12}(1-r_{12}^2)^{-1} & (1-r_{12}^2r_{23}^2)(1-r_{12}^2)^{-1}(1-r_{23}^2)^{-1} & -r_{23}(1-r_{23}^2)^{-1} & 0 & 0 \\ 0 & -r_{23}(1-r_{23}^2)^{-1} & (1-r_{23}^2r_{34}^2)(1-r_{23}^2)^{-1}(1-r_{34}^2)^{-1} & -r_{34}(1-r_{34}^2)^{-1} & 0 \\ 0 & 0 & -r_{34}(1-r_{34}^2)^{-1} & (1-r_{34}^2r_{45}^2)(1-r_{34}^2)^{-1}(1-r_{45}^2)^{-1} & \text{etc.} \\ 0 & 0 & 0 & 0 & -r_{p,-1,p}(1-r_{p,-1,p}^2)^{-1}(1-r_{p-1,p}^2)^{-1} \end{bmatrix}$$

References

Dudley JW (1993) Molecular markers in plant improvement: manipulation of genes affecting quantitative traits. Crop Sci 33: 660–668

Falconer DS (1989) Introduction to quantative genetics, 3rd edn. Longman Scientific and Technical with John Wiley and Sons, New York

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Johnston J (1984) Econometric methods, 3rd edn. McGraw-Hill Book Co, New York

Knapp SJ, Bridges WC, Birkes D (1990) Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79: 583–592

Knott SA, Haley CS (1992) Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet Res 60:139–151

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

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121: 185–199

Luo ZW, Kearsey MJ (1989) Maximum likelihood estimation of linkage between a marker gene and a quantitative locus. Heredity 63:401–408

Luo ZW, Kearsey MJ (1991) Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. II. Application to backcross and doubled haploid populations. Heredity 66:117–124

Martinez O, Curnow RM (1992) Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480–488

Mather K, Jinks JL (1982) Biometrical genetics, 3rd edn. Rutledge, Chapman and Hall, New York

Moreno-Gonzalez J (1992a) Estimates of marker-associated QTL effects in Monte Carlo backcross generations using multiple regression. Theor Appl Genet 85:423–434

Moreno-Gonzalez J (1992b) Genetic models to estimate additive and non-additive effects of marker-associated QTL using multiple regression techniques. Theor Appl Genet 85:435–444

Soller M, Brody T, Genizi MA (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor Appl Genet 47:35–39

Stuber CW, Edwards MD, Wendel JF (1987) Molecular marker-facilitated investigations of quantitative trait loci in maize. II. Factors influencing yield and its component traits. Crop Sci 27:639–648

Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627–640

Zehr BE, Dudley, JW, Chojecki MA, Maroof S, Mowers RP (1992) Use of RFLP markers to search for alleles in a maize population for improvement of an elite hybrid. Theor Appl Genet 83: 903–911