Theor Appl Genet (1992) 85:480–488

THEORETICAL AND APPLIED GENETICS

Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers

O. Martínez and R. N. Curnow

Department of Applied Statistics, University of Reading, United Kingdom

Received February, 1992; Accepted May 20, 1992
Communicated by J. W. Snape

Summary. The use of information from flanking markers to estimate the position and size of the effect of a quantitative trait locus (QTL) lying between two markers is shown to be affected by QTLs lying in neighbouring regions of the chromosome. In some situations the effects of two QTLs lying outside the flanked region are reinforced in such a way that a 'ghost' QTL may be mistakenly identified as a real QTL. These problems are discussed in the framework of a backcross using a regression model as the analytical tool to present the theoretical results. Regression models that use information obtained from three or more nearby markers are shown to be useful in separating the effects of QTLs in neighbouring regions. A simulated data set exemplifies the problem and is analysed by the interval mapping method as well as by the regression model.

Key words: Quantitative trait loci – Interval mapping – RFLPS mapping

Introduction

The concept that the inheritance of many quantitative characteristics results from the segregation of genes, each of small effect, at many loci with expression modified by the environment is now well accepted. Until recently the accurate estimation of the number, location and effect sizes of these genes was practically impossible, because the effects of the individual quantitative trait loci (QTLs) could not be identified. However, the situation has changed with the advent of molecular polymorphisms that can be used as codominant markers. When there is both such a marker map and a segregating population for a characteristic of interest, it is often possible to obtain information about the number, effects and positions of the QTLs affecting the trait (Paterson et al. 1988).

The simplest, but obviously incorrect, approach for the estimation process is to assume that the QTLs can only occur exactly at the marker loci and thus use the differences between the phenotypic means of each marker class in a given cross to infer the existence and size of the effect of the QTLs. Because linkage with the marker locus is never complete, the QTL will be incorrectly located and the size of its effect will be underestimated. These difficulties in the methodology have been discussed by Lander and Botstein (1989). To remedy them they proposed the 'interval mapping' approach in which they calculated, for each chromosome interval flanked by markers, the LOD score for the hypothesis that there is a QTL in the interval compared with the hypothesis that there is not. This score is a function of the position within the interval and the size of the effect of the hypothesized QTL. The size of the effect is estimated for all possible positions of the QTL, and the resulting maximized LOD score is plotted against position. A QTL is said to be present when this maximized LOD score exceeds some threshold value.

As Lander and Botstein (1989) mention, when the model specified for this method is correct, i.e. when there is only one QTL segregating and it lies between the markers being studied, this method has all the desired properties of maximum likelihood estimators. In contrast, if there is a QTL in a neighbouring region of the chromosome, the procedure is no longer a maximum likelihood approach and, as we shall see, can

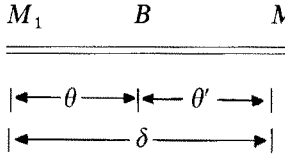
Correspondence to: O. Martínez

lead to erroneous results. An analytical study of the likelihood function of the data is difficult. We present here a regression model closely related to the LOD score approach whose performance is much easier to study.

Regression mapping

To explain regression mapping, we consider a backcross where the usual genetical and statistical assumptions are made. Thus, all of the individuals in the parental lines are homozygous for different alleles of the genes of interest and for the markers; the effects of the QTLs are additive between and within loci, and the environmental error has a normal distribution with the same variance for all the genotypes. The assumption of only additive effects within loci for the QTLs is not necessary if we consider doubled haploid lines instead of a backcross, and all the results of the model below are directly applicable to doubled haploid lines.

Let M_1, m_1 and M_2, m_2 be the alleles at the two marker loci M_1 and M_2, and B, b be the alleles of a QTL B located between these markers. This can be represented diagrammatically by



where the distance between the markers (δ) is known. We wish to estimate the distance between one of the markers and the QTL, say θ or equivalently θ', as well as the size of the effect of the B locus.

We need to consider which mapping function will be used to measure the distance between loci; the decision depends on biological considerations that take into account the degree of interference exhibited in the particular species. Two extreme situations are possible: (a) complete interference, in which no double recombination is allowed, when $\delta = \theta + \theta'$, or (b) no interference, where recombinations in neighbouring regions are independent events, when $\delta = \theta + \theta' - 2\theta\theta'$. Both situations were studied, but only the results for the more complex case of no interference (b) are presented, with some comments about the differences that arise when complete interference is assumed.

The genotypes of the parental lines will be $P_1 - m_1bm_2/m_1bm_2$ and $P_2 - M_1BM_2/M_1BM_2$, and the $F_1 - M_1BM_2/m_1bm_2$. The backcross ($B_1 = F_1 \times P_1$) will have four distinguishable marker groups, say $1 - M_1m_1M_2m_2$, $2 - M_1m_1m_2m_2$, $3 - m_1m_1M_2m_2$ and $4 - m_1m_1m_2m_2$. For simplicity of notation we will identify these marker groups by the marker alleles inherited from the F_1 parent, say $1 - M_1M_2$, $2 - M_1m_2$, $3 - m_1M_2$ and $4 - m_1m_2$.

Let Y_{ij} be the observation on the i-th plant in the j-th marker group, for $i = 1, 2, \ldots n_j$ (the number of individuals in the j-th group; $j = 1, 2, 3, 4$) and Γ_j be the probability that an allele B is present in an individual from the particular marker group j. Then, under the hypothesis of only additive effects at the B locus, say B and b, the expectation of Y_{ij} is given by

$$E[Y_{ij}] = 2b + (B - b)\Gamma_j$$
$$= \beta_0 + \beta_1\Gamma_j$$

where $\beta_0 = 2b$ and $\beta_1 = B - b$.

We consider the model

$$Y_{ij} = E[Y_{ij}] + \varepsilon_{ij}$$
$$= \beta_0 + \beta_1\Gamma_j + \varepsilon_{ij} \quad (1)$$

where ε_{ij} is normally distributed with mean zero and variance σ_e^2. Thus, we can regress the phenotypic values of the individuals on the probabilities that an allele B is inherited from the F_1 parent. In the case of no interference (b) we find

$$\Gamma_1 = P[M_1BM_2|M_1M_2] = 1 - \frac{\theta(\delta-\theta)}{(1-\delta)(1-2\theta)} = 1 - \Gamma_4$$

$$\Gamma_2 = P[M_1Bm_2|M_1m_2] = \frac{(1-\theta)(\delta-\theta)}{\delta(1-2\theta)} = 1 - \Gamma_3. \quad (2)$$

In the case of complete interference (a), the equations for the Γ's in (2) are simplified to $\Gamma_1 = 1 - \Gamma_4 = 1$ and $\Gamma_2 = 1 - \Gamma_3 = (\delta - \theta)/\delta$. In both situations, the Γ_j's are functions of an unknown parameter, θ or equivalently θ'. However, instead of (1) we can consider

$$Y_{ij} = \beta_0(t) + \beta_1(t)\Gamma_j(t) + \varepsilon_{ij} \quad (3)$$

where now $\Gamma_j(t)$ is the value of Γ_j at some hypothetical value of θ, $\theta = t$. Consider the least squares estimators of the parameters of this model with this value of θ, say $\hat{\beta}_0(t)$ and $\hat{\beta}_1(t)$. That is, for each putative value of θ, we obtain a fitted value for the phenotype,

$$\hat{Y}_{ij}(t) = \hat{\beta}_0(t) + \hat{\beta}_1(t)\Gamma_j(t). \quad (4)$$

A measure of the fit of the model is given by the residual sum of squares,

$$RSS(t) = \left\{ \sum_{j=1}^{4} \sum_{i=1}^{n_j} (Y_{ij} - \hat{Y}_{ij}(t))^2 \right\}. \quad (5)$$

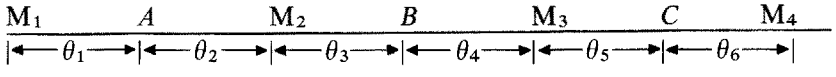
If we graph RSS(t), $0 \le t \le \delta$, against the corresponding chromosome positions t, we obtain a measure of the evidence of the presence of a QTL between the markers involved. A clear minimum value of RSS(t) would be evidence for the presence of a QTL. If we accept that a QTL is present in the interval, we can select as an estimate of its position the value of t that minimizes the RSS(t), say $\hat{\theta} = t_m$. If there is no segregation distortion in the cross, i.e. if each of the marker

groups appears in their expected proportion, then $E[\hat{\beta}_1(\theta)] = B - b$, $E[\hat{\beta}_0(\theta)] = 2b$ and t_m is a consistent estimator of θ.

The function RSS(t) behaves in a way approximately proportional to the inverse of the LOD score obtained with the Lander and Botstein (1989) method of interval mapping; a large RSS(t) corresponding to a low LOD score and conversely. The difference between the two methods arises solely because in (3) the error term (ε_{ij}) does not have a normal distribution, being a random variable from one of a number of normal distributions. The regression method proposed does use the same information as the interval mapping method, and the efficiency of the use of the information is unlikely to be substantially less than the maximum likelihood approach. The regression mapping approach does provide a tool by which to study algebraically the performance of procedures that use the information from flanking markers in situations when there may be more than a single QTL segregating in the cross. Such a study is not feasible with the LOD score procedure.

Expected value of RSS(t) under different hypothesis about the QTLS

To investigate the effect of QTLs in neighbouring regions of the chromosome, consider the following situation:



where loci *A*, *B* and *C* can contain alleles A, a; B, b and C, c, these symbols denoting the QTLs alleles affecting the characteristic as well as their effects; where, as before, each M_i is a codominant marker and where the distances in the chromosome are given by θ_i ($i = 1, 2, \ldots 6$). The B_1 segregating population is the product of the cross

$$M_1AM_2BM_3CM_4/m_1am_2bm_3cm_4 \times m_1am_2bm_3cm_4/m_1am_2bm_3cm_4.$$

We are interested in the behaviour of the estimation procedure when there is more than one QTL segregating in the cross. By modifying the values of the parameters we can study different situations, for example: (1) $A = C = 0$ and $B > 0$ – only the B QTL is present, and (2) $A = C > 0$ and $B = 0$ – two QTLs are present, and one empty flanked region ($M_2 - M_3$) exists between them.

The expected value of the residual sum of squares for the model presented in (3), averaged over the frequencies of the marker groups as well as over the data set, is given by

$$E[RSS(t)] = \sum_{j=i}^{4} p_j[V[Y_j] - 2C[Y_j, \hat{Y}_j(t)] + V[\hat{Y}_j(t)] + (E[Y_j] - E[\hat{Y}_j(t)])^2],$$

and neglecting terms resulting from the variance of the marker group means we can write

$$E[RSS(t)] \cong \sum_{j=1}^{4} p_j V[Y_j] + \sum_{j=1}^{4} p_j (E[Y_j] - E[\hat{Y}_j(t)])^2 = V_w + h(t) \quad (6)$$

where the p_j, $j = 1, 2, 3$ and 4, are the expected proportions of each marker class,

$p_1 = P[MM]$, $p_2 = P[Mm]$, $p_3 = P[mM]$, $p_4 = P[mm]$;

$E[Y_j]$ are the expected values of the quantitative characteristic in the j-th marker class and

$$E[\hat{Y}_j(t)] = E[\hat{\beta}_0(t)] + E[\hat{\beta}_1(t)]\Gamma_j(t) \quad (7)$$

for each $j = 1, 2, 3$ and 4. Thus $E[Y_j]$ is the true expectation of the observed characteristic for individuals in the j-th marker class for the given genetic situation. As an illustration, suppose that r different gametes ($G_1, G_2, \ldots G_r$) can be inherited from the F_1 parent, each one having the corresponding genetic value $g_1, g_2, \ldots g_r$ and with probabilities $P[G_i = g_i|$ markers group j] for the j-th marker class, then

$$E[Y_j] = \sum_{j=1}^{r} g_i P[G_i = g_i|\text{markers group j}]. \quad (8)$$

$E[\hat{\beta}_0(t)]$ and $E[\hat{\beta}_1(t)]$ in (7) are evaluated using (8).

The term $V_w = \sum_{j=4}^{4} p_j V[Y_j]$ in (6) represent the expected variance within each one of the four marker groups and does not depend on t but on the markers selected to be used in the model. h(t) depends on the putative position of the QTL, say t, as well as on the markers selected to be used in the model.

We shall use E[RSS(t)] to study the average performance of RSS(t) as a measure of the fit of the model for the different values of t. We shall not investigate the additional effects on the procedure of the random variation between the traits of individuals of the same marker genotype. Unfortunately h(t) is the ratio of two polynomials of degree eight in t. Therefore, the behaviour of E[RSS(t)] can only be studied numerically. The function E[RSS(t)] will now be evaluated at a grid of values for t within each flanked region. Ideally, E[RSS(t)] would be flat within flanked regions where

there is no QTL, and would show a single clear minimum at the true position of the QTL when one is present.

When analysing real data, the usual ANOVA table can be constructed for the model presented and the F test used to determine the approximate significance of any minimum value of the RSS(t). An extra degree of freedom needs to be subtracted from the degrees of freedom for the RSS(t) to account for the estimation of the position of the proposed QTL. The level of significance will only be approximate because the model is non-linear in t when no interference is assumed (see Eqs. 2), and in general the variation between plants within the same marker group is not normally distributed.

Numerical results

This section presents numerical values for E[RSS(t)] under different genetical situations. We are interested mainly in the shape of the function for different values of the parameters of interest, the effects and positions of the QTLs in the studied and in the neighbouring regions. Since an appropriate expected recombination value for the flanking markers appears to be 0.20 (Lander and Botstein 1989), we shall fix recombination between markers at that value. Two flanking markers give the least information about the position and effect of a QTL between them when the QTL is exactly halfway between the markers ($\theta = \theta'$), so that situation will be assumed. Without loss of generality we can put the effect of the alleles a, b and c equal to zero. Two cases of interest will be discussed, first (1) $A = C = 0$ and $B = 1$ – only the B QTL is present, and $B = 1$ by an arbitrary choice of scale, second, (2) $A = C = 1/2$ and $B = 0$ – two QTLs with effects of the same sign are present in different flanking regions, with their effects adding to B in (1). In both circumstances we shall study no interference and comment on the results obtained with complete interference when there is an important difference.

$A = C = 0$ and $B = 1$

Figure 1 presents the chromosome map and the graph of the function E[RSS(t)] for each point on the chromosome between the markers M_1 and M_4. The graph for each one of the three flanked regions, $M_1 - M_2$, $M_2 - M_3$ and $M_3 - M_4$ was obtained using only the information of the corresponding flanking markers and varying the value of t between 0 and 0.20 in each region. An increment of 0.01 was used, and then a smooth function was employed to interpolate between the data points. From this graph we can see that the absolute minimum (equal to $V_w = 0.055$) is achieved exactly in the chromosomal position of the QTL,

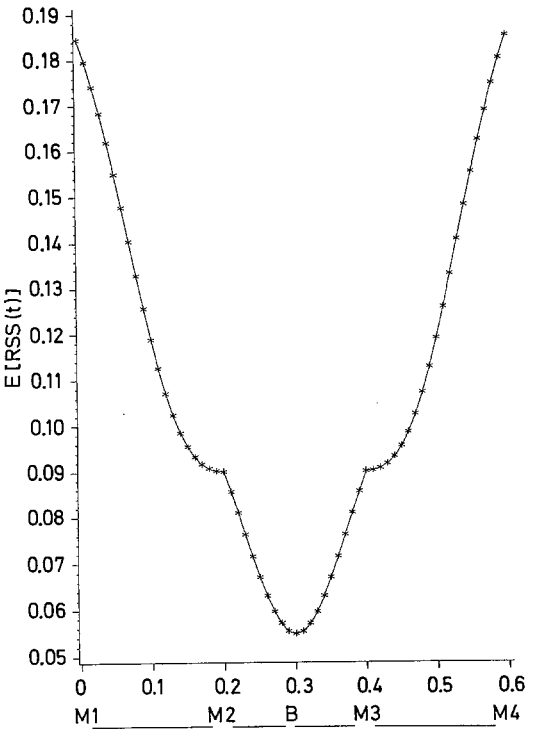


Fig. 1. E[RSS(t)] for each point between the markers M_1 and M_4. Size of the effect B = 1. No interference

t = 0.10. The corresponding size of effect for this point is $E[\hat{\beta}_1(0.1)] = 1.00$, confirming the fact that the method is consistent for both position and size of the effect. It is important to note that the function E[RSS(t)] is affected by the presence of the QTL B even when the model does not use the flanking markers between which that QTL is located (M_2 and M_3); i.e. when the model is applied to the neighbouring regions flanked by $M_1 - M_2$ and $M_3 - M_4$, the function E[RSS(t)] is not flat, but has a minimum (for each interval) at the position of the marker that is closer to the QTL B.

$A = C = 1/2$ and $B = 0$

Figure 2 presents the chromosome map and the graph of the function E[RSS(t)] for each point in the chromosome between the markers M_1 and M_4. The function E[RSS(t)] was calculated as before, but taking into account both QTLs when obtaining $E[Y_j]$. Two points of main interest can be noted from this graph. First, the method is no longer consistent for the positions, nor for the effects of the QTLs present. The minimum in the $M_1 - M_2$ region is 0.06 and is achieved at the position 0.13 (the real position of the QTL being 0.1); the minimum in the $M_3 - M_4$ region is again 0.06 and is found at distance 0.47 – the QTL being at position 0.50. Second, the absolute minimum of the function is 0.045 and is achieved in position 0.30. This is a 'ghost' QTL, with estimated effect 0.82, that results

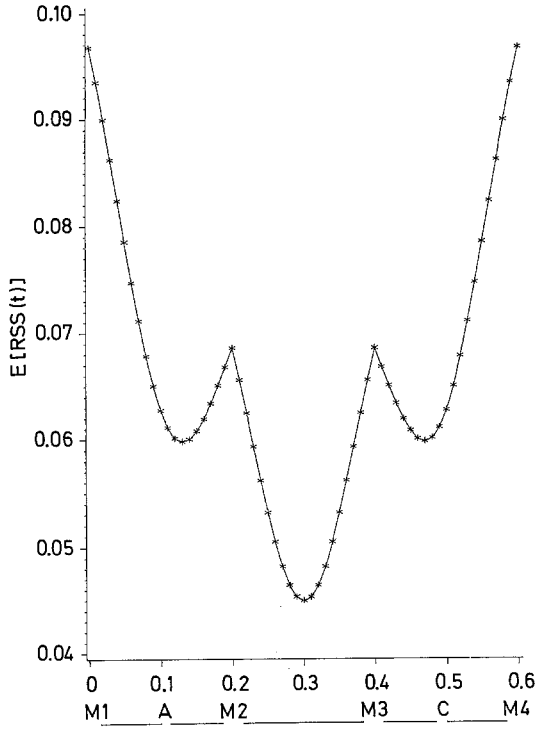


Fig. 2. E[RSS(t)] for each point between the markers M_1 and M_4. Size of the effects A = C = 1/2. No interference

from the presence of the two QTLs in the neighbouring flanking regions, each one with effect 0.5; so the 'ghost' QTL in the region $M_2 - M_3$ accounts for 82% of the total effect of the two real QTLs.

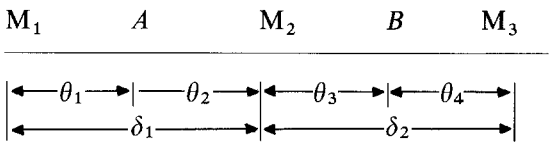
If the effects of the genes A and C are of different sign, that is if, for example, A = 1/2 and C = − 1/2, then the 'ghost' effect will not be present, but the size of the effects will be underestimated and their positions moved closer to the outside markers (M_1 and M_4 in this case).

Summarizing, in analysing real data, when only one QTL is present it affects the function RSS(t) not only in its flanked region but also in neighbouring regions; when two QTL are present, the RSS(t) function can present a global minimum somewhere in between the two QTLs. This 'ghost' effect can be wrongly interpreted as a true QTL.

Regression mapping with three markers

The results presented in the last section show that when more than one QTL is present in the same region of a chromosome, methods using information only from the flanking markers can lead to misleading conclusions. The natural way to avoid this problem is to use information from more than one pair of consecutive markers, for example, the information from three consecutive markers, say M_1, M_2 and M_3.

If we assume that two QTLs are present, one in each interval, we have



(9)

with a backcross B_1 we now have eight different markers groups, say 1-MMM, 2-MMm, 3-MmM, 4-mMM, 5-Mmm, 6-mMm, 7-mmM and 8-mmm (with the order of the markers being: 1, 2, and 3), and the model

$$Y_{ij}(t) = \beta_0(t) + \beta_1(t)\Gamma_j(t) + \beta_2(t)\Delta_j(t) + \beta_3(t)\Psi_j(t) + \varepsilon_{ij} \tag{10}$$

where j is the marker class of the Y_{ij} individual (j = 1, 2,... 8) and i = 1, 2,... n_j, the number of individuals in the j-th marker class. Now $t' = (t_1, t_2)$ is a vector with two components, t_1 the putative distance from M_1 to A; $0 \leq t_1 \leq \delta_1$, and t_2 the putative distance from the marker M_2 to B, $A \leq t_2 \leq \delta_2$, and

$$\Gamma_j(\mathbf{t}) = P[AB|\text{j-th marker class, and } \mathbf{t}], \tag{11}$$

$$\Delta_j(\mathbf{t}) = P[Ab|\text{j-th marker class, and } \mathbf{t}], \tag{12}$$

$$\Psi_j(\mathbf{t}) = P[aB|\text{j-th marker class and } \mathbf{t}]. \tag{13}$$

From the least square estimators of the parameters, $\hat{\beta}_0(t)$ to $\hat{\beta}_3(t)$, we can obtain estimates of each Y_{ij} value, say $\hat{Y}_{ij}(\mathbf{t})$, and so for each value of **t** (in the bivariate space: $0 \leq t_1 \leq \delta_1$, and $0 \leq t_2 \leq \delta_2$) we can measure the fit of the model by the RSS(**t**), defined by (5), but now summed over j = 1, 2,... 8. We select as estimators of the positions of the two assumed QTLs the values in the vector **t** that minimize RSS(**t**), say $\mathbf{t} = \mathbf{t}_m$. If the assumption of the model is true, that is there are only two QTLs segregating, one between M_1 and M_2 and the other between M_2 and M_3, the procedure is consistent with respect to both the positions and sizes of the effects of the QTLs.

Because the parameterization of this model accounts for the effect of each one of four possible QTL genotypes, aabb, AaBb, Aabb and aaBb, estimated by $\hat{\beta}_0(\mathbf{t})$ to $\hat{\beta}_3(\mathbf{t})$, respectively,

$$\hat{\beta}_0(\mathbf{t}) + \hat{\beta}_1(\mathbf{t}) - \hat{\beta}_2(\mathbf{t}) - \hat{\beta}_3(\mathbf{t})$$

will estimate a component of epistatic effects of the two loci.

As before we can use the function E[RSS(**t**)] to study the behaviour of the model when there are other QTLs present in neighbouring regions of the chromosome. The next section presents numerical results for two cases of interest; when the model is correct and there are two QTL's, one in each interval and when there is only one QTL. In each case only the results for no interference are presented. The results for complete

interference were very similar, despite the fact that the data are then much more informative about the existence of the QTLs. For example, with complete interference we can be sure that the marker class $M_1M_2M_3$ has the QTLs A and B (if the situation assumed in Eq. 9 is true), because only double recombination could produce gametes such as $M_1aM_2BM_3$. With complete interference marker classes 3-MmM and 6-mMm will not exist and the Γ functions are linear in **t**. It is important to remark that our results only indicate that the cases of no interference and complete interference are similar on average, i.e. when the sample size is very large.

$A = B = 1/2$

The map of the genes involved in this case is

M_1 *A* M_2 *B* M_3

|←0.1→| |←0.1→|

where as before the recombination between the consecutive markers is 0.20 recombination units. Figure 3 shows the graph of the E[RSS(**t**)] bivariate surface generated by values of $\mathbf{t} = (t_1, t_2)$ for t_1 between 0 and 0.19 and t_2 between 0 and 0.20. The surface in Fig. 3 has a global minimum at the point $t_1 = 0.1$, $t_2 = 0.1$; the real position of the QTLs. At that point $E[\hat{A}] = E[\hat{B}] = 1/2$, the real sizes of the QTL effects. This confirms the consistency of the procedure. At the point $t_1 = 0$, $t_2 = 0$, corresponding to assuming that A is at M_1 and B is at M_2, there is a high value of E[RSS(**t**)], 0.0362. If we vary t_1, with t_2 constant at $t_2 = 0$ there is no appreciable change in E[RSS(**t**)]. This is because when we assume that one of the QTLs (B in this case) is in the middle of the region (the position of M_2), this QTL can account for a large part of the variation in the data. The right-hand corner of the graph, the point $t_1 = 0$, $t_2 = 0.20$, corresponds to the assumption that the QTL A is at M_1 and the QTL B at M_3. At this point E[RSS(**t**)] has the highest value, 0.0524. If we set $t_1 = 0$ and change the value of t_2, the graph is very informative, giving a local minimum at $t_2 = 0.1$. The same occurs if we set $t_2 = 0.2$ and then change t_1. In the left-hand corner of the graph the value for $t_1 = 0.20$, $t_2 = 0$ is not given. An indeterminacy occurs in the function E[RSS(**t**)], at that point both QTLs are assumed to be at M_2, constituting a single QTL, and so the effects of A and B are not separately estimable.

In general we see from the graph that when the assumptions of the model are correct an accurate estimation of the positions and effects of both QTLs should be possible.

$A = 0$, $B = 1$

In this case the chromosome situation is

M_1 M_2 *B* M_3

|←0.1→|

and the results are presented graphically in Fig. 4. In this case the global minimum of E[RSS(**t**)] is reached in a line, all the points where $t_2 = 0.1$, independently of

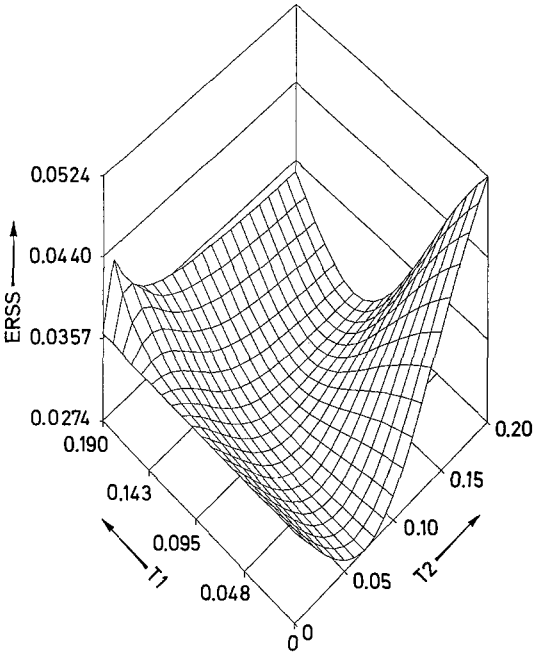


Fig. 3. Three markers regression mapping. E[RSS(**t**)] for regions M_1–M_2(*T*1) and M_2–M_3 (*T*2). No interference. Two QTLs present, one in the middle of each region; see text

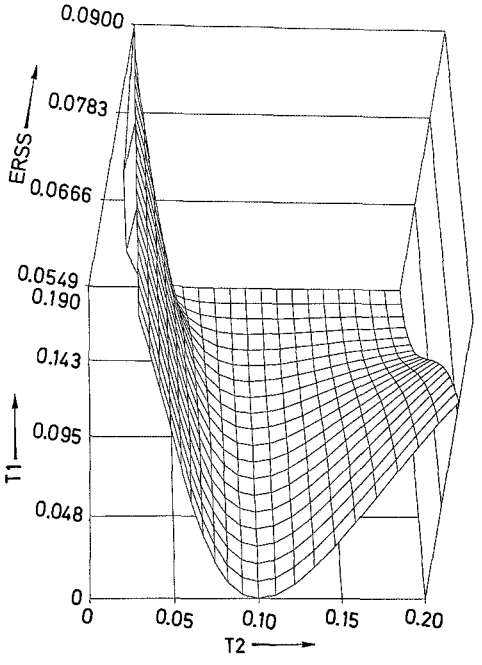


Fig. 4. Three markers regression mapping. E[RSS(**t**)] for regions M_1–M_2(*T*1) and M_2–M_3 (*T*2). No interference. One QTL present in the middle of the region M_2–M_3 (*T*2); see text

the values for t_1. Along this line the expected values for the size of the QTL effects also take their real values $E[\hat{A}] = 0$ and $E[\hat{B}] = 1$. This is because the position of a non-existent QTL is irrelevant. Figures 3 and 4 show that bivariate regression mapping with three markers can discriminate between the presence of one and two QTLs.

The regression model presented above can be generalized to use the information of any number of markers. However, simultaneous analysis of many markers is restricted because the sample sizes required for a given accuracy increase exponentially with the number of markers.

Simulation

A single simulation will be used to illustrate how a 'ghost' QTL arises when using both the Lander and Botstein (1989) interval mapping and the regression mapping approach using flanking markers only.

Because the aim of the simulation was to observe systematic biases rather than sampling variation a large sample size of 2000 observations was selected. The phenotypic scale was chosen so that the environmental variance was $\sigma_e^2 = 1$. A model with two QTLs A and C with size of effects equal to 1/2 for each locus was used. The genetical situation was simulated corresponding to the backcross,

$$\underset{M_1AM_2M_3CM_4/m_1am_2m_3cm_4}{F_1} \times \underset{m_1am_2m_3cm_4/m_1am_2m_3cm_4}{P_1}$$

where the map of the genes involved is given by

M_1	A	M_2	M_3	C	M_4
$\hat{0}$	$\hat{0.1}$	$\hat{0.2}$	$\hat{0.4}$	$\hat{0.5}$	$\hat{0.6}$

No interference was allowed. The analysis was performed by both the Lander and Botstein interval mapping (see Fig. 5) and the regression mapping procedure (Fig. 6). Figure 5 and 6 illustrate how the two measures, LOD score and RSS(t), are closely although inversely related. Table 1 presents the estimates where $\hat{\beta}_1(\hat{\theta})$ is the estimate of the genetic effect obtained by regression mapping and minimum RSS($\hat{\theta}$) denotes the local minimum of the residual sum of squares, reached at the value $t = \hat{\theta}$. The corresponding estimates of the position of the putative QTL obtained by selecting the position that maximizes the LOD score in the interval studied is denoted by $\tilde{\theta}$, and the estimated size of the effect is denoted by $\hat{G}(\tilde{\theta})$.

The interval between the markers M_2 and M_3 does not have any QTL, however the global maximum of the LOD score and the global minimum of the RSS(t) functions are both located between M_2 and M_3 and at the same position between the markers, $\hat{\theta} = 0.29$. This is a ghost QTL produced by the effects of the QTLs A

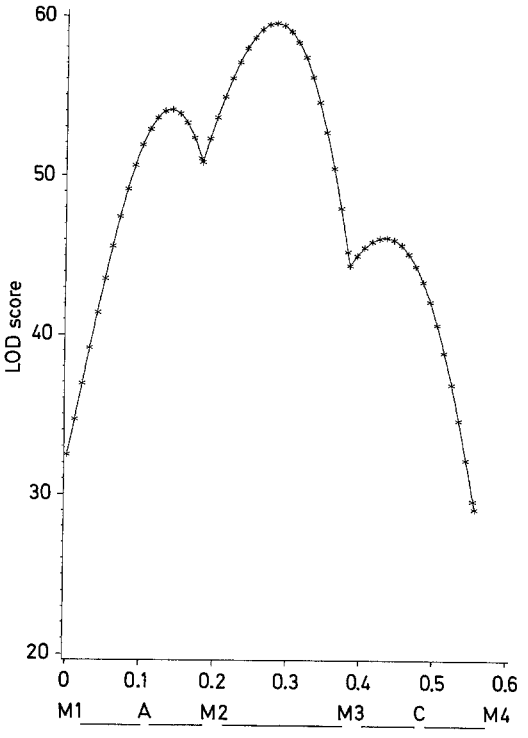


Fig. 5. Interval mapping for 2000 simulated observations. Size of the QTLs effects: A = 1/2, C = 1/2. Error variance, $\sigma_e^2 = 1$

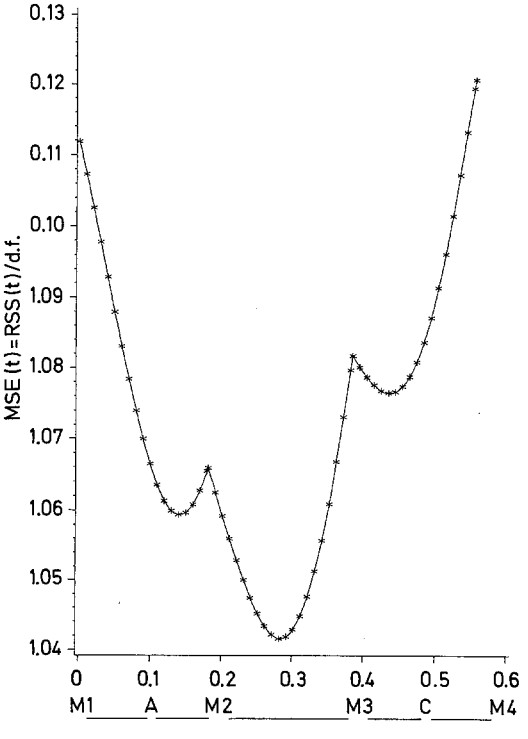


Fig. 6. Regression mapping for 2000 simulated observations. Size of the QTLs effects: A = 1/2, C = 1/2. Error variance, $\sigma_e^2 = 1$

Table 1. Estimates for the QTL sizes of effect and positions for a simulated data set using Interval Mapping and two marker regression mapping procedures

Markers	Regression mapping			Interval mapping		
	Minimum RSS($\hat{\theta}$)	$\hat{\theta}$	$\hat{\beta}_1(\hat{\theta})$	Maximum LOD	$\hat{\theta}$	$\hat{G}(\hat{\theta})$
M_1–M_2	1.0594	0.14	0.8205	54.2637	0.14	0.8216
M_2–M_3	1.0416	0.29	0.9136	58.6804	0.29	0.8847
M_3–M_4	1.0764	0.44	0.7599	49.4325	0.44	0.7493

and C reflected within the M_2-M_3 region from outside the region.

Discussion

When there is only one QTL segregating in a population the interval mapping procedure or the regression model using pairs of markers can be confidently applied to estimate the effect and the position of the QTL. Even in this simple situation problems can arise; for example, if a QTL A is at one end of the chromosome, the order of the loci being A, M_1, $M_2 \ldots$, M_r, then a local maximum of the LOD score is likely to arise between the markers M_1 and M_2, giving a mislocalization of the QTL and an under-estimation of its effect. An example of this situation appears to have happened in barley with the gene *Yield* 1 (Hyne and Snape 1991). The local minima will tend to appear in the marker position that is closest to the real QTL (see Fig. 1).

Greater problems arise if more than one QTL is segregating. We have seen that the effects of QTLs in neighbouring regions affect the estimation of the position and size of effect of a QTL between the pair of markers. Of particular interest are the results presented in Fig. 2. In this case two QTLs are segregating, each one in a different flanking region – QTL A in M_1-M_2 and QTL C in M_3-M_4. In this situation, if we assume the erroneous hypothesis that there is only one QTL segregating, the method is likely to locate a ghost QTL in the region M_2-M_3, i.e. a QTL where none exists. The problem is that the method cannot discriminate between very different hypotheses; in fact, for any threshold proposed, it is not difficult to place two QTLs in neighbour regions of a central flanked region so that only the ghost effect in the empty region will pass the proposed threshold and hence only the 'ghost' will be detected. The 'ghost' minimum may not only be a local minimum, but the global minimum of the function for all regions searched. The ghost effect is much more important than a simple 'false positive' given by random variation; the ghost effect can be present whatever the sample size.

As mentioned by Knapp (1991), if multiple QTLs affect a trait then estimates of the sizes of the effects of QTLs from individual locus models are biased. There can be ghost effects present in the models presented by Weller (1986), Jensen (1989), Luo and Kearsey (1989, 1991), Simpson (1989) and Knapp et al. (1990), because they do not take into account the possible presence of more than one QTL.

Lander and Botstein (1989) suggest that when a LOD score graph shows evidence of two QTLs, each of the QTLs should, in turn, be fixed at their estimated positions. The difference between the LOD scores calculated for the fixed QTL and a possible second QTL and the LOD scores for the fixed QTL only is then plotted against distance along the chromosome. A high peak indicates the presence of a second QTL. This procedure is incorrect in that the estimated positions and sizes of effects of the two QTL are not independent. The search must be over all possible pairs of values of the two positions and the two sizes of effect. The Lander and Botstein procedure will probably be a good approximation when the two QTLs are sufficiently well separated. However, in the case shown in Fig. 2, when the distance is not great, the procedure fails in a major way because the ghost QTL has the larger apparent effect and, when fixed, there will be little evidence of any other QTLs, their effects having already been almost completely attributed to the ghost QTL. The global minimum of the E[RSS(**t**)] function is reached in position 0.30, and the expected value of the estimates for the size of the effect of the 'ghost' QTL is 0.91. The effects of the real QTLs, A and C, have $A + C = 1$, so the ghost accounts for 91% of the effects of the QTLs. The correct approach is to search the bivariate or multivariate LOD or RSS(**t**) surface generated when taking into account the presence of all possible QTLs by using information from all the markers involved.

Taking the information from three markers into account will result in a search for one or more QTLs over a bivariate surface. This search can be done by the maximum likelihood approach, but here we will discuss the almost equivalent results that can be achieved by regression mapping, as presented in (10). When the assumption that two QTL are segregating, one in each of the neighbour regions M_1-M_2 and M_2-M_3 is true, then the regression mapping procedure is consistent, and so E[RSS(**t**)] has a simple minimum value at the position of the two QTLs (Fig. 3). Even when there is only one QTL in one of the regions, the method gives a surface that is very flat as the position of the other hypothesized but non-existent QTL varies. From Fig. 4, the surface reaches the minimum value for E[RSS(**t**)] in a line corresponding to $t_2 = 0.1$, and $0 < t_1 < 0.19$. Along this line the value for E[RSS(**t**)] is equal to $V_w = 0.055$, and the expectation of the estimator of the size of the effect $E[\hat{\beta}_1(\mathbf{t})] = 1$; i.e. the procedure is still consistent. Thus, the optimum procedure will always be to look for QTLs in the two neighbouring regions,

because even if there is none or one there the procedure still gives good results. Unfortunately using three flanking markers instead of two demands a larger sample size, because any variation in the expected proportions of the marker classes (in this case eight instead of four when using only two flanking markers) can seriously affect the accuracy of the estimation. It is important to remark that the expectations calculated in this paper to give the values of E[RSS(t)] are based on the assumption that there is no segregation distortion; if this assumption is not made, a technical problem arises because then the expectations of the estimators do not exist. This is simply because all the marker classes have a probability larger than zero of being empty. Nevertheless, if the sample size is large enough and all the eight marker classes are well represented for a given triplet of markers, procedures using three markers are recommended in any region of the chromosome that appears to have a high LOD score or a minimum RSS(t), that is, in any region in which the presence of one or more QTLs is suspected. In some cases this procedure can help to decide if an effect is a real QTL or a ghost one. For example, assume that the situation is as presented in Fig. 2; i.e. two QTLs are present, with an empty flanked region ($M_2 - M_3$) between them. In that case we saw that an univariate search will give a global minimum in the empty region, and if we use the bivariate regression method we obtain again estimates for two QTLs: the real A QTL, for which we obtain biased estimations of size and position and the B ghost, for which we will obtain a size of effect different from zero and a position between M_2 and M_3. One way to decide that the effect B is a ghost using information from three markers at the same time will be to perform regression mapping now using the information of markers M_2, M_3 and M_4 and searching in the bivariate surface generated. Doing this we obtain estimates for the putative QTL in $M_2 - M_3$, B, and for the putative QTL in $M_3 - M_4$, C. We will still obtain the ghost effect B; in this case with the same size of effect, but the position of the ghost will now be different. The fact that the B position is moved when estimated from different marker triplets and the fact that the other QTL, say C, is present will be good clues to conclude then that the apparent B QTL is not real. Another way to try to distinguish between two and three QTLs in the situation presented in Fig. 2 is to perform three-marker regression mapping, first with markers M_1, M_2 and M_4 and then with markers M_1, M_3 and M_4. In that case the estimation of both analyses will coincide, giving a consistent estimation of position and size of effect of both QTLs, A and C, and eliminating the incorrect estimation of the 'ghost' B. However, the decisive test to conclude if B is a ghost effect or a real QTL is to use the information provided by the four markers $M_1 - M_4$ at the same time. Using this information by means of (four-marker) regression mapping will show that the B effect is not real; the ghost will disappear. Unfortunately, this search can be performed only if the sample size of the segregating population is large enough to have an appropriate representation of each of the 16 different marker classes.

Haley and Knott (1992) have recently developed a similar method for F_2 populations. In their publication they compare, by simulation, regression and maximum likelihood methods and show the close correspondence of the results from the two methods. They emphasize the speed of application and generality of regression methods.

In summary, we recommend that in any region where the presence of a QTL is suspected when using any of the estimation methods available, a bivariate, or if possible multivariate, search should be performed. This will reduce the effects of QTLs that may be segregating in neighbouring regions.

Acknowledgements. We are grateful to Dr. J. W. Snape for helpful discussions based on an earlier draft of this paper and to the referees for their constructive comments.

References

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

Hyne G, Snape JW (1991) Mapping quantitative trait loci for yield in wheat. In: Biometrics in plant breeding. Proc 8th Meet Eucarpia Section, Biometrics Plant Breed. Research Institute of Agroecology and Soil Management, Brno, Czechoslovakia, pp 47–56

Jensen J (1989) Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. Theor Appl Genet 78:613–618

Knapp SJ (1991) Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbreed and doubled haploid progeny. Theor Appl Genet 81:333–338

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

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

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

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

Paterson AH, Lander E, Hewitt JD, Peterson S, Lincln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 355:721–726

Simpson SP (1989) Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. Theor Appl Genet 77:815–819

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