Heredity **79** (1997) 319–328 *Received 12 December 1996*

Model diagnostics for fitting QTL models to trait and marker data by interval mapping

CHRISTINE A. HACKETT*

Biomathematics and Statistics Scotland, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, U.K.

Diagnostic tools are presented which enable the geneticist to assess the agreement between linkage data and a fitted normal mixture model for interval mapping. The theoretical likelihood profile along a chromosome is derived for a single quantitative trait locus (QTL) segregating in a backcross population, along with upper and lower bounds. This is useful for detecting two QTLs on a chromosome. Residuals are used to indicate the need for transformation of the trait values to a different scale before analysis, and the use of an incorrect distribution is shown to reduce the maximum lod score. A strategy for the regular use of diagnostic tools for interval mapping is presented.

Keywords: backcross, diagnostics, genetic markers, interval mapping, likelihood profile, quantitative trait loci.

Introduction

Statistical methods for mapping quantitative trait loci (QTLs) have advanced rapidly over the last 20 years. The original approach was to compare trait values of the different genotypes at a marker locus (e.g. Soller *et al.*, 1976), equivalent to one-way analysis of variance with the different marker genotypes corresponding to levels of a factor. A more realistic model was proposed by Weller (1986), who modelled the trait distribution for each marker class as a mixture of distributions corresponding to the different QTL genotypes, with the mixing proportion a function of the recombination fraction between the marker and the QTL. This approach was extended by Weller (1987) and others to the case of two markers flanking a QTL.

As molecular marker technology has advanced, maps of molecular markers covering the whole genome have become available. Lander & Botstein (1989) proposed interval mapping to locate the positions of QTLs relative to all the markers on a chromosome. Again, this uses a mixture model. A lod score is calculated for each point on the chromosome as the $\log_{10}$ of the ratio of the likelihood of a QTL at that point to that of no QTL at that point. The maximum of this likelihood profile is taken to indicate the most likely position of a QTL, if this lod is above a specified threshold. Jansen (1993), Jansen & Stam (1994) and Zeng (1993, 1994) have shown how the precision of interval mapping may be increased by including additional markers as explanatory variables in the mixture model to remove genetic variation owing to QTLs in other parts of the genome.

Statistical models for quantitative traits have thus developed considerably in recent years. However, statistical modelling should be an iterative process: view data→fit model→look for departures from model→amend model→refit model, etc. Here, we look in detail at the normal mixture model fitted to map QTLs and describe some diagnostics, which might indicate that this model is inadequate. These diagnostics are then tested for their ability to detect two types of departure from the model, using simulated data. The methods will be developed for a backcross between two inbred parents, but may be modified for other types of cross.

Diagnostic tools

Diagnostics based on residuals

Many diagnostics for departures from linear models are based on the distribution of residuals, and here we can use the same approach. The interval mapping method (Lander & Botstein, 1989) postulates a single QTL at the position on the chromosome corresponding to the maximum of the ratio of the likelihood of a QTL to the likelihood of no QTL.

*Correspondence. E-mail: chacke@scri.sari.ac.uk

The QTL genotype for each individual is not observable, but the conditional probability of each QTL genotype, given the marker genotypes and the trait values, is calculated as part of the EM fitting algorithm. Let the QTL genotypes be Qq and qq, with trait means μ_Q and μ_q. Denote the full set of marker information by **M** and the trait values by Y. Then a fitted value, $\hat{Y}_i$, may be calculated for individual i as

$$\hat{Y}_i = \mu_Q P(Qq \mid \mathbf{M}, Y_i) + \mu_q P(qq \mid \mathbf{M}, Y_i)$$

and the residuals R_i as $Y_i - \hat{Y}_i$. Plots of residuals against fitted values will be investigated for their ability to detect nonconstant variance, as in a linear model. For many individuals, the conditional probabilities $P(Qq \mid \mathbf{M}, Y_i)$ and $P(qq \mid \mathbf{M}, Y_i)$ are close to 0 or 1, so the fitted values are close to μ_Q or μ_q and the problem is close to a nonmixture problem. For nonmixture data, the residuals should be approximately normally distributed, and this may be tested by the correlation, r, between the ordered residuals and the normal order statistics (Filliben, 1975). Simulation is needed to derive a lower threshold for r for the QTL mixture model.

Diagnostics based on the likelihood profile

Diagnostics based on the distribution of residuals use the parameters of the mixture model corresponding to a QTL at the peak of the likelihood profile. However, the interval mapping approach gives a complete profile for the likelihood of a QTL at every point on the chromosome. If the true situation is a single QTL on the chromosome, then the likelihood profile should peak at its position. If, however, there are two linked QTLs on the chromosome, the likelihood profile may actually peak between the two QTLs (Haley & Knott, 1992; Martinez & Curnow, 1992) and a single QTL may be erroneously deduced. The shape of the likelihood profile over the complete chromosome may enable these two situations to be distinguished.

Haley & Knott (1992) use a regression method for interval mapping of quantitative trait loci, where the likelihood at each point is derived by regressing the trait values on their expected values as functions of the distance from that point to the flanking markers. They found that the regression method gives a very similar profile to that obtained by fitting a normal mixture model using maximum likelihood. In the Appendix, there is shown to be a monotonic relationship between the likelihood profile and the profile of the regression sum of squares. A profile for the regression sum of squares is then derived for the situation of a single QTL on a chromosome, together with upper and lower confidence limits. Therefore, we can investigate whether the observed profile lies within the expected limits.

Diagnostics based on regression coefficients

The normal mixture model is fitted to every point on a chromosome: at marker locations this is equivalent to regression on that marker. Assuming the marker has a recombination fraction r_Q with a QTL with effect $\beta = \mu_Q - \mu_q$, the expected value of the regression coefficient can be shown to be $(1-2r_Q)\beta$. Departures of the regression coefficient from this value may indicate a misspecified model.

A single QTL for a trait whose variance increases with its mean

For some traits, the variance increases with the mean and a log-normal distribution may be more appropriate than a normal distribution. We seek diagnostics to indicate the need for a transformation of the trait data before modelling. A set of marker data was simulated for a backcross population with 200 individuals. Six markers were simulated on a single chromosome, with recombination fractions between adjacent markers having expected value 0.1. A QTL was simulated, lying halfway between markers 2 and 3. Four sets of trait values (A, B, C and D) were simulated using the log-normal distribution, with parameters chosen to give a low (A and B) or high (C and D) heritability, and a small (A and C) or large (B and D) ratio for the two standard deviations associated with the two QTL genotypes (see Table 1). The same set of marker data was used for each simulation. For each set, a single QTL was fitted by interval mapping, treating the trait as normally distributed, and the position of the QTL, the QTL means and the common standard deviation were estimated. One hundred simulations of a normally distributed trait were then generated, assuming a QTL at the estimated position and with the estimated means and standard deviation. This gives data sets AN, BN, CN and DN. Thus, diagnostics from the log-normal data may be compared with their corresponding distribution from 100 simulations based on normal data. A further 100 simulations were run using a log-normal distribution with the same parameters as A, B, C and D to see how frequently discrepancies between the normal model and the log-normal data were detected (data sets AO, BO, CO and DO). Finally, these 100 log-normal traits were analysed after a logarithmic transformation to see how the conclusions of the

Table 1 Power of the normal scores test to detect nonconstant variance among the residuals

Data set	Heritability (%)	Ratio of standard deviations σ_q/σ_Q^*	Original correlation (r)	Critical value (C_r)	Number of log-normal simulations with $r \geq C_r$	Number of transformed simulations with $r \geq C_r$
A	6	2	0.888	0.9935	0	96
B	6	1.2	0.964	0.9930	0	97
C	50	2	0.948	0.9935	0	94
D	50	1.2	0.992	0.9925	52	96

The critical value for the test is taken to be the 5 per cent point of the distribution of correlations from 100 normal simulations.
*Plants with QTL genotypes *Qq* and *qq* are assumed to have variances σ_Q^2 and σ_q^2, respectively.

analysis were affected (data sets AT, BT, CT and DT).

Distribution of residuals

Figure 1a shows a plot of the residuals, R, against the fitted values for the first simulation from set A, and Fig. 1b shows the corresponding normal scores plot. There were some outliers, and the normal scores plot showed marked curvature. The normal scores correlation, r, for this data set was 0.888, whereas 95/100 of the normal simulations (AN) had values of r above 0.9935. The largest value of r over the set of 100 log-normal simulations (AO) was

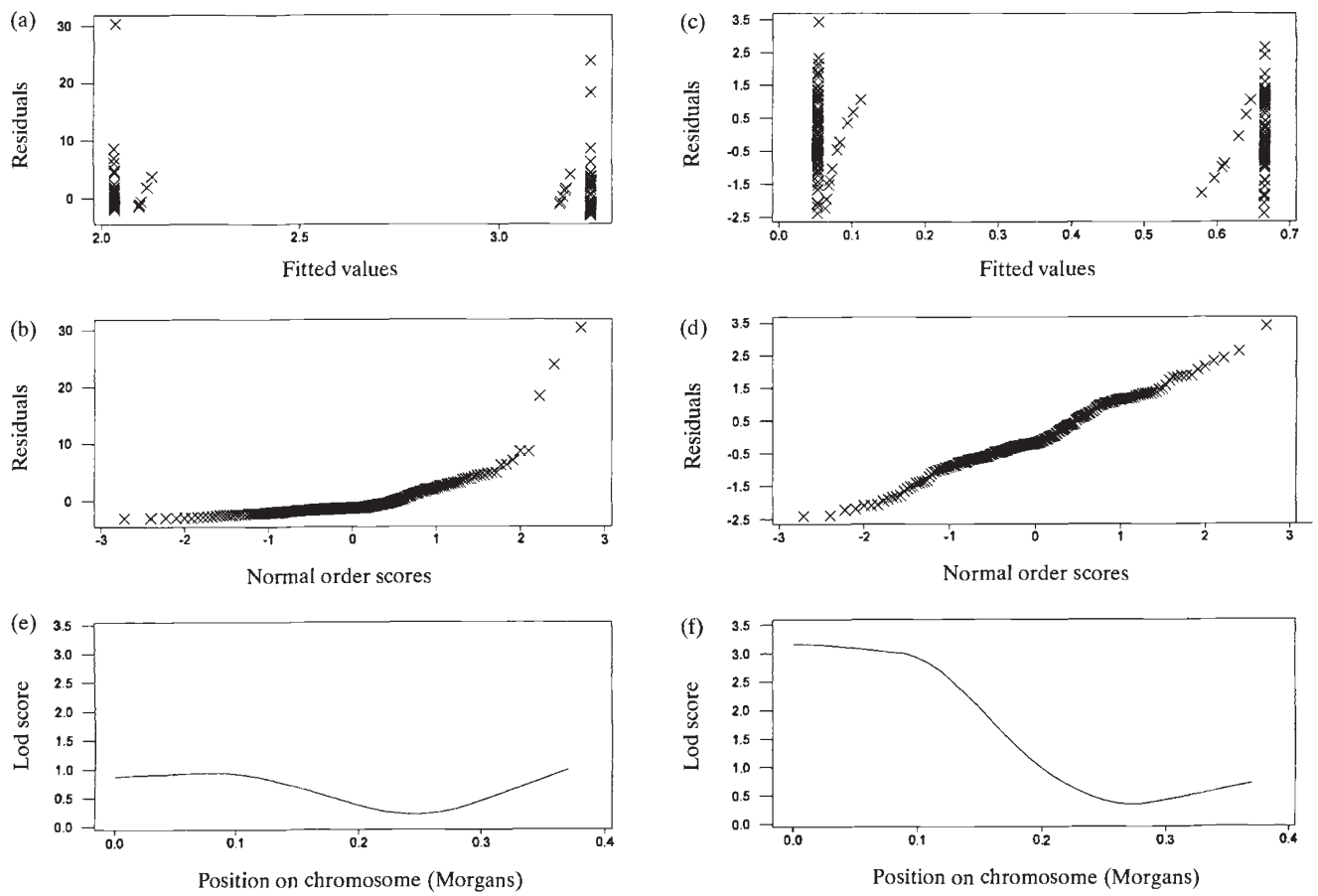

Fig. 1 Effect of a log transformation on simulated data set A. (a) Residuals against fitted values for untransformed data. (b) Normal scores plot for untransformed data. (c) Residuals against fitted values for log-transformed data. (d) Normal scores plot for log-transformed data. (e) Lod profile for untransformed data. (f) Lod profile for log-transformed data.

0.890, so a normal scores test based on a critical value of 0.9935 would reject normality for every log-normal simulation. When a logarithmic transformation was used before the QTL analysis, r was larger than 0.9935 for 96/100 simulations. Figure 1(c,d) shows the residuals against the fitted values, and the normal scores plot for the transformed data. Table 1 summarizes results on the power of the normal scores test to detect non-normality among the residuals for the four data sets. For data sets AO, BO and CO, the normal scores correlation is consistently lower than that expected in a normal population. Only for data set DO, where the heritability is high and the ratio of the true QTL standard deviations is only 1.2, would many of the log-normal simulations be accepted as normal by this criterion. The logarithmic transformation before the QTL analysis consistently improved the distribution of the residuals.

Likelihood profiles

The profiles of the regression sum of squares are calculated from the likelihood profiles using eqn A1 in the Appendix. The observed profiles generally lay close to the expected profiles and were within the 95 per cent confidence limits for at least 95 of the 100 log-normal and transformed log-normal simulations for each of data sets A, B, C and D. Therefore, the shape of the profile does not provide a useful diagnostic for detecting nonconstant variance. However, inspection of the profiles suggested that the profiles for the transformed simulations AT and CT had narrower peaks than the corresponding simulations AO and CO.

The widths of the peak may be compared by means of a support interval. For this, we return to the familiar scale of the lod profile. Lander & Botstein (1989) used a one-lod support interval (i.e. an interval bounded by the positions on the chromosome, at which the lod score has decreased to 1/10 of its maximum) to indicate the approximate position of a QTL. We will use support intervals of 1, 2 and 3 lods to compare the shape of different profiles.

Table 2 compares the support intervals for each data set. A comparison of the widths for the 100 simulated log-normal and transformed log-normal traits, using a Mann–Whitney U-test for a shift in the median, shows that the width of the likelihood peak was significantly reduced by the transformation for data sets A and C ($P<0.001$ for each support interval). For sets B and D, where the ratio of the standard deviations was 1.2 rather than 2, the widths of the support intervals were not significantly different. For all four data sets, there were significant increases in the maximum lod score for the transformed data (A: mean increase of 2.2, SE = 0.12; B: mean increase of 0.19, SE = 0.037; C: mean increase of 3.7, SE = 0.15; D: mean increase of 0.23, SE = 0.027). Figure (1e,f) shows the lod profile for the first simulation from data set A, before and after a logarithmic transformation.

Table 2 Widths (cM) of support intervals from the lod profiles for each data set

		Log-normal data			Transformed data		
Data set	Support interval	5%	50%	95%	5%	50%	95%
A	1 lod	15	23	38	10	17	31
	2 lods	23	38	38	16	27	38
	3 lods	34.5	38	38	21	38	38
B	1 lod	11.5	25	38	12	23	38
	2 lods	21	38	38	19.5	38	38
	3 lods	32	38	38	30.5	38	38
C	1 lod	6.5	8	12	5	7	9
	2 lods	9	12	18	7	10	14
	3 lods	11	15.5	21.5	9	13	18
D	1 lod	6	7	12	6	7	12
	2 lods	8	12	17	8	12	17
	3 lods	10	15	21.5	10.5	15	21.5

The table gives the 5, 50 and 95 percentiles of the width among 100 log-normal and transformed log-normal simulations.

Regression coefficients

No discrepancies were observed between the observed and expected trait coefficients for any of the simulations for data sets A, B, C and D. These coefficients are not a useful diagnostic for the need for a transformation.

Two QTLs on a chromosome

If there are two QTLs on a chromosome, the likelihood profile may have a maximum between them, and a single QTL (a 'ghost' QTL) may be deduced at the wrong location (Haley & Knott, 1992; Martinez & Curnow, 1992). Here, we seek diagnostics to identify when a linked QTL has been omitted from a model. Using the same simulated population and markers as before, seven sets of traits (E–K) were simulated to represent a range of QTL locations, sizes and signs (Table 3). In each case, a single QTL was fitted initially at the position corresponding to the maximum of the likelihood profile. One hundred traits were then generated, using the estimated QTL parameters (data sets EN–KN). A further 100 simulations were then generated, assuming two QTLs in the original positions, to see how frequently the diagnostic tests detected discrepancies (data sets EO–KO).

Distribution of residuals

The residuals from each two-QTL trait showed no curvature in the normal scores plot, and approximately 95 per cent were accepted as normally distributed using the normal scores test. The residuals were not useful for detecting that a linked QTL has been omitted from the model.

Likelihood profiles

The observed profile of regression sum of squares for the first trait simulated from set E is shown in Fig. 2a, along with the expected profile for a single QTL with the same location and parameter values, and the corresponding 95 per cent confidence interval. The observed and expected profiles are very close at the peak of the profile, but the observed curve decreases much more slowly than the expected curve and lies outside the 95 per cent confidence limits for the last 16 cM. Similar features were observed for the simulations from sets F, G and H. Even when the observed profile stayed within the 95 per cent confidence limits, it generally decreased more slowly than the expected profile. To compare the profiles, we can compute a sum of the proportional difference between the observed and expected profiles:

$$P = \Sigma \left| \frac{L_{\text{obs}} - L_{\text{exp}}}{L_{\text{upper}} - L_{\text{lower}}} \right|,$$

where L_{obs} is the observed regression sum of squares, L_{exp} is the expected sum of squares, L_{upper} and L_{lower} are the upper and lower bounds and the sum is taken over the points at which the likelihood profile is evaluated. For the original trait $P = 15.3$, whereas for the corresponding single QTL simula-

Table 3 Number of two QTL simulations, in which diagnostics indicate the single QTL model is inadequate, using expected likelihood profiles and expected regression coefficients

	First QTL		Second QTL		Likelihood profiles		Regression coefficients
Data set	Flanked by	Size	Flanked by	Size	No. outside 95% CI	No. with $P>95\%$ point	No. with t significant for 1+ marker
E	1,2	1	4,5	1	79	94	59
F	1,2	1	3,4	1	47	79	27
G	1,2	1.5	2,3	1.5	37	24	7
H	1,2	1	4,5	0.5	30	53	20
I	1,2	1	4,5	−1	90	32	99
J	1,2	1	3,4	−1	72	49	85
K	1,2	1.5	2,3	−1.5	46	10	55

The critical values for P are taken as the 95 per cent point of their distributions among 100 simulations based on a single QTL. The sizes of the QTLs are expressed relative to the environmental variance and their positions are midway between the two markers indicated.

tions (EN), the upper 95 per cent point for P was 6.2. The power of these tests to detect discrepancies from the expected profile decreases with the size and separation of the two QTLs (Table 3). For sets E, F and H, P detects more discrepancies than the profile bounds. For sets G and K, with QTLs in neighbouring intervals, neither method was particularly successful. For sets I, J and K, with QTLs of opposite signs, the profiles tended to dip below the lower bound between the QTL positions and P detected fewer discrepancies than the profile bounds. This is illustrated for the first trait from set I in Fig. 2b.

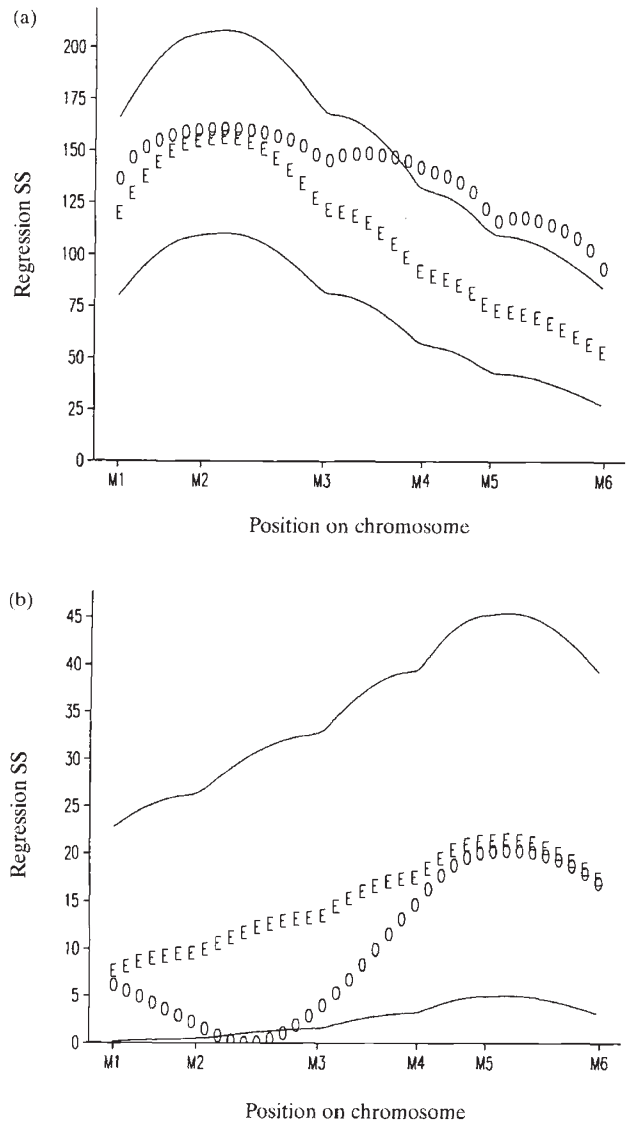

Fig. 2 Observed (O) and expected (E) profiles and 95 per cent confidence interval of the regression sum of squares. (a) Data set E — two QTLs with the same sign. (b) Data set I — two QTLs with opposite signs.

Table 4 Comparison of observed regression coefficients of trait I on the markers with those expected if a single QTL is present

Marker	Observed β	Standard error	Expected β	t
1	0.352	0.167	−0.358	4.25
2	0.202	0.168	−0.407	3.63
3	−0.271	0.167	−0.490	1.31
4	−0.549	0.164	−0.569	0.12
5	−0.638	0.163	−0.633	−0.03
6	−0.579	0.164	−0.565	−0.09

Regression coefficients

The regression coefficients of traits E, F, G and H on the markers were generally quite close to those expected for a single QTL on the chromosome. However, for sets I, J and K, differences between the observed and expected coefficients were more apparent. The likelihood profile for trait I suggested a single QTL between markers 5 and 6, with means $\mu_Q = 0.341$ and $\mu_q = -0.311$. The expected regression coefficient is negative for every marker, whereas the observed coefficients are positive for markers 1 and 2 (Table 4). Table 3 summarizes the number of simulations in which a t-test indicates a significant difference from the expected coefficient. The regression coefficients appear to be a useful diagnostic tool when the QTLs are of opposite signs. As before, the ease of detection increases with the QTL separation.

Conclusions

Studies of the inheritance of quantitative traits require substantial resources, and the statistical analysis should seek to explain variation in the data as fully as possible. This involves a careful exploration of a fitted QTL model to see whether it is consistent with the data or whether a more complicated genetic mechanism is necessary. Here, two common deviations from the usual mixture model assumptions have been investigated. There are, of course, other sources of discrepancies, such as an incorrect ordering of markers, which should be borne in mind.

It has been shown here that, if data from a log-normal distribution with variance increasing with the mean are analysed as though they were normal,

then the maximum lod score is reduced and the support intervals are wider than if a transformation is used before QTL analysis. For QTLs of small effect, this could make the difference between detection and omission. This ties in with the results of Jansen (1992), who investigated a trait with an exponential distribution and found that a QTL was detected only when the correct distribution was used. In some cases, in which examination of the residuals indicates non-normality, and especially if the trait takes discrete values, it may be preferable to use another distribution to model the traits. Jansen (1992) discusses the use of distributions other than the normal, Hackett & Weller (1995) investigate the use of an ordinal regression model for mapping ordered categorical traits and Visscher *et al.* (1996) and Xu & Atchley (1996) discuss the mapping of binary traits.

From the results of this study, the following diagnostic strategy is recommended.

1 Regression analysis of the trait on all the markers on the chromosome of interest. An examination of the residuals from this regression should indicate whether there is any need for a transformation of the trait to stabilize the variance. Most statistical software will also indicate observations with large residuals, and these should be investigated carefully, as they may influence the selection of markers linked to QTLs. The sizes of the regression coefficients should give a first indication of the likely QTL location. If the sign of the regression coefficient changes along the chromosome, there is a possibility of two linked QTLs of opposite signs.

2 Calculation of the likelihood profile along the length of the chromosome, using transformed trait values if suggested above. The residuals may be examined again, either graphically using a normal scores plot or by means of a test of the normal scores correlation coefficient.

3 Calculation of the expected likelihood profile, and upper and lower bounds. If the observed profile goes outside the bounds on the expected profile, there may be two QTLs on the chromosome. The observed profile can also stay within the bounds, but its shape may differ from that expected. If the profile decreases more slowly than expected, there may be two QTLs of the same sign, and comparison of the sum of proportional differences with that expected from simulated populations with one QTL is a more sensitive test than the profile bounds. If the profile shows an unexpected dip, there may be linked QTLs of opposite sign, and comparison of observed and expected regression coefficients is a useful test.

If linked QTLs are detected, then more sophisticated QTL mapping procedures are required, such as the 'MQM mapping' and 'composite interval mapping' methods of Jansen (1993), Jansen & Stam (1994) and Zeng (1993, 1994). These provide a powerful method for continuing the QTL analysis.

Acknowledgements

I thank Mr J. W. McNicol and Mr R. A. Kempton of BioSS for useful comments on this work, and two referees for their suggestions. The research was funded by the Scottish Office Agriculture, Environment and Fisheries Department.

References

FILLIBEN, J. J. 1975. The probability plot correlation test for normality. *Technometrics*, **17**, 111.

HACKETT, C. A. AND WELLER, J. I. 1995. Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics*, **51**, 1252–1263.

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. 1992. A general mixture model for mapping quantitative trait loci by using molecular markers. *Theor. Appl. Genet.*, **85**, 252–260.

JANSEN, R. C. 1993. Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.

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

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

MARTINEZ, 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.

PEARSON, E. S. AND HARTLEY, H. O. 1972. *Biometrika Tables for Statisticians*, Vol. II. Cambridge University Press, Cambridge.

SOLLER, M., BRODY, T. AND GENIZI, G. 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theor. Appl. Genet.*, **47**, 35–39.

VISSCHER, P. M., HALEY, C. S. AND KNOTT, S. A. 1996. Mapping QTLs for binary traits in backcross and F2 populations. *Genet. Res.*, **68**, 55–63.

WELLER, J. I. 1986. Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. *Biometrics*, **42**, 627–640.

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.

XU, S. Z. AND ATCHLEY, W. R. 1996. Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics*, **143**, 1417–1424.

ZENG, Z. B. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 10972–10976.

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

Appendix: derivation of theoretical lod profile

The traditional lod profile is calculated as $G_1(d) = \log_{10}(L_1/L_0)$, where L_1 is the likelihood of a model with a QTL at position d, and L_0 is the likelihood of a model with no QTL. This test is related to the traditional likelihood ratio test statistic $G_2(d) = 2\log_e(L_1/L_0)$, which has an asymptotic χ^2-distribution, by $G_2 = 2\log_e(10)G_1$.

Haley & Knott (1992) use a regression method for interval mapping of quantitative trait loci, in which the likelihood at each point is derived by regressing the trait values on their expected values as functions of the distance from that point to the flanking markers. They found the regression method and maximum likelihood methods gave very similar profiles, and the regression approach will therefore be used here to derive an equation for the lod profile. The likelihood ratio test may be written in terms of the regression, residual and total sum of squares (SSR, SSE and SST) as

$$G_2 = n \log_e(\mathrm{SST}/\mathrm{SSE}) = -n \log_e(1-\mathrm{SSR}/\mathrm{SST}). \quad \text{(A1)}$$

Thus, G_2 is a monotonic function of SSR and we will use the distribution of SSR to derive the expected profile and upper and lower bounds. In a conventional linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, the (mean-corrected) regression sum of squares is

$$\mathrm{SSR} = \hat{\beta}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2 = \mathbf{Y}'\mathbf{A}\mathbf{Y} \quad \text{where } \mathbf{A} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}\mathbf{1}'/n.$$

If the errors are independent and normally distributed $N(0,\sigma^2)$, and there is a single explanatory variable, SSR/σ^2 has a noncentral χ^2-distribution with expectation

$$E(\mathbf{Y}'\mathbf{A}\mathbf{Y}/\sigma^2) = 1 + E(\mathbf{Y}')\mathbf{A}E(\mathbf{Y})/\sigma^2 = 1 + \lambda.$$

We now derive $\mathbf{X}$ for a backcross situation in a similar manner to the derivation of the coefficients for an F_2 cross by Haley & Knott (1992). The cases of the QTL outside and within the marker interval of interest must be considered separately.

QTL outside the marker interval

The backcross is between genotypes *AAPPBBQQ* and *aappbbqq*, with the latter as the recurrent parent. A and B are markers, Q the QTL and P the point at which we wish to calculate the likelihood of a QTL. The recombination fractions are r_A between P and A, r_B (P and B), r (A and B), s_A (Q and A) and s_B (Q and B). The expected trait values of individuals with genotypes *Pp* and *pp* are $m-a$ and $m+a$, respectively. Table A1 gives the expected proportion of each genotype and the expected trait values. We estimate the trait parameters m and a by regression on the last column of Table A1. Assuming that the markers follow the expected segregation ratios, the $(n \times 2)$ matrix $\mathbf{X}$ of explanatory variables has its first column all ones, and its second

Table A1 Derivation of the explanatory variable for regression at locus P

Marker genotype	*Pp*	*pp*	Expected trait value	Explanatory variable
AaBb	$(1-r_A)(1-r_B)/2$	$r_A r_B/2$	$m-a(1-r_A-r_B)/(1-r)$	$-(1-r_A-r_B)/(1-r)$
Aabb	$(1-r_A)r_B/2$	$r_A(1-r_B)/2$	$m+a(r_A-r_B)/r$	$(r_A-r_B)/r$
aaBb	$r_A(1-r_B)/2$	$(1-r_A)r_B/2$	$m-a(r_A-r_B)/r$	$-(r_A-r_B)/r$
aabb	$r_A r_B/2$	$(1-r_A)(1-r_B)/2$	$m+a(1-r_A-r_B)/(1-r)$	$(1-r_A-r_B)/(1-r)$

column consists of the terms in the last column of Table A1, each repeated to correspond to the expected marker genotype frequency. Then

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{vmatrix} 1/n & 0 \\ 0 & 1/nS \end{vmatrix},$$

where
$S = (1-r_A-r_B)^2/(1-r) + (r_A-r_B)^2/r.$

Therefore, **A** is a block matrix, with blocks of size $n(1-r)/2$, $nr/2$, $nr/2$, $n(1-r)/2$:

$$\begin{vmatrix} \frac{(1-r_A-r_B)^2}{nS(1-r)^2} & \frac{-(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} & \frac{(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} & \frac{-(1-r_A-r_B)^2}{nS(1-r)^2} \\ \frac{-(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} & \frac{(r_A-r_B)^2}{nSr^2} & \frac{-(r_A-r_B)^2}{nSr^2} & \frac{(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} \\ \frac{(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} & \frac{-(r_A-r_B)^2}{nSr^2} & \frac{(r_A-r_B)^2}{nSr^2} & \frac{-(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} \\ \frac{-(1-r_A-r_B)^2}{nS(1-r)^2} & \frac{(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} & \frac{-(1-r_A-r_B)(r_A-r_B)}{nSr(1-r)} & \frac{(1-r_A-r_B)^2}{nS(1-r)^2} \end{vmatrix}$$

To calculate the expected SSR, we also need the expected trait values for each marker category, which depend on the distance from the QTL, Q. These expected values are given in the second column of Table A2. Using this:

$$\mathrm{E}(\mathbf{Y}')\mathbf{A}\mathrm{E}(\mathbf{Y}) = \frac{n(1-2r_B)^2\,(1-2s_B)^2\,(\mu_Q-\mu_q)^2}{4S}$$

and hence the SSR has a noncentral χ^2-distribution with parameter λ, where

$$\lambda = \frac{n(1-2r_B)^2\,(1-2s_B)^2\,(\mu_Q-\mu_q)^2}{4S\sigma^2}.$$

Table A2 Expected trait values for each marker category

Marker genotype	Order of loci	
	ABQ	AQB
AaBb	$(1-s_B)\mu_Q+s_B\mu_q$	$[(1-s_A)(1-s_B)\mu_Q+s_As_B\mu_q]/(1-r)$
Aabb	$s_B\mu_Q+(1-s_B)\mu_q$	$[(1-s_A)s_B\mu_Q+s_A(1-s_B)\mu_q]/r$
aaBb	$(1-s_B)\mu_Q+s_B\mu_q$	$[s_A(1-s_B)\mu_Q+(1-s_A)s_B\mu_q]/r$
aabb	$s_B\mu_Q+(1-s_B)\mu_q$	$[s_As_B\mu_Q+(1-s_A)(1-s_B)\mu_q]/(1-r)$

QTL within the marker interval

In this case, the derivation of **A** is unchanged, but the expected trait value in each category is now given by the third column of Table A2. In this case, we obtain

$$\lambda = \frac{n(\mu_Q - \mu_q)^2}{4S\sigma^2}\left[\frac{(1-r_A-r_B)\,(1-s_A-s_B)}{1-r} + \frac{(r_A-r_B)\,(s_A-s_B)}{r}\right]^2.$$

The noncentral χ_1^2-distribution has mean $1+\lambda$ and variance $2+4\lambda$. For $\lambda > 9$ a lower 2.5 per cent point may be calculated as $(\sqrt{\lambda}-1.96)^2$, and for $\lambda > 1$ an upper 2.5 per cent point is given by $(\sqrt{\lambda}+1.96)^2$ (Pearson & Hartley, 1972).