A Statistical Framework for Quantitative Trait Mapping

Śaunak Sen and Gary A. Churchill

The Jackson Laboratory, Bar Harbor, Maine 04609

Manuscript received July 29, 2000
Accepted for publication June 4, 2001

ABSTRACT

We describe a general statistical framework for the genetic analysis of quantitative trait data in inbred line crosses. Our main result is based on the observation that, by conditioning on the unobserved QTL genotypes, the problem can be split into two statistically independent and manageable parts. The first part involves only the relationship between the QTL and the phenotype. The second part involves only the location of the QTL in the genome. We developed a simple Monte Carlo algorithm to implement Bayesian QTL analysis. This algorithm simulates multiple versions of complete genotype information on a genomewide grid of locations using information in the marker genotype data. Weights are assigned to the simulated genotypes to capture information in the phenotype data. The weighted complete genotypes are used to approximate quantities needed for statistical inference of QTL locations and effect sizes. One advantage of this approach is that only the weights are recomputed as the analyst considers different candidate models. This device allows the analyst to focus on modeling and model comparisons. The proposed framework can accommodate multiple interacting QTL, nonnormal and multivariate phenotypes, covariates, missing genotype data, and genotyping errors in any type of inbred line cross. A software tool implementing this procedure is available. We demonstrate our approach to QTL analysis using data from a mouse backcross population that is segregating multiple interacting QTL associated with salt-induced hypertension.

THE problem of identifying the genetic factors underlying complex and quantitative traits has a long history. The idea of using the association between a discrete trait (marker) and a continuously variable phenotype to establish linkage of a quantitative trait locus (QTL) first appeared in the work of Sax (1923). The first known statistical approach is due to Thoday (1961). Modern analysis of quantitative trait genetics utilizes large sets of DNA-based markers to carry out genome scans that are capable of identifying multiple genetic factors associated with a trait in a mapping population. Statistical analysis of QTL mapping data is typically carried out by interval mapping (Lander and Botstein 1989) in which likelihood-ratio tests are computed on a dense grid of possible QTL locations. The interval mapping procedure is based on an expectation-maximization (EM) algorithm (Dempster *et al.* 1977), which maximizes the likelihood of a single-gene genetic model by averaging over the possible states of the unknown genotype at each possible QTL location. Despite its explicit use of a single gene model, this approach has been successfully applied to detect multiple genes that underlie a wide range of complex and/or quantitative traits. Rapp (2000) provides an excellent review of the state of the art of QTL methodology focusing on hypertension in rats.

Multiple-QTL models are an improvement over single-QTL models because of their ability to separate linked QTL on the same chromosome and to detect interacting QTL that may otherwise be undetected. They provide increased power to detect QTL and can eliminate biases in estimates of effect size and location that can be introduced by using an inappropriate single-QTL model (Schork *et al.* 1993). A variety of approaches have been proposed for mapping multiple QTL. Haley and Knott (1992) described a simple approximation to interval mapping that is based on regression, which can be applied to multiple-QTL models. Composite interval mapping (CIM) and multiple-QTL mapping (MQM; Jansen 1993; Jansen and Stam 1994; Zeng 1993, 1994) represent attempts to reduce the multidimensional search for QTL to a series of one-dimensional searches. This is achieved by conditioning on markers outside a region of interest to account for the effects of other QTL. Multiple interval mapping (MIM) proposed by Kao *et al.* (1999) extends interval mapping directly to the case of multiple QTL. An EM algorithm is used to calculate LOD scores under a multiple-QTL model. Like MIM we use an explicit multiple-QTL model but we replace the EM algorithm with a Monte Carlo algorithm. Thus we trade off exact computation for ease of computation and flexibility.

For pragmatic reasons, we chose to approach the problem of mapping multiple QTLs from a Bayesian perspective. The Bayesian framework provides a clear picture of the probabilistic structure of the QTL mapping problem. In particular, the treatment of unknown quantities (the QTL locations, the QTL genotypes, and the pheno-

Corresponding author: Gary Churchill, The Jackson Laboratory, 600 Main St., Bar Harbor, ME 04609. E-mail: garyc@jax.org

typic means and variances) is straightforward. However, we depart from a strictly Bayesian analysis at several points as indicated below. Bayesian approaches to linkage analysis and QTL mapping have been described by others (Hoeschele and VanRaden 1993; Satagopan *et al.* 1996; Uimari and Hoeschele 1997; Sillanpää and Arjas 1999). We restrict our attention to inbred line crosses and comment on Satagopan *et al.* (1996). They constructed a Markov chain Monte Carlo (MCMC) algorithm that sequentially samples from the unknown QTL locations, QTL genotypes, and genetic model parameters. An advantage of the MCMC approach is the ability to explore the model space by allowing the number of QTL to change as part of the Markov chain (Green 1995). Our computations rely on an independent sample Monte Carlo approach, using importance sampling of multiple imputed data sets. This avoids the problematic issue of mixing of the Markov chain. Furthermore the Monte Carlo error is inversely proportional to the square root of the number of imputed data sets and can be tightly controlled. The operational simplicity of our algorithm should make it more accessible to nonspecialists in MCMC methodology.

Central to our approach is the observation that, by conditioning on the unobserved QTL genotypes, the problem of QTL mapping can be divided into two simple and statistically independent parts: the genetic model part, which relates the QTL genotypes to the phenotype, and the linkage part, in which the locations of the QTL are determined. This observation is not new (Jansen 1993; Thompson 2000); it is implicit in almost all published work on QTL. However, making the QTL genotypes the central focus offers several advantages. It leads to a readily implemented algorithm and that admits a broad range of generalizations. The problems of missing marker data, genotyping errors, covariates, nonnormal and multivariate phenotypes, epistatic QTL, crossover interference, and nonstandard cross designs can all be addressed within this framework.

In the sections that follow, we outline our approach using a general notation that highlights the probabilistic structure of the QTL mapping problem. This structure is used to devise an efficient computational strategy for Bayesian inference. We point out the relationship of the log posterior distribution of QTL location (LPD) to the LOD score and show how the former leads to a simple method for constructing confidence regions for QTL locations. This is followed by a description of our software tools. After discussing possible variations of the basic theme, we demonstrate our approach with an example of a complete QTL data analysis. Many of the technical details can be found in appendices a–f.

A FRAMEWORK FOR QTL INFERENCE

Heuristic and motivation: If two inbred strains raised in a common environment show markedly different values for a phenotype, it is implicit that there is a genetic basis for the difference. Association between the genotypes and phenotypes in progeny derived from a cross between such strains will provide information about the genetic basis for the trait. However, there are at least two components of "noise" in the data that can obscure the genetic effects. First is the environmental variation that is inherent in most quantitative phenotypes. Second is the incomplete nature of the genotype information, which can only be observed at the typed markers. Marker genotypes may also contain missing values and errors. Typically, an investigator is interested in knowing how many genes contribute to the trait, where they are located, and how they act. Statistical approaches are needed to address these questions.

Consider the simple case in which the trait is affected by allelic variation at a single gene. In the presence of environmental variation, the phenotype can be viewed as a "noisy" version of a biallelic locus (the QTL) whose position in the genome is unknown. If the genotype of the QTL could be known, the effects of the QTL could be estimated by simply looking at the distribution of phenotypes within the groups of individuals defined by the QTL alleles. Furthermore, the QTL position could be localized by mapping the biallelic QTL relative to the typed markers. Thus the unknown QTL genotypes are key.

If we have complete genotype information on a dense set of markers, a one-dimensional genome scan can be performed by regressing the phenotype on each of the markers. The more a marker explains the phenotypic variance, the more likely it is to be close to a QTL. This belief can be quantified by plotting the LOD score or *F*-statistic obtained from regressing the phenotype on the marker. In practice the markers may be widely spaced across the genome and some marker genotypes will be missing or in error. Still we can imagine having access to a dense set of completely genotyped markers; we call them *pseudomarkers.* Using linkage information in the available marker data we can infer the genotypes of the pseudomarkers. There will be some uncertainty regarding the pseudomarker genotypes so we may decide to construct several, say 10, versions of this ideal genotype data each of which is consistent with the observed marker genotypes. Variation in the imputed genotypes reflects the uncertainty in our knowledge of the true complete genotypes. If the typed markers are dense, the variation will be negligible. As in the case when we had complete genotype information, we can now regress the phenotype on each of the pseudomarkers and repeat this process for each of the imputed versions. We now have not 1, but 10 sets of LOD scores. The strength of evidence in favor of a QTL being near any given pseudomarker can be quantified by averaging the 10 LOD scores. For technical reasons explained below, the average is not an arithmetic mean of the LOD scores.

This approach extends to the simultaneous mapping of multiple QTL. Suppose that there are two QTL influencing the trait. Using the same pseudomarkers, LOD scores are computed for each *pair* of pseudomarkers for a two-dimensional genome scan. Interactions between QTL can be accommodated. We implemented pairwise and triplet genome scans and illustrate their application below. Higher-order genome scans could become computationally prohibitive. We argue below that the pairwise scans should be sufficient for the analysis of data with many segregating QTL provided that interactions are limited to pairwise effects and that there are no groups of three or more tightly linked QTL. When these conditions are met, the results of single and pairwise genome scans can be combined into a single statistical model that describes the simultaneous effects of multiple QTL on a quantitative phenotype. The last step represents a compromise that is necessary because a full search for four or more QTL is computationally impractical. The best approach to search for multiple QTL models remains an open problem.

Data structures and notation: Suppose that we have data on n animals or plants derived from an inbred line cross. Denote the quantitative trait measurements by $y = (y_1, y_2, \ldots, y_n)'$, and denote the genotyping data by the $n \times k$ matrix $m = (m_{ij})$, where the rows correspond to individuals and the columns correspond to markers. The quantities y and m are the *observed data.* Assume that the chromosome of origin, order, and genetic distance between the markers is known. In practice, these quantities may have to be estimated.

The genetic model, denoted by H, is a description of the distribution of phenotypes given the QTL genotypes. If there are p contributing QTL and the trait values are normally distributed within the QTL genotype classes, a general linear model may be used to describe the relationship of the phenotype to the QTL genotypes. The parameters of the genetic model are denoted by μ. The locations of the QTL are denoted by the p-dimensional vector γ. The QTL genotypes are denoted by the $n \times p$ matrix $g = (g_{ij})$. The rows of g correspond to individuals and columns correspond to the loci. The quantities μ, γ, and g are the *unobserved data.* Note that the meaning and dimensionality of the unobserved data structures depend on the genetic model, H.

Theory: Our goal is to make inferences about the genetic model parameters (μ) and the QTL locations (γ) given the observed data. We use Bayesian statistical theory because it provides a convenient and mathematically consistent method for describing uncertainties in the form of posterior distributions. As outlined above, we combine information from the marker genotypes and the phenotypes to reconstruct the unknown QTL genotypes. Multiple imputed versions of the QTL genotypes are then used to compute approximations to the posterior densities of interest $p(\gamma|y, m)$ and $p(\mu|y, m)$. Imputed genotypes can also be used to compute the marginal probability of the data $p_H(m, y)$, which is useful for making model comparisons.

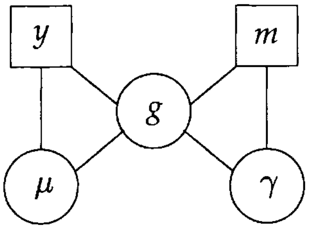

FIGURE 1.—Graph of the stochastic dependencies between the five main data structures on the QTL mapping problem. Boxes represent observed data structures and circles represent unobserved (or missing) data structures.

To develop our arguments we begin by looking at the joint distribution of all of the observed and unobserved data. Under the assumption of no ascertainment, the joint distribution can be factorized as

$$p(y, m, g, \mu, \gamma) = \big(p(y|g, \mu)p(\mu)\big)\big(p(g|m, \gamma)p(m)p(\gamma)\big). \quad (1)$$

A proof is provided in APPENDIX A. This factorization implies that, conditional on the QTL genotypes, g, the genetic model part of the problem involving (y, μ) can be solved independently from the linkage part of the problem involving (m, γ). Figure 1 represents this conditional independence and highlights the central role of the unobserved QTL genotypes.

This decomposition of the problem into two parts conditional on the unobserved QTL genotypes suggests that we should begin by obtaining the posterior distribution of the QTL genotypes. In APPENDIX B we show that the posterior distribution of the QTL genotypes after integrating out the parameters μ and γ can be expressed as

$$p(g|y, m) \propto p(y|g)p(g|m). \quad (2)$$

The first term indicates how compatible a phenotype is with the QTL genotypes. The second term measures the compatibility of the QTL genotypes with the observed marker data.

Sampling QTL genotypes: Expression (2) suggests an efficient computational approach for simulating from the posterior distribution of the QTL genotypes. We can first simulate samples from $p(g|m)$ and then weight each sampled genotype by $p(y|g)$. The idea is that the genotypes that are most compatible with the observed marker data are most likely to turn up in the simulation from $p(g|m)$. Among those genotypes, the ones that are most compatible with the phenotypes will get the largest weights. Details are provided in APPENDIX B.

We want to consider models with multiple QTL, including cases in which there is linkage and/or interaction among the QTL. If there are p QTL in the model, the location γ will have p components and the QTL genotypes will constitute an $n \times p$ matrix. Furthermore, the locations are not known *a priori* and thus we want to scan through all possible locations to search for QTL. These considerations suggest that we should simulate genotypes at all possible locations in the genome from their joint distribution given the marker data. In practice we generate genotypes on a discrete grid of locations spanning the genome, which we refer to as the pseudomarker grid. For a given p-tuple of pseudomarker locations $u = (u_1, \ldots, u_p)$, the ith realization of genotypes is an $n \times p$ matrix denoted as $r_i(u)$.

A weighted sample of QTL genotypes is generated by the following steps:

1. Select a regularly spaced grid G of pseudomarker locations and create q realizations of the pseudomarkers by sampling from the distribution $p(g|m, \gamma = G)$. The notation $\gamma = G$ is used to indicate that the entire grid of pseudomarkers is simulated as a joint distribution. We assume that there is no crossover interference and that genetic distances between the markers are known. With these assumptions, a simple Markov chain sampling scheme can be used to generate the pseudomarker genotypes (Lander and Green 1987).
2. For each p-tuple of locations u in each pseudomarker realization ($i = 1, \ldots, q$), calculate the weight under the assumed genetic model H:

$$W_H(r_i(u)) = p(y|g = r_i(u))p(\gamma = u). \qquad (3)$$

If the prior on the QTL locations is uniform, the term $p(\gamma = u)$ is constant for all locations u and can be ignored.

Note that the weight function is model dependent whereas the pseudomarker generation is not. This is convenient for exploring the space of models as it reduces the amount of computation required when a new model is considered. For normally distributed data under the assumption that the prior distribution on the genetic model parameters is of the order of one observation, the weights are approximately proportional to $n^{-v/2}\mathrm{RSS}^{-n/2}$, where n is the sample size, v is the model dimension, and RSS is the residual sum of squares obtained by regressing phenotypes on the QTL genotypes. Genotypes that explain more of the variation in the phenotype get a bigger weight. Additionally, the dimension of the genetic model is penalized. This penalty becomes important when different genetic models are compared. A derivation of normal model weights is provided in appendix c.

Estimating QTL locations: In appendix b we show that the posterior distribution of the QTL location is proportional to the average weight of all pseudomarker realizations at that location

$$p(\gamma = u|y, m) \propto \int p(y|g)p(g|m, \gamma = u)p(\gamma = u)\,dg \simeq \sum_{i=1}^{q} W_H(r_i(u)). \qquad (4)$$

In interval mapping, inference about QTL locations is based on a likelihood-ratio test, which when expressed on the scale of base 10 logarithm, is called the LOD score. The LOD score at location γ can be shown to be equal to

$$\mathrm{LOD}(\gamma) = \text{constant} + \log_{10}\left(\sup_{\mu} p(y, m|\mu, \gamma)\right). \qquad (5)$$

The logarithm (base 10) of the posterior distribution of the QTL locations (LPD) is

$$\begin{aligned}\mathrm{LPD}(\gamma) &= \log_{10}(p(\gamma|y, m)) \\ &= \text{constant} + \log_{10}\left(\int p(y, m|\mu, \gamma)p(\mu)\,d\mu\right) \\ &\quad + \log_{10}(p(\gamma)).\end{aligned} \qquad (6)$$

We expressed the LOD score and LPD in this form to illustrate their similarity. Details are provided in appendix d. By comparing (5) and (6) we see that the LOD score takes a maximum over the genetic model parameters whereas the LPD carries out an averaging operation. In most situations when a uniform prior on the QTL locations is used, LOD(γ) and LPD(γ) will be approximately equal to each other up to an additive constant. In Figure 2, on the basis of the hypertension data discussed below, we compare the LOD score, the LPD, and the Haley and Knott (1992) approximation to the LOD score. We can see that in this example the LOD score and the LPD are essentially indistinguishable. However, we note that it is possible to construct examples where the two quantities differ.

The LPD raised to the power of 10 is the posterior density of the QTL location and hence can be used to construct confidence intervals (Sen 1998; Dupuis and Siegmund 1999). In our implementation, the pseudomarker grid is discrete, and thus $p(\gamma|m, y)$ is a discrete probability distribution. This is a reasonable approximation to a continuous distribution of locations provided that the grid density exceeds our ability to resolve the QTL locations. In practice even a very coarse grid (10 cM) is quite effective. A denser grid (2 cM) is preferable for localization of QTL once they have been assigned to a chromosome. The discrete nature of the grid makes the computation of a highest posterior density (HPD) region straightforward. For example, on a chromosome with two QTL, the weights for each pair of pseudomarkers can be normalized and ranked from highest to lowest. A $1 - \alpha$ HPD region is constructed by including the pairs with highest weights in the set until the sum of the weights first exceeds $1 - \alpha$. Bayesian confidence intervals based on the LPD have the desired long-run frequency coverage in large samples. See Sen (1998) for a further discussion of those issues.

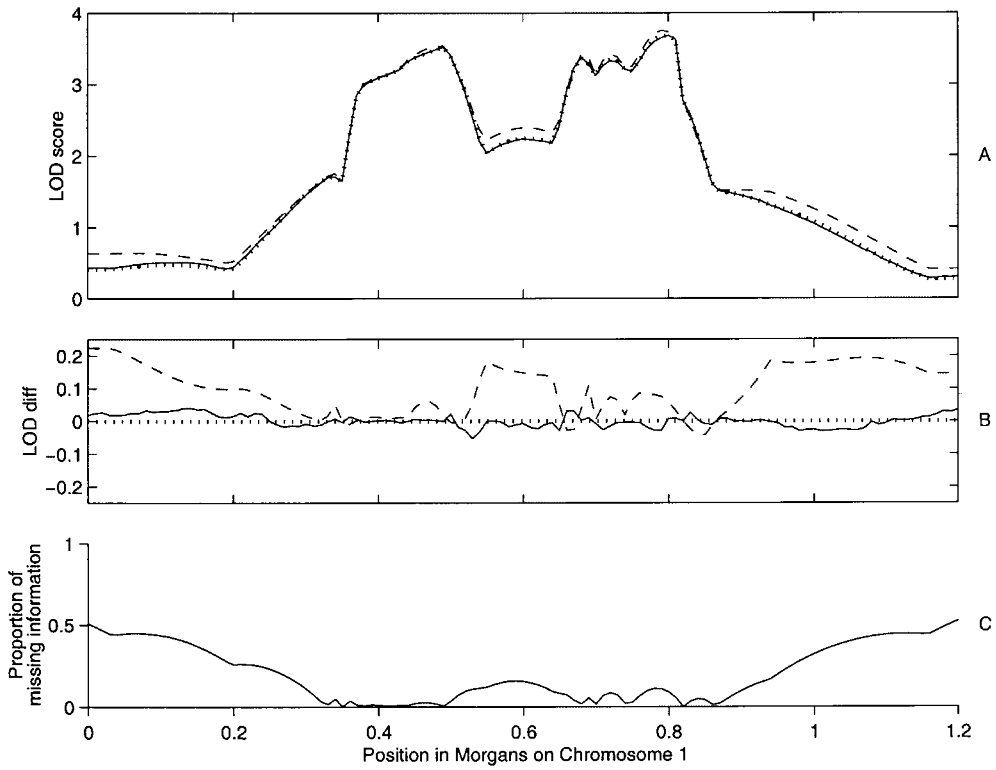

Figure 2.—Comparison of LOD score, LPD, and Haley-Knott approximation onto the LOD score on chromosome 1 of the hypertension data: (A) Plot of the LOD score calculated using the EM algorithm (solid line), the LPD (dotted line), and the approximate LOD score calculated using the Haley-Knott method (dashed line). (B) Plot of the differences between LPD-LOD (dotted line) and LPD-approximate LOD (dashed line). (C) Plot of the proportion of missing genotype information as a function of location on chromosome 1. One can see that the three methods agree with each other where the proportion of missing genotype information is small. The discrepancy increases as the proportion of missing information increases.

Substantive prior information can be incorporated into the LPD through the additive term $\log_{10}p(\gamma)$. This is analogous to the process of accumulating evidence by adding LOD scores (Morton 1955).

Estimating QTL model parameters: In appendix b we show that the posterior distribution of the model parameters can be expressed as

$$p(\mu|y, m) \propto \iint p(\mu|g, y)\, W_H(g, \gamma)\, p(g|m, \gamma)\, dg d\gamma$$
$$\simeq \sum_{i=1}^{q}\sum_{u} p(\mu|y, g = r_i(u))\, W_H(r_i(u)), \qquad (7)$$

where the summation on u is over all p-tuples of pseudomarker locations. The first term in the summation is the "complete data" posterior distribution of the model parameters given the phenotypes and the QTL genotypes. The second is the weight of the QTL genotypes. Thus (7) is a weighted mixture of complete data posterior densities. Posterior means and variances of the model parameters, $E(\mu|y, m)$ and $V(\mu|y, m)$, are computed by the method of iterated expectation as detailed in appendix b.

We depart from a strictly Bayesian approach here. Suppose that we are entertaining a model with six QTL. The summation over u in (7) would range over all sextuples of pseudomarkers in the genome and could be prohibitive to compute. In practice we take advantage of the partitioning of the genome into chromosomes. We estimate the model parameters one (or two) chromosomes(s) at a time and restrict the summation over u just to the chromosome(s) of interest. For example, if a chromosome is assumed to contain two QTL, we use a summation over all pairs of pseudomarkers on the chromosome to estimate the effects associated with the two QTL. If two unlinked QTL have an interaction term in the model, the two must be considered simultaneously and the summation will run over all pairs on the two chromosomes of interest. By estimating the parameters associated with small subsets (of size one, two, or three) separately we can significantly reduce the amount of computation with negligible effects on the results. When we are estimating effects of one set of QTL, the other QTL may be represented by including marker genotypes as covariates in the regression (see section on covariates below) as in CIM and MQM. Alternatively, we can simply ignore the other QTL. In practice both approaches yield essentially identical point estimates. This is a consequence of the independent assortment of chromosomes, which results in approximate orthogonality between unlinked locations in the genome. The standard errors will generally be smaller when conditioning.

Model scanning and model selection: In practice, the genetic model H, which includes the number of QTL and their interactions, is not known and has to be chosen on the basis of the data. The problem of selecting an appropriate model is challenging and we cannot offer a complete formal solution. The model selection problem is fundamental to multiple QTL analysis. Broman and Speed (1999) reviewed different QTL analysis methods from a model-selection point of view. They propose a criterion for model selection on the basis of a modification of the Bayesian information criterion (BIC) of Schwarz (1978), which they called BIC_δ. Kao *et al.* (1999) use a stepwise selection procedure.

The Bayes factor (Kass and Raftery 1995) is a Bayesian inferential device that can be used to support an exploratory analysis of potential models. The Bayes factor for comparing two models, H and K, is the ratio of the marginal distribution of the observed data calculated under the two models

$$B(H, K) = \frac{p_H(y, m)}{p_K(y, m)}. \tag{8}$$

The marginal probability of the data under a model H is approximately equal to the average of all the pseudomarker weights

$$p_H(y|m) \simeq \frac{1}{qs}\sum_{i=1}^{q}\sum_{u} W_H(r_i(u)),$$

(see appendix e), where s is the number of p-tuples of pseudomarker locations. In practice we limited the application of Bayes factors to making decisions about subsets of the QTL in a model. For example, we may compare a two-QTL model on a given chromosome to a single-QTL model on that chromosome. We can compare a two-QTL model with an interaction term to a two-QTL model with only additive effects. In these cases, the summation on u can be restricted to the chromosome of interest. Bayes factors can present computational difficulties (Satagopan *et al.* 1996).

Our data analysis consists of a model scanning step followed by model selection. It represents a departure from the Bayesian approach to model selection. We carry out single and pairwise genome scans and select only those regions (or pairs) that exceed stringent permutation testing thresholds (Churchill and Doerge 1994). We then fit multiple gene models that include the regions identified as being significant in the genome scans. This approach is consistent with the idea that one should report only highly significant QTL to minimize false positive results (Lander and Kruglyak 1995). There is often some fine tuning required to determine which interaction effects to include and to resolve linked QTL. These model comparisons may be carried out using Bayes factors or likelihood-ratio tests.

Permutation testing for the pairwise genome scan requires a bit of explanation. We are seeking pairs of loci that together contribute significantly to the observed phenotype distribution. Thus we base our test on a comparison of the full model, including an interaction effect, to a null model with no genetic effects. The significance of this overall test is assessed by permutation analysis. If a pair is found to be significant it is necessary to make some secondary checks to ensure that the pair is actually representing two QTL. First one examines the size of the interaction by comparing the full model to an additive model (with no interaction effects). If there is a significant interaction, the pair represents two interacting QTL. If the interaction is not significant, each member of the pair should be individually significant. In this case the pair represents two additive QTL. In Sugiyama *et al.* (2001) the secondary decisions were made using marker-regression-based tests and nominal P values. In our application we are using Bayes factors on the chromosomes on which the loci are located. Secondary decisions about the significance of the interaction do not require genomewide corrections as the pair has already been selected on the basis of a stringent criterion. One should be reasonably conservative about declaring interaction effects. We suggest that a P value of at least 0.01 or a Bayes factor of 10 is a reasonably conservative guideline.

Stepwise procedures for model selection using the BIC_δ of Broman and Speed (1999) and the F-to-enter criterion of Kao *et al.* (1999) can be carried out using our software tools. It is an area that holds promise but one we have not adequately explored. We also note that a QTL may be deemed important if it explains a substantial proportion of the variance even if it fails to achieve *statistical significance.* For nonnormally distributed phenotypes the ANOVA concept does not carry over and alternative criteria (such as those based on deviance in the case of generalized linear models) may be used. How well a QTL is localized also provides a measure as to how important the QTL may be. Some balance of judgment is required and all the evidence in support of a reported QTL should be reported.

Prior distributions: All Bayesian analyses depend on prior distributions. The influence of the prior distribution decays with increasing sample size and, for most problems, vanishes asymptotically. For sample sizes that are typical in most QTL studies (50–250 individuals), the prior distribution on the model parameters is not likely to have a large effect on the posterior distributions. However, it has a more tangible impact on the Bayes factors used for model selection. For example, the Bayes factors are not well defined if (improper) reference priors are used for the genetic model parameters.

In our analyses we used proper priors whose weight is approximately equal to that of one observation. This assumption leads to the penalty term of $n^{-v/2}$ in the weight function, where v is the dimension of the genetic model (see appendix c). Using proper priors also helps stabilize numerical computations when the phenotypes are not normally distributed.

Implementation: We implemented a basic set of computational tools in the MATLAB computing environment (The Mathworks, http://www.mathworks.com). We chose MATLAB for prototyping convenience, but any computing environment with tools for regression analysis and programming would suffice. All MATLAB functions used for this article are available at http://www.jax.org/research/churchill/ under software. The MATLAB environment provides a variety of functions that can be used for preliminary analysis and manipulation of the data.

Pseudomarker generation is implemented in the function IMPUTE, which takes marker genotypes and map positions as input and generates a three-dimensional array of imputed genotypes (the first dimension is individuals, the second dimension is pseudomarker positions, and the third dimension is replications). This array is used repeatedly in subsequent analysis steps.

Weights for imputed QTL genotypes are computed by genome scan functions. These functions can be applied to the whole genome or restricted to chromosomes of interest. A one-dimensional scan is performed using the function MAINSCAN, which produces a LPD profile assuming a single-QTL model at each pseudomarker location. MAINSCAN will produce essentially identical results to a Mapmaker/QTL analysis. A two-dimensional scan can be carried out using PAIRSCAN. This function assumes a two-QTL model and computes the weights both with and without an interaction effect. It scans through all pairs of marker loci and produces a two-dimensional LPD profile. The functions PLOTMAINSCAN and PLOTPAIRSCAN are used to plot the results from the scans. Traditionally, scanning functions have plotted the LOD score. Our functions plot the proportion of variance explained. This is approximately a linear multiple of the LOD score (Lander and Botstein 1989; Dupuis and Siegmund 1999) for models with normally distributed phenotypes and it has an intuitive appeal. Permutation tests on the one-dimensional and two-dimensional scans are performed by PERMUTEST and PERMUTEST2, respectively. A three-dimensional scan can be performed using TRIPLESCAN. We typically restrict a triple scan to a limited number of genomic regions that have already been identified in the one- and two-dimensional scans. TRIPLESCAN can be used to assess three-way interactions and is useful for localizing QTL in some situations. For example, if there are two linked QTL, one or both of which interact with a third unlinked QTL, a joint analysis of their effects will be required to provide unbiased estimates of location and effect sizes. Higher-dimensional scans may be performed using the SCAN function. This function is generally slower than ONESCAN or TWOSCAN but is more flexible.

After the genome scans have been carried out we can obtain estimates of the QTL model parameters. For a QTL that is not linked to or interacting with any other QTL, the function ONEESTIMATE computes estimates of the posterior mean and standard error of the effect size. This function, applied on one chromosome at a time, provides an estimate of the effect size of a QTL on *that chromosome.* For linked and/or interacting QTL, we provide functions TWOESTIMATE and THREE-ESTIMATE. Scanning and estimation functions for non-normally distributed phenotypes have been written and are not in the testing phase.

Bayes factors for model comparisons are obtained from the output of the scan functions since the marginal distribution of the data under a genetic model is the average of the pseudomarker weights. Bayes factors are calculated on selected genomic regions of interest. For example, to compare a single-QTL model on chromosome 1 to a two-QTL additive model on chromosome 1, we will compare the average pseudomarker weight on chromosome 1 (obtained from MAINSCAN) to the average pseudomarker weight on chromosome 1 for an additive model (obtained from PAIRSCAN).

Model selection is carried out using the permutation tests on one-dimensional and two-dimensional scans followed by secondary tests. Loci can be “flagged” by the functions FLAG and FLAG2. The former uses Bayes factors for secondary tests while the latter uses likelihood ratios.

For localization of a QTL or a pair of QTL, we construct a dense pseudomarker grid on the chromosome and repeat the imputation and scanning steps on that chromosome. Then we plot the results using LOCALIZE or PAIRLOCALIZE to plot the posterior distribution of the QTL or QTL pair.

We are currently constructing examples, including analysis scripts that illustrate the steps involved in applying these software modules. Results will be posted on our web site. In our experience we find that each QTL data set is unique and requires a tailored analysis. Thus we prefer an interactive software environment that allows the analyst to work with the data.

EXTENSIONS

We consider two general classes of extensions to the basic framework, those that alter the genetic model and those that alter the linkage model. Changes to the genetic model can be implemented by programming a new weight function. Changes to the linkage model affect only the code that simulates the pseudomarker genotypes. Modularity in software design as well as in data analysis that the framework provides is an important advantage.

Extensions of the genetic model: For normally distributed phenotypes, the weights are based on a normal regression model. More general distributions can be accommodated by calculating the weights assuming a generalized linear model (McCullagh and Nelder 1989). Generalized linear models have been mentioned by several authors, for example Jansen (1993), but they do not appear to be widely used in practice. This may be due in part to the robustness of normal regression models but is also due to lack of readily available software tools. Shepel *et al.* (1998) used a Poisson regression model on marker loci and stepwise selection using the BIC criterion to identify multiple loci. Implementing alternative distributions in a package based on the EM algorithm would require extensive reprogramming. In our software package only the weight function needs to be revised.

In general the weight function for any model is

$$W_H(y, g) = \int p_H(y|g, \mu) p(\mu) \, d\mu. \qquad (9)$$

For generalized linear models, under the assumption that the prior distribution on the genetic model parameters is of the order of a single observation, this works out to be approximately

$$\exp\left(-\frac{\text{dev}}{2}\right) n^{-v/2},$$

where dev $\simeq -2 \log(p(y|g, \mu = \hat{\mu})) - \log(p(\mu = \hat{\mu}))$ is the observed *unscaled deviance* (McCullagh and Nelder 1989), $\hat{\mu}$ is the posterior mean of μ, and v is the dimension of the genetic model. If the phenotype data follow a Poisson distribution, the weight function is

$$W_H(y, g) = \left(\prod_{i=1}^{k} \frac{n_i^{y_{i.}}}{y_i!}\right)^{-1},$$

where k is the number of genotype classes, n_i is the number of observations in the ith class, and $y_{i.}$ is the sum of the observations in the ith class. For binomial data, the weight function will be

$$W_H(y, g) = \left(\prod_{i=1}^{k} \binom{m_{i.}}{y_{i.}}\right)^{-1},$$

where $m_{i.}$ is the total of the size parameters of the binomial observations in each group and $y_{i.}$ is the sum of the observations in the ith class.

Many complex trait studies involve measurement of multiple related phenotypes. If two phenotypic measurements are affected by the same set of genes, then it can be more efficient to consider a multivariate analysis (Jiang and Zeng 1995; Korol *et al.* 1995; Ronin *et al.* 1999). If the phenotype is multivariate normal, then the appropriate weighting scheme is

$$W_H(y, g) = n^{-v/2} \det(S)^{-n/2},$$

where S is the residual covariance matrix of the multivariate ANOVA and v is the model dimension (appendix c). Models in multivariate space are more complex and some of the nice interpretations that apply to univariate normal phenotypes do not carry over. Although multivariate phenotyping comes with the promise of greater power to detect QTL, there are some costs. Unless two phenotypes are affected by the same biochemical pathway, adding a phenotype into the analysis may add genes and interactions to the list of genes affecting the (multivariate) phenotype. This may complicate the analysis of complex traits where a large number of genes are known to be affecting the trait of interest. We recommend that multivariate trait analysis be used with caution.

An interesting mutlivariate data structure is presented by Broman *et al.* (2000). They consider a time to death phenotype in which some of the animals survive beyond the observation period. The data are represented as a binary indicator of survival status (y_1) and, for those animals that died, a time to death (y_2). To reproduce the results of Broman *et al.* (2000) we implemented the weight function

$$W_H(y, g) = W_H^{\mathrm{B}}(y_1, g) W_H^{\mathrm{N}}(y_2, g), \qquad (10)$$

where W_H^{B} denotes the weight function corresponding to a binomial distribution and W_H^{N} to that of a normal distribution.

In QTL experiments covariates are often collected in addition to the phenotypes of interest. The logic behind measuring covariates is to measure and adjust for environmental factors that influence phenotypic expression. These might be blocking factors in an agricultural field trial or cage number in a mouse cross. The covariate may also be another phenotype, in which case some care must be exercised in the interpretation of results. We assume that there is no direct association between the QTL that affect the trait of interest and the covariate. If this is not the case, the phenotype and the covariate should be treated as a multivariate phenotype. We denote the covariate by x and any unknown parameters governing the distribution of x are denoted by ν. It can be shown by calculations similar to that in appendix a that the appropriate weight function is

$$W_H(y, g) = p(y|x, g) p(\gamma = u). \qquad (11)$$

The practical implication of (11) is that the weights are now based on a regression analysis of the phenotype on the covariates and the QTL genotypes.

The use of marker loci as covariates in QTL analysis was suggested by Jansen (1993) and also by Zeng (1993). When analysis of a QTL is focused on a single chromosome or other genomic region, the use of unlinked markers as covariates presents no difficulties. The weight function (11) is appropriate. This can be a useful device for reducing the complexity of QTL analysis by accounting for other segregating QTL in a cross and is easy to implement using our software tools. However, the use of *linked* markers as covariates presents some difficulties and may significantly reduce power. If there are multiple unlinked QTL in a cross, conditional analysis of one QTL with marker covariates to control for others can be an effective strategy. When there are multiple linked QTL or interactions among unlinked QTL, we recommend a joint analysis of their effects.

Extensions of the linkage model: Missing marker data are common in QTL experiments, sometimes due to difficulties with typing but also as a result of selective genotyping. Following Rubin (1976) and Schafer (1997) we argue that as long as the missingness mechanism depends on the observed data and not on the missing data values, it does not have an impact on the posterior distributions of the parameters of interest. Still, one must use techniques such as the EM algorithm or multiple imputation to take missing data into account. If there is some ascertainment bias, such as if only animals with high phenotypes are collected and the rest discarded, then this assumption would be violated. As a general rule the phenotype data from an entire cross should be included in the analysis even for individuals with no genotype information.

Let R denote the missing marker data pattern, and let m_{obs} and m_{mis} denote the observed and missing marker data, respectively. It can be shown that the form of the posterior distributions of the QTL locations γ, the genetic model parameters μ, and the QTL genotypes g given the observed data R, y, and m_{obs} is unchanged. The marginal distribution of the observed data under the genetic model H is

$$p_H(y,\ m_{\mathrm{obs}},\ R) = p(R|m_{\mathrm{obs}},\ y)p_H(y|m_{\mathrm{obs}})p(m_{\mathrm{obs}}). \qquad (12)$$

Thus to compute the Bayes factor for comparing different models, we only need to calculate $p_H(y|m_{\mathrm{obs}})$, other terms being independent of the model.

Selective genotyping is a practical device for reducing the cost of QTL mapping experiments (Lander and Botstein 1989). For example, an investigator, after measuring the phenotype of the animals from a cross, may decide to genotype only the extremes. In this case, the missing data pattern depends only on the observed data (the phenotypes). If the decision to genotype was based on a phenotype y, but the trait of interest is another phenotype z, then the appropriate analysis would use a multivariate phenotype composed of (y, z). Although it is technically not correct to analyze z as a univariate trait when data have been selected using y, we do not anticipate a serious bias if the univariate analysis is used.

Genotyping errors were considered by Lincoln and Lander (1992). Their formulation was based on a simple model of the probability of a typing error. We note that this falls into the missing data framework presented above and can be handled by modifying the pseudomarker simulation code.

Backcross and intercross designs have been widely utilized in quantitative trait mapping studies. However, there are alternative approaches and there is interest in developing new cross designs that might improve the resolution of QTL mapping studies. Designs that utilize recombinant inbred lines, congenic or isogenic lines, repeated backcrossing, and advanced intercross lines are some examples (Darvasi 1998). In principle, any crossing design can be accommodated into our procedure provided that one can simulate a pseudomarker genotype conditional on observed marker data.

Finally, we note that it is possible to incorporate models of crossover interference into the pseudomarker simulation module. This may substantially increase the computational burden because the conditional independence assumptions that simplify the computations under a no-interference model cannot be used (Zhao *et al.* 1995).

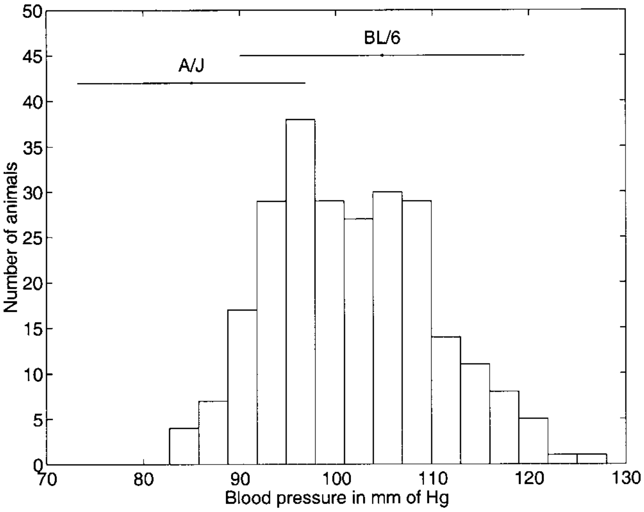

Figure 3.—Distribution of blood pressure values. The histogram of the blood pressures of 250 mice in the backcross reported in Sugiyama *et al.* (2001) is shown. Also displayed on top of the histogram are the means and two standard deviation error bars of the two parental strains. Eight mice in the BL/6 strain and 10 mice in the A/J strain were measured.

EXAMPLE

We illustrate the application of our approach by a reanalysis of a hypertension cross described in Sugiyama *et al.* (2001). Blood pressure measurements were obtained on 250 mice from a backcross between strain C57BL/6J (high blood pressure) and strain A/J (low blood pressure). We analyze blood pressure data assuming a normal model. Figure 3 shows the distribution of the blood pressure of the backcross individuals compared with the parental lines. The mean blood pressure was ~101.6 mm of Hg and the standard deviation of the blood pressure was ~8.4 mm of Hg. A total of 174 markers were typed. Initially only the extremes of the backcross population were genotyped in a standard selective genotyping design. Following an initial analysis of the data, additional genotyping was carried out in regions of interest on all mice.

We used a 10-cM pseudomarker grid with 16 imputations for the initial scanning step. The model refinement and localization were done using a 2-cM pseudomarker grid with 256 imputations. The initial genome scan using MAINSCAN revealed two significant QTL on chromosomes 1 and 4 and a suggestive peak on chromosome 15. In Figure 4 the proportion of the variance explained is graphed by location on the genome. Also shown is the 5% cutoff based on a permutation test. QTL on chromosomes 1 and 4 explain ~6 and 13.5% of the variance, respectively. The next biggest QTL is on chromosome 15, which explains ~5.5% of

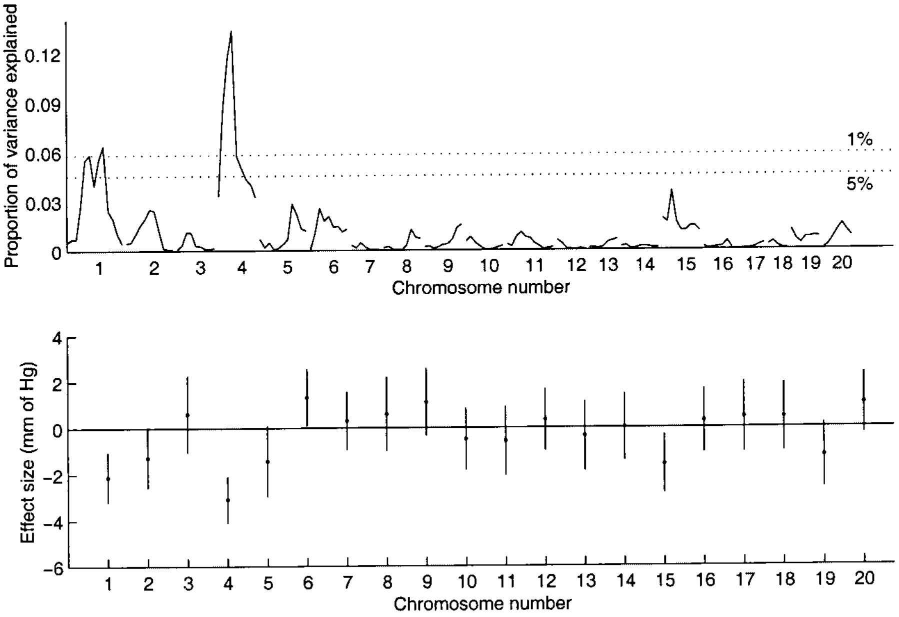

Figure 4.—Results of one-dimensional scan. The top shows the proportion of the variance explained as a function of QTL location. The pseudomarker spacing used was 10 cM and 16 sets of imputations were used. The two horizontal lines correspond to 1 and 5% critical values based on 1000 permutations. Strong evidence for QTL on chromosomes 1 and 4 and weak evidence for a QTL on chromosome 15 are indicated. There is a hint of two QTL on chromosome 1. The bottom shows a plot of estimated effects and 95% confidence intervals based on a single-QTL model on each chromosome. The effect size is defined as half of the effect of substituting an A/J allele in place of a C57BL6/J allele.

the variance. The Bayes factor estimates for comparing the null model of no genetic effects to a single-QTL model are <1 for all chromosomes except chromosomes 1, 4, and 15 (the respective values are 37.3, 1.1 × 10^5, and 1.7). Figure 5 shows the result of a two-dimensional scan using two-QTL models fitted using PAIRSCAN. We find that the QTL on chromosomes 1 and 4 together explain ∼22% of the variance. There is evidence for interaction between loci on chromosomes 6 and 15. The Bayes factor for interaction *vs.* no interaction is 20.4. The interaction explains ∼6% of the variance above an additive model. There is also evidence for interacting QTL on chromosomes 7 and 15, which explains ∼5% of the variance over an additive model (the Bayes factor is 11.4). The Bayes factor indicates that the evidence for the 6 × 15 interaction is stronger than that of the 7 × 15 interaction even though the size of the estimated effects is about the same (see Figure 8). A look at the localization plots (see Figure 7) reveals that the localization information is stronger with the 6 × 15 interaction than with the 7 × 15 interaction. The combined results of these scans suggest that we should look closely at chromosomes 1 and 4 to determine if there may be multiple QTL on either of these

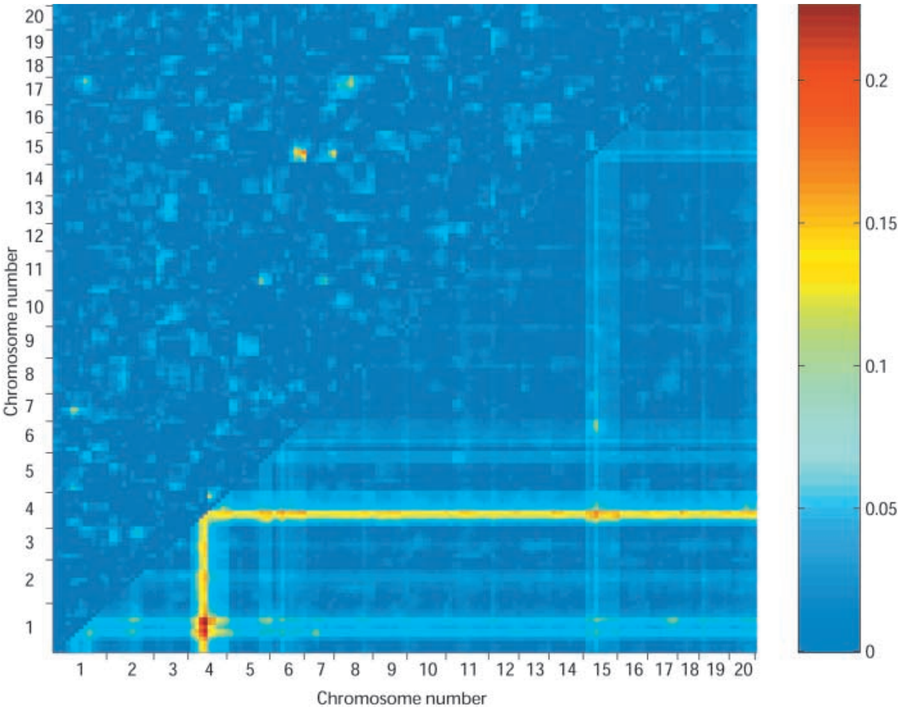

Figure 5.—Results of two-dimensional scan. The proportion of variance explained by a two-QTL model with interactions is plotted below the diagonal. The difference of the variance explained by a two-QTL model with interactions and an additive model is shown above the diagonal. The values above the diagonal were inflated by a factor of 3 to enhance visibility. The 5% permutation-derived threshold level based on 100 permutations for the full model (below the diagonal) corresponds to 8.7% variance explained. The tick marks correspond to the *ends* of the labeled chromosomes.

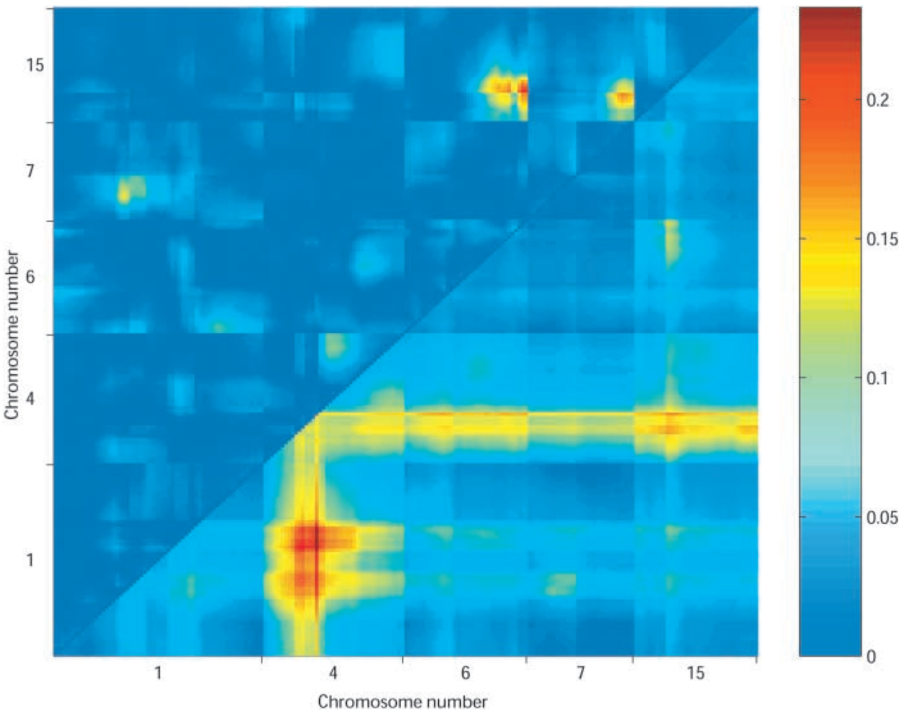

FIGURE 6.—Close-up of two-dimensional scan on chromosomes 1, 4, 6, 7, and 15. This suggests that there may be two QTL on chromosome 1. The interaction effects between loci on chromosomes 6 and 15 and those on 7 and 15 are apparent in the top diagonal. This was produced using 256 sets of imputations on a 2-cM pseudomarker grid.

chromosomes and that we should examine the simultaneous effects of loci on chromosomes 6, 7, and 15 to sort out the nature and extent of the interactions.

We carried out a TRIPLESCAN restricted to chromosomes 6, 7, and 15. There is no evidence for a three-way interaction among these markers. The Bayes factor for the three-way interaction model *vs.* a model with all three two-way interactions is 0.2. There is also no evidence for an interaction between chromosomes 6 and 7. Thus we conclude that the two pairwise interactions detected using PAIRSCAN are sufficient to describe the joint effect of these loci. The 6 × 15 interaction was reported by SUGIYAMA *et al.* (2001). The 7 × 15 interaction was not detected in their analysis, which may be due to large intermarker spacings on chromosome 7.

Now we turn our attention to chromosome 1, which showed some evidence for two QTL on the basis of the output from MAINSCAN and PAIRSCAN. The Bayes factor comparing a two-locus additive model on chromosome 1 to a single-locus model on chromosome 1 is

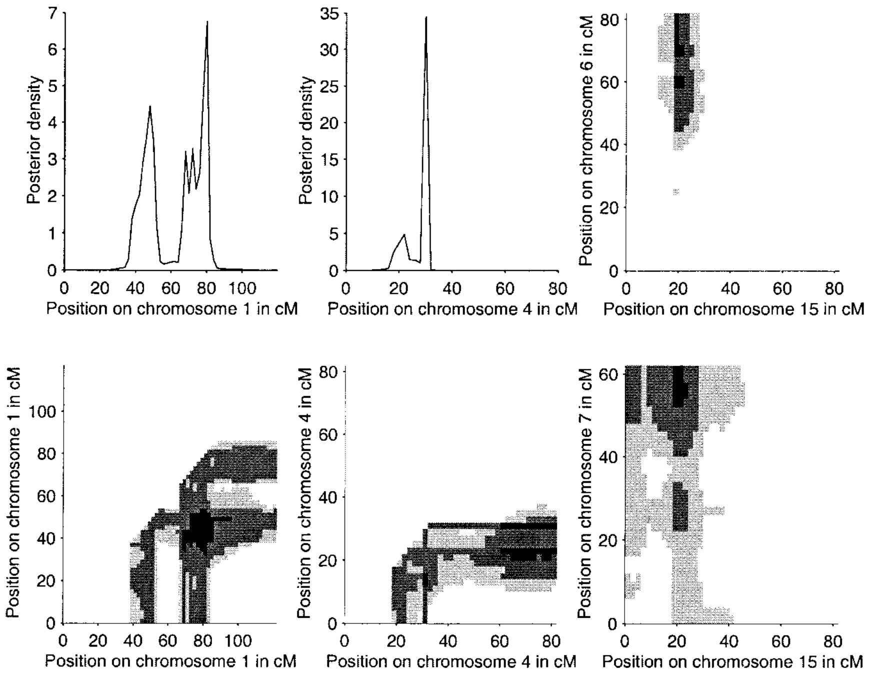

FIGURE 7.—Summary of localization information of QTL on chromosomes 1, 4, and 15. For locus pairs, we plot the 50, 95, and 99% confidence intervals in successively lighter shades. The top left shows the posterior distribution of the QTL on chromosome 1. There is a hint of two QTL here, but this was not supported very strongly by the Bayes factor. The bottom left shows the posterior distribution according to a two-QTL model on chromosome 1. The 95% confidence interval is pretty big, reflecting moderate but not convincing evidence for two QTL on that chromosome. The posterior distribution of the QTL on chromosome 4 is shown in the top middle. The bottom middle shows the posterior distribution assuming a two-QTL model on chromosome 4. The top right shows the joint posterior distribution of the interacting QTL on chromosomes 6 and 15; the bottom right shows the joint posterior distribution of the interacting QTL on chromosomes 7 and 15. Note that the QTL on chromosome 15 is in the same place in both.

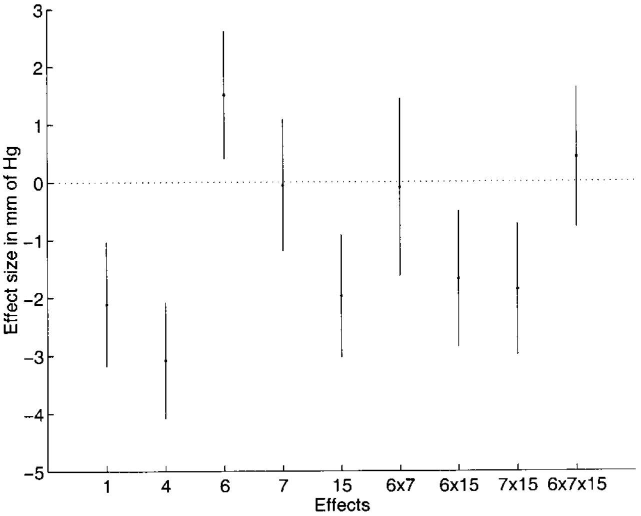

Figure 8.—Estimated effects of QTL and their interactions in the hypertension cross. The estimated effects are identified with the chromosome the QTL are in. Also shown are the 95% confidence intervals for the effects based on the posterior variances of the effects. The sizes of the interaction effects 6 × 15 and 7 × 15 are comparable to the main effects of 1 and 15. By adding up the squares of the effect sizes we find that the plotted QTL effects explain ∼41% of the total variance. For comparison, we included the interaction effects 6 × 7 and 6 × 7 × 15, which are not in the final model.

0.41. This is inconclusive evidence but it favors the single-QTL model. The localization plot assuming a two-QTL model on chromosome 1 seems to indicate two QTL, but not overwhelmingly. Additional experimentation may be needed to provide stronger evidence for the presence of two QTL on chromosome 1. We note that Sugiyama *et al.* (2001) conclude that there are two QTL on chromosome 1. Our reanalysis suggest weaker evidence in favor of two QTL than their original analysis. A localization scan of chromosome 4 suggests the possibility of multiple QTL but again there is not strong support provided by the Bayes factor, which is 0.15.

The final model that we arrived at through this analysis includes five loci on chromosomes 1, 4, 6, 7, and 15. The localization information for these loci is given in Figure 7. The main effects for QTL explain, respectively, 6, 13.5, 3, 0, and 5.5% of the variance. There are two two-way interactions in the model. The interaction between QTL on chromosomes 6 and 15 explains ∼6% and the one between QTL on chromosomes 7 and 15 explains ∼6.5% of the variance. Taken together the model explains ∼41% of the variance. The effect sizes are summarized in Figure 8.

DISCUSSION

We analyzed the stochastic dependencies between the different observed and unobserved data structures in the QTL mapping problem. By conditioning on the unobserved QTL genotypes it is possible to decompose the QTL mapping problem into the linkage part and the genetic model part. This decomposition is used to develop a computational strategy that uses multiple imputations of a pseudomarker grid to approximate integrals needed to perform Bayesian inference. Our software provides a set of flexible extensible tools suitable for the analysis of QTL data.

The problem of model selection remains a thorny issue in theoretical statistics and presents a serious challenge in the analysis of QTL data. For this reason we emphasized exploratory tools over formal model selection procedures in our analysis. We provide a visual representation of two-QTL models and two-way interactions in the QTL mapping problem.

Most of the QTL analysis methods proposed so far have relied on linear (or generalized linear) regression models. Recently, classification and regression trees (CART) models were proposed for QTL data (T. Speed, personal communication) because they have a richer interaction space. We note here that by using appropriate weight functions, $W_H(y, g)$, where the model, H, can be a regression tree, we can include CART in our framework.

Multiple-QTL models were considered by Haley and Knott (1992), who presented a simple regression approximation to the LOD score. The LOD score, the Haley-Knott approximation, and the LPD will coincide at locations where there is complete genotype information. At locations with incomplete genotype information, the three versions of the LOD score will differ. This can be seen in Figure 2. The regression approximation to the LOD score deviates substantially from the LOD score and the LPD in regions where the proportion of missing information is high. The bias in the Haley and Knott (1992) approximation to the LOD score has been investigated in detail by Kao (2000).

We have already mentioned the close connection between the LOD score and the LPD. Notwithstanding their similarities, the two can be numerically divergent. There is an important *conceptual* difference between the LOD score and the LPD. The LOD score is designed to be a *test for linkage* wherein the location of the QTL (γ) is a nuisance parameter and μ is the parameter of interest. Indeed, the *maximum* of the LOD score is used as the test statistic for linkage. The LPD is designed with the reciprocal purpose of *localization* where the location of the QTL (γ) is the parameter of interest. The parameter μ is the nuisance parameter and is integrated out.

The Monte Carlo error in approximating the posterior distributions by the pseudomarker algorithm is inversely proportional to the square root of the number of imputations. Unlike MCMC methods, we do not have to worry about proper mixing of the Markov chains. Accuracy of the calculations also depends on the density of the pseudomarker grid. For initial genome scans we find that 20 pseudomarker sets at a spacing of ∼10 cM are adequate. For fine mapping, we recommend a 2-cM pseudomarker map on selected chromosomes with ~200 replicated data sets. Adjustment for individual situations is quite easy. More replications may be required when the proportion of missing genotypes is high. The investigator will have to vary the density of the pseudomarker map and the number of replicated data sets according to his/her needs. The best way to know what is reasonable for the map is to start with a sparse pseudomarker set and repeat the analysis. If the pictures differ markedly, more replications are needed. To control Monte Carlo fluctuations we also adopted some additional devices explained in greater detail in APPENDIX F. Another option to controlling Monte Carlo noise (which we have not currently implemented) would be to use smoothing techniques on the LPD since it is approximately piecewise quadratic (KONG and WRIGHT 1994). The variation in the imputed genotypes at a pseudomarker location can be used as an estimate of the amount of missing genotype data for that location. This may be used to decide if more genotyping needs to be done in a particular region of the genome; we implemented a graphical diagnostic tool called PLOTMISSINGPROP to identify regions of the genome that might benefit from additional genotyping.

We presented a QTL mapping framework in the context of analyzing data from inbred line crosses. However, the idea of using multiple imputation of complete genotype information is applicable to more general gene-mapping situations, including complex human pedigrees. How it performs in practice requires further investigation.

Finally, it is pertinent to point out some of the inherent limitations of QTL mapping studies. The number of QTL that we can reliably detect in a cross depends on the number of individuals available, the numbers of segregating QTL, and the strength of their effects. If there are interactions between QTL that are ignored, then we may fail to find those QTL. For example, with 100 progeny from a backcross, we can expect to observe ~25 individuals in each genotype class for two unlinked QTL. With three unlinked QTL, the number is ~12. If the three QTL are linked, some genotype classes may be unrepresented in the data or may have very small numbers. Thus, it may be impossible to detect and characterize interactions of order 3 or higher and genes that act through high-order interactions may not be detected. It has been shown that model misspecification can lead to misleading results (WRIGHT and KONG 1997). We also acknowledge that the localization of QTL in any cross of a reasonable size is limited by the number of crossover events. QTL mapping studies will generally not be adequate to identify the polymorphic genes or regulatory regions that are affecting a trait of interest. The availability of complete genome sequences will expand the list of candidate genes for a QTL and will facilitate efforts to identify the genes responsible for the observed effects. Despite these limitations, we feel that QTL mapping studies will continue to contribute to our understanding of the nature of quantitative trait genetics. However, to achieve this goal we must acknowledge the complexity of quantitative inheritance, which is often determined by multiple interacting genetic loci, and we must develop and apply analytic methods that are appropriate for this problem.

We thank Drs. B. Paigen, F. Sugiyama, and K. Broman for sharing data with us. This work has benefited enormously from the critical comments of Dr. K. Broman and two anonymous referees. The American Heart Association provided financial support through a grant to G.A.C.

LITERATURE CITED

BROMAN, K. W., and T. SPEED, 1999 A review of methods for identifying QTLs in experimental crosses, pp. 114–142 in *Statistics in Genetics and Molecular Biology*, Vol. 33 of *IMS Lecture Notes—Monograph Series*, edited by F. SEILLER-MOISEIWITSCH. Institute of Mathematical Statistics, Hayward, CA.

BROMAN, K. W., V. L. BOYARTCHUK and W. F. DIETRICH, 2000 Mapping time-to-death quantitative trait loci in a mouse cross with high survival rates. Technical Report MS00-04, Department of Biostatistics, Johns Hopkins University, Baltimore.

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

DARVASI, A., 1998 Experimental strategies for the genetic dissection of complex traits in animal models. Nat. Genet. **18:** 19–24.

DEMPSTER, A., N. LAIRD and D. RUBIN, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B **39:** 1–22.

DUPUIS, J., and D. SIEGMUND, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics **151:** 373–386.

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

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

HOESCHELE, I., and P. VANRADEN, 1993 Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor. Appl. Genet. **85:** 946–952.

JANSEN, R., 1993 A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. **85:** 252–260.

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

JIANG, C., and Z-B. ZENG, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140:** 1111–1127.

KAO, C.-H., 2000 On the difference between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. Genetics **156:** 855–865.

KAO, C.-H., Z-B. ZENG and R. D. TEASDALE, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

KASS, R., and A. RAFTERY, 1995 Bayes factors. J. Am. Stat. Assoc. **90:** 773–795.

KONG, A., and F. A. WRIGHT, 1994 Asymptotic theory for gene mapping. Proc. Natl. Acad. Sci. USA **91:** 9705–9709.

KOROL, A. B., Y. I. RONIN and V. M. KIRZHNER, 1995 Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics **140:** 1137–1147.

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

LANDER, E., and L. KRUGLYAK, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. **11:** 241–247.

LANDER, E. S., and P. GREEN, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

Lee, P. M., 1997 *Bayesian Statistics*, Ed. 2. Arnold Publishers, London.

Lincoln, S. E., and E. S. Lander, 1992 Systematic detection of errors in genetic linkage data. Genomics **14:** 604–610.

McCullagh, P., and J. Nelder, 1989 *Generalized Linear Models*, Ed. 2. Chapman & Hall, London/New York.

Morton, N., 1955 Sequential tests for the detection of linkage. Am. J. Hum. Genet. **7:** 277–318.

Rapp, J. P., 2000 Genetic analysis of inherited hypertension in the rat. Physiol. Rev. **80:** 131–172.

Ronin, Y. I., A. B. Korol and E. Nevo, 1999 Single- and multiple-trait mapping analysis of linked quantitative trait loci: some asymptotic analytical approximations. Genetics **151:** 387–396.

Rubin, D., 1976 Inference and missing data. Biometrika **63:** 581–592.

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

Sax, K., 1923 The association of size difference with seed-coat pattern and pigmentation in *phaseolus vulgaris*. Genetics **8:** 552–560.

Schafer, J., 1997 *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London/New York.

Schork, N., M. Boehnke and J. Terwilliger, 1993 Two-trait-locus linkage analysis: a powerful strategy for mapping complex genetic traits. Am. J. Hum. Genet. **53:** 1127–1136.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. **6:** 461–464.

Sen, S., 1998 Confidence intervals for gene location: the effect of model misspecification and smoothing. Ph.D. Thesis, Department of Statistics, University of Chicago.

Shepel, L. A., H. Lan, J. D. Haag, G. M. Brasic, M. E. Gheen *et al.*, 1998 Genetic identification of multiple loci that control breast cancer susceptibility in the rat. Genetics **149:** 289–299.

Sillanpää, M., and E. Arjas, 1999 Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics **151:** 1605–1619.

Sugiyama, F., G. A. Churchill, D. C. Higgins, C. Johns, K. P. Makaritsis *et al.*, 2001 Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. Genomics **71:** 70–77.

Thoday, J., 1961 Location of polygenes. Nature **191:** 368–370.

Thompson, E. A., 2000 *Statistical Inference From Genetic Data on Pedigrees*, Vol. 6 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics, Beachwood, OH.

Uimari, P., and I. Hoeschele, 1997 Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics **146:** 735–743.

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

Zeng, Z-B., 1993 Theoretical basis of separation of multiple linked gene effects on 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.

Zhao, H., T. P. Speed and M. S. McPeek, 1995 Statistical analysis of crossover interference using the chi-square model. Genetics **139:** 1045–1056.

Communicating editor: S. Tavaré

APPENDIX A: FACTORIZATION OF THE JOINT DISTRIBUTION

The joint distribution of the observed and missing data structures is

$$\begin{aligned} p(y, m, g, \mu, \gamma) &= p(y|m, g, \mu, \gamma)p(m, g, \mu, \gamma) \\ &= p(y|g, \mu)p(g|m, \mu, \gamma)p(m, \mu, \gamma) \\ &= p(y|g, \mu)p(g|m, \gamma)p(m, \mu, \gamma) \\ &= (p(y|g, \mu)p(\mu))(p(g|m, \gamma)p(m)p(\gamma)). \qquad \text{(A1)} \end{aligned}$$

The first equality is an application of the definition of conditional probability and the second follows from the assumption that the phenotype depends on the location of the gene only through the genotype of the QTL and the genetic model. The third equality follows from the assumption that the distribution of the QTL genotype depends on the genetic model only through the position of the QTL relative to the markers. In the final equality we assume that the distribution of the marker genotypes is independent of the location of the QTL and the disease model and regroup the terms. This final assumption will be true if there is no ascertainment. If individuals have been selected based on their phenotype and the phenotypes of nonselected individuals are not included in the analysis this result will not hold.

APPENDIX B: POSTERIOR DISTRIBUTIONS

An expression for the distribution of the QTL genotypes given the observed data can be derived from the full joint distribution using elementary operations as

$$\begin{aligned} p(g|m, y) &\propto \iint p(y, m, g, \mu, \gamma)\, d\mu d\gamma \\ &= \int\left[\int p(y|g, \mu)p(\mu)\, d\mu\right] \\ &\quad \times \left[p(g|m, \gamma)p(m)p(\gamma)\right] d\gamma \\ &\propto \int p(y|g)p(g|m, \gamma)p(\gamma)\, d\gamma \\ &= p(y|g)p(g|m). \qquad \text{(B1)} \end{aligned}$$

The expression (B1) suggests that we obtain a representation of the posterior distribution of g by sampling from $p(g|m, \gamma)$ and then weighting the samples by $W(g) = p(y|g)$. The weights are importance sampling weights and the idea is that $p(g|m)$ has a flatter distribution than $p(g|m, y)$ and therefore is a good proposal distribution to use. Summing the weights is a numerical approximation to integration with respect to $p(\gamma)d\gamma$. It is possible to use an irregularly spaced list of γ values. In that case we will have to weigh each point suitably as in a numerical trapezoidal integration formula. A numerical integration over the QTL location space is reasonable because the shape of the log-likelihood in large samples is piecewise quadratic and hence smooth (Kong and Wright 1994).

The posterior distribution of the QTL locations is

$$\begin{aligned} p(\gamma|m, y) &\propto \iint p(y, m, g, \mu, \gamma)\, dg d\mu \\ &= \iint p(y|g, \mu)p(\mu)p(g|m, \gamma)p(m)p(\gamma)\, dg d\mu \\ &= \int\left[\int p(y|g, \mu)p(\mu)\, d\mu\right]p(g|m, \gamma)p(\gamma)p(m)\, dg \qquad \text{(B2)} \\ &\propto \int p(y|g)p(g|m, \gamma)p(\gamma)\, dg \qquad \text{(B3)} \\ &\approx \sum_{i=1}^{q} W_H(r_i(\gamma))p(\gamma). \qquad \text{(B4)} \end{aligned}$$

The last step, summation of the weights at location γ, is a Monte Carlo approximation to the integration with respect to dg. If a flat prior on γ is used it reduces to a plain sum; otherwise it is a weighted sum. Note that γ and $r_i(\gamma)$ will have dimensions $1 \times p$ and $p \times n$ if there are p QTL in the model.

The posterior distribution of the model parameters, μ, is

$$\begin{aligned} p(\mu|m, y) &= \iint p(\mu, \gamma, g|m, y)\, dg d\gamma \\ &= \iint p(\mu, \gamma|g, m, y) p(g|m, y)\, d\gamma\, dg \\ &= \int\left(\int p(\gamma|g, m, y)\, d\gamma\right) p(\mu|g, m, y) p(g|m, y)\, dg \\ &= \int p(\mu|g, y) p(g|m, y)\, dg \\ &\approx \sum_{i=1}^{q}\sum_{u} W_H(r_i(u)) p(\mu|g = r_i(u)). \qquad \text{(B5)} \end{aligned}$$

This is a weighted average of the complete data posterior densities $p(\mu|g, y)$ over the unknown genotypes. In practice we do not use this density directly. Instead we compute and report its mean and variance. For large samples the posterior density will be approximately normal and confidence regions for posterior means can be computed in the usual way.

The posterior mean, $E(\mu|m, y)$, is the mean of the conditional mean of $E(\mu|y, m, g)$ averaged, over $p(g|m, y)$. This is achieved by taking the samples of g weighted with $p(y|g)p(\gamma)$ as in the previous section. For each of the samples, we calculate the posterior distribution $p(\mu|y, g)$. This is then averaged with weights for the sampled g. To compute a posterior mean, we use the iterated expectation formula

$$\begin{aligned} \mathbf{E}(\mu|y, m) &= \mathbf{EE}(\mu|y, g, m) \\ &= \int \mathbf{E}(\mu|g, y) p(g|y, m)\, dg \\ &\approx \sum_{i=1}^{q}\sum_{u} \mathbf{E}(\mu|y, g = r_i(u)) W_H(r_i(u)). \qquad \text{(B6)} \end{aligned}$$

The summation over locations, indexed by u, requires a bit of attention. In principle, for a model with p QTL, the summation on u should run through all p-tuples in the genome. However when $p > 3$ this is not practical. To reduce the computation and thus admit larger models, we restrict the summation to one chromosome at a time or to pairs of chromosomes in the case of unlinked interacting QTL. When we focus on the estimation of effects of a subset of QTL, the others may be fixed by including linked markers as covariates or they may simply be ignored. The latter approach works because unlinked QTL are approximately orthogonal to one another in the genotype space as a result of the random segregation of chromosomes. This obviously does not apply to linked QTL. Although this is a departure from the strict Bayesian approach, the effect is minimal because the QTL location densities tend to be highly concentrated on specific chromosomes (or chromosome pairs).

Standard errors of estimated QTL effects can be obtained as the square root of the posterior variance. Again we use the concept of iterated expectation. Thus

$$\begin{aligned} \mathbf{V}(\mu|y, m) &= \mathbf{EV}(\mu|g, y, m) + \mathbf{VE}(\mu|g, y, m) \\ &\approx \sum_{i=1}^{q}\sum_{u} W_H(r_i(u)) V(\mu|y, g = r_i(u)) \\ &\quad + \sum_{i=1}^{q}\sum_{u} W_H(r_i(u)) (E(\mu|y, g = r_i(u)) \\ &\quad - E(\mu|y, m))^2. \qquad \text{(B7)} \end{aligned}$$

Similar comments on the summation with respect to u apply here. Despite the intimidating appearance of these expressions, the computation is quite simple. The variance estimate is composed of two terms. The first term is the weighted mean of the conditional variances $V(\mu|y, m, g)$. The second term is the weighted variance of the conditional means $E(\mu|y, m, g)$.

APPENDIX C: DERIVATION OF WEIGHTS FOR NORMAL DATA

Suppose we have n observations and that the relationship of y given x, g, and μ is described by a normal linear regression model. Let X be the $n \times p$ model matrix corresponding to the model and let $\mu = (\beta, \phi)$ be the parameters in the model.

$$y|X, \mu \sim N(X\beta, \phi I_n). \qquad \text{(C1)}$$

Assume conjugate priors for β and ϕ. We assume that the prior distribution of β conditional on ϕ is normally distributed with prior mean β_0 and variance ϕQ^{-1}. The prior distribution of ϕ is inverse χ^2 with parameters S_0 and ν_0. The interpretations of S_0 and ν_0 are that they are the prior error sum of squares and prior degrees of freedom, respectively. For Bayesian calculations of this type see LEE (1997).

Thus we have

$$p(y|X, \beta, \phi) = (2\pi\phi)^{-n/2}\exp\left(-\frac{1}{2\phi}(y - X\beta)'(y - X\beta)\right)$$

$$p(\beta|\phi) = (2\pi\phi)^{-p/2}\det(Q)^{1/2}\exp\left(-\frac{1}{2\phi}(\beta - \beta_0)'Q(\beta - \beta_0)\right)$$

$$p(\phi) = \Gamma(\nu_0/2)^{-1}\left(\frac{1}{2}S_0\right)^{\nu_0/2}\exp\left(-\frac{1}{2\phi}S_0\right)\phi^{-\nu_0/2-1}.$$

Hence

$$p(y, \beta, \phi|X) = (2\pi\phi)^{-n/2-p/2}\det(Q)^{1/2}\Gamma(\nu_0/2)^{-1}\left(\frac{1}{2}S_0\right)^{\nu_0/2} \times \exp\left(-\frac{1}{2\phi}S\right), \qquad \text{(C2)}$$

where

$S = S(\beta) = (y - X\beta)'(y - X\beta) + (\beta - \beta_0)'Q(\beta - \beta_0) + S_0.$

It can be shown that

$$S(\beta) = S(\hat{\beta}) + (\beta - \hat{\beta})'(X'X + Q)(\beta - \hat{\beta}),$$

where $\hat{\beta} = (Q + X'X)^{-1}(X'y + Q\beta_0)$. Hence integrating (C2) with respect to β we get

$$p(y, \phi|X) = \frac{(2\pi)^{-n/2}}{\Gamma(\nu_0/2)}\left(\frac{\det(Q)}{\det(Q + X'X)}\right)^{1/2}\left(\frac{1}{2}S_0\right)^{\nu_0/2} \times \exp\left(-\frac{1}{2\phi}S_1\right)\phi^{-\nu_1/2-1}, \tag{C3}$$

where $\nu_1 = \nu_0 + n$ and $S_1 = S(\hat{\beta})$. Integrating (C3) with respect to ϕ we get

$$p(y|x, g) = p(y|X) = (2\pi)^{-n/2}\frac{\Gamma(\nu_1/2)}{\Gamma(\nu_0/2)}\left(\frac{\det(Q)}{\det(Q + X'X)}\right)^{1/2}\frac{(S_0/2)^{\nu_0/2}}{(S_1/2)^{\nu_1/2}}. \tag{C4}$$

This is the weight function we use for normally distributed phenotypes. Note that we can interpret S_1 as the *posterior* residual sum of squares and ν_1 as the *posterior* degrees of freedom. If we make the additional assumption that the prior variance matrix of β is proportional to $(X'X)^{-1}$, *i.e.*, $Q = \alpha X'X$ for some α, then the above is proportional to

$$\left(\frac{\alpha}{1 + \alpha}\right)^{p/2} S_1^{-\nu_1/2} \simeq \left(\frac{\alpha}{1 + \alpha}\right)^{p/2} \mathrm{RSS}^{-n/2}$$

(since n, ν_1, ν_0, and S_0 do not depend on the model matrix X), for large sample sizes, where RSS is the residual sum of squares of the regression of y on X. Additionally, if we make the assumption that

$$\frac{\alpha}{1 + \alpha} \simeq \frac{1}{n},$$

which says that the posterior variance matrix of β is approximately n^{-1} times the prior variance matrix of β, then the weight function becomes approximately $\mathrm{RSS}^{-n/2}n^{-p/2}$.

Note that on the log scale this is identical to the BIC of Schwarz (1978) given by

$$\mathrm{BIC} = \log(\mathrm{RSS}) - p\log(n)/n.$$

Broman and Speed (1999) recommend using a modified BIC-type criterion of the form

$$\mathrm{BIC}_\delta = \log(\mathrm{RSS}) - \delta p\log(n)/n,$$

for some δ that they pick to be somewhere between 2 and 3. Note that this corresponds to weak prior information of the order of $n^{-\delta}$. Another interpretation would be to suggest that the higher penalty is a correction for multiple testing.

Similar calculations using an inverse Wishart distribution as the prior for the covariance matrix lead to a weighting function for the multivariate normal phenotype,

$$(\det(S))^{-n/2}n^{-pt/2},$$

where t is the number of phenotypic traits being measured and S is the residual sum of squares and products matrix (analogous to the residual sum of squares for the univariate case).

APPENDIX D: LOD SCORE AND THE LPD

The LOD score at a given location of the QTL is defined as

$$\begin{aligned}\mathrm{LOD}(\gamma) &= \log_{10}\left(\frac{\sup_\mu p(y, m|\mu, \gamma)}{\sup_\mu p(y, m|\mu = 0, \gamma)}\right)\\ &= \text{constant} + \log_{10}\left(\sup_\mu p(y, m|\mu, \gamma)\right)\\ &= \text{constant} + \log_{10}\left(\sup_\mu \int p(y|g, \mu)p(g|m, \gamma)\,dg\right),\end{aligned} \tag{D1}$$

where $\mu = 0$ is understood to mean the case when there is no effect of the genotype on the phenotype. This definition generalizes that proposed by Lander and Botstein (1989) and is equivalent the LOD score for MIM (Kao *et al.* 1999).

The LPD is defined as the logarithm (base 10) of the posterior distribution of the QTL location. It is, like the LOD score, a function of location and is defined for both single and multiple QTL models. Using (B2) and (B3) it is equal to

$$\begin{aligned}\mathrm{LPD}(\gamma) &= \log_{10}(p(\gamma|y, m))\\ &= \text{constant}\\ &\quad + \log_{10}\left(\iint p(y|g, \mu)p(g|m, \gamma)p(\gamma)p(\mu)\,dg\,d\mu\right)\\ &\approx \text{constant} + \log_{10}\left(\sum_{i=1}^{q} W_H(r_i(u))\right).\end{aligned} \tag{D2}$$

The LPD replaces the supremum operation in the definition of the LOD score with an integration. In practice, the posterior distribution of the QTL effects is sufficiently sharp that there is little difference in the two.

APPENDIX E: BAYES FACTORS

To calculate Bayes factors, we need to calculate under each model, H,

$$\begin{aligned}p_H(y, m) &= \iiint p_H(y, m, g, \mu, \gamma)\,dg\,d\mu\,d\gamma\\ &= \iiint \left(p_H(y|g, \mu)p(\mu)\right)\left(p(g|m, \gamma)p(m)p(\gamma)\right)\\ &\quad \times dg\,d\mu\,d\gamma\end{aligned}$$

$$= \int\left(\left(\int p_H(y|g,\ \mu)p(\mu)\,d\mu\right)\left(\int p(g|m,\ \gamma)p(\gamma)\,dg\right)\right.$$

$$\left.\times\ d\gamma p(m)\right)$$

$$= \left(\int\int p_H(y|g)p(g|m,\ \gamma)p(\gamma)dgd\gamma\right)p(m) \qquad \text{(E1)}$$

$$\approx \sum_u \sum_{i=1}^{q} W_H(r_i(u)). \qquad \text{(E2)}$$

Since $p(m)$ is the same no matter what genetic model we use, we have to calculate only the term inside the brackets, which is just the average of the weights for all the samples g drawn from $p(g|m,\ \gamma)$ for a regularly spaced grid of locations.

APPENDIX F: NUMERICAL ISSUES

Recall from (4) that the posterior density of the QTL at a particular pseudomarker location is obtained by averaging the weights at the pseudomarker. These weights tend to be very highly skewed in practice, which means that occasionally some extreme-valued weights can completely distort the average value. To prevent this, we decided to use a robustified version of the average. It is based on the observation that even though the weights are quite skewed, on the log scale the weights are more or less symmetrically distributed, albeit somewhat more heavily tailed than the Gaussian distribution. If we pretend that the weights, W, are lognormally distributed, *i.e.*, $L = \log(W)$ is normally distributed with some mean μ and variance σ^2, then

$$E(W) = E(\exp(L)) = \exp\left(\mu + \frac{1}{2}\sigma^2\right)$$

$$> \exp(\mu) = \exp(E(L)).$$

This means that we can estimate the mean of W by estimating the mean and variance of $\log(W)$. We do just that except that we estimate the mean and variance by throwing out $\log_2(q)$ weights that are most extreme. As $q \uparrow \infty$ the proportion of weights discarded goes to 0. If 16 pseudomarker realizations are used, we discard four weights (the two largest and the two smallest). If 256 pseudomarker imputations are used we discard the eight most extreme weights.