Heredity **77** (1996) 23–32 *Received 27 July 1995*

On the mapping of QTL by regression of phenotype on marker-type

J. C. WHITTAKER*, R. THOMPSON† & P. M. VISSCHER‡

Department of Applied Statistics, University of Reading, PO Box 240, Whiteknights Road, Reading RG6 2FN, †Rothamsted Experimental Station, Harpenden, Herts AL5 2JQ and ‡Institute of Ecology and Resource Management, University of Edinburgh, Edinburgh EH9 3JG, U.K.

We consider the properties of the regression of phenotype on marker-type in F_2 and backcross populations. We show that this regression provides exactly the same information about the location and effect of QTL as conventional regression mapping. For certain QTL configurations this information is insufficient to map the QTL. Where the QTL can be mapped, the position and effect of QTL can be estimated directly from the coefficients of the regression of phenotype on marker-type. This requires much less computational effort than conventional regression mapping. Examples are given to illustrate the development of the theory.

Keywords: genetic markers, inbred lines, interval mapping, QTL, regression.

Introduction

Much work has now been carried out on the theoretical aspects of mapping quantitative trait loci (QTL). In particular, attention has focused recently on the problems of mapping multiple QTL: the problems that occur when multiple QTL are mapped one by one using standard interval mapping techniques (Lander & Botstein, 1988) have been documented by Haley & Knott (1992) and Martínez & Curnow (1992), whereas both Jansen (1994a,b) and Zeng (1994) have developed methods where the estimates of a QTL's location and effect are improved by including a number of markers in the model as cofactors to absorb the effects of QTL other than the one under study. Jansen's method involves the maximization of a likelihood function by the EM algorithm; the other methods require that the residual sum of squares from a regression model be minimized by a numerical search procedure. Estimation of a QTL's location and effect in the regression models is based on the marker class means for the markers flanking the QTL, written in terms of the location and effect of the QTL; the maximum likelihood methods use these means together with information about the distribution of phenotypes within the marker classes. It has been shown by Haley & Knott (1992) that the two approaches provide virtually identical results, which implies that nearly all the useful information about the QTL is contained in the marker class means. The use of marker class means to locate QTL was first suggested by Mather & Jinks (1977) for a single pair of markers in a backcross population. Regression of phenotype on marker-type has been suggested as a tool for QTL mapping by several authors, notably Stam (1991) and Wright & Mowers (1994). Wright & Mowers (1994) considered what we shall describe as *isolated* QTL; i.e. they assumed that marker intervals contain at most one QTL and that any interval containing a QTL is flanked by intervals which are devoid of QTL. With this model they developed an estimate for the additive effect of a QTL based on the regression coefficients of the markers flanking the QTL in the multiple regression of phenotype on marker-type. In F_2 populations this estimate is asymptotically unbiased when there is complete interference and only slightly biased with no interference provided that the markers are close together.

In this paper we show that in F_2 and backcross populations with QTL having additive effects the regression methods of Haley & Knott (1992) and Martínez & Curnow (1992) are exactly equivalent to regression of phenotype on marker-type. We show how the effect and location of a single QTL can be estimated from the regression coefficients of the markers flanking that QTL without resorting to iterative numerical optimization, and examine the effect of dominance and epistasis on the model. Finally, we note that there are interesting restrictions on the

*Correspondence.

inferences that can be made when two or more adjacent intervals both contain QTL. We shall assume Haldane's (1919) mapping function throughout.

Regression of phenotype on marker-type

Consider an F_2 population resulting from 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. For each individual in the F_2 we have the phenotype y and the marker-type $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ where x_i is 1 if the individual has M_iM_i at the ith marker locus, -1 if the individual has m_im_i at the ith marker locus, and 0 if the individual is heterozygous at the ith marker locus. The QTL genotype $\mathbf{g} = (g_1, g_2, \ldots, g_n)$, where g_i labels the number of Q_i alleles at the ith QTL locus as 1, 0, -1 for the Q_iQ_i homozygote, the Q_iq_i heterozygote and the q_iq_i homozygote, respectively, is unknown, as is the genetic value z. We shall assume initially that the QTL combine additively between and within loci, so $z = \Sigma_{i=1}^{n}\, a_i g_i$ where a_i is the effect of the ith QTL. Nonadditive QTL are discussed in the Dominance and epistasis section.

Expected values of marker-class means

For an F_2 population the expected genotype at a QTL, conditional on the genotype of the flanking markers, can be calculated as a function of r_L and r_R, the recombination fractions between the QTL and the left and right flanking markers, and θ, the recombination fraction between the two flanking markers. Regarding θ as known and writing r_R as a function of θ and r_L we can write the expected genotype as $E(g \mid x_L x_R, r_L)$, a function of the flanking marker-type $x_L x_R$ and the QTL location r_L. For an interval containing a single additive QTL of effect a, the mean of the marker-class with marker-type at the left and right flanking markers x_L and x_R, respectively, is therefore $aE(g \mid x_L x_R, r_L)$. This is easily calculated for an F_2 population (see for example table 1 in Haley & Knott (1992)). Table 1 in our paper contains $E(g \mid x_L x_R, r_L)$ in a simplified form from that used there.

Regression mapping of QTL

Suppose we wish to examine the evidence that a QTL exists in the interval between markers x_i and x_{i+1}. *Regression mapping* (Haley & Knott, 1992; Martínez & Curnow, 1992) uses the differences between means of flanking marker-classes to do this. We hypothesize a single QTL at a given position in the interval, say at recombination fraction r_L from the left-hand flanking marker. Coding the number of alleles from the first line at this hypothetical locus as $h = 1, 0, -1$ as above we can use Table 1 to fit the linear model

$$E(Y) = \beta_0 + \beta_1 E(h \mid x_i x_{i+1}, r_L) \qquad (1)$$

by least squares to estimate β_1, the additive effect of the hypothetical QTL. We now consider the expected value of h conditional on marker-type, $E(h \mid x_i x_{i+1}, r_L)$, as a function of the location of the hypothetical QTL, r_L. Examination of the residual sum of squares obtained by fitting the above model for a number of locations allows the calculation of the most likely position for the QTL. It is then possible to test the hypothesis that a QTL exists in the interval against the null hypothesis that no QTL exists. It should be noted that better estimates will be obtained if a suitable set of markers, **S**, is included in the model as cofactors, to account for the influence of other QTL in the genome (Jansen, 1994a,b; Zeng, 1994); i.e. if we fit the model

$$E(Y) = \beta_0 + \beta_1 E(h \mid x_i x_{i+1}, r_L) + \sum_{j \in \mathbf{S}} \beta_j x_j. \qquad (2)$$

An alternative formulation

It is easily checked (Appendix A) that putting $\lambda = E(g \mid x_L = 1,\ x_R = 0,\ r_L)$ and $\rho = E(g \mid x_L = 0,\ x_R = 1,\ r_L)$ gives $E(g \mid x_L x_R, r_L) = \lambda x_L + \rho x_R$ for any x_L and x_R, so that eqn 1 and

$$E(Y) = \beta_0 + a\lambda x_L + a\rho x_R \qquad (3)$$

are equivalent. Note that here three linear parameters replace the two linear (β_0 and β_1) and one

Table 1 $E(g \mid x_L x_R, r_L)$ for an F_2 population: r_L and r_R are the recombination fractions between the QTL and the left and right flanking markers, and θ is the recombination fraction between the two flanking markers

x_L	x_R	$E(g \mid x_L x_R, r_L)$
1	1	$(1-r_L-r_R)/(1-\theta)$
1	0	$[r_R(1-r_R)(1-2r_L)]/\theta(1-\theta)$
1	−1	$(r_R-r_L)/\theta$
0	1	$[r_L(1-r_L)(1-2r_R)]/\theta(1-\theta)$
0	0	0
0	−1	$[r_L(1-r_L)(2r_R-1)]/\theta(1-\theta)$
−1	1	$(r_L-r_R)/\theta$
−1	0	$[r_R(1-r_R)(2r_L-1)]/\theta(1-\theta)$
−1	−1	$(-1+r_L+r_R)/(1-\theta)$

nonlinear (r_L) parameters of eqn 1. We shall explain in the section Isolated QTL how to derive r_L and a directly from λ and ρ, so removing the need to search a sequence of points for the most likely value of r_L.

Extending this to n QTL it follows that

$$E(z \mid \mathbf{x}) = \sum_{i=1}^{n} a_i E(g_i \mid x_{i_L} x_{i_R} r_{i_L}) = \sum_{i=1}^{n} a_i(\lambda_i x_{i_L} + \rho_i x_{i_L}).$$

Defining the set of QTL flanked on the left by the jth marker as $L(j) = \{i : i_L = j\}$ and the set of QTL flanked on the right by the jth marker as $R(j) = \{i : i_R = j\}$ and writing

$$\beta_j = \sum_{i \in L(j)} \lambda_i a_i + \sum_{i \in R(j)} \rho_i a_i$$

we obtain

$$E(z \mid \mathbf{x}) = \sum_{j=1}^{m} \beta_j x_j.$$

We have shown that $E(z \mid \mathbf{x})$, the function of $\mathbf{x}$ with maximal covariance with z, is linear in x. The coefficients of the linear regression of phenotype on marker-type are chosen so as to give the linear function of $\mathbf{x}$ with maximal covariance with phenotype and therefore with genetic value z. It follows that the coefficients β_j are coefficients of the linear regression of phenotype on marker-type. Also, all the information about the QTL that is present in the marker-group means is included in the regression coefficients. In the rest of this paper we shall examine the consequences of this result for QTL mapping.

Isolated QTL

It is known (Stam, 1991) that if a marker interval contains a single QTL, with the intervals on either side of this interval devoid of QTL, i.e. the QTL is isolated, the regression coefficients of the markers flanking the interval containing the QTL depend only on the QTL within the interval. This property can also be deduced easily from the above, for if the ith QTL is isolated, and flanked by markers j and $j+1$, $\beta_j = \lambda_i a_i$ and $\beta_{j+1} = \rho_i a_i$, so that the appropriate regression coefficients depend only on QTL i. Furthermore, we known from Table 1 that

$$\beta_j = \frac{a_i r_R (1-r_R)(1-2r_L)}{\theta(1-\theta)} \quad \text{and}$$

$$\beta_{j+1} = \frac{a_i r_L (1-r_L)(1-2r_R)}{\theta(1-\theta)}$$

where θ, r_L and r_R are the recombination fractions between markers j and $j+1$, QTL i and marker j and QTL i and marker $j+1$, respectively. Using $\theta = r_L + r_R(1-2r_L)$ to eliminate r_R gives

$$\beta_j = \frac{a_i(\theta - r_L)}{\theta(1-\theta)} \left(\frac{1-\theta-r_L}{1-2r_L} \right)$$

$$\beta_{j+1} = \frac{a_i r_L (1-r_L)}{\theta(1-\theta)} \left(\frac{1-2\theta}{1-2r_L} \right).$$

Dividing β_{j+1} by β_j and rearranging shows that r_L is a root of the quadratic

$$[\beta_{j+1} + \beta_j(1-2\theta)] r_L(r_L - 1) + \beta_{j+1}\theta(1-\theta) = 0.$$

Knowing that $r_L \in (0, 0.5)$, so that only one of the roots is a feasible solution, we see that

$$r_L = 0.5 \left[1 - \sqrt{1 - \frac{4\beta_{j+1}\theta(1-\theta)}{[\beta_{j+1} + \beta_j(1-2\theta)]}} \right].$$

We have shown that given the regression coefficients of the two markers flanking an isolated QTL it is possible to locate that QTL without resort to iterative numerical procedures. Furthermore, a little manipulation gives

$$a^2 = \frac{[\beta_j + (1-2\theta)\beta_{j+1}][\beta_{j+1} + (1-2\theta)\beta_j]}{1\text{-}2\theta}.$$

It is worth noting that the r_L depends only on the ratio of β_j and β_{j+1}, and that both β_j and β_{j+1} must have the same sign as a.

We can therefore reproduce the conventional regression mapping approach for a single QTL by regressing phenotype on each pair of adjacent markers in turn, selecting, from the pairs in which both markers have regression coefficients of the same sign, the pair giving the smallest residual sum of squares and solving the above equations to obtain estimates of the location and effect of the QTL. This will in general give a considerable saving in effort. Note that a pair of adjacent markers with regression coefficients of opposite sign arises when the data are incompatible with the presence of a single QTL between the two markers. We must conclude that either there are two QTL of opposite sign within the interval or there is none, and the regression coefficients are nonzero by chance or because of the effects of QTL in adjoining intervals. In this situation the plot of RSS for this interval used in the conventional regression mapping approach would show a minimum at one of the flanking markers.

Example

We simulated a sample of 300 F_2 individuals using a genome with a single QTL. Phenotypic variance was 1, with the QTL responsible for 10 per cent of the phenotypic variance, which implies $a = 0.447$, the recombination fraction between the QTL and the left-hand flanking marker was 0.08 and the inter-marker recombination fraction was 0.1967. The graph of RSS produced by the conventional regression mapping approach for the interval containing the QTL is given in Fig. 1. Regressions were performed at 10 points equally spaced along this interval. All models fitted a mean term in addition to the QTL effect. The minimum RSS is 308.95 at point 4, and this corresponds to an estimated recombination fraction between the QTL and left flanking marker of 0.09. Regression of phenotype on marker-type for this interval gave regression coefficients of 0.3176 and 0.1732 for the left and right flanking markers, respectively, so on solving as above we estimate $\hat{r}_L = 0.081$ and $\hat{a} = 0.50$. The residual sum of squares at this point is 308.90. To test for significance of this QTL, we would fit a model containing just a mean term, compute the usual F-statistic and compare with the $F_{2,297}$ distribution; here this is highly significant. Note that this is easier than the construction of tests for conventional regression mapping because we have removed the search procedure and so have the usual degrees of freedom for the test.

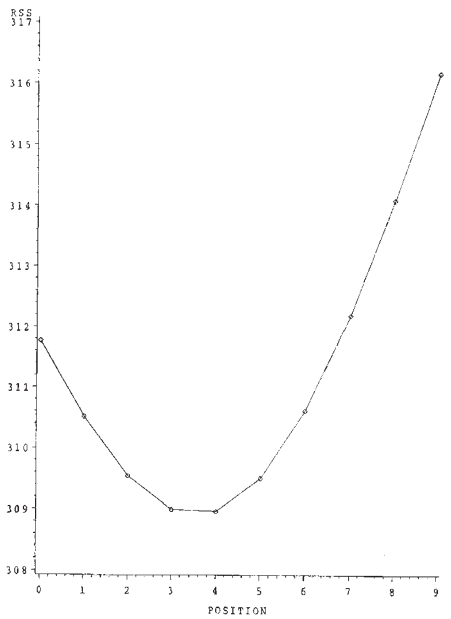

Fig. 1 Residual sum of squares against position for the example in the section on Isolated QTL.

Backcross populations

Suppose that a backcross population is produced such that the possible marker-types are *MM* and *Mm* at each marker. Coding these by the contribution of the gamete from the heterozygous parent, so that *MM* is coded as 0.5 and *Mm* as -0.5, gives the marker-group means given in Table 2. Note that $g = 1$ for *QQ* individuals and $g = 0$ for *Qq* individuals. Defining

$$\lambda = \frac{(\theta - r_L)}{\theta(1-\theta)}\left(\frac{1-\theta-r_L}{1-2r_L}\right) \text{ and } \rho = \frac{r_L(1-r_L)}{\theta(1-\theta)}\left(\frac{1-2\theta}{1-2r_L}\right)$$

as for an F_2 population, we find that $E(g \mid x_L x_R, r_L) = 0.5\ [1 + \lambda x_L + \rho x_R]$. Thus this method extends easily to backcross populations.

Nonisolated QTL

Suppose that QTL i is between the $(j-1)$th and jth markers and QTL $i+1$ is between the jth and $(j+1)$th markers, with no QTL between the $(j-2)$th and $(j-1)$th and $(j+1)$th and $(j+2)$th markers.

Two methods of mapping these QTL using regression mapping have been suggested. The simplest is to treat each QTL as isolated in turn, i.e. use the means of the marker groups x_{j-1}, x_j to map the first QTL by fitting the model

$$E(Y) = \beta_0 + \beta_1 E(h \mid x_{j-1} x_j, r_{(i-1)(j-1)}) + \sum_{k \in \mathbf{S}} \beta_k x_k$$

and the means of the marker groups x_j, x_{j+1} to map the second QTL by fitting the model

$$E(Y) = \beta_0 + \beta_1 E(h \mid x_j x_{j+1}, r_{ij}) + \sum_{k \in \mathbf{S}} \beta_k x_k,$$

searching over a number of putative QTL positions (r_{ij} is the recombination fraction between the ith QTL and the jth marker) to find the minimum value of the residual sum of squares in each case. That this method leads to biased estimates has now been recognized (Haley & Knott, 1992, Martínez & Curnow, 1992): in the language of the section on Isolated QTL this is because in mapping the first QTL we assume that the effect of the second QTL on marker j results from the first QTL. For example, it is shown in Appendix B that for two QTL of effect a located at the mid-points of intervals of recombination fraction 0.18, so that $r_{i(j-1)} = r_{ij} = r_{(i+1)j} = r_{(i+1)(j+1)} = 0.1$, we would estimate (in an infinite population) $\hat{r}_{i(j-1)} = 0.1283$ and $\hat{a} = 1.496a$.

A more sophisticated method, three-marker regression mapping, was developed in Haley & Knott (1992) and Martínez & Curnow (1992) in an attempt to eliminate this bias. This three-marker technique uses the means of the marker groups x_{j-1}, x_j, x_{j+1}, to estimate the position of both QTL simultaneously by fitting the model

$$E(Y) = \beta_0 + \beta_1 E(h_1 \mid x_{j-1} x_j, r_{j-1}) + \beta_2 E(h_2 \mid x_j x_{j+1}, r_j) + \sum_{k \in \mathbf{S}} \beta_k x_k. \quad (4)$$

for a range of r_{j-1} and r_j. The minimum of the two-dimensional residual sum of squares surface produced by this process is then taken as an estimate of the location of the QTL. But we can see from eqn 3 that eqn 4 is equivalent to

$$E(Y) = \beta_0 + \beta_{j-1} x_{j-1} + \beta_j x_j + \beta_{j+1} x_{j+1} + \sum_{k \in \mathbf{S}} \beta_k x_k,$$

so that all the information we can obtain about the location and effect of the QTL i and $i+1$ is contained in the regression coefficients β_{j-1}, β_j, β_{j+1}. This is clearly insufficient to map the two QTL: we cannot estimate the four parameters required from the three pieces of information available. Therefore, the residual sum of squares surface produced by eqn 4 cannot have a unique minimum. Writing $\beta_{j-1} = \lambda_i$, $\beta_j = \rho_i + \lambda_{i+1}$ and $\beta_{j+1} = \rho_{i+1}$ we see that any λ_i, λ_{i+1}, ρ_i, ρ_{i+1} satisfying these equations should be a minimum of the residual sum of squares. It is reasonably easy to specify the set of solutions of these equations: we find that the set of solutions is a line satisfying eqn 6 in Appendix C. This is in contrast to the results reported in Martínez & Curnow (1992), where a minimum in the RSS appeared to be found using a numerical search procedure. This suggested minimum was an artefact of searching over a limited grid of points: the grid of points for which the value of the RSS is calculated will usually be constructed in such a way that the only point on this line of minima that is included in that grid is the actual QTL location, and so this appears to be a unique minimum.

It is worth noting that, as in the isolated QTL case, not all values of β_{j-1}, β_j and β_{j+1} are compatible with a model fitting a QTL in each interval. In particular, β_j must have the same sign as either β_{j-1} or β_{j+1}. Finally, this result extends easily to backcross populations.

Table 2 $E(g \mid x_L x_R, r_L)$ for a backcross population: r_L and r_R are the recombination fractions between the QTL and the left and right flanking markers, and θ is the recombination fraction between the two flanking markers

x_L	x_R	$E(g \mid x_L x_R, r_L)$
0.5	0.5	$0.5[1+(1-r_L-r_R)/(1-\theta)]$
0.5	−0.5	$0.5[1+(r_R-r_L)/\theta]$
−0.5	0.5	$0.5[1+(r_L-r_R)/\theta]$
−0.5	−0.5	$0.5[1-(1-r_L-r_R)/(1-\theta)]$

Example

We now give a simple example to show how the methods discussed in this paper might be applied in practice. A single sample of 2000 F_2 individuals was generated using the genome tabulated in Table 3. This has three chromosomes of length 1 M, each with five evenly spaced markers. Markers are therefore 25 cM apart, which gives a recombination fraction between markers of 0.1967. We have numbered the chromosomes 1, 2 and 3, and the markers are numbered from 0 to 14. Additive QTL were located between markers 0 and 1, 3 and 4, 6 and 7, 12 and 13 and 13 and 14, so that we have three isolated QTL and a pair of nonisolated QTL. Heritability was set to 0.5 and all QTL effects were of equal magnitude, scaled so as to give a phenotypic variance of 1. The first and third QTL effects were negative and the remainder positive.

The regression of phenotype on all markers for this data set gives a residual sum of squares of 1238.9, with regression coefficients

(0, −0.2966), (1, −0.1422), (2, 0.0221), (3, 0.2209),
(4, 0.1956), (5, −0.0189), (6, −0.1922),
(7, −0.2404), (8, 0.0100), (9, 0.0108), (10, −0.0254),
(11, 0.0371), (12, 0.3019), (13, 0.2644),
(14, 0.3370).

This immediately suggests the presence of QTL in intervals (0,1), (3,4), (6,7), (12,13) and (13,14). Regressing on these markers gives a RSS of 1240.0 with coefficients

(0, −0.2975), (1, −0.1323), (3, 0.2296), (4, 0.1962),
(6, −0.2047), (7, −0.2377), (12, 0.3145), (13, 0.2640),
(14, 0.3355). (5)

The small change in RSS suggests that the omitted markers do not flank QTL. Omitting each of the markers 0, 1, 3, 4, 6, 7, 12, 13, 14 from this model in turn results in a considerable increase in RSS. The smallest increase is given by omitting marker 1 to give a RSS of 1249.8 with coefficients

Table 3 Genome used in simulation

Left-hand flanking marker	Chromosome	d	r_L	a
0	1	0.0754	0.07	−0.4472
3	1	0.1250	0.1106	0.4472
6	2	0.1250	0.1106	−0.4472
12	3	0.0754	0.07	0.4472
13	3	0.1782	0.15	0.4472

d is the distance in cM between the QTL and its left-hand flanking marker, r_L the corresponding recombination fraction and a the QTL effect.

(0, −0.3735), (3, 0.1977), (4, 0.1955), (6, −0.2086),

(7, −0.2356), (12, 0.3192), (13, 0.2598), (14, 0.3398).

This difference in RSS is highly significant, so we can conclude that any subset of 0, 1, 3, 4, 6, 7, 12, 13, 14 fits the data signficantly less well than the model including all nine markers. Note also that for the model omitting marker 1, markers 0 and 3 have opposite sign. We pointed out in the section on Isolated QTL that the regression coefficients of markers flanking a single QTL must have the same sign, so this suggests that a marker flanking a QTL has been omitted from the model.

We shall now use eqn 5 to map the QTL. QTL in the intervals (0,1), (3,4) and (6,7) are isolated, so from the section on Isolated QTL we can use their regression coefficients to map the QTL. We get

(0.0720, −0.4413), (0.1030, 0.4391), (0.1176, −0.4562)

as estimates for the recombination fraction between a QTL and its left-hand flanking marker, respectively, markers 0, 3 and 6, and the QTL effect. We know that QTL in intervals (12,13) and (13,14) are not isolated and so we cannot estimate without bias their location and effect: the best we could do would be to obtain the line of locations consistent with these regression coefficients. Note that treating these intervals as isolated gives esimates for location and effect of (0.1022, 0.5965) and (0.1218, 0.6181) for intervals (12,13) and (13,14), respectively, so the bias caused by ignoring nonisolation is considerable.

In conventional regression mapping, to find positions for the five suggested QTL would have required a five-dimensional search using n^5 different combinations with n positions for each marker. Our analysis shows that there are many equivalent models for the nonisolated QTL in intervals (12,13) and (13,14) and that we can map the other three QTL algebraically. This clarfies the identification of nonisolated QTL and reduces considerably the computational effort. The example is idealized: we have a small genome, a large population size and a high heritability, and this makes the analysis rather straightforward. The same method is applicable in more complicated cases, although deciding which markers to include in the model obviously becomes much more difficult, particularly if the population size is too small to allow simultaneous estimation of regression coefficients for all markers with reasonable accuracy. Selecting a 'best' subset of variables to include in a regression is a much studied statistical problem; see for example Miller (1990). Here there is a further complication in that the regression coefficients are to be used to obtain estimates of the underlying parameters, the QTL locations and effects. This imposes certain restraints on the subsets of variables that should be considered: for example, if marker k is included in the model, marker $(k-1)$ or $(k+1)$ should be also. An exception to this rule might be when markers are fitted as cofactors to absorb the effect of QTL which, although too small to be mapped individually, contribute a significant portion of genetic variance.

It should also be noted that there is a particular problem with QTL of small additive effect but large dominance or epistatic effect, for if we select markers according to their additive effects such loci may be excluded from the model. Whether this is important depends on context: for marker-assisted selection, for instance, we may only be interested in additive effects.

Dominance and epistasis

The move from additive to nonadditive QTL has interesting consequences. We consider isolated and nonisolated QTL in turn.

Estimation of dominance and epistasis for isolated QTL

Wright & Mowers (1994) state that in F_2 populations the regression of phenotype on marker-type is unaffected by dominance effects. This is easily seen by considering a single QTL with additive effect a and dominance effect d flanked by markers x_L and x_R. Then $\mathrm{cov}(x_L, y) = \mathrm{cov}(x_L, ag + d\delta)$, where δ is one if $g = 0$ and zero otherwise. Hence

$$\begin{aligned}\mathrm{cov}(x_L, y) &= a\,\mathrm{cov}(x_L, g) + d\,\mathrm{cov}(x_L, \delta)\\ &= a\,\mathrm{cov}(x_L, g) + d[E(\delta x_L) - E(x_L)E(\delta)]\\ &= a\,\mathrm{cov}(x_L, g),\end{aligned}$$

because $p(x_L = 1,\ g = 0) = p(x_L = -1,\ g = 0)$, and this shows that $\mathrm{cov}(x_L, y)$ is unaffected by d; it

follows that the regression coefficient of x_L is also unaffected by d. The argument clearly extends to any number of loci. It should be stressed that this is an asymptotic result in the sense that any finite sample will have, because of chance effects, nonzero covariance between additive and dominance effects.

Thus we can estimate the location and additive effect of QTL by regression on marker-type as in the section on Isolated QTL and then fit the model

$$E(Y) = \sum_{i=1}^{m} \beta_i x_i + \sum_{i=1}^{n} d_i p(g_i = 0 \mid \mathbf{x})$$

to estimate the dominance effects d_i, because $p(g_i = 0 \mid \mathbf{x})$, the probability of heterozygosity given marker-type $\mathbf{x}$, can be calculated given an estimate of the position of the ith QTL. Epistatic effects can obviously be dealt with in the same way.

This may, however, not be the most efficient method of mapping nonadditive QTL: we are essentially using information about the additive effect to estimate QTL location and then using this estimate of location to estimate dominance effects. Better estimates should be obtained in finite samples by using information about additive and dominance effects together, as in the usual regression mapping approach to mapping QTL with dominance.

Mapping nonisolated QTL with dominance

We have seen that in the absence of dominance or epistasis it is impossible to map nonisolated QTL. It might be expected that with dominance the situation becomes even worse, because we have another parameter to estimate for each QTL. Surprisingly, this is *not* so: in Appendix D we present a method of mapping two QTL in adjacent intervals when at least one of the QTL has nonzero dominance effect. It follows that the Martínez & Curnow (1992) three marker regression method will work for this situation. It is possible to map nonisolated QTL in the presence of dominance effects because the contributions to the means of the marker groups x_{j-1}, x_j, x_{j+1}, for $x_{j-1}\,x_j\,x_{j+1} = 1, 0, -1$, arising from dominance result in those marker groups containing more information about the location of the QTL than is present in the absence of dominance, and this more than offsets the extra parameters that must be estimated. These dominance effects also mean that $E(z \mid \mathbf{x})$ is now a nonlinear function of $\mathbf{x}$.

We have seen that the regression coefficients β_i are unaffected by dominance effects so that it is only possible to restrict the position of two QTL in adjacent intervals to a line of possible solutions using the regression coefficients. Thus it is impossible to add dominance effects to the regression model as we did for isolated QTL in the section on Estimation of dominance and epistasis for isolated QTL. The best we could do would be to fit the model

$$E(Y) = \beta_{j-1}x_{j-1} + \beta_j x_j + \beta_{j+1}x_{j+1} + d_i p(g_i = 0 \mid x_{j-1}x_j) + d_{i+1} p(g_{i+1} = 0 \mid x_j x_{j+1})$$

for values of $p(g_i = 0 \mid x_{j-1}x_j)$, $p(g_{i+1} = 0 \mid x_j x_{j+1})$ values to find the $r_{i(j-1)}$, $r_{(i+1)(j+1)}$ minimizing the residual sum of squares. This is equivalent to the usual three marker regression method. A computionally simpler approach would be to fit the additive model to get estimates of β_{j-1}, β_j, β_{j+1} and then calculate the RSS for models including dominance terms along the line of QTL locations compatible with these coefficients. This has the advantage that we search over one dimension instead of the two required by Martínez & Curnow, but as in the section on Estimation of dominance and epistasis for isolated QTL this two-stage process may not make full use of available information. The two approaches will be identical if and only if the minimum obtained by the two-dimensional search lies on the line obtained from the regression coefficients of the additive model, and this may not be true in finite populations.

Discussion

We have presented a method of mapping QTL based on the regression of phenotype on marker-type. The method removes the need for a numerical search procedure as used in conventional regression mapping and allows unbiased estimates of all isolated QTL to be obtained from a single regression. We have assumed that no marker observations are missing, but it should be easy to deal with missing marker observations using the methods of Martínez & Curnow (1994).

The expression of marker-group means as the sum of contributions from the right and left flanking markers is informative in stressing that the QTL we find are really covariances between a marker and phenotype. There is an infinite number of QTL configurations that would result in the same marker group means. The fact that we have only enough information to fit one QTL in the interval does not mean that only one QTL exists. In the absence of further information we should perhaps regard the estimated QTL positions as representing the 'centre of gravity' of loci within the interval that affect the trait. We have also seen that the marker group

means do not provide sufficient information to map additive, nonisolated QTL. Maximum likelihood methods extract slightly more information from the data than do regression methods, but produce almost identical results for isolated QTL. We would expect that this additional information would be sufficient to allow estimates of effect and position for nonisolated QTL but that these estimates would be very imprecise. The situation is analogous with attempting to map a QTL using single marker methods: maximum likelihood can in theory estimate both position and effect, whereas regression cannot.

We have shown that it is impossible to locate nonisolated QTL within their intervals. For the same reasons it may be impossible to distinguish isolated from nonisolated QTL. For example, consider a chromosome in which every other marker interval contains a single QTL, with all the QTL effects having the same sign, say positive: the QTL are therefore isolated and can be mapped. However, all markers will have positive regression coefficients and so we cannot tell from the data that the QTL are isolated; the data could equally well come from a number of models, including one in which every interval contains a QTL. To know that a QTL is isolated we require that the marker to the left of the left-hand flanking marker has a regression coefficient which is either of opposite sign to that of the flanking marker or zero, and the marker to the right of the right-hand flanking marker has a regression coefficient which is either of opposite sign to that of the flanking marker or zero.

These problems may not be important in some applications. In particular, $E(z \mid \mathbf{x})$, the expected genetic value conditional on marker-type, depends only on the regression coefficients β, so that in F_2 populations marker-assisted selection (MAS) can be performed using β with the same efficiency as if the QTL had been mapped. It should be stressed that this is only true for F_2 populations: in subsequent generations the situation is more complicated, although it is easy to show that the result holds for infinite populations in the absence of selection. Also, computer simulations to compare MAS based on regression of phenotype on marker-type with a method more akin to regression mapping showed little difference between the two methods over a 20 generation time span (Whittaker *et al.*, 1995).

The fact that nonisolated QTL with dominance can be mapped is intriguing, but probably not very useful. A limited amount of computer simulation of QTL with dominance has been performed and this suggests that although hypothesis tests for the significance of dominance terms have reasonable power, the estimates of location obtained are poor. It should also be noted that, if a dominance effect is included when mapping nonisolated additive QTL, a minimum will be found in the RSS surface because of chance variations in the values of the marker group means. The difficulties of mapping nonisolated QTL cannot be overemphasized.

Finally, we would expect that epistasis between two nonisolated QTL would allow these QTL to be mapped in the same way as dominance effects, although this has not yet been investigated.

Acknowledgements

J.C.W. was funded by the BBSRC and P.M.V. by the Marker Assisted Selection Consortium of the U.K. pig industry (Cotswold Pig Development, J.S.R. Farms, National Pig Development Company, Newsham Hybrid Pigs, Pig Improvement Company, and the Meat and Livestock Commision). R.T. acknowledges support from MAFF. We thank Chris Haley and Robert Curnow for helpful discussions of this work.

References

HALDANE, J. B. S. 1919. The combination of linkage values and the calculation of distance between loci of linked factors. *J. Genet.*, **8**, 299–309.

HALEY, C. S. AND KNOTT, S. A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69**, 315–324.

JANSEN, R. C. AND STAM, P. 1994a. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.

JANSEN, R. C. 1994b. Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, **138**, 871–881.

LANDER, E. S. AND BOTSTEIN, D. 1988. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

MARTÍNEZ, O. AND CURNOW, R. N. 1992. Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theor. Appl. Genet.*, **85**, 480–488.

MARTÍNEZ, O. AND CURNOW, R. N. 1994. Missing markers when estimating quantitative trait loci using regression mapping. *Heredity*, **73**, 198–206.

MATHER, K. AND JINKS, J. L. 1977. *Introduction to Biometrical Genetics*. Chapman and Hall, London.

MILLER, A. J. 1990. *Subset Selection in Regression*. Chapman and Hall, London.

STAM, P. 1991. Some aspects of QTL analysis. *Proceedings of the VIIIth meeting of the Eucarpia Section Biometrics in Plant Breeding*. Brno.

WHITTAKER, J. C., CURNOW, R. N., HALEY, C. S. AND THOMP-

SON, R. 1995. Using marker-maps in marker-assisted selection. *Genet. Res.*, **66**, 255–265.

WRIGHT, A. J. AND MOWERS, R. P. 1994. Multiple regression for molecular-marker, quantitative trait data from large F_2 populations. *Theor. Appl. Genet.*, **89**, 305–312.

ZENG, Z.-B. 1994. Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.

Appendix A: linearity of $E(g \mid x_L x_R, r_L)$

As in the section on Expected values of marker-class means, let $\lambda = E(g \mid x_L = 1, \; x_R = 0, \; r_L)$ and $\rho = E(g \mid x_L = 0, x_R = 1, r_L)$. From Table 1 this is

$$\lambda = \frac{r_R(1-r_R)(1-2r_L)}{\theta(1-\theta)} \text{ and } \rho = \frac{r_L(1-r_L)(1-2r_R)}{\theta(1-\theta)},$$

so we see immediately that $E(g \mid x_L = -1, \; x_R = 0, \; r_L) = -\lambda$ and $E(g \mid x_L = 0, x_R = -1, r_L) = -\rho$. Also,

$$\lambda + \rho$$

$$= \frac{r_R(1-r_R)(1-2r_L) + r_L(1-r_L)(1-2r_R)}{\theta(1-\theta)}$$

$$= \frac{r_R(1-r_L)(1-r_R-r_L) + r_L(1-r_R)(1-r_L-r_R)}{\theta(1-\theta)}$$

$$= \frac{1-r_L-r_R}{1-\theta}$$

$$= E(g \mid x_L = 1, x_R = 1, r_L) = -E(g \mid x_L = -1, \; x_R = -1, r_L)$$

and $\lambda - \rho$

$$= \frac{r_R(1-r_R-2r_L+2r_Rr_L) - r_L(1-r_L-2r_R+2r_Rr_L)}{\theta(1-\theta)}$$

$$= \frac{(r_R-r_L)[1-r_R(1-r_L)-r_L(1-r_R)]}{\theta(1-\theta)}$$

$$= E(g \mid x_L = 1, x_R = -1, r_L) = -E(g \mid x_L = -1, \; x_R = 1, r_L)$$

so that $E(g \mid x_L, x_R, r_L) = \lambda x_L + \rho x_R$ as required.

Appendix B: bias when nonindependence of QTL is ignored

Suppose that markers 1 and 2 flank QTL 1 and markers 2 and 3 flank QTL 2, with no QTL to the left of marker 1 or right of marker 3. Then treating QTL 1 and 2 as isolated and supposing the recombination fraction between markers 1 and 2 to be θ we would estimate r_{11}, the recombination fraction between marker 1 and QTL 1, by

$$\hat{r}_{11} = 0.5\left[1 - \sqrt{1 - \frac{4\beta_2\theta(1-\theta)}{[\beta_2 + \beta_1(1-2\theta)]}}\right],$$

where $\beta_1 = \lambda_1$, $\beta_2 = \rho_1 + \lambda_2$ and $\beta_3 = \rho_2$. Suppose that the QTL are of equal effect, a, and that $r_{11} = r_{12} = r_{22} = r_{23} = 0.1$, which implies that $\theta = 0.18$. Using the usual formulae for λ_i, ρ_i, we get $\lambda_1 = \rho_1 = \lambda_2 = \rho_2 = 0.49878a$ and substitution into the above gives $\hat{r}_{11} = 0.1283$. Thus $\hat{r}_{12} = 0.0696$ and $\hat{a} = 1.496a$. Note that the bias is considerable, despite the fact that marker 3 has been fitted as a cofactor.

Appendix C: mapping nonisolated additive QTL

We find the minima of the RSS surface supposing that two adjacent intervals both contain QTL, as in the section on Nonisolated QTL. We shall suppose for simplicity that the markers are equally spaced, with θ the intermarker recombination fraction and r_{ij}^* the recombination fraction between the actual QTL i and marker j. Then we showed in the same section that any λ_i, λ_{i+1}, ρ_i, ρ_{i+1} satisfying $\beta_{j-1} = \lambda_i$, $\beta_j = \rho_i + \lambda_{i+1}$ and $\beta_{j+1} = \rho_{i+1}$ should give a minimum of the residual sum of squares surface. Rewriting these equations in terms of recombination fractions and QTL effects we see that any hypothetical pair of QTL with effects a_i, a_{i+1} and position described by $r_{i(j-1)}$, $r_{i(j+1)}$ satisfying

$$\beta_j = \frac{\beta_{(j-1)}r_{i(j-1)}(1-r_{i(j-1)})(1-2r_{ij})}{r_{ij}(1-r_{ij})(1-2r_{i(j-1)})} + \frac{\beta_{(j+1)}r_{(i+1)(j+1)}(1-r_{(i+1)(j+1)})(1-2r_{(i+1)j})}{r_{(i+1)j}(1-r_{(i+1)j})(1-2r_{(i+1)(j+1)})}$$

is consistent with β_{j-1}, β_j, β_{j+1}. Eliminating r_{ij} in the first term on the right-hand side gives

$$\frac{\beta_{(j-1)}r_{i(j-1)}(1-r_{i(j-1)})(1-2r_{ij})}{r_{ij}(1-r_{ij})(1-2r_{i(j-1)})} = \frac{\beta_{(j-1)}r_{i(j-1)}(1-r_{i(j-1)})(1-2\theta)}{(\theta-r_{i(j-1)})(\theta-r_{i(j-1)})(1-\theta-r_{i(j-1)})}$$

and, calculating the second term by symmetry we get

$$\frac{\beta_j}{1-2\theta} = \frac{1}{(\theta-r_{i(j-1)})(1-\theta-r_{i(j-1)})(\theta-r_{(i+1)(j+1)})(1-\theta-r_{(i+1)(j+1)})}$$

$$\times[\beta_{j-1}r_{i(j-1)}(1-r_{i(j-1)})(\theta-r_{(i+1)(j+1)})(1-\theta-r_{(i+1)(j+1)})$$
$$+\beta_{(j+1)}r_{(i+1)(j+1)}(1-r_{(i+1)(j+1)})(\theta-r_{i(j-1)})(1-\theta-r_{i(j-1)})].$$

Writing $x=r_{i(j-1)}(1-r_{i(j-1)})$, $y=r_{(i+1)(j+1)}(1-r_{(i+1)(j+1)})$ and simplifying shows that x, y satisfy

$$0=\beta_j\theta^2(1-\theta)^2+[\beta_j+(1-2\theta)(\beta_{j-1}+\beta_{j+1})]xy$$
$$-\theta(1-\theta)\{[\beta_j+\beta_{j-1}((1-2\theta)]x+[\beta_j+\beta_{j+1}((1-2\theta)]y\}. \quad (6)$$

Thus, given the regression coefficients β_{j-1}, β_j, β_{j+1}, the set of possible locations of the two QTL is the set of $r_{i(j-1)}\in(0,\theta)$, $r_{(i+1)(j+1)}\in(0,\theta)$ satisfying this equation. This can be easily computed for any β_{j-1}, β_j, β_{j+1}.

Appendix D: mapping nonisolated QTL with dominance

Again suppose that markers 1 and 2 flank QTL 1 and markers 2 and 3 flank QTL 2, with no QTL to the left of marker 1 or right of marker 3. Let the recombination fraction between QTL i and marker j be r_{ij}, the recombination fraction between markers i and j be θ_{ij} and the additive and dominance effects of QTL i be a_i and d_i, respectively. Then we can write the means of the marker classes as

$\lambda_1+\rho_1+\lambda_2+\rho_2+d^1_{11}+d^2_{11}$ for the class $x_1=1$, $x_2=1, x_3=1$

$\lambda_1+\rho_1+\lambda_2+d^1_{11}+d^2_{10}$ for the class $x_1=1, x_2=1$, $x_3=0$

$\lambda_1+\rho_1+\lambda_2-\rho_2+d^1_{11}+d^2_{1-1}$ for the class $x_1=1$, $x_2=1, x_3=-1$

and so on, where $d^1_{ij}=d_1p(g_1=0\,|\,x_1=i,\ x_2=j)$, $d^2_{ij}=d_2p(g_2=0\,|\,x_2=i,\ x_3=j)$ and $p(g_1=0\,|\,x_1=i,\ x_2=j)$ has been tabulated in Table 4. Note the following relations:

Table 4 $p(g_1=0\,|\,x_1=i,x_2=j)$ for F_2 populations

x_1	x_2	$p(g_1=0\,\|\,x_1=i,x_2=j)$
1	1	$2r_{11}(1-r_{11}r_{12}(1-r_{12})/(1-\theta_{12})^2$
1	0	$r_{11}(1-r_{11})[1-2r_{12}(1-r_{12})]/\theta_{12}(1-\theta_{12})$
1	−1	$2r_{11}(1-r_{11}r_{12}(1-r_{12})/\theta^2_{12}$
0	1	$r_{12}(1-r_{12})[1-2r_{11}(1-r_{11})]\theta_{12}(1-\theta_{12})$
0	0	$\{[r^2_{11}+(1-r_{11})^2][r^2_{12}+(1-r_{12})^2]\}/[\theta^2_{12}+(1-\theta_{12})^2]$
0	−1	$r_{12}(1-r_{12})[1-2r_{11}(1-r_{11})]/\theta_{12}(1-\theta_{12})$
−1	1	$2r_{11}(1-r_{11}r_{12}(1-r_{12})/\theta^2_{12}$
−1	0	$r_{11}(1-r_{11})[1-2r_{12}(1-r_{12})]/\theta_{12}(1-\theta_{12})$
−1	−1	$2r_{11}(1-r_{11}r_{12}(1-r_{12})/(1-\theta_{12})^2$

$$p(g_1=0\,|\,x_1=1,x_2=1)=p(g_1=0\,|\,x_1=-1,x_2=-1)$$
$$p(g_1=0\,|\,x_1=1,x_2=1)=\frac{\theta^2_{12}}{(1-\theta_{12})^2}p(g_1=0\,|\,x_1=1,$$
$$x_2=-1)$$
$$p(g_1=0\,|\,x_1=1,x_2=-1)=p(g_1=0\,|\,x_1=-1,x_2=1)$$
$$p(g_1=0\,|\,x_1=1,x_2=0)=p(g_1=0\,|\,x_1=-1,x_2=0)$$
$$p(g_1=0\,|\,x_1=0,x_2=1)=p(g_1=0\,|\,x_1=0,x_2=-1)$$

We now show that given the 27 marker group means m_{ijk} for i, j, and k equal to 0 or 1, it is possible to map the two QTL, in spite of their nonindependence. We have

$$m_{111}=\lambda_1+\rho_1+\lambda_2+\rho_2+d^1_{11}+d^2_{11}$$
$$m_{11-1}=\lambda_1+\rho_1+\lambda_2-\rho_2+d^1_{11}+\frac{(1-\theta_{23})^2}{\theta^2_{23}}d^2_{11}$$

and we know that $\beta_1=\lambda_1$, $\beta_2=\rho_1+\lambda_2$ and $\beta_3=\rho_2$ can be found by regression of phenotype on marker-type, because the β_i are independent of the dominance terms. Thus we can find

$$d^1_{11}+d^2_{11}=m_{111}-\beta_1-\beta_2-\beta_3$$
$$d^1_{11}+\frac{\theta^2_{23}}{(1-\theta_{23})^2}d^2_{11}=m_{11-1}-\beta_1-\beta_2+\beta_3$$

and subtracting gives

$$\frac{1-2\theta_{23}}{\theta^2_{23}}d^2_{11}=m_{11-1}-m_{111}+2\beta_3$$

where the right-hand side is known; hence we can estimate d^2_{11}, and therefore d^1_{11}. Similarly, the eqn $d^1_{01}+d^2_{11}=m_{011}+\beta_2+\beta_3$ allows d_{01} to be estimated. From Table 4

$$\frac{d_{11}}{d_{01}}=\frac{2\theta_{12}r_{11}(1-r_{11})}{(1-\theta_{12})[1-2r_{11}(1-r_{11})]},$$

which implies that $-2[d^1_{11}(1-\theta_{12})+d^1_{01}\theta_{12}]r_{11}(1-r_{11})+d^1_{11}(1-\theta_{12})=0$, and this quadratic can be solved for r_{11}. Substitution into the appropriate equations now allows the estimation of r_{21}, a_1, a_2, d_1 and d_2. (Note that we need only $d_1\neq0$ or $d_2\neq0$ for this method to work.)

We stress that this is not the optimal method of mapping nonisolated QTL, because it ignores some of the available information. It is presented to show that the means of the marker classes provide sufficient information to map nonisolated QTL with dominance, and that therefore the usual three marker regression method (Haley & Knott, 1992; Martínez & Curnow, 1992), which does use all available information, can be used to map such QTL.