A REVIEW OF METHODS FOR IDENTIFYING QTLS IN EXPERIMENTAL CROSSES

By Karl W. Broman[1] and T. P. Speed

Marshfield Medical Research Foundation and University of California, Berkeley

Identifying the genetic loci contributing to variation in traits which are quantitative in nature (such as the yield from an agricultural crop or the number of abdominal bristles on a fruit fly) is a problem of great importance to biologists. The number and effects of such loci (called quantitative trait loci or QTLs) help us to understand the biochemical basis of these traits, and of their evolution in populations over time. Moreover, knowledge of these loci can aid in designing selection experiments to improve the traits.

There are a large number of different methods for identifying the QTLs segregating in an experimental cross. Little has been written critically comparing the methods, and there have been few studies comparing their performance; we make an attempt at this.

1. Introduction. Identifying the genetic loci contributing to variation in traits which are quantitative in nature (such as the yield from an agricultural crop or the number of abdominal bristles on a fruit fly) is a problem of great importance to biologists. The number and effects of such loci (called quantitative trait loci or QTLs) help us to understand the biochemical basis of these traits, and of their evolution in populations over time. Moreover, knowledge of these loci can aid in designing selection experiments to improve the traits.

There are a large number of different methods for identifying the QTLs segregating in an experimental cross. Little has been written critically comparing the methods, and there have been few studies comparing their performance; we make an attempt at this.

There are three major points which we would like to make in this paper: first, identifying QTLs is best seen as a model selection problem; most current methods do not view the problem in this way. Second, it is important to critically compare the different approaches to any problem; more refined analyses and more complicated algorithms do not necessarily lead to improved results. Finally, different situations may require different methods; with each new experiment, one must reconsider the available approaches, as each new problem may require a new method.

---

[1] This work was part of the author's Ph.D. dissertation at the University of California, Berkeley.

AMS 1991 *subject classifications.* Primary 92D10; secondary 62J05.

Key words and phrases. QTL, interval mapping, model selection.

We will focus on a backcross experiment, and will assume that the QTLs act additively. Identifying interactions between loci is a much more difficult problem; considering only the simple additive case will lead to greater clarity. We concentrate almost exclusively on detecting QTLs, considering the estimation of the QTLs' effects and precise locations of secondary importance.

In the remaining part of this section, we will describe the backcross experiment, the resulting data, some statistical models and the goals of the QTL experiment and its analysis. In Section 2, we will review the major approaches to identifying QTLs. In Section 3, we will describe the results of some simulations to compare the performance of a number of the most important methods.

1.1. *Experiments.* Most experiments aimed at identifying quantitative trait loci (QTLs) begin with two pure-breeding lines which differ in the trait of interest. We will call these the low (L) and high (H) parental lines. The lines are the result of intensive inbreeding, so that each is essentially homozygous at all loci (meaning that, at each locus, they received the same allele from each of their two parents). Crossing these two parental lines gives the first filial (or F_1) generation. The F_1 individuals receive a copy of each chromosome from each of the two parental lines, and so, wherever the parental lines differ, they are heterozygous. All F_1 individuals will be genetically identical, just as the individuals in each of the parental lines were.

In a backcross, the F_1 individuals are crossed to one of the two parental lines, for example, the low line. The backcross progeny, which may number from 100 to over 1000, receive one chromosome from the low parental line, and one from the F_1. Thus, at each locus, they have genotype either LL or HL. As a result of crossing over during meiosis (the process during which gametes or sex cells are formed), the chromosome received from the F_1 parent is a mosaic of the two parental chromosomes. At each locus, there is half a chance of receiving the allele from the low parental line (L) and half a chance of receiving the allele from the high parental line (H). The chromosome received will alternate between stretches of L's and H's.

The goal is to look for an association between the phenotype of an individual and whether it received the L or H allele from the F_1 parent at various marker loci.

We use the backcross as our chief example, because of its simplicity. At each locus in the genome, the backcross progeny have one of only two possible genotypes. The strategies developed for analyzing backcross experiments will generally work for other experimental crosses as well.

1.2. *Data.* In an experiment like a backcross, each of the progeny is scored for one or more traits. (We will consider only one trait.) In addition, the progeny are typed at a number of genetic markers: at each marker, it is determined whether the allele an individual received from the F_1 parent was that from the low or high parental line. Thus, at each of these marker loci, we determine, for

each of the progeny, whether its genotype is LL or HL.

A genetic map specifying the relative locations of the markers may be known, or will be estimated using the data from the current experiment. Such a map gives the linear order of the markers on the various chromosomes. The distance between markers in a genetic map is given by genetic distance, in the units centiMorgans (cM). Two markers are separated by d cM, if d is the expected number of crossovers between the markers in 100 meiotic products.

Generally, we will write y_i for the phenotype (trait value) of individual i, and $x_{ij} = 1$ or 0 according to whether individual i has genotype HL or LL at the jth marker.

Typical experiments involve 100 to 1000 progeny, and use between 100 and 300 genetic markers.

1.3. *Models.*

1.3.1. *Model for recombination.* A diploid organism has two copies of each chromosome, one from its mother and one from its father. During the formation of gametes (sex cells), in the process of meiosis, the two homologous copies of a chromosome may undergo exchanges, called crossovers. Each of the gametes formed contains one copy of each chromosome, and each of these will be a mosaic of the two original homologs.

The locations of the crossovers along a chromosome are often modelled as a Poisson process (the assumption of "no interference"), with the processes in different individuals and on different chromosomes in one individual being independent. Moreover, at each locus, there is an equal chance that the allele is either paternally or maternally derived.

Consider a chromosome with k markers, and let $x_{ij} = 1$ or 0 if the ith individual has genotype HL or LL, respectively, at the jth marker. Then $\Pr(x_{ij} = 1) = \Pr(x_{ij} = 0) = 1/2$, for all i, j, and letting $x_j = (x_{ij})$, the x_j form a Markov chain.

Consider markers j_1 and j_2, separated by a distance of d cM (so that d is the expected number of crossovers between these two markers in 100 meioses). If an odd number of exchanges occur between these markers, then $x_{ij_1} \neq x_{ij_2}$. This event is called a recombination. Let $r = \Pr(x_{ij_1} \neq x_{ij_2})$. Then $r = \frac{1}{2}(1 - e^{-2d/100})$. This is called the Haldane map function (Haldane 1919).

1.3.2. *Model connecting genotype and phenotype.* Let y denote the phenotype for an individual derived from a backcross experiment. Let g be a vector giving its genotype at all loci. Let $\mu_g = \mathrm{E}(y|g)$, the average phenotype for individuals with genotype g, and $\sigma_g^2 = \mathrm{Var}(y|g)$, the variance of the phenotypes of individuals with genotype g. In principle, these could be arbitrary functions of g. But imagine that there are a small number, p, of loci which affect the trait. Let $(g_1, \ldots, g_p)$ denote the genotypes of the individual at these loci. Then

$$\mathrm{E}(y|g) = \mu_{g_1 \ldots g_p}$$

$$\text{and } \operatorname{Var}(y|g) = \sigma^2_{g_1 \ldots g_p}.$$

Often, we assume that the trait is homoscedastic—that the variance is constant within the genotype groups:

$$\operatorname{Var}(y|g) = \sigma^2.$$

There are 2^p different possible genotypes at the p QTLs. Each genotype could have a distinct trait mean. But it is often assumed that the loci act additively. Let $z_j = 1$ or 0, according to whether $g_j = \text{HL}$ or LL. We imagine that

$$\operatorname{E}(y|g) = \mu + \sum_{j=1}^{p} \beta_j z_j .$$

Deviation from additivity (i.e. interactions between the QTLs) is called epistasis.

Most current methods use this assumption of additivity. Pairwise interactions are occasionally included, but few studies have found significant effects when using such an approach (Tanksley 1993), possibly because of the enormous number of pairwise interactions which must be considered. Strong evidence for epistasis has been demonstrated in one of the most studied quantitative traits, the number of abdominal bristles in *Drosophila* (Shrimpton and Robertson 1988; Long et al. 1995). Thus one should not discount the importance of epistasis.

It is important to note that additional cofactors, such as sex and treatment effects, may also be included in the above model, though we do not discuss that issue here.

A further often used assumption is that, given the genotypes at the QTLs, the trait y follows a normal distribution. Thus, if we group the backcross progeny according to their genotypes at the p QTLs, the phenotypes within each group will be normally distributed. The phenotypes for the backcross progeny, considered as a whole, will follow a mixture of normal distributions.

In this paper, we will focus on the case of strict additivity, with the further assumption of normality. This is not because we feel that it is the best approach, but rather because this simple case is still not well solved. In comparing the major approaches to identifying QTLs, the important differences will stand out most clearly if we avoid the added difficulties which accompany a search for epistasis.

1.4. *Goals.* Consider a backcross giving n progeny. For individual i, we obtain the phenotype, y_i, and the genotype at a set of M markers. Let $x_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the jth marker.

We imagine that there are a set of p QTLs, and write $z_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the jth QTL. Let

$$y_i = \mu + \sum_{j=1}^{p} \beta_j z_{ij} + \epsilon_i$$

where the ϵ_i are independent and identically distributed (iid) normal$(0, \sigma^2)$.

The ultimate goal is to estimate the number of QTLs, p, the locations of the QTLs, and their effects, β_j. In estimating the number and locations of the QTLs, we may make two errors: we may miss some of the QTLs, and we may include additional, extraneous loci.

In practice, a scientist may be satisfied with finding a few QTLs with large effect. In QTL experiments aimed at improving an agricultural crop, one seeks only the major QTLs, which may then be introgressed from one line into another. Furthermore, with a few major QTLs in hand, it may be possible to design experiments which identify the other QTLs segregating in a cross.

How one chooses to balance the two errors, of missing important loci and of including extraneous loci, depends on the goals of the scientists who designed the cross. In some cases, one may wish to find as many of the QTLs as possible and be undeterred by the possibility that several of the identified loci are, in fact, extraneous ones, of no effect. In other situations, one may be satisfied with identifying only a few major QTLs, in order to avoid including extraneous ones.

2. Major approaches. There are a large number of different methods for identifying the QTLs segregating in an experimental cross. In this section, we describe most of the proposed methods and discuss their advantages and disadvantages. It is best to distinguish between methods which model a single QTL at a time from those which attempt to model the effects of several QTLs at once. In Section 2.1, we review the single QTL methods, and in Section 2.2, we review the multiple QTL methods.

2.1. *Single QTL methods.* We will consider five basic single QTL methods: analysis of variance at a single marker, maximum likelihood using a single marker, interval mapping (i.e., maximum likelihood using flanking markers), an approximation to interval mapping called "regression mapping," and a further method which gives results approximating interval mapping, called "marker regression." Each of these methods includes a so-called "genome scan." The loci are considered one at a time, and a significance test for the presence of a single QTL is performed at each. Generally, the significance level used for the tests is adjusted to account for the multiple tests performed. Areas of the genome which give significant results are indicated to contain a QTL.

2.1.1. *Analysis of variance.* Analysis of variance (ANOVA) is the simplest method for identifying QTLs (see Soller et al. 1976). Consider a single marker locus, and group the progeny according to their genotypes at that marker. To test for the presence of a QTL, we look for differences between the mean phenotype for the different groups using ANOVA. If a QTL is tightly linked to the marker, then grouping the progeny according to their marker genotypes will be nearly the same as grouping them according to their (unknown) QTL genotypes, with recombinants being placed in the wrong groups.

Consider a backcross with a single segregating QTL. Suppose that the progeny with QTL genotype HL have mean phenotype μ_H, and that progeny with QTL genotype LL have mean phenotype μ_L, so the QTL has effect $\beta = \mu_H - \mu_L$.

Consider a marker locus which is a recombination fraction r away from the QTL. Of the individuals with marker genotype HL, a fraction $(1-r)$ of them have QTL genotype HL, while the remainder have QTL genotype LL, and so these individuals have mean phenotype $(1-r)\mu_H + r\mu_L = \mu_H - \beta r$. The individuals with marker genotype LL have mean phenotype $(1-r)\mu_L + r\mu_H = \mu_L + \beta r$. Thus the mean difference between the two marker genotype groups is $(\mu_H - \beta r) - (\mu_L + \beta r) = \beta(1-2r)$. And so a non-zero mean difference between the marker genotype groups indicates linkage between the marker and a QTL.

There are two drawbacks to this method. First, we do not receive separate estimates of the location of the QTL relative to the marker (r) and its effect (β). QTL location is indicated only by looking at which markers give the greatest differences between genotype group means. Second, when the markers are widely spaced, the QTL may be quite far from all markers, and so the power for detection will decrease, since the difference between marker genotype means decreases linearly as the recombination fraction between the marker and the QTL increases.

2.1.2. *Maximum likelihood with a single marker.* To get around the problems with ANOVA, several authors have proposed to explicitly model the location of the QTL with respect to the marker, and then use maximum likelihood (ML), or an approximation to ML, to estimate the distance between the marker and the QTL as well as the QTL's effect (Weller 1986, 1987; Simpson 1989). This method makes use of the fact that the marker genotype groups are not quite the same as the QTL genotype groups.

Consider again the backcross discussed in the previous section. Suppose that the individuals who are HL at the QTL have phenotypes which are normal(μ_H, σ^2), and the individuals who are LL at the QTL have phenotypes which are normal(μ_L, σ^2). Then at a marker which is a recombination fraction r away from the QTL, the phenotype distribution for individuals who are HL is a mixture of two normals, with density

$$f_1(y; \mu_H, \mu_L, \sigma, r) = (1-r)\phi\left(\frac{y-\mu_H}{\sigma}\right) + r\phi\left(\frac{y-\mu_L}{\sigma}\right),$$

where ϕ is the density of the standard normal distribution. The phenotype distribution for individuals who are LL at the marker has density

$$f_2(y; \mu_H, \mu_L, \sigma, r) = (1-r)\phi\left(\frac{y-\mu_L}{\sigma}\right) + r\phi\left(\frac{y-\mu_H}{\sigma}\right).$$

Let $x_i = 1$ or 0, according to whether individual i has marker genotype HL or LL. Let y_i denote the phenotype for individual i. Then the likelihood under this model, letting θ denote the vector of parameters (μ_H, μ_L, σ), is

$$L(\theta, r; y, x) = \prod_i [f_1(y_i; \theta, r)]^{x_i} [f_2(y_i; \theta, r)]^{1-x_i}$$

Maximizing this function over θ, using, for example, the EM algorithm (Dempster et al. 1977), gives the maximum likelihood estimates. This is done for a particular

value of the recombination fraction r. We then maximize the likelihood over r to obtain $\hat{r}$.

Linkage between the marker and the QTL is tested by performing a likelihood ratio test, comparing the above model, with a single QTL linked to the marker, to the null hypothesis of no segregating QTLs, where the individuals are assumed to have phenotypes which are normal(μ, σ^2).

The likelihood under the null hypothesis, letting $\theta_0 = (\mu, \sigma)$ is

$$L_0(\theta_0; y) = \prod_i \phi\left(\frac{y_i - \mu}{\sigma}\right).$$

The likelihood ratio test is performed by calculating the likelihood ratio, or, as seems to be preferred by geneticists, the LOD score, which is the log (base 10) likelihood ratio

$$\mathrm{LOD}(r) = \log_{10}\left[\frac{\max_\theta L(\theta, r; y, x)}{\max_{\theta_0} L_0(\theta_0; y)}\right]$$

and comparing it to the distribution of the maximum LOD score under the null hypothesis (that is, under the assumption that no QTLs are segregating).

This method has the advantage of giving separate estimates of the QTL's location with respect to a marker and its effect. One disadvantage is the great increase in computation associated with maximizing the likelihood function to obtain parameter estimates. But a bigger problem involves combining the information for different markers to give a single estimate of the QTL location; it is not at all clear how this can be done.

2.1.3. *Interval mapping.* Lander and Botstein (1989) improved on the previous single marker approaches by considering flanking markers. Their method has been called "interval mapping," and is currently one of the most commonly used methods for identifying QTLs in experimental crosses. (Note that Mather and Jinks (1977) proposed a similar approach, using the method of moments with flanking markers.)

Again, they assume that there is a single segregating QTL, and that backcross individuals have phenotypes which are normally distributed with mean μ_H or μ_L, according to whether their QTL genotype is HL or LL, and common variance σ^2. Further, they use the assumption of no crossover interference, and require a genetic map specifying the locations of the markers.

Consider two markers which are separated by d cM, corresponding to a recombination fraction of $r = \frac{1}{2}(1 - e^{-2d/100})$, and a putative QTL located d_L cM from the left marker, corresponding to a recombination fraction of $r_L = \frac{1}{2}(1 - e^{-2d_L/100})$. The recombination fraction between the QTL and the right marker is then $r_R = (r - r_L)/(1 - 2r_L)$. There are four possible sets of genotypes at the two marker loci; for each, we can calculate the conditional probability for each of the two QTL genotypes, given the marker genotypes. These are displayed in Table 1. Note that, with fully informative markers, the flanking markers provide all of the information about the QTL genotypes.

TABLE 1
Conditional probabilities for the QTL genotype given the genotypes at two flanking markers

marker genotype		QTL genotype	
left	right	HL	LL
HL	HL	$(1-r_L)(1-r_R)/(1-r)$	$r_L r_R/(1-r)$
HL	LL	$(1-r_L)r_R/r$	$r_L(1-r_R)/r$
LL	HL	$r_L(1-r_R)/r$	$(1-r_L)r_R/r$
LL	LL	$r_L r_R/(1-r)$	$(1-r_L)(1-r_R)/(1-r)$

For each of the four sets of marker genotypes, we can now write down the conditional phenotype density, which has the form of a mixture of two normal distributions, similar to those seen in Section 2.1.2. Thus we can obtain the likelihood for our four parameters, $(\mu_H, \mu_L, \sigma, r_L)$.

Lander and Botstein (1989) proposed to maximize this likelihood, for fixed r_L, using the EM algorithm (Dempster et al. 1977). They then calculated the LOD score, which is the log (base 10) likelihood ratio comparing the hypothesis of a single QTL at the current locus (i.e., the current value of r_L) to the null hypothesis of no segregating QTLs (meaning that the individuals' phenotypes follow a normal(μ, σ^2) distribution). The two likelihoods in this ratio must be maximized over their respective parameters.

The procedure outlined above is performed for each locus in the genome. The likelihood under the null hypothesis is calculated just once. The likelihood for the hypothesis of a single QTL must be calculated at each locus in the genome (or, really, just every 1 cM or so), and so the EM algorithm must be performed at each locus.

The LOD score is then plotted against genome location, and is compared to a genome-wide threshold. Whenever the LOD curve exceeds the threshold, we infer the presence of a QTL. The point at which the LOD is maximized is used as the estimate of the QTL location. A one- or two-LOD support interval, the region around the inferred QTL in which the LOD score is within one or two of its maximum, is used as an interval estimate for QTL location.

The genome-wide threshold, used to indicate the significance of a peak in the LOD curve, is obtained by finding the 95th percentile of the maximum LOD score, across the entire genome, under the null hypothesis of no segregating QTLs.

Figure 1 gives an example of a LOD curve calculated using interval mapping. We simulated 200 backcross progeny, having a single chromosome of length 100 cM with 11 equally spaced markers, using a model with a single QTL located 35 cM from the left of the chromosome. The effect of the QTL (the difference between the means for HL versus LL individuals) was 0.75σ, giving a heritability, the proportion of the total phenotypic variance due to the QTL, of 0.36. The dots plotted on the curve point out the locations of the marker loci. Using a LOD

threshold of 2.5, the observed peak is significant. The inferred QTL is estimated to be at 37 cM, with a maximum LOD score of 3.4. The one-LOD support interval covers the region from 27 cM to 47 cM, which does indeed include the actual location of the simulated QTL.

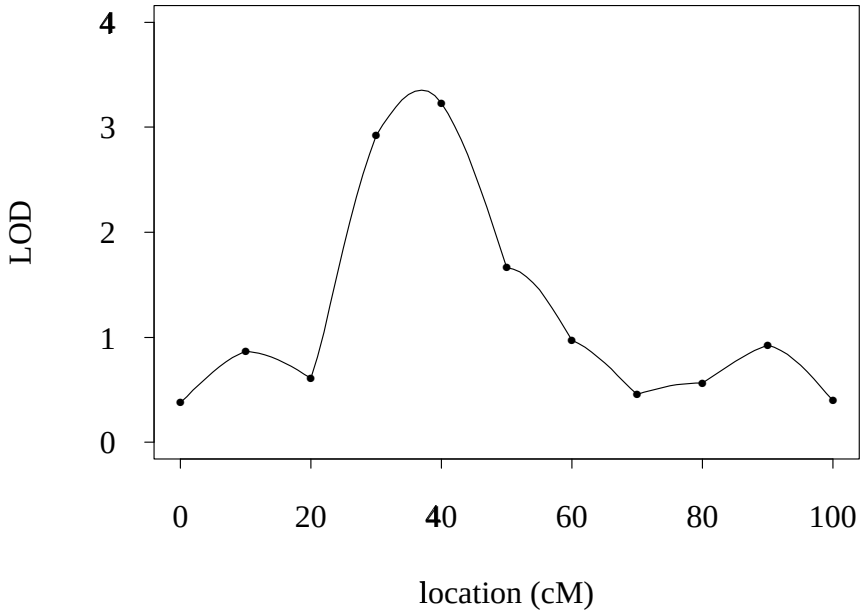

FIG. 1. *An example LOD curve, calculated using interval mapping, for some simulated data.*

A great deal of effort has been expended in trying to understand the appropriate LOD threshold to use. Lander and Botstein (1989) performed simulations to estimate the threshold for various different genome sizes and marker densities. They gave analytical calculations for the case of a very dense marker map. These guidelines should suffice for most uses. If one is concerned, additional simulations, conforming to the particular case under study, can be performed quite easily, or one can use a permutation test (Churchill and Doerge 1994), which has the advantage of avoiding the assumption of normally distributed environmental variation.

A number of studies have assessed the performance of interval mapping in comparison to ANOVA (van Ooijen 1992; Knott and Haley 1992; Darvasi et al. 1993; Rebaï et al. 1995; Hyne et al. 1995). The chief benefit of interval mapping is that it gives more precise estimates of the location and effect of a QTL. It does not give an appreciable increase in the power for detecting QTLs, and it requires a great deal more computational effort than does single marker ANOVA.

Hyne et al. (1995) stated that when a QTL is located very near one end of a linkage group, its estimated location, as given by interval mapping, will be

biased, since if one looks for QTLs only within the two extreme markers on the linkage group, its estimated location will never be outside of the last marker. It is possible to extend the LOD curves beyond the most extreme markers, however; outside of these markers, we can use the single marker maximum likelihood method, described in the previous section. Doing this should eliminate the bias problem. (Of course, a slight increase in variance, and a slight decrease in power, will accompany this approach.)

Look again at Figure 1. The dots on the LOD curve are at the marker loci. At these points, interval mapping is really just ANOVA, since the genotypes there are known exactly. If we performed only ANOVA, we would get exactly those points on the LOD curve. Interval mapping links these points together, and indicates that the best estimate for the QTL position is at 37 cM. But the markers at both 30 and 40 cM are within the one-LOD support interval.

2.1.4. *Regression mapping.* Knapp et al. (1990), Haley and Knott (1992), and Martínez and Curnow (1992) independently developed a method which approximates interval mapping very well, but requires much less computation. The method has come to be called "regression mapping." The presentation in Haley and Knott (1992) is by far the best.

Consider again the model of the previous section, with two markers separated by a recombination fraction r, and a putative QTL located between them, at a recombination fraction r_L from the left marker. The conditional expected value of the phenotype for an individual, given its genotypes at the flanking markers, is

$$\mathrm{E}(y|\text{marker gen.}) = \mu_L + (\mu_H - \mu_L)\,\Pr(\text{QTL gen. is HL}|\text{marker gen.}),$$

where Pr(QTL gen. is HL|marker gen.) is as shown in Table 1.

In regression mapping, we regress the individuals' phenotypes on their conditional probabilities for having the genotype HL at the putative QTL, given their marker genotypes. The log likelihood is calculated assuming that

$$y|\text{marker gen.} \sim \text{normal}(\tilde{y}, \sigma^2)$$

where $\tilde{y} = \mathrm{E}(y|\text{marker gen.})$. This gives the LOD score

$$\mathrm{LOD} = \frac{n}{2}\log_{10}\left(\frac{\mathrm{RSS}_0}{\mathrm{RSS}}\right)$$

where n is the number of progeny, RSS is the residual sum of squares from the above regression, $\sum_i (y_i - \hat{y}_i)^2$, and RSS_0 is the residual sum of squares under the null hypothesis of no segregating QTLs, $\sum_i (y_i - \bar{y})^2$.

Like interval mapping, the LOD score is calculated at each locus in the genome, but here, we need only calculate a single regression at each locus, rather than perform the EM algorithm at each locus, which requires a number of iterations, each containing a regression. Thus, there is a great savings in computation time. Also, because regression mapping requires only simple regression calculations, it is much easier to include additional effects into the analysis, such as sex or treatment effects. This may translate into large increases in performance.

Figure 2 displays the difference between the LOD curves calculated by regression mapping and interval mapping, for the data used in the previous section. The difference between the two curves is very subtle, being less than 0.1 in absolute value. Regression mapping gives results every bit as good as interval mapping, with a great deal less computation.

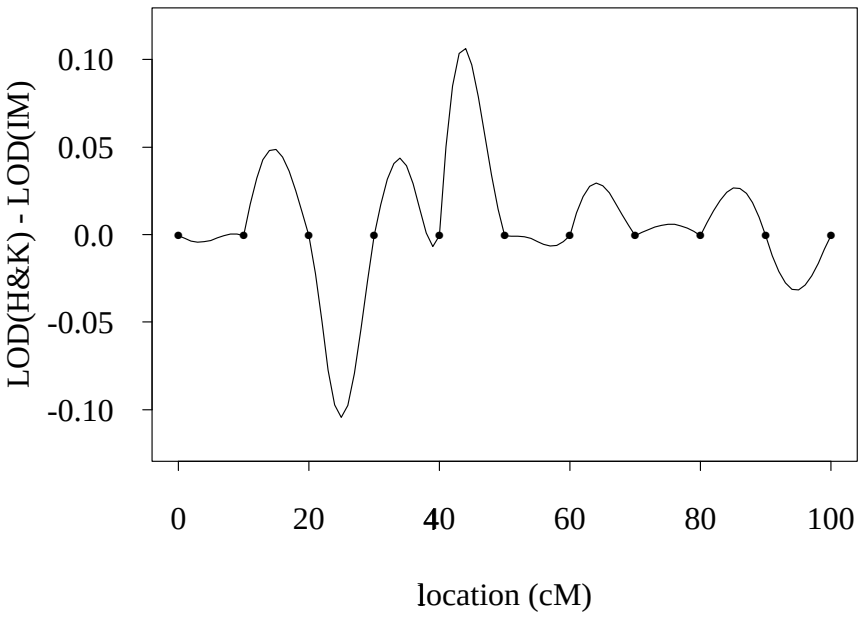

FIG. 2. *The difference between the LOD curves calculated using regression mapping and interval mapping for some simulated data.*

2.1.5. *Marker regression.* Kearsey and Hyne (1994) and Wu and Li (1994) independently developed a further method, which seems to approximate interval mapping quite well, with less intensive computation. But this method, which Kearsey and Hyne call "marker regression," seems more awkward and less adaptable than Haley and Knott's "regression mapping," and has not been shown to provide any further benefits.

Consider a linkage group with M markers, and fix the location for a putative QTL. Let r_j be the recombination fraction between the QTL and the jth marker. Group the individuals according to whether they have genotype HL or LL at marker j. Let $\hat{\beta}_j$ be the difference between the phenotype means for these two groups. As shown in Section 2.1.1,

$$\mathrm{E}(\hat{\beta}_j) = \beta(1 - 2r_j),$$

where $\beta = \mu_H - \mu_L$, the effect of the QTL.

Kearsey and Hyne (1994) suggest regressing the $\hat{\beta}_j$ for the M markers on the values $(1-2r_j)$, without an intercept. This is performed for each locus on the linkage group; we seek the locus giving the minimum residual sum of squares in this regression.

Wu and Li (1994) point out that the $\hat{\beta}_j$ do not have constant variance. The variance of $\hat{\beta}_j$ is approximately $4[\sigma^2 + r_j(1-r_j)\beta^2]/n$, where n is the number of progeny, and σ^2 is the environmental variance. They suggest using weighted least squares, using weights inversely proportional to the variances of the $\hat{\beta}_j$. But since σ and β are not known, it is not clear how to do this, unless one were to use a form of iteratively re-weighted least squares.

Wu and Li (1996) further point out that the $\hat{\beta}_j$ are correlated, and recommend using general least squares using an estimate of the covariance matrix.

We applied the method of Kearsey and Hyne (1994) to the simulated data analyzed in Sections 2.1.3 and 2.1.4. Figure 3 displays the residual sum of squares curve. The minimum is realized at 42 cM.

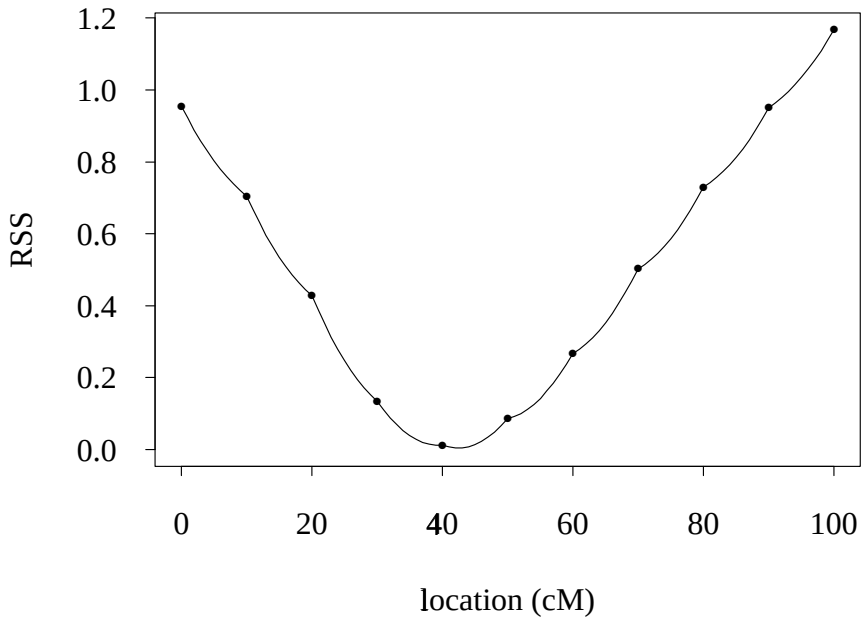

FIG. 3. *The residual sum of squares curve calculated using the marker regression method for some simulated data.*

Kearsey and Hyne (1994) gave a small number of simulations which suggested that marker regression performs as well as interval mapping. But they have not made a case for real improvements, aside from ease of computation. The method seems to have no advantages over regression mapping.

2.2. *Multiple QTL methods.* Recent efforts in developing methods to identify QTLs have focused on multiple QTL methods. There are three principal reasons for modelling multiple QTLs: to increase sensitivity, to separate linked QTLs, and to estimate epistatic effects (i.e., interactions between alleles at different QTLs).

When several QTLs are modelled, one can control for much of the genetic variation in a cross, and thus individual QTLs can be more clearly seen. In contrast, when one models a single QTL at a time, the genetic variation due to other segregating QTLs is incorporated into the "environmental" variation. When two QTLs are linked, single QTL methods, such as interval mapping, often view them as a single QTL. Searches which allow multiple QTLs do a better job of separating the two loci, and identifying them as distinct. The presence of epistasis can only be detected and estimated using models which include multiple QTLs. Incorporating epistatic effects into multiple QTL models will be very difficult, however. If one were to include all possible pairwise interactions, the number of parameters in the model would quickly explode. The methods discussed here all neglect the possibility of epistasis.

In this section, we discuss four important methods which explicitly consider multiple QTLs: multiple regression, interval mapping type methods using either forward selection or multi-dimensional searches, composite interval mapping (also called MQM mapping), and Markov chain Monte Carlo using a full Bayesian model.

2.2.1. *Multiple regression.* The obvious extension of analysis of variance is multiple regression. We attempt to form a model which includes a number of different marker loci, rather than looking at the markers one at a time. Let M be the number of markers, let $x_{ij} = 1$ or 0, according to whether individual i had genotype HL or LL at the jth marker, and let y_i be the phenotype for individual i. We write

$$\mathrm{E}(y_i|x_i) = \mu + \sum_{j=1}^{M} \beta_j x_{ij}$$

where $x_i = (x_{ij})$. We presume that most of the markers have $\beta_j = 0$. We seek the set of markers, S, with non-zero coefficients, β_j, so that

$$\mathrm{E}(y_i|x_i) = \mu + \sum_{j \in S} \beta_j x_{ij}.$$

The markers in S are indicated to be near QTLs.

There are two problems associated with this method. First, we must find a way to search through the set of possible models, in order to seek good ones. In an experiment with 100 genetic markers, there are $2^{100} \approx 10^{30}$ possible models to consider; it will be impossible to look at each of them. Second, we must form a criterion for choosing from these models. For models that include the same number of markers, one generally picks the one with the smallest residual sum of squares. The difficulty is in choosing between models of different sizes: what

change in the residual sum of squares must we see before we will accept an additional marker into the model?

Cowen (1989) discussed using stepwise selection and backward deletion, and using Mallows' C_p and the adjusted-R^2 criteria, when using multiple regression to identify QTLs. More recently, Doerge and Churchill (1996) described using forward selection, with permutation tests to determine the appropriate size of the model. Wright and Mowers (1994) and Whittaker et al. (1996) described the relationship between the partial regression coefficients, obtained by regressing a trait on a set of marker loci, and the locations and effects of a set of QTLs, but they did not provide a procedure for using this information to identify QTLs.

Broman (1997) discussed the use of model selection procedures in regression to identify QTLs. A number of different methods of searching through the space of models were compared: forward selection, backward selection, and Markov chain Monte Carlo (MCMC). Forward selection was found to perform as well as the other search methods, while it requires much less extensive calculations. Further discussion focussed on the criteria for choosing a model. The usual approaches to model selection focus on minimizing prediction error, and, as a result, standard criteria for choosing models, such as Mallows' C_p and adjusted-R^2, tend to choose models with a large number of extraneous variables. With some modification, the Schwartz's BIC (Schwartz 1978) performs much better. This criterion has the form BIC-$\delta = \log \mathrm{RSS} + q\delta \log n/n$, where RSS is the residual sum of squares for the model, q is the number of markers in the model, and n is the number of progeny. With this method, one attempts to find the model which minimizes the above criterion. The parameter δ must be chosen to balance the error of missing important QTLs with the error of including too many extraneous markers; a value between 2 and 3 may be appropriate in many situations.

2.2.2. *Interval mapping revisited.* Lander and Botstein (1989) briefly mentioned a method for distinguishing linked loci. If, when performing interval mapping, the LOD curve for a linkage group shows two peaks, or a single very broad peak, Lander and Botstein recommended to fix the position of one QTL at the location of the maximum LOD, and then search for a second QTL on that linkage group. In the model selection literature, this method is generally called forward selection (Miller 1990). Though some authors (Haley and Knott 1992; Satagopan et al. 1996) have interpreted this method as applying interval mapping to the residuals from the best fit of one QTL, it is best to estimate the effects of both QTLs simultaneously, using the original data (cf Dupuis et al. 1995).

We fix the location of the first QTL, and vary the location of the second QTL along the linkage group. At each location for the second QTL, we calculate a LOD score, comparing the maximum likelihood under the hypothesis of two QTLs at these locations, to that with a single QTL, located where the first QTL was placed. Each individual's contribution to the likelihood has the form of a mixture of four normal distributions, the four components corresponding to the four possible QTL genotypes. The EM algorithm can again be used to obtain the maximum likelihood estimates and the corresponding LOD score. (One could also apply the "regression mapping" method.)

Several authors have criticized this method (Haley and Knott 1992; Martínez and Curnow 1992), pointing to the phenomenon of "ghost QTLs." When two or more QTLs are linked in coupling (meaning that their effects have the same sign), interval mapping often gives a maximum LOD score at a location in between the two QTLs.

Consider, for example, a 60 cM segment of a chromosome, with four equally spaced markers (20 cM spacing). Consider a backcross with QTLs located at 15 and 45 cM, acting additively and having equal additive effect 0.5σ. The solid line in Figure 4 gives the expected LOD (ELOD) curve for this situation, when using 200 progeny. (Since there is no closed-form expression for the ELOD curve, it was estimated by performing 1000 simulations of the above situation and averaging the LOD curves obtained. We also used the fact that the ELOD curve is symmetric about the 30 cM point, and so averaged the pairs of points on the curve which are symmetric about 30 cM.) Note that the ELOD curve is maximized at 30 cM, even though the simulated QTLs were at 15 and 45 cM. This gives rise to the term "ghost QTL." Forward selection here would give bad results. We would generally pinpoint the first QTL at around 30 cM, and then search for a second QTL, and so would be completely mistaken.

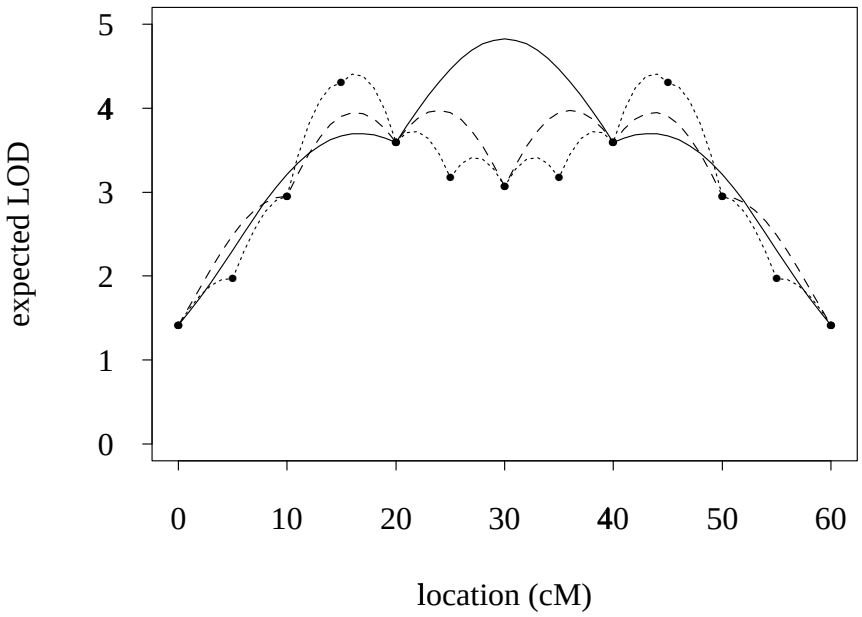

FIG. 4. *Expected LOD curves, with two QTLs located at 15 and 45 cM. The solid line, dashed line and dotted line correspond to using equally-spaced markers at spacings of 20, 10 and 5 cM, respectively.*

But this "ghost QTL" problem turns out to be an artifact of interval mapping.

The dashed and dotted lines in Figure 4 are the ELOD curves for the above example, using marker spacings of 10 and 5 cM, respectively. When the markers are more tightly spaced, the ghost QTL disappears. The ELOD curves are not maximized exactly at the true QTL locations, but things do get better as marker density increases. Note that if one considered only the marker loci, one would not be so misled. The marker loci at which the LOD is maximized are those closest to the true QTLs. Similar observations were made by Whittaker et al. (1996) and Wright and Kong (1997).

As an alternative to forward selection, several authors have recommended performing a full two-dimensional search for QTLs (Haley and Knott 1992; Martínez and Curnow 1992; Hyne and Kearsey 1995; Whittaker et al. 1996; Wu and Li 1994, 1996). Instead of fixing the location of one QTL and then searching for an additional one, the locations of both QTLs are allowed to vary simultaneously. A great deal more computation must be performed. Extending this method to more than two QTLs, as recommended by Wu and Li (1996), is possible in principle, but the computation requirements would very quickly become prohibitive.

One problem that these authors have not discussed carefully is the question of when to add an additional QTL: how large an increase in LOD should we require before allowing an additional QTL? Such guidelines are necessary, if one is to use these methods in practice.

2.2.3. *Composite interval mapping and MQM mapping.* Jansen and Zeng independently developed a method which attempts to reduce the multi-dimensional search for identifying multiple QTLs to a one-dimensional search (Jansen 1993; Jansen and Stam 1994; Zeng 1993, 1994). This is done using a hybrid of interval mapping and multiple regression on marker genotypes. By including other markers (on the same chromosome and on different chromosomes) as regressors while doing interval mapping, one hopes to control for the effects of QTLs in other intervals, so that there will be greater power in detecting a QTL, and so that the effects of the QTLs will be estimated more precisely. Jansen called the method MQM mapping (short for "marker-QTL-marker" or "multiple QTL models"); Zeng called it composite interval mapping.

The method is performed as follows. We choose a set of markers, S, to control for background genetic variation. Then, we perform a genome scan, as in interval mapping. At each locus in the genome, we hypothesize the presence of a QTL, and we write

$$y \sim \text{normal}(\mu + \beta z + \sum_{j \in S^\star} \beta_j x_j, \sigma^2),$$

where y is the phenotype, $z = 1$ or 0, according to the whether the genotype at the putative QTL is HL or LL, $x_j = 1$ or 0, according to whether the genotype at the jth marker is HL or LL, and $S^\star$ is a subset of our set of markers, S, where we exclude any markers that are within, say, 10 cM of the putative QTL. Under this model, the contribution of each individual to the likelihood has the form of a mixture of two normal distributions with means $\mu + \beta + \sum_{j \in S^\star} \beta_j x_j$ and $\mu + \sum_{j \in S^\star} \beta_j x_j$, with mixing proportions equal to the conditional probabilities

of the individual having QTL genotype HL and LL, given its marker genotypes. The EM algorithm, or a variant called the ECM algorithm (Meng and Rubin 1993), can be used to maximize the likelihood function.

As in interval mapping, at each locus, a likelihood ratio or LOD score is calculated, comparing the likelihood assuming that there is a QTL at that locus, to the likelihood assuming that there is not a QTL there, in which case we imagine that all progeny have phenotypes which are normally distributed with mean $\mu + \sum_{j \in S^*} \beta_j x_j$ and variance σ^2. The LOD score is plotted as a function of genome position, and is compared to a genome-wide threshold. As in interval mapping, areas of the genome for which the LOD curve exceeds the threshold are said to contain a QTL.

The genome-wide threshold is obtained by considering the distribution of the maximum LOD score under the hypothesis of no segregating QTLs anywhere in the genome. This distribution should take into account the selection of the set of marker regressors, S. The distribution can be estimated by simulating a set of data under the hypothesis of no segregating QTLs, performing the entire procedure, and calculating the maximum LOD curve obtained, and then repeating the process a number of times. The 95th percentile of these maximum LOD scores is used as the threshold.

The key problem in this method is the choice of which markers to use as regressors: using too many markers will increase the variance of the LOD score, and thus will decrease the power for detecting QTLs. Jansen (1993) and Jansen and Stam (1994) used backward deletion, with Akaike's Information Criterion (AIC) (Akaike 1969) or a slight variant, to pick the subset of markers. Zeng (1994) recommended using either all markers, dropping those within 10 cM of the putative QTL, or using all markers that are not linked to the putative QTL. Basten et al. (1996), in a manual for the program QTL Cartographer, recommended using forward selection up to a fixed number of markers, say five, and then dropping any markers that are within 10 cM of the putative QTL.

We have found that the methods that Zeng (1994) originally recommended, using all markers or all markers not linked to the putative QTL, work very badly. Including so many markers increases the corresponding LOD threshold to such a large value that power is reduced to almost zero. Only QTLs with extremely large effect will be found by this method.

The performance of the other methods for choosing the set of marker regressors depends on how many markers are chosen. And once we have found a way to choose this set, the task of identifying QTLs is essentially done: the best set of markers to use is exactly the set of markers which are closest to the underlying QTLs. In Section 3, we present some simulation studies which assess the performance of these methods.

2.2.4. *Markov chain Monte Carlo.* Satagopan et al. (1996) have applied the Markov chain Monte Carlo (MCMC) method to the problem of identifying QTLs. MCMC is a very popular approach to solving very complex statistical problems, especially those which include a large amount of missing information. Gelman et al. (1995) gives a very good introduction to the subject.

Consider again a backcross. Satagopan et al. (1996) consider a single linkage group. (The method can be extended to several linkage groups in a straightforward way.) Consider n progeny. Let y_i be the phenotype for individual i. Suppose there are M markers, at locations $D = (D_1, D_2, \ldots, D_M)$, in cM, from the left end of the linkage group. Let $x_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the jth marker.

Let S be the number of segregating QTLs, and let $\lambda = (\lambda_1, \ldots, \lambda_S)$ be their locations, in cM, from the left end of the linkage group. Let $z_{ij} = 1$ or 0, according to whether individual i has genotype HL or LL at the jth QTL. Let β_j be the effect of the jth QTL, and assume that the environmental variation is normally distributed, with variance σ^2. Let μ be the mean of individuals for whom $z_{ij} = 0$ for all j.

As shorthand, we will write $y = (y_1, \ldots, y_n)$, $x_i = (x_{i1}, \ldots, x_{iM})$, $x = (x_1, \ldots, x_n)$, and similarly for z_i, z and β. Also, let $\theta = (\mu, \beta, \sigma)$.

We have

$$y_i | z_i, \theta \sim \text{normal}(\mu + \sum_{j=1}^{S} \beta_j z_{ij}, \sigma^2).$$

This gives the likelihood

$$L(\lambda, \theta | y, x, D) = \prod_{i=1}^{n} \sum_{q} f(y_i | z_i = q, \theta) \Pr(z_i = q | \lambda, x_i, D)$$

where the sum over q is over the 2^S possible QTL genotypes for individual i and where f is the conditional (normal) density for y.

Satagopan et al. (1996) use a full Bayesian framework, meaning that they assign a prior probability distribution to the unknown parameters (λ, θ), say $p(\lambda, \theta)$, and then look at their posterior distribution, given the data, $p(\lambda, \theta | y, x, D)$.

The goal of the MCMC method is thus to estimate the posterior distribution of the unknown parameters. This is done by creating a Markov chain whose stationary distribution is the desired posterior distribution.

Simulating from this chain gives a sequence $(\lambda_0, \theta_0), (\lambda_1, \theta_1), \ldots (\lambda_N, \theta_N)$. Estimates of the desired parameters, such as the QTL effects, β_j, are obtained by averaging over these samples. Interval estimates for the QTL locations can be obtained by looking at the smallest intervals which contain, say, 95% of the samples.

In order to determine the number of QTLs, S, Satagopan et al. (1996) run separate chains for different values of S, and use Bayes factors. In brief, for each value of S, they use their samples to estimate the probability of the data given the model, $p(y, x | S)$. They estimate the number of QTLs to be the value of S for which this estimated probability is large. If one were willing to give a prior on the number of QTLs, say $\Pr(S = s)$, the posterior distribution for S could be calculated

$$\Pr(S = s | y, x) = \frac{p(y, x | S = s) \Pr(S = s)}{\sum_s p(y, x | S = s) \Pr(S = s)}$$

The estimated number of QTLs would then simply be the value of S with the largest posterior probability.

A later report (Satagopan and Yandell 1996), using an idea developed by Green (1995), describes how to allow the unknown number of QTLs, S, to be included as an unknown parameter, so that a single Markov chain can be used to estimate S along with the other parameters. Doing this requires placing a prior distribution on the number of QTLs.

We have skipped all of the details of the MCMC method. The difficulties in applying this approach are entirely in those details. First, you need to create a Markov chain which has your posterior distribution as its stationary distribution. There are a number of standard ways to do this, such as the Gibbs sampler (Geman and Geman 1984) and the Metropolis-Hastings algorithm (Hastings 1970). The most important characteristic in the chain is that it mixes well: that it moves around the parameter space rather easily, and that it very quickly reaches its stationary distribution. Forming good Markov chains, and monitoring their behavior, is a delicate and sophisticated art.

The other important problem is in the determination of the number of QTLs. Whether we assign a prior to the unknown number of QTLs or use Bayes factors, we must make choices which balance the problem of missing real QTLs with that of including extraneous loci.

3. Simulations. In this section, we present the results of a small simulation study aimed at comparing several different methods for identifying QTLs. Our focus is on identifying QTLs, and so we look only at whether the methods detect the simulated QTLs, and not at the estimated effects and the precision with which the location is estimated. Simulations are necessary, because the methods for identifying QTLs are too complex to be assessed by analytical means, at least in the situations in which they would be used in practice.

Most authors have used simulations to demonstrate their methods for finding QTLs. Many have presented the results of applying their method to a single data set (Jansen 1993; Knapp 1991; Lander and Botstein 1989; Zeng 1994), a practice which precludes a true assessment of the method's performance. Others consider only very simple situations, such as simulating only one or two chromosomes with one or two segregating QTLs (Haley and Knott 1992; Kearsey and Hyne 1994). In practice, most QTL studies involve a search over ten or more chromosomes, and very often there is evidence for at least a moderate number of segregating QTLs (from three or four to as many as a dozen). A method's ability to detect QTLs in simulation studies which use very limited searches and in which only a small number of QTLs are allowed will say little about its performance in the more complex situations where the method is anticipated to be used.

Also missing from the literature is a careful comparison of the performance of the many methods available for identifying QTLs. It is surprising that such comparisons are not a routine part of the presentation of a new method. Before dropping a simple approach in favor of a more complex one, we should have evidence that the complexities of the new approach will be accompanied by a

real improvement in performance.

We compared four different methods for identifying QTLs: analysis of variance (ANOVA) at the marker loci, the method of Zeng (1994), forward selection using a BIC-type criterion, and forward selection using a permutation test at each stage (Doerge and Churchill 1994). These methods were described in Section 2.

Interval mapping (IM) was ignored, because it provides no improvement over simple ANOVA when using a relatively dense marker map (10 cM spacing or less) and a small or moderate number of progeny (500 or less), at least when it comes to identifying QTLs. This can easily be seen when inspecting the one- or two-LOD support intervals which accompany any application of IM: they invariably span several markers. The benefit of IM is in providing more precise estimates of QTL location and effects.

For Zeng's method, we used forward selection up to either 3, 5, 7 or 9 markers to obtain the set of regressors, and limited the search for QTLs to marker loci. With ANOVA and Zeng's method, we obtained genome-wide thresholds by performing 1000 simulations under the hypothesis of no segregating QTLs: the estimated threshold was the 95th percentile of the maximum LOD score across all markers. In addition, for these two methods, we required that the LOD dropped by at least 2.2 in base 10 (corresponding to 5 in base e) between "peaks" before we declared that two QTLs were identified. This value was obtained empirically (in other words, by trial and error).

The BIC-type criterion used is $\log \mathrm{RSS} + \delta q \log n/n$, where RSS is the residual sum of squares, n is the number of progeny, q is the number of markers in the model, and δ is either 2, 2.5 or 3. We use BIC-2, BIC-2.5 and BIC-3 to identify these criteria. For the permutation method, at each stage we used the 95th percentile of 500 permutations to determine whether to add another marker.

In the study described in this section, we simulated 250 backcross progeny, obtained from inbred lines, with nine chromosomes, each of length 100 cM and having 11 equally spaced markers per chromosome (thus at a 10 cM spacing). The recombination process was assumed to exhibit no interference. The environmental variation followed a normal distribution with standard deviation $\sigma = 1$.

We modelled three QTLs with equal additive effect 0.5. One QTL was located at the center of chromosome 1, and two QTLs were located on chromosome 2 at 30 and 70 cM. The linked QTLs were either in coupling (effects of equal sign) or repulsion (effects of opposite sign). The QTLs were assumed to act additively. The heritability for these models (defined as the ratio of the genetic variance to the total phenotypic variance) was 0.20 and 0.12 when the linked QTLs were in coupling and repulsion, respectively. Note that all QTLs were located exactly at marker loci.

For each QTL model we performed 1000 simulations. The result of the application of each method was a set of marker loci indicated to be at or near QTLs. In assessing the results, we defined a chosen marker to be correctly identifying a QTL if it was within 20 cM of a QTL; otherwise it was deemed incorrect. If more than one chosen marker were within 20 cM of the same QTL, one was called correct and the others were called incorrect.

The estimated genome-wide LOD (base 10) thresholds for ANOVA and Zeng's method (using forward selection up to 3, 5, 7 and 9 markers) are displayed in Table 2. The estimated standard errors for the thresholds, obtained using a bootstrap (Venables and Ripley 1994), are approximately 0.1. For ANOVA, the threshold corresponded closely to the threshold in Figure 4 of Lander and Botstein (1989). For Zeng's method, the threshold increased with the number of regressors used.

TABLE 2

Estimated genome-wide LOD thresholds for a backcross with 250 progeny and nine 100 cM chromosomes each containing 11 equally-spaced markers

	Zeng			
ANOVA	3	5	7	9
2.5	3.3	3.6	3.8	4.0

In Table 3, we display the joint distribution, across the 1000 simulations, of the numbers of correctly and incorrectly chosen markers for the case of three QTLs with two QTLs linked in coupling, and using 250 progeny. The four columns labelled "Zeng" correspond to Zeng's method using forward selection up to either 3, 5, 7 or 9 markers. The three columns labelled "BIC" correspond to forward selection using the BIC-2, BIC-2.5 and BIC-3 criteria. The column "permu" gives the results for using forward selection with a permutation test at each stage. The second-to-last row in the table includes all simulations with two or more incorrectly chosen markers. The last row in the table gives the number of simulations in which at least one incorrect marker was chosen.

TABLE 3

Distribution of the numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing three QTLs with two QTLs linked in coupling, and using 250 progeny

#	#		Zeng				BIC			
cor	incor	ANOVA	3	5	7	9	2	2.5	3	permu
3	0	69	31	25	19	13	180	65	19	133
2	0	526	412	315	240	199	509	496	395	539
1	0	332	429	443	430	421	199	395	554	246
0	0	1	97	184	281	334	0	2	9	0
3	1	4	0	0	0	0	7	0	0	6
2	1	26	6	7	5	6	59	13	5	37
1	1	40	18	12	18	13	35	28	17	34
0	1	0	6	11	5	12	0	0	0	0
other		2	1	3	2	2	11	1	1	5
≥ 1 wrong		72	31	33	30	33	112	42	23	82

ANOVA nearly always found at least one QTL, and often found two, but it had

difficulty in separating the two linked QTLs. ANOVA added incorrect markers about 7% of the time. Zeng's method did worse than ANOVA in this situation. It suffered from low power for detection, and the power decreased sharply as the number of markers used as regressors increased; using three markers as regressors worked best in this case. Forward selection using BIC-2 did a better job of detecting the QTLs, but included incorrect markers 11% of the time—much more often than the other methods. The use of a larger multiplier helped to avoid this problem, but at the expense of a lower power for detection. Forward selection using a permutation test did well: it detected more QTLs than ANOVA and Zeng's method, while including incorrect markers only 8% of the time.

Table 4 shows which of the QTLs were correctly identified by the different methods. The first three columns, labelled "model," correspond to the three QTLs: first the QTL on chromosome 1, and then the two linked QTLs on chromosome 2. A one in these columns indicates that the QTL was correctly identified; a zero indicates that it was not found. Note that in this table, we ignore the markers which were incorrectly identified. For example, in the column labelled "ANOVA," the model "1 1 1" was identified 73 times out of 1000 simulations; this includes 69 times in which no extraneous markers were included, and 4 times in which one extraneous marker was included (see Table 3).

TABLE 4

Models identified in 1000 simulations of the model containing three QTLs with two QTLs (represented in the second and third columns) linked in coupling, and using 250 progeny

		Zeng				BIC			
model	ANOVA	3	5	7	9	2	2.5	3	permu
111	73	31	25	19	13	187	65	19	139
110	254	189	140	102	83	258	239	189	260
101	257	191	144	112	92	264	241	192	274
011	41	39	40	33	31	52	30	20	45
100	3	115	173	182	180	1	3	8	1
010	200	161	136	131	120	131	218	293	152
001	171	171	147	135	135	107	202	270	129
000	1	103	195	286	346	0	2	9	0

When forward selection and ANOVA identified just one QTL, it was almost always one of the two linked QTLs, but Zeng's method often picked only the QTL on chromosome 1. When two QTLs were identified, all of the methods tended to pick the QTL on chromosome 1 and one of the two linked QTLs. Note that the two linked QTLs on chromosome 2 were chosen at approximately equal frequencies by all of the methods, as expected by symmetry: the models "1 1 0" and "1 0 1" were chosen nearly the same number of times, as were the models "0 1 0" and "0 0 1."

Table 5 displays the joint distribution, across the 1000 simulations, of the numbers of correctly and incorrectly chosen markers when the linked QTLs are

in repulsion.

The methods did not perform as well when the linked QTLs were in repulsion; ANOVA and forward selection suffered much more than Zeng's method. The number of incorrectly chosen markers showed little change from the case of coupling, for all of the methods. But the number of correctly identified QTLs, in comparison to coupling, was halved for ANOVA and forward selection. Zeng's method, on the other hand, showed very little change in its ability to identify QTLs, with the result that here his method worked better than ANOVA.

TABLE 5

Distribution of the numbers of correctly and incorrectly chosen markers in 1000 simulations of a model containing three QTLs with two QTLs linked in repulsion, and using 250 progeny

#	#		Zeng				BIC			
cor	incor	ANOVA	3	5	7	9	2	2.5	3	permu
3	0	4	102	83	60	46	174	80	27	78
2	0	123	222	225	203	168	135	79	46	99
1	0	572	402	412	385	381	426	458	398	524
0	0	231	230	242	314	372	156	338	507	219
3	1	0	0	1	2	2	25	5	0	6
2	1	7	7	6	10	10	21	6	2	12
1	1	30	19	20	19	13	29	12	4	28
0	1	32	18	10	4	7	24	22	16	31
other		1	0	1	3	1	10	0	0	3
≥ 1 wrong		70	44	38	38	33	109	45	22	80

It is interesting to see that whereas Zeng's approach performed quite poorly when the QTLs were linked in coupling, even in comparison to ANOVA, it performed somewhat better than all of the other methods when the QTLs were in repulsion. The reason that Zeng's method is more successful in teasing out a pair of QTLs linked in repulsion, may be that such QTLs look more important when both are included in the model. Zeng's method forces the fit of the larger model, whereas forward selection considers the markers one at a time. This difference is best illustrated in Table 4. When identifying just one QTL, ANOVA and forward selection generally pick one of the two linked QTLs, whereas Zeng's method picks from the three QTLs at nearly equal proportions.

Whereas the simulations presented here considered only cases with three QTLs, Broman (1997) also performed simulations with five QTLs; the results were similar.

4. Conclusions and discussion. Current methods for identifying QTLs focus on interval mapping: inferring the location of a QTL between marker loci. Yet interval mapping and its approximations have been shown to provide little improvement in power over simple ANOVA at the marker loci. When we dispense with interval mapping, we are left only with ANOVA and multiple regression; the use of these more simple methods for identifying QTLs has been neglected.

In addition, most current methods use multiple tests of hypotheses. The problem of identifying QTLs is best viewed as a problem in model selection. Having discarded interval mapping, we then seek to choose a set of marker loci which are at or near QTLs. The problem is not the standard one in model selection, where attention has been on minimizing prediction error. Still, the model selection literature has much to say about our current problem. Clearly, the single-QTL methods, such as ANOVA and interval mapping, will perform poorly when multiple linked QTLs are segregating in a cross. An appropriate approach is difficult to prescribe. In the simulations in Section 3, one method (forward selection) performed best in the case of QTLs linked in coupling, while another (Zeng's approach) performed best in the case of QTLs linked in repulsion. A more refined method, such as Markov chain Monte Carlo, does not necessarily lead to improved results. For example, for the data in Satagopan et al. (1996), interval mapping seemed to give nearly identical results to MCMC.

A number of decisions must be made when performing any model selection procedure. First, one must choose a criterion. For Zeng's method, one must choose how many variables to use as initial regressors; for the BIC-δ criterion, one must choose the value of the parameter δ. Second, one must decide how to search through the space of models: will forward selection suffice, or would a more extensive search, as provided by MCMC, give improved results? The choices that one makes will depend upon the experiment being performed: whether it is a backcross or an intercross, the number of progeny, the density of marker loci, the underlying genetic structure of the trait, and the ultimate goal of the experiment. When making these choices, one will need to perform multiple simulation experiments, using scenarios that seem reasonable, and using criteria for determining the performance of an approach which correspond to the goals of the study.

Simulation studies of the kind we mention can be helpful for designing a strategy, but, after the data are obtained, further analysis must be carried out. We see the need for research on the use of resampling and bootstrap methods of analysis, to complement the randomization approach of Churchill and Doerge (1994), which focuses on null models.

REFERENCES

AKAIKE, H. (1969) Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics* **21** 243–247.

BASTEN, C. J., WEIR, B. S. and ZENG, Z.-B. (1996) *QTL Cartographer: A reference manual and tutorial for QTL mapping.* Program in Statistical Genetics, Department of Statistics, North Carolina State University.

BROMAN, K. W. (1997) Identifying quantitative trait loci in experimental crosses. PhD dissertation, Department of Statistics, University of California, Berkeley.

CHURCHILL, G. A. and DOERGE, R. W. (1994) Empirical threshold values for quantitative trait mapping. *Genetics* **138** 963–971.

COWEN, N. M. (1989) Multiple linear regression analysis of RFLP data sets used in mapping QTLs. Pages 113–116 in *Development and application of molecular markers to prob-*

lems in plant genetics, edited by HELENTJARIS, T. and BURR, B. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

DARVASI, A., WEIREB, A., MINKE, V., WELLER, J. I. and SOLLER, M. (1993) Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics* **134** 943–951.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B **39** 1–38.

DOERGE, R. W. and CHURCHILL, G. A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142** 285–294.

DUPUIS, J., BROWN, P. O. and SIEGMUND, D. (1995) Statistical methods for linkage analysis of complex traits from high-resolution maps of identity by descent. *Genetics* **140** 843–856.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (1995) *Bayesian data analysis.* Chapman and Hall, New York.

GEMAN, S. and GEMAN, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6** 721–741.

GREEN, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82** 711–732.

HALDANE, J. B. S. (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8** 299–309.

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

HASTINGS, W. F. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

HYNE, V., KEARSEY, M. J., PIKE, D. J. and SNAPE, J. W. (1995) QTL analysis: unreliability and bias in estimation procedures. *Molecular Breeding* **1** 273–282.

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.

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

KEARSEY, M. J. and HYNE, V. (1994) QTL analysis: a simple 'marker-regression' approach. *Theoretical and Applied Genetics* **89** 698–702.

KNAPP, S. J. (1991) Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. *Theoretical and Applied Genetics* **81** 333–338.

KNAPP, S. J., BRIDGES, W. C., JR. and BIRKES, D. (1990) Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics* **79** 583–592.

KNOTT, S. A. and Haley, C. S. (1992) Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. *Genetics Research* **60** 139–151.

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

LONG, A. D., MULLANEY, S. L., REID, L. A., FRY, J. D., LANGLEY, C. H. and MACKAY, T. F. C. (1995) High resolution mapping of genetic factors affecting abdominal bristle number in *Drosophila melanogaster*. *Genetics* **139** 1273–1291.

MARTÍNEZ, O. and 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.

MATHER K. and JINKS, J. L. (1977) *Introduction to biometrical genetics.* Chapman and Hall, London.

Meng, X.-L. and Rubin, D. B. (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* **80** 267–278.

Miller, A. J. (1990) *Subset selection in regression.* Chapman and Hall, New York.

Rebaï, A., Goffinet, B. and Mangin, B. (1995) Comparing power of different methods for QTL detection. *Biometrics* **51** 87–99.

Satagopan, J. M. and Yandell, B. S. (1996) Estimating the number of quantitative trait loci via Bayesian model determination. Special contributed paper session on genetic analysis of quantitative traits and complex diseases, Biometrics section, Joint Statistical Meetings, Chicago, Illinois.

Satagopan, J. M., Yandell, B. S., Newton, M. A. and Osborn, T. C.(1996) A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144** 805–816.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.

Shrimpton, A. E. and Robertson, A. (1988) The isolation of polygenic factors controlling bristle score in *Drosophila melanogaster*. I. Allocation of third chromosome sternopleural bristle effects to chromosome sections. *Genetics* **118** 437–443.

Simpson, S. P. (1989) Detection of linkage between quantitative trait loci and restriction fragment length polymorphisms using inbred lines. *Theoretical and Applied Genetics* **77** 815–819.

Soller, M., Brody, T. and Genizi, A. (1976) On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and Applied Genetics* **47** 35–39.

Tanksley, S. D. (1993) Mapping polygenes. *Annual Review of Genetics* **27** 205–233.

van Ooijen, J. W. (1992) Accuracy of mapping quantitative trait loci in autogamous species. *Theoretical and Applied Genetics* **84** 803–811.

Venables, W. N. and Ripley, B. D. (1994) *Modern applied statistics with S-Plus.* Springer-Verlag, New York.

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 *Locypersicon* (tomato) with the aid of genetic markers using approximate maximum likelihood methods. *Heredity* **59** 413–421.

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

Wright, A. J. and Mowers, R. P. (1994) Multiple regression for molecular-marker, quantitative trait data from large F_2 populations. *Theoretical and Applied Genetics* **89** 305–312.

Wright, F. A. and Kong, A. (1997) Linkage mapping in experimental crosses: the robustness of single-gene models. Genetics **146** 417–425.

Wu, W.-R. and Li, W.-M. (1994) A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. *Theoretical and Applied Genetics* **89** 535–539.

Wu, W.-R. and Li, W.-M. (1996) Model fitting and model testing in the method of joint mapping of quantitative trait loci. *Theoretical and Applied Genetics* **92** 477–482.

Zeng, Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. USA* **90** 10972–10976.

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

Center for Medical Genetics
Marshfield Medical Research Foundation
1000 N Oak Av
Marshfield WI 54449
BromanK@cmg.mfldclin.edu

University of California
Department of Statistics
367 Evans Hall # 3860
Berkeley, California 94720–3860
speed@stat.berkeley.edu