Heredity **69** (1992) 315–324

Received 26 November 1991

A simple regression method for mapping quantitative trait loci in line crosses using flanking markers

C. S. HALEY & S. A. KNOTT*

*AFRC Institute of Animal Physiology and Genetics Research, Edinburgh Research Station, Roslin, Midlothian EH25 9PS, and *Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, West Mains Road, Edinburgh EH9 3JT, U.K.*

The use of flanking marker methods has proved to be a powerful tool for the mapping of quantitative trait loci (QTL) in the segregating generations derived from crosses between inbred lines. Methods to analyse these data, based on maximum-likelihood, have been developed and provide good estimates of QTL effects in some situations. Maximum-likelihood methods are, however, relatively complex and can be computationally slow. In this paper we develop methods for mapping QTL based on multiple regression which can be applied using any general statistical package. We use the example of mapping in an F_2 population and show that these regression methods produce very similar results to those obtained using maximum likelihood. The relative simplicity of the regression methods means that models with more than a single QTL can be explored and we give examples of two linked loci and of two interacting loci. Other models, for example with more than two QTL, with environmental fixed effects, with between family variance or for threshold traits, could be fitted in a similar way. The ease, speed of application and generality of regression methods for flanking marker analyses, and the good estimates they obtain, suggest that they should provide the method of choice for the analysis of QTL mapping data from inbred line crosses.

Keywords: inbred lines, interval mapping, maximum likelihood, QTL, regression.

Introduction

The development of genetic maps of markers based upon DNA polymorphisms is beginning to provide the experimental geneticist and the plant and animal breeder with powerful tools for the study of quantitative genetic variation. The use of markers to detect individual loci responsible for quantitative genetic variation (quantitative trait loci or QTL) provides much greater power than segregation analysis without marker information (Knott & Haley, 1992). The use of pairs of flanking markers for 'interval mapping' with maximum-likelihood analysis of the data (Lander & Botstein, 1989) provides little extra power for the detection of QTL close to a marker but gives much more accurate parameter estimates than analyses using only a single marker (Knott & Haley, 1992).

There have been several publications detailing maximum-likelihood methods for the analysis of flanking marker data from one or other of the segregating generations derived from a cross between inbred lines (e.g. Weller, 1987; Lander & Botstein, 1989; Knapp *et al.*, 1990; Knapp, 1991; Paterson *et al.*, 1991). Although these methods can provide accurate estimates of parameters and allow hypothesis testing, custom-written software is required for their implementation. Furthermore, iterative numerical methods are required to maximize the likelihood and thus maximum-likelihood becomes increasingly intractable as the model becomes more complex, for example when it is desired to analyse the data for the presence of two or more linked or interacting QTL. Methods based on least squares lack some of the attractive properties of maximum-likelihood, but have been shown to have similar power for the detection of QTL in single marker analyses (Lander & Botstein, 1989; Haley, 1991). Least squares methods can also be readily implemented using one of the many computer statistical packages available. Here we develop regression methods which can be implemented in a standard com-

Correspondence: C. S. Haley.

puter statistical package to perform flanking marker analyses for the detection of QTL. We show that these methods can be used in the same way as maximum-likelihood methods and give very similar estimates. We also show how the regression method can be used to analyse data in which two linked or interacting QTL are present.

Methods

One QTL model

The model applied assumes a QTL (Q) lying between two co-dominant flanking markers (A and B) and is developed for mapping in the F_2 generation of a cross between two inbred lines which carried different alleles for all three loci. We assume that the variance within the three QTL genotypes is the same and is normally distributed, or can be transformed to be so. The genotypes of the two inbred lines crossed are $A_1A_1Q_1Q_1B_1B_1$ and $A_2A_2Q_2Q_2B_2B_2$. The genotypic effects of the three QTL genotypes possible in the F_2 are set at $m+a$, $m+d$ and $m-a$ for Q_1Q_1, Q_1Q_2 (or, equivalently, Q_2Q_1) and Q_2Q_2, respectively, where m is the mid-parent (mean of the homozygotes) and a and d are the additive and dominance deviations, respectively. The recombination fraction between A and Q is r_A and that between Q and B is r_B. In our methods (as in those of Lander & Botstein, 1989 or Knapp *et al.*, 1990) the recombination fraction between the flanking markers is assumed known and fixed at r, this fraction may be estimated from the marker data prior to QTL analyses (e.g. Knott & Haley, 1992). For all the analyses we assume no interference, thus we expect $r = r_A + r_B - 2r_Ar_B$ and we use Haldane's (1919) mapping function to convert distances in Morgans into recombination fractions. The absence of interference is assumed in the methods of Lander & Botstein (1989) and Paterson *et al.* (1991), but Knapp *et al.* (1990) assume complete interference in applying their model.

The expected mean in terms of the putative QTL for each F_2 marker genotype can readily be derived. For example, the gamete $A_1Q_1B_1$ has expected frequency $(1-r_A)(1-r_B)/2$ and the gamete $A_1Q_2B_1$ has expected frequency $r_Ar_B/2$. The homozygous marker genotype $A_1A_1B_1B_1$ has an expected frequency of $(1-r)^2/4$ in the F_2 and the expected frequencies of the three possible QTL genotypes with this marker genotype are $(1-r_A)^2(1-r_B)^2/4$, $2(1-r_A)(1-r_B)r_Ar_B/4$ and $r_A^2r_B^2/4$ for the QTL genotypes Q_1Q_1, Q_1Q_2 and Q_2Q_2, respectively. Summing over QTL genotypes and scaling for the expected frequency of the marker genotype, the expected mean performance of an F_2 individual of homozygous marker genotype $A_1A_1B_1B_1$ is thus:

$$m + a[(1-r_A)^2(1-r_B)^2 - r_A^2r_B^2]/(1-r)^2$$
$$+ d[2(1-r_A)(1-r_B)r_Ar_B]/(1-r)^2.$$

The coefficients of a and d in terms of recombination fractions for each of the nine flanking marker genotypes possible in an F_2 population are given in Table 1. The expectations for other segregating generations or collections of inbred lines derived from an inbred line cross could be easily derived in a similar manner.

The expectations in Table 1 can now be used to fit a and d by multiple regression. To do this for a given interval between two markers, numerical values for the

Table 1 Expectations for the mean genotypic effect of a QTL for all possible flanking marker genotypes in an F_2 population

Marker genotype	Expectation in terms of:	
	a (additive genetic deviation)	d (dominance genetic deviation)
$A_1A_1B_1B_1$	$[(1-r_A)^2(1-r_B)^2 - r_A^2r_B^2]/(1-r)^2$	$[2r_A(1-r_A)r_B(1-r_B)]/(1-r)^2$
$A_1A_1B_1B_2$	$[(1-r_A)^2r_B(1-r_B) - r_A^2r_B(1-r_B)]/r(1-r)$	$[r_A(1-r_A)(1-r_B)^2 + r_A(1-r_A)r_B^2]/r(1-r)$
$A_1A_1B_2B_2$	$[(1-r_A)^2r_B^2 - r_A^2(1-r_B)^2]/r^2$	$[2r_A(1-r_A)r_B(1-r_B)]/r^2$
$A_1A_2B_1B_1$	$[r_A(1-r_A)(1-r_B)^2 - r_A(1-r_A)r_B^2]/r(1-r)$	$[(1-r_A)^2r_B(1-r_B) + r_A^2r_B(1-r_B)]/r(1-r)$
$A_1A_2B_1B_2$	0	$[r_A^2r_B^2 + r_A^2(1-r_B)^2 + (1-r_A)^2r_B^2 + (1-r_A)^2(1-r_B)^2]/[r^2+(1-r)^2]$
$A_1A_2B_2B_2$	$[r_A(1-r_A)r_B^2 - r_A(1-r_A)(1-r_B)^2]/r(1-r)$	$[(1-r_A)^2r_B(1-r_B) + r_A^2r_B(1-r_B)]/r(1-r)$
$A_2A_2B_1B_1$	$[r_A^2(1-r_B)^2 - (1-r_A)^2r_B^2]/r^2$	$[2r_A(1-r_A)r_B(1-r_B)]/r^2$
$A_2A_2B_1B_2$	$[r_A^2r_B(1-r_B) - (1-r_A)^2r_B(1-r_B)]/r(1-r)$	$[r_A(1-r_A)(1-r_B)^2 + r_A(1-r_A)r_B^2]/r(1-r)$
$A_2A_2B_2B_2$	$[r_A^2r_B^2 - (1-r_A)^2(1-r_B)^2]/(1-r)^2$	$[2r_A(1-r_A)r_B(1-r_B)]/(1-r)^2$

coefficients a and d for each marker genotype can be calculated for a putative QTL at several positions (e.g. one centiMorgan (cM) intervals) between two markers. Multiple regression is used to fit m, a and d for each position separately using the numerical values as coefficients for a and d. This provides estimates of a and d, as well as giving regression and residual sums of squares and mean squares allowing the calculation of the regression variance (F) ratio and thus a test for a and d. The position which gives the best fitting model (i.e. produces the smallest residual mean square) gives the most likely position of a QTL and the best estimates of its effect. These operations (calculation of the numerical values for the coefficients of a and d for the marker genotype of each individual, fitting the regression, iteration for different points between the flanking marker, etc.) can all be written in the language of a general computer statistical package. We performed all calculations using the package GENSTAT, carrying out regression analyses using the FIT directive (GENSTAT 5 Committee, 1989). For example, the coefficients of a and d for each of the nine possible marker genotypes for a putative QTL mid-way between two markers 20 cM apart (i.e. $r = 0.1648$; $r_A = r_B = 0.0906$) are shown in Table 2. In order to fit a QTL at this position, individual plants or animals would be given expectations in terms of a and d according to their marker genotype and the model fitted in GENSTAT with the command 'FIT a + d' (a constant, equivalent to m, is fitted by default in GENSTAT).

Comparison of regression and maximum-likelihood methods

In maximum-likelihood analyses, a combined test for the presence of p parameters can be obtained from the maximized likelihood (L_1) of the model in which the p parameters are estimated compared with the maximised likelihood (L_0) from which the parameters are omitted (or set at some value). Then $2\log_e(L_1/L_0)$ provides a test statistic (the likelihood ratio test) which should be asymptotically distributed as a χ^2 with p degrees of freedom (Wilks, 1938). (N.B. Lander & Botstein, (1989) employ the equivalent test $\log_{10}(L_1/L_0)$ — the LOD score.) Regression is maximum likelihood when errors are independent and normally distributed (e.g. Draper and Smith, 1966). In this case the likelihood ratio test can be written in terms of the residual sum of squares of the full model (fitting the regression), and the reduced model (omitting the regression, and the number of observations (RSS_{full}, $RSS_{reduced}$ and n, respectively):

$$\text{likelihood ratio test} = n \log_e(RSS_{reduced}/RSS_{full})$$

(Aitkin *et al.*, 1989).

Equating the sums of squares to products of mean squares and their degrees of freedom in the full model and approximating using the Taylor expansion:

$$\text{likelihood ratio test} \approx pMS_{regression}/MS_{residual}$$

or

$$\approx pF_{regression}$$

where p parameters are fitted. In the methods outlined above segregation of the QTL within marker genotype classes causes failure of the assumptions that the residual errors are normally distributed within marker genotype classes and are the same across classes. Nonetheless, as we show below, the value of $[n \log_e (RSS_{reduced}/RSS_{full})]$ provides a very close approximation to the likelihood ratio test and, furthermore, the parameter estimates from the two methods are very similar.

In order to explore the properties of the regression method we have compared its performance in the analysis of simulated data directly with the maximum likelihood method. The data analysed were a subset of those analysed by Knott & Haley (1992). Briefly, data were simulated for 50 replicates of 1000 F_2 individuals from a cross between two inbred lines. The genome of each individual consisted of a pair of chromosomes 100 cM in length carrying marker loci at the ends and at 10-cM intervals (i.e. 11 markers in total). Intervening markers were omitted in some analyses giving more widely spaced markers. In the first analyses a single QTL at 25 cM from one end of the chromosome was used, with additive deviation (a) of 0.0, 0.125, 0.25, or 0.5 residual phenotypic standard deviations (d was set at 0). A QTL of this magnitude would explain, respectively, approximately 0.0, 0.8, 3.0 or 11.1 per cent of the total phenotypic variance in the F_2.

Analyses of the data were by regression as described above or by maximum likelihood using flanking

Table 2 The coefficients of a and d for each of the nine possible marker genotypes for a putative QTL mid-way between two markers 20 cM apart (i.e. $r = 0.1648$; $r_A = r_B = 0.0906$)

Marker genotype	Expectation in terms of:	
	a	d
$A_1A_1B_1B_1$	0.9803	0.0195
$A_1A_1B_1B_2$	0.4902	0.5
$A_1A_1B_2B_2$	0.0	0.5
$A_1A_2B_1B_1$	0.4902	0.5
$A_1A_2B_1B_2$	0.0	0.9625
$A_1A_2B_2B_2$	−0.4902	0.5
$A_2A_2B_1B_1$	0.0	0.5
$A_2A_2B_1B_2$	−0.4902	0.5
$A_2A_2B_2B_2$	−0.9803	0.0195

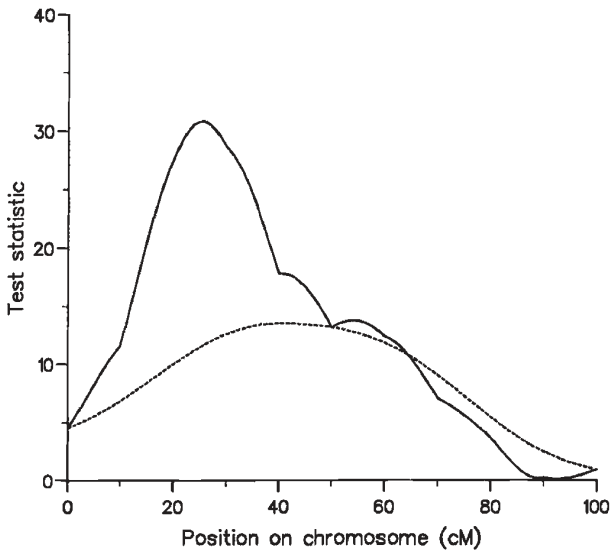

Fig. 1 Curves produced by the analysis of a single set of data using the regression or maximum likelihood methods to fit a model for a single QTL. The data were generated with a single QTL with additive deviation $a = 0.25$ (one half the F_2 residual standard deviation between homozygotes) 25 cM along the chromosome. This QTL explains approximately 3 per cent of the phenotypic variance in the F_2. Eleven markers on a 100 cM chromosome were simulated and in the analyses either all 11 markers (at 10-cM intervals, ——) or three markers (at 50-cm intervals, – – –) were used. In both analytical methods the putative QTL is positioned sequentially at 1-cM intervals along the chromosome and the model fitted at each point. The height of the curves is given by the test statistic which for the maximum-likelihood method is $2\log_e$ of the ratio of the likelihoods (QTL in that position/no QTL) and for the regression method is $[n \log_e(RSS_{reduced}/RSS_{full})]$. On the scale used the curves produced by the two methods are indistinguishable.

markers as outlined by Paterson *et al.* (1991) and detailed by Knott & Haley (1992). For a single set of data, evidence for a QTL can be plotted graphically as shown by Lander & Botstein (1989). In these graphs the likelihood ratio test statistic is plotted at regular (e.g. 1-cM) intervals along the chromosome, with the peak value representing the most likely position of a QTL. (N.B. Lander & Botstein (1989) chose to plot $\log_{10}$ of the ratio, whereas we prefer to plot $2\log_e$ of the ratio in order to facilitate comparison with the χ^2 distribution and with the regression method). The analogous graph for the regression method is to plot the value of $[n \log_e(RSS_{reduced}/RSS_{full})]$ or of $(pF_{regression})$ for each point along the chromosome. An example of the curves produced by the two methods for a single set of data analysed with two marker densities is shown in Fig. 1. Inspection of Fig. 1 shows that the values of the likelihood ratio test from the maximum-likelihood method and of $[n \log_e(RSS_{reduced}/RSS_{full})]$ from the regression method are so similar that the two curves cannot be distinguished. The curve for $(pF_{regression})$ would be very similar to those for the other two test statistics, with the greatest deviation between the curves being at higher values.

For the purposes of summarizing the analyses over the 50 replicates simulated for each combination of parameters, only the interval in which the QTL was placed was analysed. In the maximum likelihood analyses the distance in centiMorgans of the QTL from the first marker (cM_A) was estimated along with the other parameters (i.e. a, d and $\sigma^2_{residual}$). In the regression analyses the model was fitted at 1-cM intervals and the best fitting model selected to provide estimates of cM_A, a, d and $MS_{residual}$. In order to compare the two types of analysis, the correlation and the regression of cM_A and a estimated in the regression model, on the same parameters estimated by maximum-likelihood, were calculated as were the correlation and the regression of $\sqrt{MS_{residual}}$ on $\sigma_{residual}$ and of $[n \log_e(RSS_{reduced}/RSS_{full})]$ on the likelihood ratio test. These statistics are shown together with the mean values of the estimated parameters in Table 3. With both types of analysis the estimates of d were close to 0 and no more were significant than would be expected due to chance; therefore statistics for this parameter are not given.

Table 3 shows that the estimates from the two methods are very similar overall and estimates for the same set of data are very closely related, with both correlations and regressions often close to unity. The lowest correlations and regressions were for the estimated position when no QTL was simulated or its effect was small, even here the lowest correlations and regressions between estimates were about 0.72. The correlation and regression between test statistics from the two methods never fell below 0.96. Inspection of the results showed that the lower correlations for estimated position, when the simulated QTL was of small or no effect, were due to a few of the replicates (at most three out of 50) for which the two methods gave very different estimates of position (at opposite ends of the interval) for a QTL of small estimated effect. Reanalysis of these datasets by maximum-likelihood, using the regression estimates as initial values, resulted in maximum-likelihood estimates close to those from regression and a slightly increased test statistic, indicating that the maximum-likelihood method had reached a local maximum in the initial analyses. With these new estimates correlations between the two methods were greatly improved, for example, increasing from 0.721 to 0.996 for position for the datasets simulated with $a = 0.125$ and 50-cM spaced markers. Note that the correlations between estimated parameters and test

Table 3 Comparison of regression and maximum-likelihood analyses of simulated data. For each combination of simulated parameters 50 replicates were analysed with either 10 or 50 cM spaced markers (QTL 5 or 25 cM from first marker, respectively). The table shows mean estimates of the additive genetic deviation (a) of the QTL and its distance from the first marker (cM_A), the residual standard deviation and the test statistic (with the standard deviation of the estimates over replicates in parentheses). The correlation between parameter estimates from the two methods and the slope of the regression of the estimate from the regression method on that from the maximum-likelihood method are also shown

		10 cM spaced markers				50 cM spaced markers			
Simulated QTL effect		a	cM_A	Residual s.d.	Test statistic	a	cM_A	Residual s.d.	Test statistic
$a=0.0$	Mean (ML method)	−0.010	4.81	1.002	2.87	0.002	29.67	1.001	3.06
		(0.055)	(4.59)	(0.025)	(2.43)	(0.061)	(20.66)	(0.025)	(1.77)
	Mean (regression method)	−0.010	4.90	1.003	2.88	−0.001	27.80	1.003	3.12
		(0.055)	(4.60)	(0.025)	(2.42)	(0.061)	(20.89)	(0.025)	(1.72)
	Correlation	0.998	0.900	1.000	1.000	0.977	0.766	0.971	0.991
	Slope	0.993	0.902	1.002	0.994	0.987	0.775	0.983	0.962
$a=0.125$	Mean (ML method)	0.130	4.22	1.003	10.10	0.123	24.75	1.002	7.42
		(0.044)	(3.97)	(0.022)	(4.75)	(0.061)	(15.36)	(0.021)	(4.65)
	Mean (regression method)	0.130	4.74	1.005	10.15	0.121	25.42	1.006	7.41
		(0.042)	(4.04)	(0.022)	(4.69)	(0.063)	(15.49)	(0.021)	(4.62)
	Correlation	0.987	0.836	1.000	0.998	0.970	0.721	0.992	0.998
	Slope	0.943	0.845	0.997	0.986	0.987	0.727	0.988	0.991
$a=0.25$	Mean (ML method)	0.258	5.56	0.997	33.14	0.247	24.18	0.996	20.53
		(0.047)	(3.04)	(0.020)	(11.85)	(0.069)	(10.56)	(0.020)	(8.83)
	Mean (regression method)	0.258	5.58	0.999	33.18	0.247	24.02	1.001	20.49
		(0.048)	(3.06)	(0.020)	(11.95)	(0.068)	(10.58)	(0.020)	(8.80)
	Correlation	0.999	0.990	0.999	1.000	0.997	0.997	0.976	0.999
	Slope	1.022	0.999	0.999	1.008	0.990	1.008	0.938	0.996
$a=0.5$	Mean (ML method)	0.500	5.00	0.995	110.00	0.497	25.57	0.994	64.76
		(0.047)	(1.74)	(0.022)	(19.59)	(0.061)	(4.76)	(0.025)	(15.28)
	Mean (regression method)	0.500	5.04	1.002	109.8	0.496	25.52	1.025	64.34
		(0.047)	(1.75)	(0.021)	(19.44)	(0.062)	(4.80)	(0.022)	(15.23)
	Correlation	0.994	0.984	0.995	0.998	0.990	0.996	0.950	0.997
	Slope	0.997	0.987	0.980	0.991	1.000	1.004	0.866	0.994

statistics from the two methods are likely to be underestimated. This is because the regression was fitted at fixed 1-cM points along the chromosome and so accuracy in this dimension is restricted to the nearest centiMorgan, whereas the recombination fraction was estimated along with other parameters in the maximum-likelihood analyses.

Two QTL model

The extension for the analysis of two or more linked QTL is trivial and simply consists of replicating the method used for a single QTL in two or more dimensions. Thus, for example, the regression model now fits a putative QTL between the third and fourth markers on a chromosome at the same time as fitting one between the first and second markers. The chromosome of interest is searched in two dimensions (each representing one QTL) to the desired precision to find the positions for the two QTL that give the best fitting model. This process can be visualized graphically by extension of the method for one QTL into an additional dimension. Figure 2 shows the curve produced by regression analysis fitting only a single QTL to a set of data produced with two linked QTL, 25 and 75 cM from one end of the chromosome (heights given by $n \log_e(RSS_{reduced}/RSS_{full})$]. Figure 3 shows the surface produced fitting two QTL to the same set of data. Inspection of Fig. 2 might lead to the conclusion that a single QTL is located at the centre of the chromosome. Lander & Botstein (1989) suggest that, if two QTL are suspected on a chromosome, the position and effect of

one should be fixed whilst estimating the position and effect of a second. This strategy would be ineffective for the data used here, because the obvious place to fix the first QTL would be at the centre of the chromosome. Figure 3 reflects more accurately the data generated, with the maximum suggesting two QTL located approximately 30 and 90 cM from one end of the chromosome. Note that Fig. 2 is effectively a slice through the three-dimensional surface shown in Fig. 3 down the leading diagonal and this is equivalent to placing the two QTL in the same position.

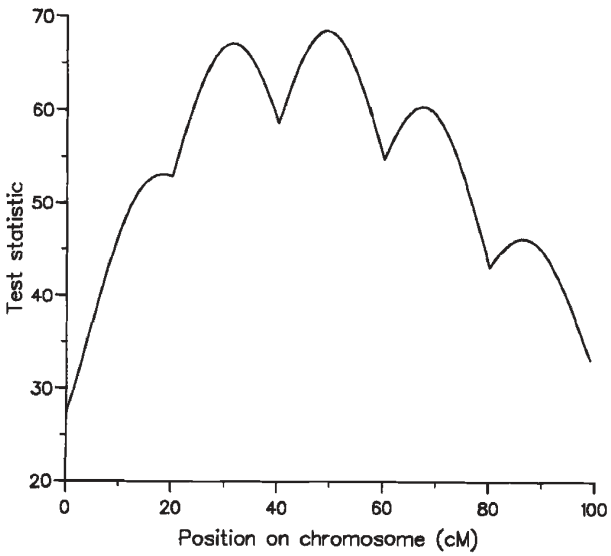

Fig. 2 Curve produced by the analysis of a single set of data using the regression method to fit a model for a single QTL. The data were generated with two QTL each with additive deviation $a = 0.25$ (one half of the F_2 residual standard deviation between homozygotes) positioned 25 and 75 cM along the chromosome. The QTL were in association in the parental lines. Six markers on a 100-cM chromosome were simulated and used in the analyses.

To evaluate the fitting of two QTL by regression the analysis of simulated data was again employed. Data were generated for a chromosome 100 cM in length with markers at 20-cM intervals and two QTL each with additive deviation $a = 0.25$. Three situations were explored. First, where the two QTL were placed 25 and 75 cM from one end of the chromosome and the parental inbred lines were in association (i.e. one carried both increasing alleles and the other both decreasing alleles). Secondly, with the QTL also placed at 25 and 75 cM but where the parental lines were in dispersion (i.e. each carried one increasing and one decreasing allele). Thirdly, where the two QTL were placed 25 and 45 cM from one end of the chromosome and the parental lines were in association. For each situation 20 replicates were generated and analysed. Analysis consisted of initially fitting a single QTL and selecting the best fitting regression followed by fitting two QTL and selecting the best fitting regression, in both cases additive effects only were fitted and resolution was to 1 cM. The computational costs of fitting two QTL were eased by finding an initial maximum using resolution to 5 cM then using a resolution of 1 cM around this point.

Table 4 gives the mean estimates of the QTL effects and positions from the 20 replicate analyses for each

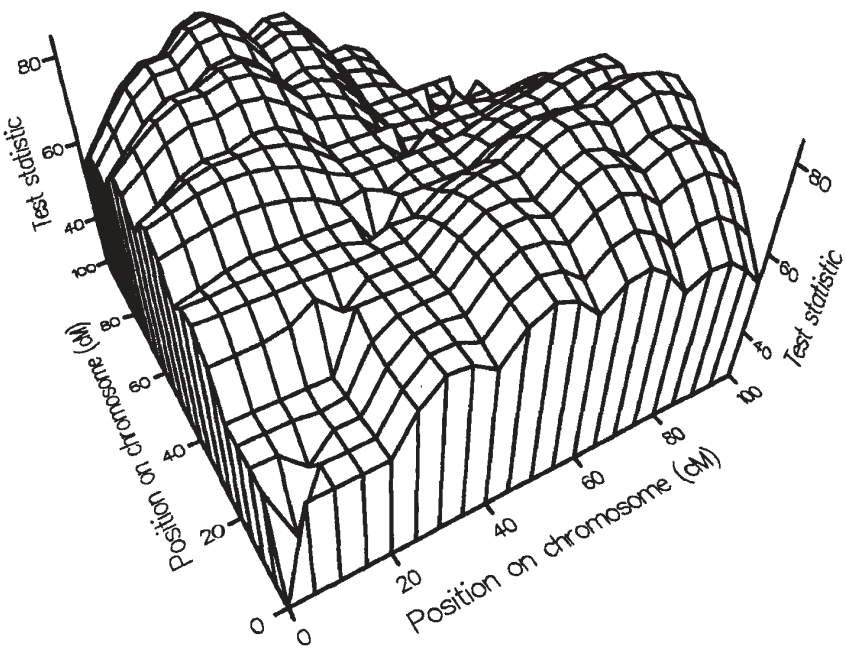

Fig. 3 A three-dimensional surface produced by regression analysis fitting two QTL to the same data as used to produce Fig. 2. Each QTL was moved 5 cM at a time to cover the whole area and at each point $[n \log_e(RSS_{reduced}/RSS_{full})]$ gives the height of the surface.

Table 4 Results from regression analyses of simulated data containing two linked QTL. For each combination of simulated parameters 20 replicates of 1000 F_2 individuals were analysed. The genome consisted of a 100 cM chromosome with 20 cM spaced markers. Simulated QTL were with additive deviation $a = 0.25$ at 25 and 75 cM in association in model 1, at 25 and 75 cM in dispersion in model 2 and at 25 and 45 cM in association in model 3. Model 4 was the same as model 1 except for the presence of interaction between the two QTL ($aa = 0.5$). The table shows mean parameter estimates and test statistics [$n \log_e(RSS_{reduced}/RSS_{full})$] for the presence of one versus no QTL, two versus one QTL and two locus interaction. The standard deviations of the estimates over replicates are given in parentheses

	Test			First QTL		Second QTL		
Model	1 versus 0 QTL	2 versus 1 QTL	Interaction	Position	Effect	Position	Effect	Interaction
1	62.21	15.89	—	22.35	0.263	78.90	0.249	—
	(14.27)	(4.80)	—	(5.13)	(0.061)	(8.16)	(0.045)	—
2	16.76	20.01	—	22.30	0.260	76.80	−0.262	—
	(5.92)	(6.70)	—	(5.10)	(0.060)	(6.86)	(0.039)	—
3	83.07	5.67	—	25.65	0.325	65.10	0.173	—
	(21.92)	(5.58)	—	(7.51)	(0.183)	(16.75)	(0.188)	—
4	58.44	13.05	—	25.10	0.256	76.50	0.245	—
	(17.74)	(6.24)	—	(5.66)	(0.060)	(6.31)	(0.062)	—
4	—	—	47.09	24.90	0.248	74.85	0.246	0.522
	—	—	(15.59)	(4.20)	(0.048)	(4.07)	(0.055)	(0.090)

Table 5 Model for two locus epistasis in terms of the additive and dominance deviations of each QTL (a and d, respectively) and additive–additive (aa), additive–dominance (ad and da) and dominance–dominance (dd) components (e.g. Mather & Jinks, 1982)

	Genotype for second QTL		
	Q_1Q_1	Q_1Q_2	Q_2Q_2
Genotype for first QTL			
Q_1Q_1	$m + a_1 + a_2 + aa$	$m + a_1 + d_2 + ad$	$m + a_1 - a_2 - aa$
Q_1Q_2	$m + d_1 + a_2 + da$	$m + d_1 + d_2 + dd$	$m + d_1 - a_2 - da$
Q_2Q_2	$m - a_1 + a_2 - aa$	$m - a_1 + d_2 - ad$	$m - a_1 - a_2 + aa$

situation. Table 4 also gives the mean test statistic [$n \log_e(RSS_{reduced}/RSS_{full})$] showing the improvement in fitting a single QTL over fitting no QTL and the test statistic showing the improvement in fitting two QTL over fitting a single QTL (here RSS_{full} and $RSS_{reduced}$ are the values from the two models being compared). The mean statistic for the test of one versus no QTL suggests the presence of a QTL in all models, although, compared to data with a single QTL of effect $a = 0.25$, test statistics where QTL in association were simulated were inflated and those with QTL in dispersion were deflated due to the covariance between linked QTL. Including a second QTL in the model gives a marked improvement in fit for the two models where the QTL were 50 cM apart and on average the estimates of parameters were very good. QTL which are only 20 cM apart were difficult to separate, with some replicates showing no improvement in fit when a second QTL was included in the model, and in consequence the parameters were poorly estimated.

Interactions between QTL

Regression can also be used to fit models allowing for interactions between QTL. The principle is identical in that genetic expectations are calculated for each marker genotype and the model is fitted by regression. This is somewhat more difficult to carry out with interactions between loci as two or more pairs of flanking markers must be considered simultaneously in calculating the genetic expectations for any individual. Nonetheless, with the simple common model for two

locus interactions (e.g. Mather & Jinks, 1982) shown in Table 5, the epistatic expectations for each individual can be rapidly calculated as the product of the two single QTL expectations (given in Table 1) for an individual of that marker genotype. For example, the expectation for a particular marker genotype for the coefficient of additive–additive epistasis (*aa*), for two QTL at different given positions in the same marker interval, or in two separate marker intervals, is the product of the coefficeint of *a* for a QTL at the first position and the expectation for *a* for a QTL at the second position.

Analysis of simulated data was again used to evaluate the use of regression when epistatic interactions were present. Data were generated for a chromosome 100 cM in length with markers at 20 cM intervals and two QTL each with additive deviation $a = 0.25$ and additive–additive epistasis of $aa = 0.5$. The two QTL were at 25 and 75 cM from one end of the chromosome and the parental inbred lines were in association. For this situation 20 replicates were generated and analysed. Analysis consisted of initially fitting a single QTL and selecting the best fitting regression followed by fitting two non-interacting QTL and selecting the best fitting regression, followed by fitting two QTL with additive–additive interaction.

Table 4 give the mean estimates of the QTL effects and positions from the 20 replicates for the analyses performed both with and without the interaction term, as well as the mean test statistic showing the improvement in fitting two QTL over fitting a single QTL and that showing the improvement due to the inclusion of an interaction term. On average, QTL effects were well estimated whether or not an interaction term was included in the model, this is probably because the mean effect of each individual QTL was not affected by the interaction model used in this example. Inclusion of the interaction term improved the fit of the model markedly and also slightly reduced the standard deviation of the estimates over replicates.

Other extensions

The method has been developed for the analysis of data from crosses between inbred lines where the F_2 can be considered a single homogeneous population (e.g. a plant population in a single randomized plot). In other cases, although F_2 families may be genetically homogeneous there may be environmental variation (e.g. in crosses between inbred lines of animals). Alternatively, the cross may be between sufficiently diverse lines for them to be fixed for different alleles at many or all markers and QTL, even though the lines crossed were not inbred. In both of these cases the regression method can be used for analysis. Between plot or site environmental variation can be removed by the inclusion of a fixed effect for each site in the model and other factors or covariates could also be included without greatly increasing the difficulty of fitting the model. Between family variation can be removed either by the inclusion of an effect for each family or by fitting the regression in the context of a restricted maximum-likelihood model (Patterson & Thompson, 1971) in which a between family variance component is fitted.

The regression and maximum-likelihood interval mapping methods are suitable for the analysis of normally distributed quantitative data (or data that can be transformed to normality). In a generalized linear model context, the principles of the regression method could be applied to the analysis of data of other types (McCullagh & Nelder, 1983). An example would be threshold data where the population falls into two classes (e.g. died/survival, susceptible/resistant) but it is supposed that there is an underlying continuous distribution which is made up of contributions from the environment as well as a number of QTL. For analysis of these data the expectations of individual marker genotypes would be exactly the same as for normally distributed data, but now these would be expectations for the underlying distribution. Instead of fitting the model using ordinary multiple regression with a normal error it could be fitted using a binomial error. Again this model could be fitted for points along the chromosome using a general statistical package such as GENSTAT (GENSTAT 5 Committee, 1989) and the evidence for the presence of a QTL displayed graphically as a likelihood curve (or surface for two QTL).

Discussion

We have shown here that it is possible to fit flanking marker models for the detection of QTL by regression and that the regression method gives very similar results to the maximum-likelihood method. The regression method provides good estimates for the positions and the effects of QTL. The use of regression not only eases the analysis of experimental data but also allows thorough study of the power of flanking marker methods both through simulation and theoretically. The latter would be possible because the expectations, such as those in Table 1, can be used to predict the magnitude of the regression (and thus the test statistic) for a QTL of given size in a given position and this can be used to predict the power through the use of non-central F or χ^2.

Our results show that $[n \log_e(RSS_{reduced}/RSS_{full})]$ is a very close approximation to the likelihood ratio test. The close similarity between both statistics and para-

meter estimates from the two methods indicates that the great majority of the information is contained in mean differences between marker genotypes with little coming from the within genotype distribution. Theoretically, the regression method should suffer from the failure of the assumption of normality within marker genotype due to the segregation of the QTL but in practice this does not seem to be important for QTL of realistic size.

The values of $[n \log_e(RSS_{reduced}/RSS_{full})]$ or of $(F_{regression})$ provide for hypothesis testing, but some comment is required on the degrees of freedom of the test and on the level of significance required when multiple tests are being performed. For a single QTL only one (for an additive deviation) or two (for additive and dominance deviations) effects are fitted, and the procedure requires finding the position along the chromosome that produces the best fit. This suggests that the test uses the number of genetic effects fitted (p) plus one (for the position) degrees of freedom. This is borne out by the mean value of approximately three found for $[n \log_e(RSS_{reduced}/RSS_{full})]$ for analyses where no QTL was simulated and an additive and dominance effect fitted, three being the expected value under the null hypothesis for a test distributed as a χ^2 with 3 d.f. Thus the divisor of $SS_{regression}$ should be adjusted accordingly if $(F_{regression})$ is being used for hypothesis testing. For each additional QTL included in the model further degrees of freedom will be used equal to the number of genetic effects fitted plus one for the estimation of the position of the QTL.

The question of the appropriate level at which to set the significance when testing for the presence of QTL in many intervals is problematic. The LOD threshold of 3, customary in analysis of human genetic linkage (Morton, 1955), is equivalent to a likelihood ratio test statistic of 13.8 with 1 d.f. or a significance level of approximately 0.0002. The appropriate test statistic for significance in the context of interval mapping has been discussed by Lander & Botstein (1989) who suggest a LOD threshold of between 2 and 3 (corresponding to values of the likelihood ratio test of between 9.2 and 13.8) depending upon genome size and marker density [although it is not clear to these authors how many degrees of freedom Lander & Botstein (1989) consider this test to have]. Lander & Botstein (1989) quote a LOD threshold of 2.4 as being appropriate for the tomato genome, this is approximately a significance level of 0.001 for a 1 d.f. test. The 0.001 significance level is appropriate if an overall significance level of 0.05 is required and 50 independent tests are being performed, and this may provide a reasonable benchmark for situations where more than 50 tests are being performed but tests for adjacent regions are correlated due to linkage. In light of this uncertainty, and in the absence of large scale simulation studies mimicking each experimental situation, it may be test to treat the likelihood ratio test or equivalently the LOD score or the regression variance ratio as test statistics, large values of which support the hypothesis that a QTL is present.

In summary, the regression method we describe provides a relatively simple way of fitting flanking marker models to data derived from inbred line crosses. We have demonstrated the use of the methods by reference to any segregating generation or collection of recombinant inbred lines derived from a cross between inbred lines. There seems little advantage to be gained from resort to maximum likelihood methods for the analysis of these types of data. The relative simplicity and computational rapidity of the regression method makes it easier to fit models for two or more linked and/or interacting QTL, and these can also give good estimates of QTL effects. It would also be possible to extend these methods to traits which did not have a normal error structure, such as threshold traits, by using the same principles in a generalized linear model context. All steps of the regression analysis can be performed using one of a number of general computer statistical packages without resort to specialist software, putting the analyses within the grasp of any quantitative geneticist.

Note added in proof

We have recently become aware that Martinez & Curnow [*Theor. Appl. Genet.*, (in press)] have developed the use of regression for mapping QTL in a backcross experiment. Their results confirm our own (this paper and Knott & Haley, 1992) in demonstrating the biases that can be introduced when the possibility of two linked loci is ignored.

Acknowledgements

This work is jointly supported by the AFRC and MAFF in the U.K. and by the BRIDGE programme of the Commission of the European Communities. We thank Robin Thompson for helpful suggestions and a prompt.

References

AITKIN, M., ANDERSON, D., FRANCIS, B. AND HINDE, J. 1989. *Statistical Modelling in GLIM*. Oxford University Press, Oxford.

DRAPER, N. R. AND SMITH, H. 1966. *Applied Regression Analysis* John Wiley and Sons, New York.

GENSTAT 5 COMMITTEE, 1989. *Genstat 5 Reference Manual*, Clarendon Press, Oxford.

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

HALEY, C. S. 1991. Use of DNA fingerprints for the detection of major genes for quantitative traits in domestic species. *Anim. Genet.*, **22**, 259–277.

KNAPP, S. J. 1991. Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. *Theor. App. Genet.*, **81**, 333–338.

KNAPP, S. J., BRIDGES, W. C. Jr. AND BIRKES, D. 1990. Mapping quantitative trait loci using molecular marker linkage maps. *Theor. App. Genet.*, **79**, 583–592.

KNOTT, S. A. AND HALEY, C. S. 1992. Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genet. Res.* **60**, (in press).

LANDER, E. S. AND BOTSTEIN, D. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.

MATHER, K. M. AND JINKS, J. L. 1982. *Biometrical Genetics* 3rd edn. Chapman and Hall, London.

MCCULLAGH, P. AND NELDER, J. A. 1983. *Generalized Linear Models*. Chapman and Hall, London.

MORTON, N. E. 1955. Sequential tests for the detection of linkage. *Am. J. Hum. Genet.*, **7**, 277–318.

PATERSON, A. H., DAMON, S., HEWITT, J. D., ZAMIR, D., RABINOWITCH, H. D., LINCOLN, S. E., LANDER, E. S. AND TANKSLEY, S. D. 1991. Mendelian factors underlying quantitative triats in Tomato: Comparison across species, generations and environments. *Genetics*, **127**, 181–197.

PATTERSON, H. D. AND THOMPSON, R. 1971. The recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

WELLER, J. I. 1987. Mapping, and analysis of quantitative trait loci in *Lycopersicon* (tomato) with the aid of genetic markers using approximate maximum likelihood methods. *Heredity*, **59**, 413–421.

WILKS, S. S. 1938. The large sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.*, **9**, 60–62.