Genet. Res., Camb. (2002), **79**, *pp.* 97–106. *With* 3 *figures.* © 2002 Cambridge University Press
DOI: 10.1017/S0016672301005420 Printed in the United Kingdom

On the mapping of quantitative trait loci at marker and non-marker locations

GRANT A. WALLING[1]*, CHRIS S. HALEY[1], MIGUEL PEREZ-ENCISO[2]†, ROBIN THOMPSON[1,3] AND PETER M. VISSCHER[4]
[1] *Roslin Institute* (*Edinburgh*), *Roslin*, *Midlothian EH25 9PS*, *UK*
[2] *IRTA*, *Ctr UdL*, *Rovira Roure 177*, *Lleida 25198*, *Spain*
[3] *IACR-Rothamsted*, *Harpenden*, *Hertfordshire AL5 2JQ*, *UK*
[4] *Institute of Cell*, *Animal and Population Biology*, *University of Edinburgh West Mains Road*, *Edinburgh EH9 3JG*, *UK*

(*Received 4 January 2001 and in revised form 17 July and 13 September 2001*)

Summary

Previous studies have noted that the estimated positions of a large proportion of mapped quantitative trait loci (QTLs) coincide with marker locations and have suggested that this indicates a bias in the mapping methodology. In this study we predict the expected proportion of QTLs with positions estimated to be at the location of a marker and further examine the problem using simulated data. The results show that the higher proportion of putative QTLs estimated to be at marker positions compared with non-marker positions is an expected consequence of the estimation methods. The study initially focused on a single interval with no QTLs and was extended to include multiple intervals and QTLs of large effect. Further, the study demonstrated that the larger proportion of estimated QTL positions at the location of markers was not unique to linear regression mapping. Maximum likelihood produced similar results, although the accumulation of positional estimates at outermost markers was reduced when regions outside the linkage group were also considered. The bias towards marker positions is greatest under the null hypothesis of no QTLs or when QTL effects are small. This study discusses the impact the findings could have on the calculation of thresholds and confidence intervals produced by bootstrap methods.

1. Introduction

To understand the nature of quantitative genetic variation and to utilize this variation efficiently in artificial selection programmes in plant and livestock populations, methods have been developed to map quantitative trait loci (QTLs) which underlie this variation. These methods use information from multiple genetic markers to estimate the position of QTLs and their effects on the traits of interest. Lander & Botstein (1989) used maximum likelihood to map QTLs using sets of flanking markers and named their method 'interval mapping'. Haley & Knott (1992) showed that a simple linear regression method gives results that are very similar to the more complicated maximum likelihood methods. For simple population structures, results from the two methods are nearly identical with respect to the power and estimates of parameters.

Several authors have recently noted that a large proportion of mapped QTLs are estimated to be at the same position as markers, and have questioned whether there is a bias in the regression method (e.g. Spelman *et al.*, 1996; Walling *et al.*, 1998). When the null hypothesis of no QTLs is true, it would be expected that the estimated position of putative QTLs, determined by the position of the largest test statistic or Lod score, would be randomly distributed on the chromosome. In this case the proportion of estimated putative QTL positions at any one point on the

* Corresponding author. Tel. +44 (0)131 527 4325. Fax: +44 (0)131 440 0434. e-mail: Grant.Walling@bbsrc.ac.uk
† Current Address: Institut National de la Recherche Agronomique, Station d'amélioration Génétique des Animaux, BP 27, 31326, Castanet-Toloson Cedex, France.

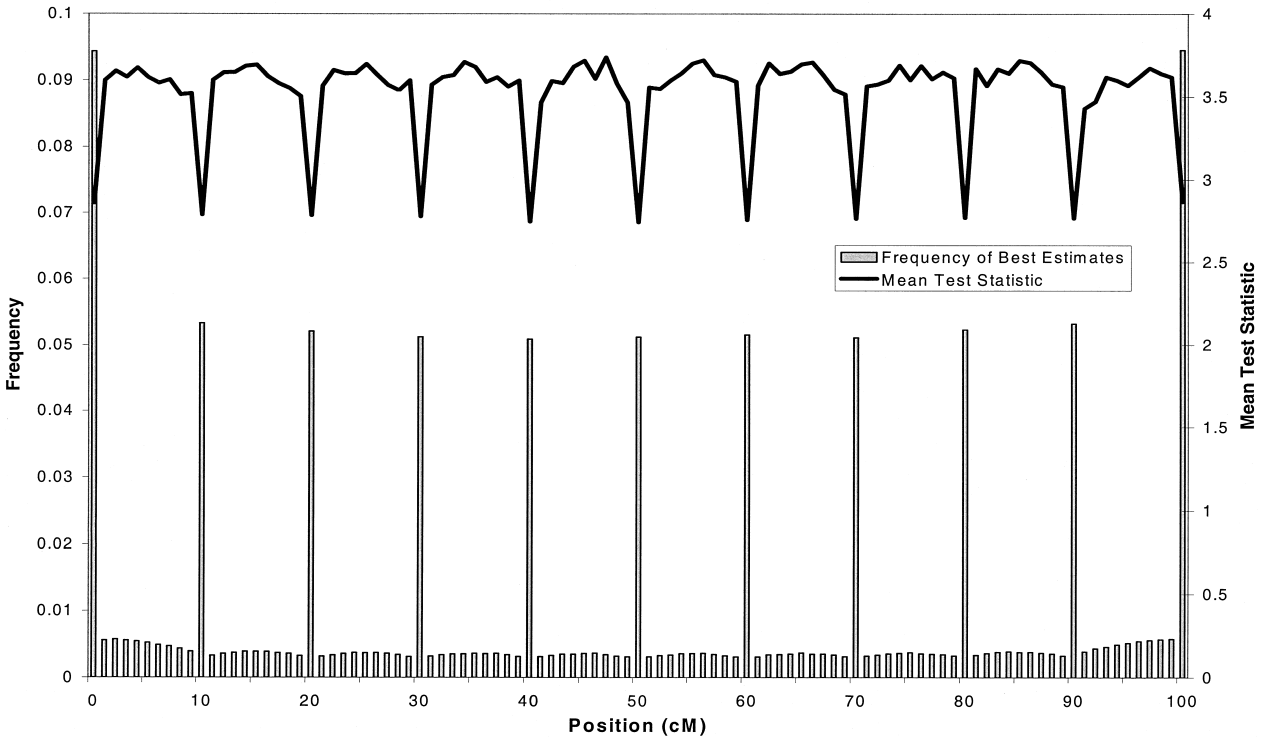

Fig. 1. Distribution of the position of the largest test statistic across a chromosome of 100 cM. Results are based on 1 million simulations of a backcross population of 200 individuals, using 11 evenly spaced fully informative markers. Also shown is the mean test statistic when the position is the estimated location of a putative QTL.

chromosome would be $1/n$, where n is the number of locations tested. Walling *et al.* (1998) showed by simulation that this was not observed, and that the QTL location was estimated more often at positions of flanking markers in comparison with non-marker locations. An example from Walling *et al.* (1998) is

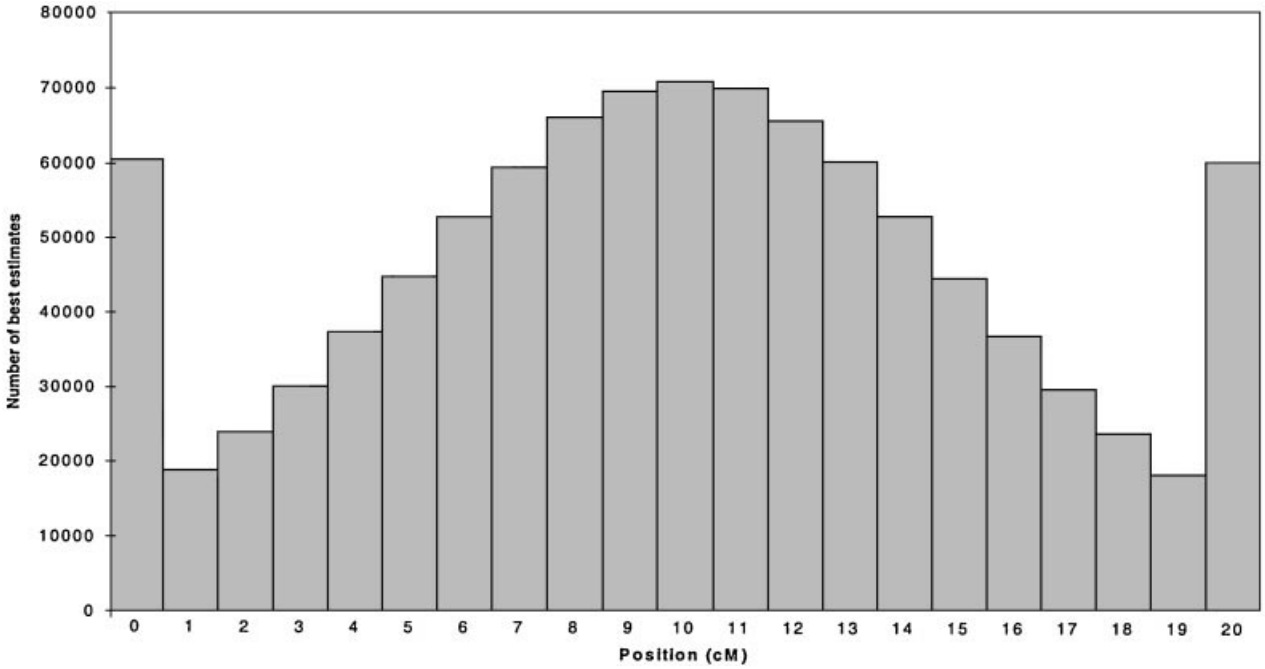

Fig. 2. Histogram showing the position of estimated QTL locations from 1 million replicates of a backcross population of 200 individuals, using a single 20 cM interval with a QTL ($h^2 = 10\,\%$) at 10 cM.

given in Fig. 1, which clearly shows that under the null hypothesis of no QTLs, a large proportion of QTLs are estimated to be at a marker position.

When a large QTL is located within an interval, the distribution of the estimated location might be expected to be approximately normal with the mean at the true location of the QTL. Fig. 2 demonstrates this is not observed in analyses of simulated data even when the power of detecting the QTL is high. A large proportion of estimates indicate that the position of the putative QTL is at the location of one of the flanking markers. Indeed only within 3 cM of the true QTL location does the frequency of estimates at an individual position outnumber the frequency at either flanking marker.

The apparent bias of the interval mapping method towards placing the estimated location of QTL at a marker position could affect the results of several types of studies. Bootstrapping is a relatively simple method of producing confidence intervals for a position of a QTL (Visscher *et al*., 1996) but, as demonstrated by Walling *et al*. (1998), the results of a bootstrap analysis could be misleading in the presence of inherent positional bias. In addition, Fig. 1 also highlights another trend. The mean test statistic is significantly lower when the estimated position for a putative QTL is located at a marker. Significance thresholds in QTL analysis are often calculated by studying the distribution of the test statistic when no QTL is present by simulation or permutation. The accumulation of estimated positions at marker locations and subsequently lower estimates of the test statistic may decrease thresholds below their correct value.

The aim of this study was to examine the accumulation of estimated QTL positions at the location of markers and to estimate through simulation and theory the proportion of estimated positions for putative QTLs at the site of a marker.

2. Linear regression

(i) *Single interval*

For a single interval, results from Whittaker *et al*. (1996) can be used to predict the probability that the highest test statistic will be at either of the flanking markers. In this study we consider only a backcross population derived from inbred lines, but the general results also apply to other population structures.

Model, assumptions and notation

$$y = \mu + \beta_L x_L + \beta_R x_R$$

with x_L and x_R the marker scores pertaining to the flanking markers. For fully informative markers in a backcross population, x_L and x_R can have two (arbitrary) values only, for example 0 and 1. Whittaker *et al*. (1996) showed that β_L and β_R are simple non-linear equations of the unknowns α and d, i.e. the QTL effect and location. For a backcross population, the QTL effect α is defined as the difference between the heterozygous and homozygous genotypes at the QTL. Hence, a simple transformation of the regression coefficients, which are estimated by a multiple regression of the phenotypes on the marker scores, gives the estimates of the QTL effect and location, thus making a grid search within the interval unnecessary. Whittaker *et al*. (1996) also showed that, if the signs of the two regression coefficients were not the same, the results were not consistent with a QTL in that interval. In terms of the equivalent grid search, this would result in the estimate of the QTL position being at the location of a flanking marker. The latter observation can be used to predict the probability of a QTL being placed at a marker because it corresponds to the probability that the sampled regression coefficients are of unequal sign. Hence, the means and (co)-variances of the two regression coefficients need to be calculated.

For a QTL at position d from the first marker, with recombination rate r_L and r_R from the first and second marker respectively, the expected values of the regression coefficients were shown by Whittaker *et al*. (1996) to be:

$$\beta_L = [r_R(1-r_R)(1-2r_L)/(r(1-r))]\alpha \quad (1)$$

$$\beta_R = [r_L(1-r_L)(1-2r_R)/(r(1-r))]\alpha \quad (2)$$

with r the recombination rate between the flanking markers.

The two estimated regression coefficients $\mathbf{b}' = [b_L\ b_R]$, conditional on a set of fixed $x_L + x_R$, follow a bivariate normal distribution, with mean $\boldsymbol{\beta}' = [\beta_L\ \beta_R]$, and covariance matrix,

$$\mathrm{var}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

with $\mathbf{X}$ being the design matrix for the regression model. We have assumed that the sample size is large so that the mean (μ) is estimated without error and hence the sampling covariance between the two regression coefficients due to the estimation of the mean can be ignored. If we now consider that the regressors are random variables, then elements of $\mathbf{X}'\mathbf{X}$ contain multiples ($\{N-3\}$, since 3 degrees of freedom are lost in the estimation procedure) of the variances and covariances of the regressors (x). This result follows from standard regression theory: when y and x are bivariately normally distributed, the variance of the regression coefficient for the model $y = \mu + bx$, is $\sigma^2/(n-3)$, with σ^2 the variance of $y|x$, and n the sample size (Kendall & Stuart, 1963). For a backcross

population, the (co)variances of x_L and x_R are easy to calculate:

$$\mathrm{var}(x_L) = \mathrm{var}(x_R) = \tfrac{1}{4},$$
$$\mathrm{cov}(x_L, x_R) = \tfrac{1}{4}(1-2r).$$

Assuming, without loss of generality, that $\sigma^2 = 1$, this gives an approximation of var(**b**) as,

$$\mathrm{var}(\mathbf{b}) \approx \frac{1}{(N-3)r(1-r)} \begin{bmatrix} 1 & -(1-2r) \\ -(1-2r) & 1 \end{bmatrix}. \quad (3)$$

Using equations (1) to (3), we have the mean and approximate (co)variances of the two regression coefficients. Hence, we can predict the probability that the estimates will be of opposite sign, using a standard bivariate normal distribution.

(ii) *No QTL model*

The no QTL model is the easiest to address, because the mean values of the regression coefficients are zero, so only the correlation between the regression coefficients, $-(1-2r)$, determines the area of the bivariate normal distribution for which the two variables are of opposite sign. The two variables exhibit symmetry around zero, hence the variances (and therefore population size) of these values are irrelevant. There is a single solution for this case that does not require numerical integration (Kendall & Stuart, 1963):

$$\begin{aligned}\text{Prob(QTL at either marker)} &= 0{\cdot}5 - \arcsin(-(1-2r))/\pi \\ &= 0{\cdot}5 + \arcsin(1-2r)/\pi. \end{aligned} \quad (4)$$

Equation [4] was checked by simulation of intervals of length 10, 20, 40 and 100 cM, and the predicted proportion of QTLs which were placed at the markers was very close to proportions observed through simulation (absolute differences $< 0{\cdot}001$). For marker spacings of 10, 20, 40 and 100 cM, the predictions from equation [4] are 0·805, 0·734, 0·648 and 0·543, respectively. Hence, for an interval of 20 cM, the probability that a QTL pertaining to the largest test statistic in this linkage group is placed at either marker is 0·734.

(iii) *Single QTL in interval*

The next level of complexity introduces a QTL into the single interval, in which case numerical integration, or some approximation, has to be used to predict the area under the bivariate normal surface for which the two variates are of opposite sign. Two approaches were used to estimate the relevant integrals: (i) A 7-order Taylor series expansion, and (ii) the algorithm of Mendell & Elston (1974). The two methods gave nearly identical results, and the results from the latter approach are shown in Table 1. In this table, the proportion of estimated QTL locations at either flanking marker was calculated using theoretical prediction and simulation, for population sizes of 50, 200 and 500. The QTL was placed in the middle of the interval, and explained 10% of the phenotypic variation in the backcross population. Although the difference between the prediction and simulation is larger than in the no QTL scenario, they remain in close agreement. There appears to be a small underestimation in the prediction.

Table 1. *Predicted and simulated proportions of estimated QTL locations that were placed at markers*

	Marker spacing (cM)			
n	10	20	40	100
50	0·540	0·424	0·347	0·394
	0·514	0·401	0·332	0·380
200	0·209	0·104	0·064	0·143
	0·182	0·089	0·053	0·128
500	0·045	0·010	0·003	0·021
	0·034	0·007	0·002	0·017

Upper values, observed values from 1 000 000 simulations of a backcross population; lower values, predictions from the algorithm of Mendell & Elston (1974). The simulated QTL was in the middle of the interval, with $h^2 = 0{\cdot}10$.

(iv) *Two intervals and no QTL*

For two intervals and three markers, there are four regression coefficients whose joint distribution determines where a QTL is placed, with the regression coefficient pertaining to the middle marker appearing twice: once in combination with the first marker, and once in combination with the third marker. Let the regression coefficients be β_1, β_{21}, β_{23} and β_3, where β_{21} is the regression coefficient of the second marker in combination with marker 1, and β_{23} is the regression coefficient of marker 2 in combination with marker 3. Again the signs of the four estimated regression coefficients (b_1, b_{21}, b_{23} and b_3) determine in which interval or at which marker the QTL is placed (Whittaker *et al.*, 1996).

The probability of obtaining a particular combination of signs can be predicted if we know the joint distribution of the four estimated regression coefficients. If $\mathbf{b}_1$ is a vector containing b_1 and b_{21}, with corresponding incidence matrix $\mathbf{X}_1$, and $\mathbf{b}_2$ is a vector with elements b_{23} and b_3, and incidence matrix $\mathbf{X}_2$,

then the covariances between the regression coefficients can be calculated as:

$$\mathrm{cov}(\mathbf{b}_i, \mathbf{b}_j) \propto (\mathbf{X}_i'\mathbf{X}_i)^{-1}(\mathbf{X}_i'\mathbf{X}_j)(\mathbf{X}_j'\mathbf{X}_j)^{-1}. \quad (5)$$

Analogous to the case of a single marker interval, the elements of the complete covariance matrix for the four regression coefficients were calculated using equation (5). For $i = j$, equation (5) reduces to (3). For $i \neq j$, i.e. for the sampling covariance between b_1, b_{21} and b_{23}, b_3, the only non-zero covariance is between b_{21} and b_{23}, and its value is, approximately, $4/(N-3)$. The result of zero sampling covariances for non-adjacent markers has been demonstrated in previous publications (e.g. Visscher, 1996).

For a large population size, it was assumed that the joint distribution of the regression coefficients is multivariate normal. This assumption was validated by a subsequent simulation study. Four variables were sampled from a multivariate normal distribution with a mean of zero and the covariance matrix as in equation (6), for values of the marker intervals, d_1 and d_2. Equation (6) was derived assuming Haldane's mapping function and by pre- and post-multiplying the 4×4 covariance matrix obtained from equation (5) by a diagonal matrix $\mathbf{D}$ with elements:

$$\mathbf{D} = diag\{[\tfrac{1}{4}(1-e^{-4d_1})]^{1/2}, [\tfrac{1}{4}(1-e^{-4d_1})]^{1/2}\}, \{[\tfrac{1}{4}(1-e^{-4d_2})]^{1/2}, [\tfrac{1}{4}(1-e^{-4d_2})]^{1/2}\}$$

For each sample of four values, the signs of the four coefficients were recorded, for a total of 1 million replicates. The observed average pattern of signs of the coefficients (e.g. $[+\,+\,+\,+]$ or $[-\,-\,+\,-]$) was compared with simulation results where markers in a backcross population were sampled. The results from simulating a backcross population (population size = 100, 200 and 500; heritability of QTL = 1%, 5% and 10%; marker spacing = 10, 20, 50 and 100 cM) were from 1000 replicates. The resulting patterns were not significantly different from each other, suggesting that equation (6), and assuming multivariate normality, is a good approximation for the distribution of the four regression coefficients. Hence, to determine the probability of obtaining any of the 16 sign configurations of the regression coefficients, for a particular distance between the markers, a simple and quick simulation was performed using a multivariate normal distribution.

The final part of the prediction is to translate the pattern of the signs of the regression coefficients into the probability that a QTL is placed at one of the three markers. This is relatively straightforward for

$$\mathrm{var}\begin{pmatrix} b_1 \\ b_{21} \\ b_{23} \\ b_3 \end{pmatrix} \propto \begin{bmatrix} 1 & -e^{-2d_1} & 0 & 0 \\ & 1 & [(1-e^{-4d_1})(1-e^{-4d_2})]^{1/2} & 0 \\ & & 1 & -e^{-2d_2} \\ symm & & & 1 \end{bmatrix}. \quad (6)$$

Table 2. *Patterns of the signs of regression coefficients for two marker intervals, and the conditional probabilities that a QTL is placed at one of the three markers*

Group	Pattern	Probabilities		
		$P(M_1)$	$P(M_2)$	$P(M_3)$
1	$-\,-\,-\,-$	0	0	0
2	$-\,-\,-\,+$	0	0	$\frac{1}{2}r_2$
3	$-\,-\,+\,-$	0	0	1
4	$-\,-\,+\,+$	0	0	0
5	$-\,+\,-\,-$	1	0	0
6	$-\,+\,-\,+$	$\frac{1}{2}$	0	$\frac{1}{2}$
7	$-\,+\,+\,-$	$r_1/(2r_1+r_{12})$	$r_{12}/(r_1+r_2+r_{12})$	$r_2/(2r_2+r_{12})$
8	$-\,+\,+\,+$	$\frac{1}{2}r_1$	0	0
9	$+\,-\,-\,-$	$\frac{1}{2}r_1$	0	0
10	$+\,-\,-\,+$	$r_1/(2r_1+r_{12})$	$r_{12}/(r_1+r_2+r_{12})$	$r_2/(2r_2+r_{12})$
11	$+\,-\,+\,-$	$\frac{1}{2}$	0	$\frac{1}{2}$
12	$+\,-\,+\,+$	1	0	0
13	$+\,+\,-\,-$	0	0	0
14	$+\,+\,-\,+$	0	0	1
15	$+\,+\,+\,-$	0	0	$\frac{1}{2}r_2$
16	$+\,+\,+\,+$	0	0	0

r_1 is the recombination fraction between the markers 1 and 2, r_2 between markers 2 and 3, and r_{12} is the recombination fraction between markers 1 and 3.
$P(M_i)$ is the probability that a QTL is placed at marker M_i.

Table 3. *Probability of QTL being placed at markers M_1, M_2 and M_3, for two adjacent intervals under the null hypothesis of no QTL effect: observed results from simulation (linear regression) and predicted results from sampling from a multivariate normal distribution*

	Observed				Predicted			
Δ	M_1	M_2	M_3	Sum	M_1	M_2	M_3	Sum
5	0·317	0·200	0·312	0·829	0·321	0·191	0·324	0·836
10	0·290	0·179	0·290	0·759	0·294	0·172	0·296	0·761
20	0·248	0·156	0·256	0·660	0·255	0·149	0·259	0·663
40	0·212	0·129	0·219	0·560	0·215	0·125	0·213	0·553
100	0·174	0·098	0·171	0·443	0·164	0·095	0·164	0·424

Δ, marker spacing (cM).

some combinations, but not for others. For example, given the pattern [+ + + +], the first pair of regression coefficients have the same sign, indicating the QTL could be within the first interval. The second pair of regression coefficients also have the same sign, indicating the QTL could be within the second interval. Under either scenario the estimated location of the QTL would not be located at a marker. Unfortunately not all combinations are conclusive: for a pattern such as [+ + + −], sometimes the QTL will be inside the first interval and sometimes it will be placed at the third marker. The other less obvious pattern is [+ − − +]. In this case the QTL can be placed at either the first, second or third marker.

The proportions were calculated theoretically, using the recombination fractions between the markers as parameters, and results are shown in Table 2. These were checked extensively using simulations, and found to be reasonable approximations (Table 3). For example, for a marker spacing of 20 cM, the probability that the largest test statistic is at one of the three markers was 0·660 and its prediction 0·663. This prediction includes simulation results from the multivariate normal distribution.

3. Maximum likelihood

It could be argued that the results obtained so far are an artefact of the method of analysis used, i.e. linear regression. The regression method does not take into account that genotypes at locations in between markers are not known precisely and variances may be heterogeneous within marker classes. In addition, there is no information used in the regression method that can distinguish between a QTL at the end marker of a linkage group, and a linked QTL outside the end marker. Using regression, the test statistic beyond the end marker is the same as the test statistic at that marker. Hence, these 'end-effects' may cause part of the accumulation of positional estimates observed at markers, i.e. that a large proportion of QTLs are placed at the extreme marker positions.

We simulated a backcross population with no QTLs and analysed the data using maximum likelihood (ML). For a particular marker interval, the test statistic was calculated at 1 cM positions in the linkage group and compared with the results from the regression method (REG). To allow comparisons between the results both methods (ML and REG)

Table 4. *Probability of QTL being placed at markers using maximum likelihood (ML) or linear regression (REG) on the same data for a single marker interval, using simulation results from 1000 replicates of a backcross population of 200 individuals with no QTL. A separate set of 1000 replicates was evaluated for extending the chromosome by 100 cM from both markers (MLEX). Prediction results presented from sampling from a multivariate normal distribution*

	Probability of the QTL being located at either marker			
Δ	ML	REG	MLEX	Prediction
5	0·903	0·877	0·320	0·860
10	0·828	0·827	0·284	0·805
20	0·750	0·754	0·238	0·734
40	0·669	0·662	0·233	0·648
100	0·540	0·538	0·175	0·543

Δ, marker spacing (cM).

Table 5. *Probability of a QTL being placed at markers M_1, M_2 or M_3, and the sum of these probabilities (Sum), using maximum likelihood (ML) or linear regression (REG) on the same data for two marker intervals, using simulation results from 1000 replicates of a backcross population of size 200 with no QTL. A separate set of 1000 replicates was evaluated for extending the chromosome by 100 cM from both extreme markers (MLEX)*

	ML				MLEX				REG			
Δ	M_1	M_2	M_3	Sum	M_1	M_2	M_3	Sum	M_1	M_2	M_3	Sum
5	0·319	0·236	0·322	0·877	0·113	0·169	0·108	0·390	0·314	0·232	0·317	0·863
10	0·311	0·191	0·299	0·801	0·099	0·154	0·110	0·363	0·310	0·190	0·302	0·802
20	0·250	0·176	0·266	0·692	0·084	0·144	0·088	0·316	0·247	0·173	0·267	0·687
40	0·224	0·142	0·209	0·575	0·085	0·121	0·068	0·274	0·230	0·136	0·219	0·585
100	0·192	0·087	0·160	0·439	0·073	0·075	0·068	0·216	0·175	0·093	0·191	0·459

Δ, marker spacing (cM).

used a grid search approach at 1 cM intervals. In a separate set of simulations, the test statistic was also calculated by extending the chromosome to 100 cM to the left of the leftmost marker, and to the right of the rightmost marker (MLEX). This was done because maximum likelihood can, in principle, detect estimated locations for a putative QTL that are outside the linkage group.

Results are shown in Tables 4 and 5. In Table 4, simulation results are shown from a single marker interval for ML, MLEX and REG. Results for the proportion of QTLs placed at markers are very similar for ML and REG when the estimated position was chosen from the given linkage group. This was expected, given previous results from the many studies demonstrating the close similarity between maximum likelihood and regression methods (e.g. Haley & Knott, 1992). When the chromosome was extended, the proportion of QTLs placed at either marker decreased significantly. For example, for a marker interval of 20 cM, the proportion of QTLs placed at either marker was 0·750 for a search over the 20 cM, and 0·238 when the search was extended to 220 cM. Extending the search even wider did not significantly reduce the proportion of highest test statistics at markers (results not shown). Hence, although the 'end-effect' gives an inflated proportion of QTLs that are placed at markers, the proportion of estimates at marker locations is still significantly higher than at non-marker locations even when the search is extended.

In Table 5, results are shown for two marker intervals. Results for REG in Table 5 are slightly larger than the equivalent points in Table 3. This is because the grid search method at 1 cM intervals rounds the results to the nearest centimorgan. Hence, with a marker at 10 cM, any result between 9·5 cM and 10·5 cM would give an estimate at the location of the marker. The equations of Whittaker *et al.* (1996) give a precise answer and differentiate between an estimate at the location of a marker and an estimate $< 0{\cdot}5$ cM away from a marker. The results for ML and REG are very similar when the search is confined to the linkage group. The proportion of estimates that are placed at the middle marker are also very similar for both methods. When the chromosome search is extended by 200 cM, the proportion of QTLs that are placed at outermost markers decreases. However, a substantial proportion of the replicates still have the estimated location of a QTL at the position of a marker. In addition, the largest proportion is now found at the central marker. For example, a total of 31·6% of positional estimates were placed at the three markers, with 14·4% at the location of the middle marker.

4. Discussion

We have shown that under the null hypothesis of no QTL the probability that the largest test statistic is found at a marker locus is high. When using linear regression to analyse QTL data from a backcross population, the proportions were predicted analytically for single and double intervals. The predictions were extended to a single interval containing a QTL located midway between two flanking markers. The extension to multiple intervals is feasible, although it would be tedious and is unlikely to increase our understanding of interval mapping. The simulations and predictions were performed for a backcross population derived from inbred lines, but the general conclusions are also valid for other designs, such as F2 and half-sib populations. The accumulation of estimated locations at the position of markers occurs when using either maximum likelihood or linear regression. In practice, more sophisticated methods to map multiple QTL are used (Jansen, 1993, 1994;

Zeng, 1993, 1994). However, for a particular marker interval, these methods essentially use interval mapping and a bias towards locating estimated positions of a QTL at a marker are also to be expected.

This work demonstrates a bias in estimated QTL position, but we should be careful how the estimated QTL position is defined. There are at least two definitions that have been discussed in the literature:

(i) The average position of a QTL from repeated sampling. If we consider the average estimated location of a QTL in a replicated experiment, it will be biased towards the middle of the chromosome, because of the impact of false positives (which are distributed symmetrically around the mid-point on any one chromosome). However, the average test statistic across experiments is largest at the true QTL position (Haley & Knott, 1992). This implies that there is a relationship between the size of the largest test statistic and its estimated position. We also found this from our simulations under the null hypothesis (Fig. 1), in that the highest test statistic had, on average, a lower value when it was placed at a marker compared with when it was in between markers.

(ii) The distribution (rather than just the point estimate) of the estimated position of the QTL, which is equivalent to investigating the distribution of the maximum test statistics along a chromosome region. The distribution of the highest test statistic, i.e. with variation in the position of a QTL across repeated sampling, does not follow a χ^2 distribution (Mangin *et al.*, 1994). In the present study we have focused on the distribution of the estimated QTL position. In particular, we have investigated the behaviour of the estimated QTL position expressed as the proportion of QTLs that are found at marker loci. For example, from our results in Table 1 we can conclude that 'if we repeatedly map a QTL with $h^2 = 0{\cdot}10$ in a single interval of 20 cM, using a very low significance threshold and a population size of $N = 200$, the probability that the QTL location is estimated at one of the flanking markers is approximately 0·10 when the actual QTL is in the middle of the chromosome'.

With reference to the above point, it is worth noting that the behaviour of the test statistic at a particular location has been studied by several authors (e.g. Haley & Knott, 1992; Mangin *et al.*, 1994). These studies demonstrated that for a particular location in the genome (i.e. both marker and non-marker locations), the test statistic asymptotically follows a central or non-central χ^2 distribution in the absence and presence of a QTL, respectively. However, these findings are useful only if we wish to focus on a particular location, e.g. a candidate gene; they do not describe the distribution of the maximum test statistic when a search along a chromosome region is performed.

Most of our results were derived for the null hypothesis of no QTL in any marker interval. It could be argued that in practice we should set a significant threshold and the bias may disappear. However, Walling *et al.* (1998) showed that even when setting an arbitrary threshold, the accumulation of positional estimates at marker locations remains. This is not surprising, because our simulations were performed with an arbitrary low threshold of zero. We illustrate this in Fig. 3, which shows the probability of a QTL being placed at a flanking marker, as a function of the

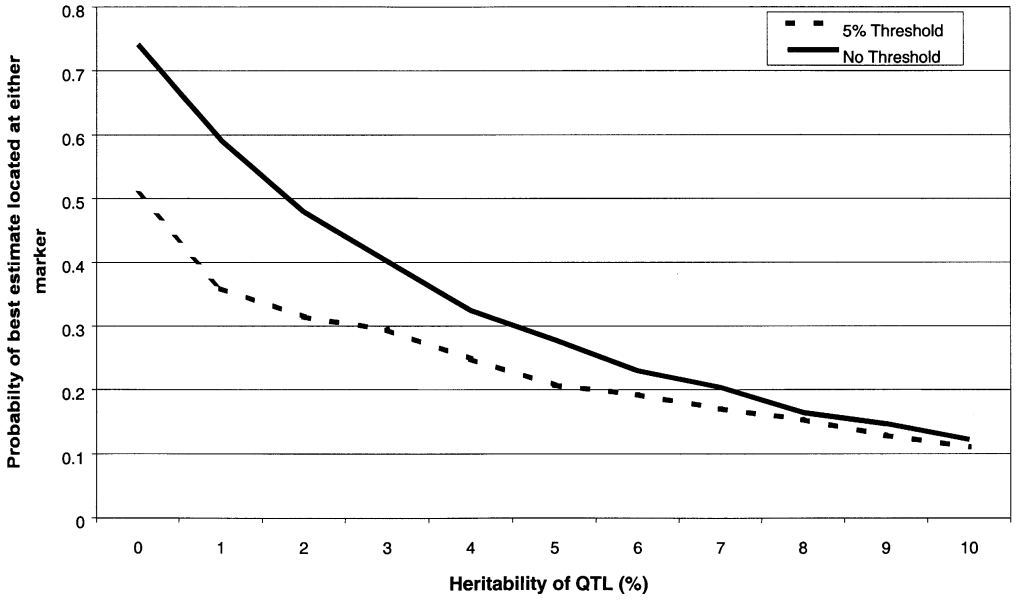

Fig. 3. Probability of QTL placed at either flanking marker of a 20 cM single interval as a function of QTL effect for the case of no threshold or a threshold corresponding to a 5% type I error rate.

true QTL heritability and the threshold for the test statistic. Fig. 3 shows that whether or not a threshold is imposed, the probability of the QTL being mapped at a marker is relatively large for QTLs of small effects.

In practice, the consequence of the bias is likely to be small. In many mapping experiments a single population is used with a stringent threshold. Under these conditions the accumulation of estimated QTL locations at the positions of markers is unlikely to occur. Areas that are affected are those using replication, especially in populations with no QTL or a QTL with small effect. The application of resampling techniques such as the bootstrap (Visscher *et al*., 1996) has consistently shown large peaks at positions of markers (e.g. Walling *et al*., 1998; Knott *et al*., 1998). These observed results are a consequence of the bias that was shown in this paper and could elongate or truncate confidence intervals that would otherwise have ended before or beyond the marker location respectively. However, the non-parametric bootstrap has been shown to perform to expectation, i.e. a 90% confidence interval contains the QTL in 90% of all cases, or be slightly conservative under certain parameter combinations (Walling *et al*., 1998). In comparison the parametric bootstrap is inaccurate and clearly influenced by the effects demonstrated in this study.

Two methods – a 'difference' method and a 'weighted' method – that directly correct for marker bias have been suggested for calculating confidence intervals for QTL locations from the non-parametric bootstrap (Bennewitz *et al*., 2000). Both methods base the correction on the distribution of the bootstrap estimates along the chromosome when a QTL may be present but not in a linkage phase with any marker. This distribution is obtained through simulation using the permutation approach. The theory presented in this paper would allow the correction to be based on the true theoretical distribution and would remove the sampling error present in the simulation approach. Bootstrapped confidence intervals using the corrected approaches are shorter without a loss in accuracy, i.e. the proportion of 90% confidence intervals that contains a QTL is still 0·90.

The results from this study may affect the calculation of suitable thresholds using the permutation test (Churchill & Doerge, 1994) or simulation. Permutation methods collate the maximum test statistic from random permutations of the phenotypic records relative to the genotypes, breaking any marker–QTL association, and hence are assumed to be sampling from the distribution of the maximum test statistic under the null hypothesis. This study demonstrates that the distribution of the maximum test statistic differs between marker and non-marker locations. Currently the application of these methods means that many of the maximum test statistics are located at the position of a marker; but as shown in Fig. 1 these are, on average, lower than the test statistics from maxima at non-marker locations. Thresholds are then calculated from these test statistics which, because of the unequal sampling from marker and non-marker locations, produce lower thresholds than would have been produced from more proportionate sampling. Simulation studies demonstrate that thresholds derived by the permutation approach give the anticipated type I thresholds at the level of the whole linkage group or genome. Nonetheless, future studies should investigate whether thresholds can be adjusted on a within-chromosome positional basis to account for the accumulation of estimates of QTL position at markers.

The authors thank Sara Knott for her helpful comments on the manuscript. This work was partly supported by an Acciones Integradas Grant, the Biotechnology and Biological Sciences Research Council (BBSRC), the Department for Environment, Food and Rural Affairs (DEFRA) and the Meat and Livestock Commission (MLC).

References

Bennewitz, J., Reinsch, N. & Kalm, E. (2000). Proposals for improved bootstrap confidence intervals in QTL mapping. *Book of Abstracts of the 51st Annual Meeting of the European Association for Animal Production* **6**, 6.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

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

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

Jansen, R. C. (1994). Controlling the type-I and type-II errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

Kendall, M. G. & Stuart, A. (1963). *The Advanced Theory of Statistics*, Vol. 1, *Distribution Theory*, 2nd edn. London: Charles Griffin.

Knott, S. A., Marklund, L., Haley, C. S., Andersson, K., Davies, W., Ellegren, H., Fredholm, M., Hansson, I., Hoyheim, B., Lundstrom, K., Moller, M. & Anderson, L. (1998). Multiple marker mapping of quantitative trait loci in a cross between outbred Wild Boar and Large White pigs. *Genetics* **149**, 1069–1080.

Lander, E. S. & Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Mangin, B., Goffinet, B. & Rebai, A. (1994). Constructing confidence intervals for QTL location. *Genetics* **138**, 1301–1308.

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

Mendell, N. R. & Elston, R. C. (1974). Multifactorial qualitative traits: genetic analysis and prediction of recurrence risks. *Biometrics* **30**, 41–57.

Spelman, R. J., Coppieters, W., Karim, L., van Arendonk, J. A. M. & Bovenhuis, H. (1996). Quantitative trait loci for five milk production traits on chromosome six in the Dutch Holstein–Friesian population. *Genetics* **144**, 1799–1808.

Visscher, P. M. (1996). Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. *Journal of Heredity* **87**, 136–138.

Visscher, P. M., Thompson, R. & Haley, C. S. (1996). Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**, 1013–1020.

Walling, G. A., Visscher, P. M. & Haley, C. S. (1998). A comparison of bootstrap methods to construct confidence intervals in QTL mapping. *Genetical Research* **71**, 171–180.

Whittaker, J. C., Thompson, R. & Visscher, P. M. (1996). On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77**, 23–32.

Zeng, Z.-B. (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the USA* **90**, 10972–10976.

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