J. R. Statist. Soc. B (2002)
64, *Part* 4, *pp.* 641–656

A model selection approach for the identification of quantitative trait loci in experimental crosses

Karl W. Broman

Johns Hopkins University, Baltimore, USA

and Terence P. Speed

University of California, Berkeley, USA, and Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section on '*Statistical modelling and analysis of genetic data*' *on Wednesday, May 22nd, 2002*, Professor D. Firth *and* Professor R. A. Bailey *in the Chair*]

Summary. We consider the problem of identifying the genetic loci (called quantitative trait loci (QTLs)) contributing to variation in a quantitative trait, with data on an experimental cross. A large number of different statistical approaches to this problem have been described; most make use of multiple tests of hypotheses, and many consider models allowing only a single QTL. We feel that the problem is best viewed as one of model selection. We discuss the use of model selection ideas to identify QTLs in experimental crosses. We focus on a back-cross experiment, with strictly additive QTLs, and concentrate on identifying QTLs, considering the estimation of their effects and precise locations of secondary importance. We present the results of a simulation study to compare the performances of the more prominent methods.

Keywords: Bayesian information criterion; Composite interval mapping; Markov chain Monte Carlo methods; Model selection; Quantitative trait loci; Regression

1. Introduction

The identification of the genetic loci that are responsible for variation in traits that are quantitative in nature (such as the yield from an agricultural crop, the number of abdominal bristles on a fruit-fly and the survival time of a mouse following an infection) is a problem of great importance to biologists. The number and effects of such loci (called quantitative trait loci (QTLs)) help us to understand the biochemical basis of these traits, and of their evolution in populations over time. Moreover, knowledge of these loci may aid in the design of selection experiments to improve these traits.

Repeated sibling mating (or, in plants, selfing) of experimental organisms has led to the establishment of panels of well-defined strains. The process of inbreeding has fixed a large number of biomedically (or agriculturally) relevant traits in these strains. If two strains, raised in a common environment, show consistent differences in a trait, we may be confident that the difference has a genetic basis. The genetic loci contributing to such a trait difference may be revealed by performing a series of experimental crosses, of which the simplest is the back-cross.

Address for correspondence: Karl W. Broman, Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, USA.
E-mail: kbroman@jhsph.edu

 1369–7412/02/64641

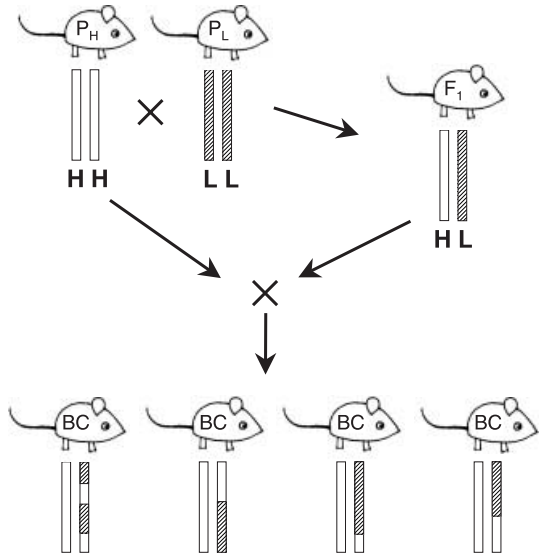

Fig. 1. A back-cross experiment begins with two inbred strains that differ in the trait of interest: the two strains are crossed to produce the F_1-generation, which is then crossed back to one of the parental strains to obtain the back-cross generation, the back-cross generation exhibits genetic variation

In a back-cross (Fig. 1), an investigator chooses two inbred strains that differ in the trait of interest (we shall call these the high (H) and low (L) parental strains). All individuals within an inbred strain are genetically identical and are homozygous at all loci. The two parental strains are crossed to form the first filial (F_1-) generation. The F_1-individuals are also genetically identical, and are heterozygous at loci at which the parental strains differ. The F_1-individuals are crossed to one of the two parental strains (e.g. the H-strain) to obtain the back-cross generation. The back-cross individuals receive one chromosome from the H-strain and one from the F_1. Thus, at each locus, they have genotype either HL or HH. The chromosome received from the F_1-parent is a mosaic of the two grandparental chromosomes, as a result of recombination during meiosis.

The investigator produces a number of back-cross progeny and determines the quantitative phenotype for each individual. Each individual is genotyped at a number of genetic markers (generally 100–300), chosen to cover the genome uniformly. At each marker and for each individual, it is observed whether the F_1-parent transmitted the H- or the L-allele. A genetic map for the marker loci will either be known or estimated on the basis of the current experiment. Such a map specifies the linear order of the marker loci along each chromosome and the distances between markers, measured in genetic distance. The genetic distance between two loci is d centimorgans (cM), if d is the average number of crossovers (points of exchange) in the intervening interval, in 100 products of meiosis.

The objective of the experiment is to identify the genomic regions for which there is an association between the phenotype of a back-cross individual and whether it received the H- or L-allele from the F_1-parent in the region. Although other experiments, such as the intercross, are more commonly used in practice, we focus on the back-cross for simplicity.

Consider a back-cross with n individuals. Let y_i denote the phenotype (trait value) of individual i, and let $x_{ij} = 1$ or $x_{ij} = 0$, according to whether individual i has genotype HH or HL respectively at marker j.

The locations of crossovers in meiosis are often modelled as a Poisson process (an assumption of *no crossover interference*). In this case, the x_{ij} for each chromosome form a Markov chain, with transition probabilities $\Pr(x_{i,j+1} = 1|x_{ij} = 0) = \Pr(x_{i,j+1} = 0|x_{ij} = 1) = r_j$, where r_j is called the *recombination fraction* between markers j and $j + 1$. We further assume that $\Pr(x_{ij} = 1) = \Pr(x_{ij} = 0) = \frac{1}{2}$, in accordance with Mendel's rules.

Imagine that there is a reasonably small number, p, of genetic loci (QTLs) that influence the trait. Let us temporarily suspend the index i for individuals and consider the relationship between an individual's genotypes at the QTLs and its phenotype (trait value). Let $z = (z_1, \ldots, z_p)$, with $z_j = 1$ or $z_j = 0$, according to whether the individual has genotype HH or HL respectively at the jth QTL. In principle, $E(y|z) = \mu_z$ and $\text{var}(y|z) = \sigma_z^2$ are arbitrary functions of z. Generally, we assume that the trait is homoscedastic—that the variance is constant within genotype groups, $\text{var}(y|z) = \sigma^2$. It is often further assumed that the residual variation is normally distributed, that $y|z \sim N(\mu_z, \sigma^2)$.

There remains the possibility that each of the 2^p possible genotypes has a distinct trait mean. However, often it is assumed that the QTLs act additively; we imagine that $E(y|z) = \mu + \Sigma_{j=1}^{p} \beta_j z_j$. Deviation from additivity (i.e. interactions between the QTLs) is called *epistasis* (Frankel and Schork, 1996). Many studies have provided strong evidence for the presence of interactions between QTLs (e.g. Shrimpton and Robertson (1988), Roberts *et al.* (1999) and Shimomura *et al.* (2001)). In this paper, however, we shall focus on the case of strict additivity. This is not because we feel that it is the best approach, but rather because this simple case is still not well solved.

With the assumption of additivity, the aim of QTL mapping is to identify the number and locations of the QTLs. One may further seek interval estimates of QTL locations and estimates of QTL effects; although these are both clearly important, we consider them of secondary interest and focus on the identification of QTLs. In the following section, we describe the current approaches to this problem. In Section 3, we frame the problem as one of model selection and describe an approach for QTL mapping that makes use of a modified version of the Bayesian information criterion BIC (Schwarz, 1978). In Section 4, we present the results of a large computer simulation to assess the performance of several major approaches to QTL mapping.

2. Current approaches

In this section, we describe the commonly used approaches for QTL mapping. For a more extensive review of the statistical methods for QTL mapping, see Doerge *et al.* (1997), Lynch and Walsh (1998) or Broman and Speed (1999).

The simplest approach to identifying QTLs, with data on an experimental cross, is to perform analysis of variance (ANOVA) at each of the marker loci (see Soller *et al.* (1976)). At each genetic marker, we split the back-cross progeny into two groups, according to their genotypes at the marker, and compare the two group phenotype means, by a t-test. Geneticists often prefer to report a LOD score, defined as the (base 10) log-likelihood ratio comparing the hypotheses

(a) the phenotypes in the two groups are normally distributed with distinct means but a common variance and
(b) the phenotypes for all individuals follow a common normal distribution, independent of genotype.

Marker loci giving large LOD scores are indicated to be linked to a QTL.

This approach has several weaknesses. First, if a QTL is not located exactly at a marker, its

effect will be attenuated as a result of recombination between the marker and the QTL. Second, at each genetic marker, we must discard individuals whose genotypes are missing. Third, when the markers are widely spaced, a QTL may be quite far from all markers, and so the power for QTL detection will decrease. Fourth, the approach considers only one locus at a time; in the presence of several QTLs, approaches that model multiple QTLs will give greater power for QTL detection, better separate linked QTLs and allow the examination of interactions between QTLs (though such interactions will not be considered here).

Lander and Botstein (1989) developed *interval mapping*, which overcomes the first three weaknesses of ANOVA at marker loci, described above. The method, which continues to be the most popular approach for QTL mapping, makes use of a genetic map of the typed markers, and, like ANOVA, assumes the presence of a single QTL. Each location in the genome is posited, one at a time, as the location of the putative QTL.

Given the marker genotype data (and assuming no crossover interference), one may calculate the probability that an individual has genotype HH (or HL) at a putative QTL. These QTL probabilities depend only on the genotypes at the flanking markers and may be found in Table 2 of Doerge *et al.* (1997). In interval mapping, one assumes that, given the QTL genotype, the phenotype follows a normal distribution with mean μ_H or μ_L, according to whether the QTL genotype is HH or HL respectively, and common standard deviation σ. Given the genotypes at the markers flanking the QTL, the conditional phenotype distribution is then a mixture of the two normal distributions, with the conditional QTL genotype probabilities, given the marker genotype data, as mixing proportions. At each position in the genome (or, in practice, at steps of 0.5 cM), one may use a version of the EM algorithm (Dempster *et al.*, 1977) to estimate the three parameters, μ_H, μ_L and σ, and may calculate a LOD score: the (base 10) log-likelihood ratio, comparing the hypothesis that there is a single QTL at the given location with the hypothesis that there is no QTL anywhere in the genome. The LOD score, as a function of chromosome position, forms a profile log-likelihood. Genomic regions for which the LOD score is large are indicated as harbouring QTLs.

The advantages of interval mapping, over ANOVA at marker loci, are that it makes more complete use of the marker genotype data (making proper allowance for missing data), and it considers positions between markers as putative locations for a QTL, thus providing increased power in the case of widely spaced markers, as well as improved estimates of QTL effects. However in the case of dense genetic markers and relatively complete marker genotype data, interval mapping provides little advantage over ANOVA. Moreover, interval mapping, like ANOVA, makes use of a single-QTL model and so is not ideal in the presence of multiple (especially linked) QTLs.

Both ANOVA at marker loci and interval mapping make use of multiple tests of hypotheses and so require some adjustment for test multiplicity. Much effort has been expended on this problem, the aim being to obtain an approximate genome-wide LOD threshold, defined as the 95th percentile of the distribution of the maximum LOD score, genome wide, under the hypothesis that there are no QTLs (i.e. that the phenotypes are simply normally distributed, independent of the marker data). Lander and Botstein (1989) performed extensive computer simulations to estimate the appropriate LOD threshold for various genome sizes and marker densities, and gave analytical calculations for the case of a very dense marker map. Another approach is to perform a permutation test (Churchill and Doerge, 1994).

As mentioned above, methods that make use of multiple-QTL models can provide increased sensitivity, better separate linked QTLs and allow the examination of interactions between QTLs. The simplest multiple-QTL method is multiple regression, the obvious extension of ANOVA at marker loci. Cowen (1989) appears to be the first to have recommended the use

of multiple regression in this context (see also Whittaker *et al.* (1995)). We shall defer further discussion of this approach to the next section.

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. One includes other markers (on the same chromosome and on different chromosomes) as regressors while performing interval mapping, in an effort to control for the effects of QTLs in other intervals, so that there will be greater power for QTL detection, and so that the effects of the QTLs will be estimated more precisely. Zeng called this approach composite interval mapping (CIM).

The method is performed as follows. We choose a subset 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 write $y = \mu + \beta z + \Sigma_{j \in S^*} \beta_j x_j + \varepsilon$ where y is the phenotype, $z = 1$ or $z = 0$ according to whether the genotype at the putative QTL is HH or HL, $x_j = 1$ or $x_j = 0$ according to whether the genotype at the jth marker is HH or HL and S^* is a subset of the marker regressors, S, where we exclude any markers that are within, say, 10 cM of the putative QTL. The residual, ε, is assumed to be distributed $N(0, \sigma^2)$.

As in interval mapping, at each locus, a LOD score is calculated, comparing the hypothesis that there is a QTL at the putative locus with the hypothesis that there is not a QTL there, in which case we imagine that all progeny have phenotypes which are normally distributed with mean $\mu + \Sigma_{j \in S^*} \beta_j x_j$ and variance σ^2. The LOD score is plotted as a function of genome position and compared with a genome-wide threshold. (Such a threshold should take into account the selection of the set of marker regressors, S.) Areas of the genome for which the LOD score exceeds a genome-wide threshold are said to contain a QTL.

The key problem with CIM is the choice of the set of markers to use as regressors: using too many markers will increase the variance of the LOD score and thus will decrease the power for QTL detection. Jansen (1993) and Jansen and Stam (1994) used backward elimination with Akaike's information criterion (Akaike, 1969), or a slight variant, to pick the subset of markers. Basten *et al.* (2000), in a manual for the program QTL Cartographer, recommended using forward selection up to a fixed number of markers, and then dropping any markers that are within 10 cM of the putative QTL.

More recently, Kao *et al.* (1999) proposed multiple-interval mapping (see also Zeng *et al.* (1999)), which is much like CIM, but the additional regressors are not required to reside at marker loci. In multiple-interval mapping, Kao *et al.* (1999) have adopted a more standard model selection approach, making use of stepwise selection.

Several other methods have been described, including Bayesian methods (e.g. Satagopan *et al.* (1996), Sillanpää and Arjas (1998), Ball (2001) and Sen and Churchill (2001)) and the use of genetic algorithms (e.g. Carlborg *et al.* (2000)). These approaches are more in line with our view that QTL mapping is a model selection problem.

3. Model selection

We consider a back-cross and assume that the genotype data are complete, and that the genetic markers are sufficiently dense, so that we may dispense with interval mapping, considering only the marker loci as putative locations for QTLs. Let y_i denote the phenotype of individual i, and let $x_{ij} = 1$ or $x_{ij} = 0$ according to whether individual i has genotype HH or HL respectively, at

marker j. We assume the linear model $y_i = \mu + \Sigma_{j=1}^{M} \beta_j x_{ij} + \varepsilon_i$, where the ε_i are independent and identically distributed $N(0, \sigma^2)$.

The problem of identifying QTLs in an experimental cross is one of model selection: in the above linear model, we seek to identify the subset of markers for which $\beta_j \neq 0$. By viewing the problem in this way, we may hope to take advantage of the extensive literature on subset selection in regression. However, much of the model selection literature has focused on the minimization of prediction error, whereas we are not so much interested in prediction as in the identification of an appropriate model.

We split the model selection problem into four distinct parts:

(a) select a class of models,
(b) compare models,
(c) search through the space of models and
(d) assess the performance of a model selection procedure.

We focus here on the class of additive models, though one might also consider linear models with pairwise interactions, or regression trees. The inclusion of the assessment of a procedure's performance as part of the model selection problem may be viewed as unusual but is clearly integral to the problem. Whether we choose to minimize the prediction error or to maximize the number of correctly identified QTLs while controlling the rate of inclusion of extraneous loci at a fixed level, a clearly stated objective is a prerequisite for making informed choices on a model selection procedure.

3.1. Model comparison

Consider the case of a linear model with normally distributed residual variation. Let Γ denote the set of models, with $\gamma \in \Gamma$ written as an M-vector with jth element 1 or 0 according to whether the jth marker is included in the model. Let $|\gamma|$ denote the number of markers in model γ, and let $\mathrm{RSS}(\gamma)$ denote the residual sum of squares after fitting γ by least squares. Imagine that we can fit all possible models.

For models with the same number of regressors k, we choose that with the smallest RSS. We write $\gamma_k = \arg\min_{\gamma:|\gamma|=k}\{\mathrm{RSS}(\gamma)\}$. Thus γ_M is the full model, with all markers included, and γ_0 is the model including no markers. $\mathrm{RSS}(\gamma_k)$ must be non-increasing in k. The key problem is to determine the decrease in RSS that must accompany the inclusion of an additional regressor. Our aim is to balance the errors of excluding important loci and of including extraneous loci.

Classical criteria for choosing the appropriate size of the model include Mallows's C_p and adjusted R^2 (Miller, 1990). In our experience, these criteria tend to include a large number of extraneous regressors and so are unsatisfactory for our purposes.

Two more modern approaches for choosing subsets of regressor variables include cross-validation and the bootstrap. In both of these approaches, an estimate of the mean-squared error of prediction is obtained. The chosen model has the smallest estimated mean-squared error of prediction. Because we are interested in identifying a reasonable model rather than minimizing the prediction error, we have not studied the performance of these approaches.

An additional approach for model comparison is the use of sequential permutation tests, appropriate in the context of a nested sequence of models, such as would be obtained by forward selection (see Doerge and Churchill (1996)). One works from the null model γ_0 to the full model γ_M, performing a permutation test at each step, testing whether the inclusion of an additional regressor is accompanied by a statistically significant decrease in the RSS. The first time that the null hypothesis is not rejected, one stops.

The approach that we favour is to minimize a criterion of the form

$$\Phi(\gamma) = \log\{\mathrm{RSS}(\gamma)\} + |\gamma| D(n)/n$$

where $D(n)$ is some function of the sample size n. (This is equivalent to maximum likelihood with a penalty on the model complexity, since in the case of normally distributed residuals $-(n/2)\log\{\mathrm{RSS}(\gamma)\}$ is the log-likelihood for the model γ.) The choice $D(n) = 2$ gives Akaike's information criterion (Akaike, 1969), whereas $D(n) = \log(n)$ gives BIC (Schwarz, 1978), and $D(n) = \log\{\log(n)\}$ gives the criterion of Hannan and Quinn (1979).

Minimization of $\Phi(\gamma)$ is approximately equivalent to the use of a threshold on the conditional LOD score $(n/2)\log_{10}\{\mathrm{RSS}(\gamma_{k-1})/\mathrm{RSS}(\gamma_k)\}$, the threshold being $D(n)/2\log(10)$. Consider our sequence of models $\gamma_0, \gamma_1, \ldots, \gamma_M$. In the case that $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ is strictly increasing in k, minimization of $\Phi(\gamma)$ is equivalent to choosing the largest value of k for which $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ is greater than $-\exp\{D(n)/n\}$. Note that it is sufficient, but not necessary, that the ratios $\mathrm{RSS}(\gamma_k)/\mathrm{RSS}(\gamma_{k-1})$ be strictly increasing, for this equivalence.

When viewed in this way, the criterion appears quite reasonable. Further support lies in the consistency of the resulting procedures. With a fixed number of possible regressors (i.e. genetic markers), and provided that $D(n)/n \to 0$ and $D(n)/\log\{\log(n)\} \to \infty$, the criterion $\Phi(\gamma)$ gives a consistent estimate of the underlying model, meaning that, as the sample size increases, the probability that the correct model is chosen converges to 1 (Rao and Wu, 1989).

We have concentrated on the case $D(n) = \delta\ \log(n)$, which we call BIC_δ:

$$\mathrm{BIC}_\delta(\gamma) = \log\{\mathrm{RSS}(\gamma)\} + \delta|\gamma|\log(n)/n.$$

Letting $\delta = 1$, this gives BIC. We have found that $\delta = 1$ performs poorly, including far too many extraneous regressor variables. A larger value of δ can give improved results, as a greater penalty on the size of the model leads to the inclusion of fewer extraneous regressors. We shall discuss the choice of δ in Section 3.3.

A further approach to the model selection problem is to place prior probabilities on each of the possible models, as well as on the model parameters, and to use Bayes's theorem to calculate the posterior distribution of the models given the data. If the goal were to pick out just one model, we could choose that which gives the largest posterior probability.

As an example, consider the priors discussed in Smith (1996). We let $y|\gamma, \beta_\gamma, \sigma^2 = X_\gamma\beta_\gamma + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$, and use the prior $\beta_\gamma \sim N\{0, c\sigma^2(X_\gamma' X_\gamma)^{-1}\}$, $p(\sigma^2|\gamma) \propto 1/\sigma^2$, $p(\gamma) \propto (c/d)^{|\gamma|/2}$. Let $c \to \infty$, resulting in a diffuse improper prior, and integrate out β_γ and σ^2. Smith (1996) showed that the resulting posterior for γ gives $-(2/n)\log\{p(\gamma|y)\} = \log\{\mathrm{RSS}(\gamma)\} + |\gamma|\log(d)/n$. Taking $D(n) = \log(d)$, we see that the model with maximum posterior is that which minimizes the above-described criterion, $\Phi(\gamma)$. We may consider this as further support for the use of the criterion $\Phi(\gamma)$. The only real justification for a criterion, however, is its performance. We shall study the performance of this criterion in Section 4.

3.2. Search of model space

The number of possible additive models is very large. If there are more than around 40 genetic markers, it will be infeasible to fit each of the $2^{40} \approx 10^{12}$ possible models. Thus, we must form a strategy for searching this large space of models, hopefully so that we may identify the good ones—those that would have been chosen if we could fit all possible models.

In the case that the number of markers is only marginally large, we may use a branch-and-bound procedure to pick out the best subsets of each size, without actually fitting all possible subsets (Miller, 1990), thus gaining considerable savings in computation over an exhaustive

search. However, with many markers, this type of procedure is still not feasible. We are thus led to techniques such as forward selection and backward elimination.

In forward selection, one begins with the null model and builds a nested sequence of models of increasing size; at each step, one adds the marker that gives the greatest decrease in the RSS. In backward elimination, one begins with the full model and builds a nested sequence of models of decreasing size; at each step, one drops the marker that gives the smallest increase in the RSS. These two sequences of models may be quite different.

Forward selection and backward elimination provide great savings in computation, since only a small fraction of the possible models are fitted. This saving is also a cost, however: we see only a fraction of the possible models, and we might not see the good ones. With forward selection, once a regressor has been included, it will be retained in all further models. With backward elimination, once a regressor has been dropped, it will be excluded from all further models.

Stepwise selection procedures, which iteratively add or subtract regressors, are commonly used for subset selection in regression. In such procedures, the 'stopping rule', for choosing the appropriate model size, is generally intertwined with the search through the model space. We prefer to keep separate the criteria for model comparison and the procedures for model search.

Forward selection has a particularly bad reputation. One can find quite simple situations in which forward selection will miss the correct model, even when the sample size is extremely large. This occurs as a result of collinearity in the regressor variables, where a regressor that does not belong in the model mimics a set of regressors that do. Backward elimination does not suffer from this problem, at least with large samples. An and Gu (1985) showed that, when using BIC, and in the case of a fixed number of regressors, the backward elimination procedure is consistent, meaning that, as the sample size increases, the probability of choosing the correct model converges to 1. The result also applies to BIC_δ. Forward selection, however, is *overconsistent*; in the limit, the selected model will contain the true model, but may also include additional, extraneous, regressors.

However, in the situation considered here, the regressors are genetic markers that, under the assumption of no crossover interference, form a Markov chain. Given the genotypes at any one marker, the genotypes at markers to its left are conditionally independent of the genotypes at markers to its right. This suggests that the sort of collinearity among regressors that may cause forward selection to include extraneous regressors, even with large samples, will not be a problem in the context of QTL mapping. Indeed, Broman (1997) showed that, in the case of a strictly additive QTL model, forward selection with BIC_δ is consistent. In computer simulations, Broman (1997) found that forward selection also worked reasonably well in samples of more typical size. We shall see below, however, that forward selection can still suffer from the inclusion of extraneous loci.

A different approach to searching the space of models is to use a randomized algorithm, such as a Markov chain Monte Carlo (MCMC), simulated annealing or a genetic algorithm. We shall consider only the MCMC method, in which one places a prior on each model and on the model parameters, and then forms a Markov chain whose stationary distribution is the posterior distribution of the models given the data. Simulations of the Markov chain give a sequence of models (a sort of walk through the space of models) which will, eventually, spend more time at models that have a high posterior probability. Whereas this method is usually used to obtain an approximation of the posterior distribution, and especially to find the region with highest posterior, here we consider it simply as a method for searching the space of models.

There are several standard ways to form a Markov chain with the desired stationary distribution. With the prior discussed above (Section 4.1), Smith (1996) used a Gibbs sampler to obtain

a Markov chain whose stationary distribution satisfies $-(2/n)\log\{p(\gamma|y)\} = \log\{\mathrm{RSS}(\gamma)\} + |\gamma|\log(d)/n$. The method, which is much like stepwise selection, is as follows. First, pick an initial model $\gamma^{(0)}$ (e.g. the null model or the model obtained by forward selection). Then, at step t, cycle through the M different markers; for each $j = 1, \ldots, M$, draw $\gamma_j^{(t)}$ from the distribution $p(\gamma_j|\gamma_{-j}^{(t)}, y)$ where $\gamma_{-j}^{(t)}$ is composed of all the elements of γ, except for γ_j, at their current values. For $i < j$, it contains the γ_i for the current step t and, for $i > j$, it contains the γ_i for the previous step $t - 1$. For the posterior written above,

$$\Pr(\gamma_j = 1|\gamma_{-j}, y) = \frac{\mathrm{RSS}(\gamma_1, \ldots, \gamma_{j-1}, 1, \gamma_j, \ldots, \gamma_M)^{-n/2}}{\mathrm{RSS}(\ldots, 1, \ldots)^{-n/2} + \sqrt{d}\,\mathrm{RSS}(\ldots, 0, \ldots)^{-n/2}}.$$

The most important characteristic for the Markov chain is that it mixes well—that it travels through the space of models with relative ease, not becoming stuck in local modes. We have implemented the above MCMC sampler and have found that it works well. In 1000 steps of the chain, it will visit around 300–500 distinct models and will almost always visit the best of those models (i.e. that giving the largest posterior probability) within the first 100 steps.

3.3. Recommended approach

It is best to consider model comparison and model search separately. One should devote the greatest effort to the formulation of a criterion for model comparison, as this is the most difficult aspect of model selection. It is helpful to imagine that we could examine all possible models. In choosing between them, we must balance the errors of excluding important regressors and including extraneous ones. The appropriate balance of these errors will vary according to the goals of the experiment, and so the appropriate criterion for comparing models should also vary.

We prefer the BIC_δ criterion, for its simplicity and its reasonable interpretability. One approach for choosing an appropriate δ is through the connection between BIC_δ and conditional LOD scores: we may choose the value of δ that corresponds to a genome-wide LOD threshold for interval mapping or ANOVA at marker loci. Let L denote such a threshold (the 95th percentile of the maximum LOD score, genome wide, under the hypothesis that there are no QTLs); then we may let $\delta = 2L/\log_{10}(n)$. Use of the derived BIC_δ criterion should, in the case of no QTLs, result in the selection of one or more extraneous loci, approximately 5% of the time. In the presence of QTLs, the rate at which extraneous loci are included is not necessarily under control, though we show in the next section, through computer simulations, that it performs adequately. Of course, such a choice of δ results in a procedure that is not consistent, as the rate of inclusion of extraneous loci will continue to be 5%, in spite of increasing sample size. If one desires a smaller false positive rate, a larger value of δ should be chosen.

The search of model space is a matter of exhausting or repetitive work. More extensive searches are better, though the improvement may not be sufficient to compensate for the increased computation. Forward selection and backward elimination are quick and simple to implement. The MCMC sampler described above is also simple to implement, and the increase in computation may be sufficiently small to justify its use.

4. Simulations

Computer simulation studies are crucial for understanding the relative performance of different model selection procedures, because such procedures aretoo complex to be assessed by analy-

tical means, at least in the situations in which they would be used in practice. It is unfortunate that large scale computer simulations are not routinely included in statistical methodological papers on QTL mapping. Many researchers have used simulations to illustrate methods for finding QTLs, but most have either presented the results on a single simulation replicate or data set or considered only very simple situations. In some cases, the value of a new approach has simply been declared on the basis of increased complexity.

Any simulation study is necessarily incomplete and artificial. Real QTL experiments do not have equally spaced markers and exhibit complex patterns of missing genotype data. The number, effects and locations of QTLs are not known; the QTLs have effects of varying size, and the QTLs may interact in complex ways. The simulation study reported here includes a small number of additively acting QTLs located exactly at marker loci and having equal-sized effects; the genetic markers were equally spaced and the genotype data were complete. Although this study may be criticized as not being sufficiently realistic, we believe that it is among the most complete and realistic such studies, and that the results are of considerable value for the assessment of the performance of the QTL mapping methods included.

4.1. Methods

We simulated a back-cross obtained from inbred lines, composed of 100, 250 or 500 progeny, with nine chromosomes, each of length 100 cM and having 11 equally spaced markers (at a spacing of 10 cM). The recombination process was assumed to exhibit no crossover interference. The marker data were complete and without errors. For each sample size, we performed 2000 simulation replicates.

We considered a model with seven QTLs of equal effect, 0.76, with all QTLs positioned exactly at marker loci. Two QTLs were located at markers 4 and 8 on chromosome 1 (separated by 40 cM), linked in *coupling* (i.e. their effects had the same sign). Two QTLs were located at markers 4 and 8 on chromosome 2, linked in *repulsion* (i.e. their effects had opposite signs). Three further QTLs were located at markers 6, 4 and 1, on chromosomes 3, 4 and 5 respectively. Four chromosomes contained no QTLs. The environmental variation followed a normal distribution with standard deviation $\sigma = 1$. As a result, the *heritability* of the trait (the proportion of the phenotypic variance attributable to the QTLs) was 50%.

We compared seven methods for identifying QTLs: ANOVA at marker loci, a simplified version of CIM, forward selection with permutation tests and the BIC_δ criterion with forward selection, backward elimination, forward selection followed by backward elimination, and MCMC sampling. Interval mapping was not considered, because it provides little improvement in power over simple ANOVA in the case of a relatively dense marker map and a moderate number of progeny, and because it would require a great increase in computation time.

For CIM, we used forward selection up to either 3, 5, 7, 9 or 11 markers to obtain the set of regressors, and we limited the search for QTLs to marker loci. With both ANOVA and CIM, we obtained genome-wide LOD thresholds (specific for the case of nine chromosomes of length 100 cM with 11 equally spaced markers on each chromosome) by performing 50 000 simulations under the null hypothesis of no QTLs. The estimated thresholds were obtained as the 95th percentile of the maximum LOD score across all markers and appear in Table 1. In addition, for these methods, we required that the LOD score dropped by at least 1.5 between 'peaks' before we declared that two QTLs were identified. This value was obtained empirically and may not be ideal. Note that this prevents these methods from identifying adjacent markers as QTLs.

The value of δ for the BIC_δ criterion was chosen to correspond to the LOD threshold for ANOVA in Table 1: $\delta = 2\ \mathrm{LOD}/\log_{10}(n)$. For $n = 100, 250, 500$, the value of δ was 2.56, 2.10

Table 1. Estimated LOD thresholds, based on 50 000 simulation replicates, for a back-cross with nine chromosomes, each 100 cM long and containing 11 equally spaced markers†

n	*ANOVA*	*Thresholds from CIM for the following numbers of markers:*				
		3	*5*	*7*	*9*	*11*
100	2.56	3.50	4.12	4.64	5.13	5.60
250	2.52	3.23	3.56	3.77	3.95	4.09
500	2.50	3.15	3.38	3.51	3.60	3.67

†Standard errors are approximately 0.01.

and 1.85 respectively. The permutation tests used 1000 replicates with $\alpha = 0.05$. In the use of forward selection, a maximum of 25 markers were considered. Backward elimination was begun at the full model. We further applied forward selection up to a model with 25 markers followed by backward elimination; the model with the minimum value of BIC_δ, among all fitted models, was chosen. For the MCMC method, we used 1000 steps of the sampler described in Section 3.2 and chose the model giving the minimum BIC_δ value. In the first 1000 of the 2000 simulation replicates performed, the MCMC sampler was started at the null model; in the second 1000 replicates, the sampler was started at the model obtained by forward selection with BIC_δ. The results were indistinguishable and thus were pooled.

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 10 cM of a QTL (i.e. if the marker was at or adjacent to the QTL); otherwise it was deemed extraneous. If more than one chosen marker were within 10 cM of the same QTL, one was called correct and the others were called extraneous.

4.2. Results

The results of the simulations are displayed in Figs 2 and 3. In terms of the number of QTLs correctly identified (upper panels in Fig. 2), MCMC sampling with the BIC_δ criterion performed best, though it was only slightly better than forward selection, and it was essentially indistinguishable from forward selection followed by backward elimination. Forward selection with BIC_δ was slightly better than with permutation tests. Backward elimination performed poorly at the smallest sample size. CIM performed slightly worse than forward selection with BIC_δ. CIM performed best when the number of markers used as regressors was 7, the number of simulated QTLs; a considerable attenuation of power was accompanied by a choice of too many or too few markers to serve as regressors in CIM. ANOVA, as might be expected, performed rather poorly for this model of multiple QTLs.

Fig. 3 provides greater detail on the number of QTLs that are correctly identified, giving separate results on the QTLs linked in coupling (upper panels), the QTLs linked in repulsion (centre panels) and the three other QTLs (lower panels). The inferior performance of ANOVA and of CIM with three or five markers serving as regressors, in the cases $n = 250$ or $n = 500$, was due largely to their poor ability to detect the QTLs linked in repulsion. In the case $n = 250$, forward selection with permutation tests also performed poorly on the QTLs linked in repulsion, because, for this method, forward selection was stopped when the first test in the

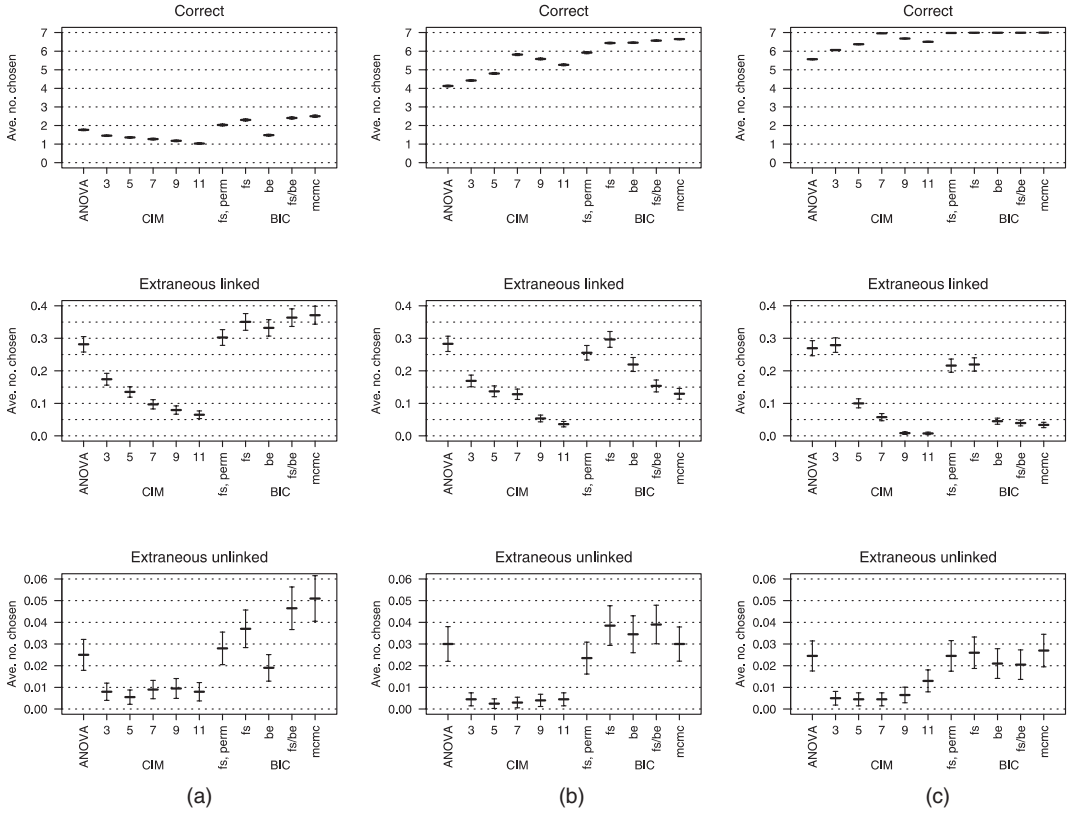

Fig. 2. Results of simulations of a back-cross with (a) $n = 100$, (b) $n = 250$ or (c) $n = 500$ individuals, under a seven-QTL model, including two QTLs linked in coupling (effects of the same sign) and two QTLs linked in repulsion (effects of opposite sign) (the upper panels indicate the average number of QTLs that are correctly identified; the centre panels indicate the average number of extraneous loci that were linked to a QTL; the lower panels indicate the average number of extraneous loci that were not linked to a QTL): the methods considered were ANOVA, CIM, preceded by forward selection up to a fixed number of loci, forward selection with permutation tests and the BIC_δ criterion with forward selection, backward elimination, forward selection followed by backward elimination, and MCMC sampling

sequence was not rejected, while the loci in repulsion appear important only when considered jointly.

All the methods except CIM chose a rather high proportion of extraneous loci linked to a QTL (centre panels in Fig. 2). For the MCMC sampling, backward elimination, and forward selection followed by backward elimination, this effect went away at high sample sizes, but ANOVA and forward selection continued to include a high proportion of extraneous linked loci even at $n = 500$. For the case $n = 100$, these extraneous linked loci were largely imprecisely localized (but correctly identified) QTLs. If a QTL was considered to be correctly identified when a marker within 20 cM was chosen (*versus* the 10 cM criterion used to create Fig. 2), the proportion of extraneous linked loci was reduced from around 30% to around 10%. For the case $n = 500$, however, these loci were truly extraneous. Forward selection identified all the QTLs but also included additional marker loci; if a more complete search of the model space was undertaken (as in the MCMC method or by following forward selection with backward elimination), these additional loci were excluded.

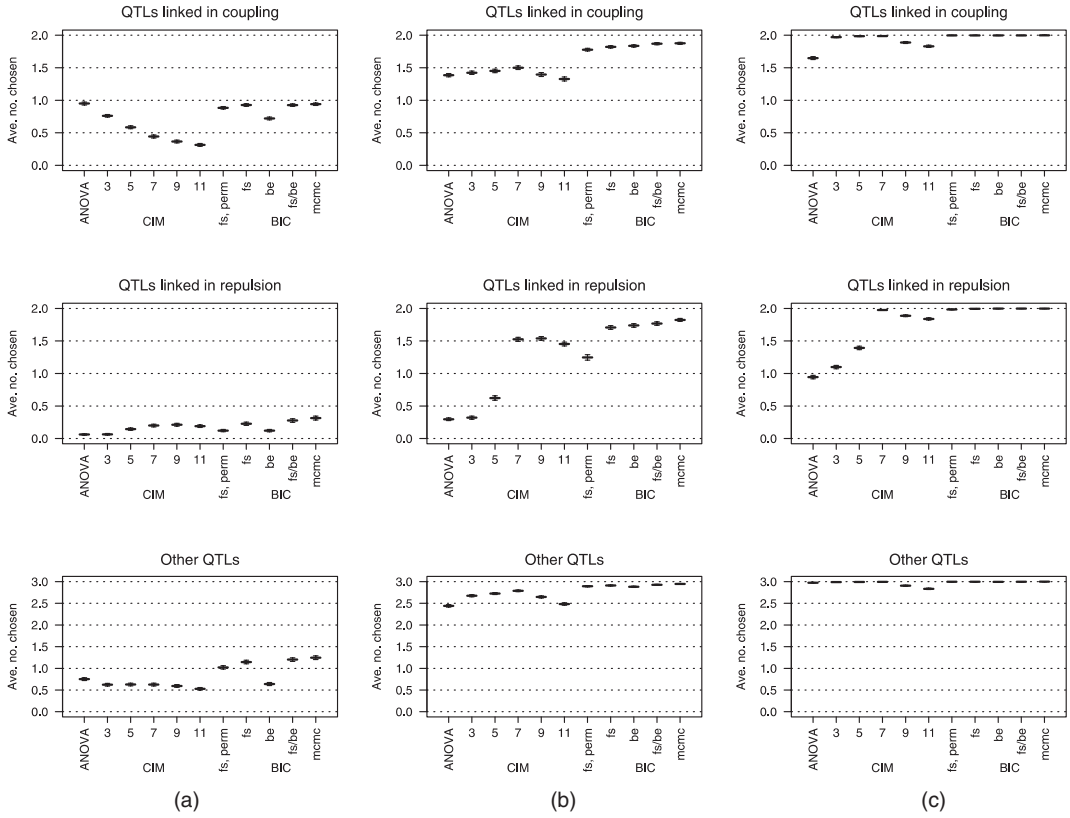

Fig. 3. Detailed results on the upper panels of Fig. 2, displaying the average number of QTLs correctly identified in a seven-QTL model, including two QTLs linked in coupling and two QTLs linked in repulsion (the upper, centre and lower panels indicate the average number of QTLs correctly identified, for the two QTLs linked in coupling, the two QTLs linked in repulsion and the three other QTLs respectively): (a) $n = 100$; (b) $n = 250$; (c) $n = 500$

The lower panels in Fig. 2 display the proportion of extraneous loci not linked to a QTL. CIM was quite conservative, delivering only about 0.5% of such extraneous unlinked loci, whereas the other methods all delivered approximately 2–3% of such extraneous unlinked loci (which might be expected, given that a 5% genome-wide threshold was used, and four out of the nine chromosomes contained no QTLs). Under the null hypothesis of no QTLs, all these methods will identify at least one extraneous QTL, 5% of the time. These results illustrate that the performance of a procedure in the presence of QTLs may be rather different from what might be expected, given its behaviour under the null hypothesis of no QTLs.

In summary, MCMC sampling with the BIC_δ criterion performed best. The key advantage of MCMC sampling over forward selection was the elimination of extraneous linked loci, which forward selection included at a reasonably high rate at $n = 500$. The same benefit could be obtained by following forward selection with backward elimination. CIM performed only slightly worse than forward selection and did not suffer from the inclusion of extraneous loci, but a correct choice of the number of markers included as regressors was extremely important; the use of too few or too many such marker regressors was accompanied by a loss of power.

5. Discussion

There are four key points that we wish to make in this paper. First, QTL mapping is best viewed as a problem of model selection. Second, the comparison of models is the most difficult part of the model selection problem. Third, large scale computer simulation studies are important for understanding the relative performance of different model selection procedures, and they should be routinely included in papers describing new approaches for QTL mapping. Fourth, more refined procedures will not necessarily provide sufficiently improved results to justify their added complexities and increased computational requirements; the choice of adopting such procedures should be based on honest estimates of the gains in performance that they provide.

We have focused on the problem of identifying QTLs. Although we have not considered the precise localization of QTLs and the estimation of QTL effects, it cannot be denied that these are important (and not always straightforward) problems in practice. However, the selection of the number and approximate locations of QTLs is a prerequisite, and so we are justified in considering only this essential part of the problem. In the case of inbred strains of mice, the step following localization of a QTL is frequently the creation of a congenic strain incorporating the relevant region onto a desired background. This is done in the hope of recovering the phenotype in this strain; only when this happens will the search for actual genes begin.

We have discussed back-cross designs because they are the simplest, but of course our study could and should be repeated for other common designs such as the intercross. Although we expect the broad conclusions to be similar to the case considered here, the work needs to be done.

We have also focused on the unrealistic situation in which QTLs are located exactly at marker loci, and in which markers are densely and regularly spaced and exhibit no missing genotype data. This was done to make plain the essence of the QTL mapping problem, and to illustrate the performance of several approaches to the problem. It may happen that our quantitative conclusions change if the markers are not densely and regularly spaced, though we do not expect this. In the case of missing genotype data and/or gaps between markers, the multiple-regression approach that we have considered will not be appropriate. One may confront this missing data problem by multiple-interval mapping (Kao *et al.*, 1999; Zeng *et al.*, 1999) or multiple imputation (Ball, 2001; Sen and Churchill, 2001); the model selection issues that we have discussed remain the essence of the problem.

We considered the case of QTLs acting strictly additively. Of course, one cannot know in advance that this will be appropriate, and a growing number of experiments provide strong evidence for the presence of interactions between QTLs. Thus we recommend that, in practice, one pursues the possibility of interactions. This may be done by the inclusion of pairwise interactions in a linear model, or the consideration of tree-based models. The BIC_δ criterion will probably remain useful in this situation, though a larger value for δ (a larger penalty for model complexity) may be required, and one may wish to place different penalties on main effects and interactions. More complex, randomized search algorithms, such as an MCMC sampler, may be especially valuable for the search of these expanded spaces of models.

We have recommended the use of the BIC_δ criterion, with the value of δ chosen by the approximate correspondence between BIC_δ and a genome-wide threshold on the LOD score. We hope that this is not interpreted as a recommendation for strict adherence to thresholds. In particular, 5% significance thresholds may not be in accordance with the goals of the experimenter. A consideration of the models selected with larger and smaller values of δ provides valuable information regarding the strength of evidence for QTLs.

Our computer simulations demonstrate the value of the BIC_δ criterion. The MCMC sampler performed best. Forward selection was nearly as good, and its tendency to include extraneous loci could be alleviated by following forward selection with backward elimination. CIM performed reasonably well, though it has the disadvantage of requiring a choice of the number of markers to serve as regressors. The sensitivity of the results of CIM to this choice suggests that, although its conversion of a multidimensional into a single-dimensional search is enviable, the approach should not be recommended. There are various schemes for selecting variables in CIM, and it may be true that one of these, different from the one that we have used, gives generally better results and invalidates this conclusion.

The improved performance of these multiple-QTL approaches, over ANOVA at marker loci, is clear but is not nearly as fantastic as we might have hoped. It is difficult to deny that a genome scan by interval mapping can give quite reasonable results. The advantages of multiple-QTL methods are the better separation of linked QTLs and the ability to examine interactions between QTLs.

Acknowledgements

Dursun Bulutoglu and Śaunak Sen generously provided comments to improve the manuscript.

References

Akaike, H. (1969) Fitting autoregressive models for prediction. *Ann. Inst. Statist. Math.*, **21**, 243–247.
An, H. and Gu, L. (1985) On the selection of regression variables. *Acta Math. Appl. Sin.*, **2**, 27–36.
Ball, R. D. (2001) Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics*, **159**, 1351–1364.
Basten, C. J., Weir, B. S. and Zeng, Z.-B. (2000) *QTL Cartographer, Version 1.14.* Raleigh: North Carolina State University.
Broman, K. W. (1997) Identifying quantitative trait loci in experimental crosses. *PhD Dissertation*. Department of Statistics, University of California, Berkeley.
Broman, K. W. and Speed, T. P. (1999) A review of methods for identifying QTLs in experimental crosses. *IMS Lect. Notes Monogr. Ser.*, **33**, 114–142.
Carlborg, O., Andersson, L. and Kinghorn, B. (2000) The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, **155**, 2003–2010.
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. In *Development and Application of Molecular Markers to Problems in Plant Genetics* (eds T. Helentjaris and B. Burr), pp. 113–116. Cold Spring Harbor: Cold Spring Harbor Laboratory.
Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* 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.
Doerge, R. W., Zeng, Z.-B. and Weir, B. S. (1997) Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statist. Sci.*, **12**, 195–219.
Frankel, W. N. and Schork, N. J. (1996) Who's afraid of epistasis? *Nat. Genet.*, **14**, 371–373.
Hannan, E. J. and Quinn, B. G. (1979) The determination of the order of an autoregression. *J. R. Statist. Soc.* B, **41**, 190–195.
Jansen, R. C. (1993) Interval mapping of multiple quantitative trait loci. *Genetics*, **135**, 205–211.
Jansen, R. C. and Stam, P. (1994) High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136**, 1447–1455.
Kao, C.-H., Zeng, Z.-B. and Teasdale, R. D. (1999) Multiple interval mapping for quantitative trait loci. *Genetics*, **152**, 1203–1216.
Lander, E. S. and Botstein, D. (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121**, 185–199.
Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*, ch. 15. Sunderland: Sinauer.
Miller, A. J. (1990) *Subset Selection in Regression*. New York: Chapman and Hall.

Rao, C. R. and Wu, Y. (1989) A strongly consistent procedure for model selection in a regression problem. *Biometrika*, **76**, 369–374.
Roberts, L. J., Baldwin, T. M., Speed, T. P., Handman, E. and Foote, S. J. (1999) Chromosomes X, 9, and the H2 locus interact epistatically to control Leishmania major infection. *Eur. J. Immunol.*, **29**, 3047–3050.
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. *Ann. Statist.*, **6**, 461–464.
Sen, Ś. and Churchill, G. A. (2001) A statistical framework for quantitative trait mapping. *Genetics*, **159**, 371–387.
Shimomura, K., Low-Zeddies, S. S., King, D. P., Steeves, T. D., Whiteley, A., Kushla, J., Zemenides, P. D., Lin, A., Vitaterna, M. H., Churchill, G. A. and Takahashi, J. S. (2001) Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res.*, **11**, 959–980.
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.
Sillanpää, M. J. and Arjas, E. (1998) Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, **148**, 1373–1388.
Smith, M. S. (1996) Nonparametric regression: a Markov chain Monte Carlo approach. *PhD Dissertation*. University of New South Wales, Sydney.
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. *Theoret. Appl. Genet.*, **47**, 35–39.
Whittaker, J. C., Curnow, R. N., Haley, C. S. and Thompson, R. (1995) Using marker-maps in marker-assisted selection. *Genet. Res.*, **66**, 255–265.
Zeng, Z.-B. (1993) Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natn. Acad. Sci. USA*, **90**, 10972–10976.
———(1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457–1468.
Zeng, Z.-B., Kao, C.-H. and Basten, C. J. (1999) Estimating the genetic architecture of quantitative traits. *Genet. Res.*, **74**, 279–289.