Heredity (2008) 100, 240–252

www.nature.com/hdy

SHORT REVIEW

Advances in Bayesian multiple quantitative trait loci mapping in experimental crosses

N Yi and D Shriner
Section on Statistical Genetics, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, AL, USA

Many complex human diseases and traits of biological and/or economic importance are determined by interacting networks of multiple quantitative trait loci (QTL) and environmental factors. Mapping QTL is critical for understanding the genetic basis of complex traits, and for ultimate identification of genes responsible. A variety of sophisticated statistical methods for QTL mapping have been developed. Among these developments, the evolution of Bayesian approaches for multiple QTL mapping over the past decade has been remarkable. Bayesian methods can jointly infer the number of QTL, their genomic positions and their genetic effects. Here, we review recently developed and still developing Bayesian methods and associated computer software for mapping multiple QTL in experimental crosses. We compare and contrast these methods to clearly describe the relationships among different Bayesian methods. We conclude this review by highlighting some areas of future research.
Heredity (2008) **100,** 240–252; doi:10.1038/sj.hdy.6801074; published online 7 November 2007

Keywords: Bayesian methods; complex traits; experimental crosses; Markov chain Monte Carlo algorithms; quantitative trait loci

Introduction

The variation of most complex traits results from interacting networks of multiple quantitative trait loci (QTL) and environmental factors (Reifsnyder *et al*., 2000; Carlborg and Haley, 2004; Moore 2005; Stylianou *et al*., 2006; Valdar *et al*., 2006; Wang *et al*., 2006). The main goal of mapping QTL is to find regions or loci of a genome that are strongly associated with a phenotype measured in an experimental cross or other types of segregating populations. This is largely a model selection issue (Broman and Speed, 2002; Sillanpää and Corander, 2002; Yi, 2004): what is the genetic architecture, in terms of genomic regions, gene action and possible interactions, that is best supported by the data? Identification of multiple interacting QTL has been a formidable challenge for geneticists and statisticians, mainly due to numerous possible variables associated with hundreds or thousands of genomic loci (markers and/or loci within marker intervals) that lead to a huge number of possible models (for example, Yi *et al*., 2005). The problem is further complicated by the facts that the genomic loci on the same chromosome are highly correlated and the genotypes at many loci are unobserved.

Non-Bayesian QTL mapping approaches have dominated QTL mapping theory and practice for most of the past two decades. Traditional non-Bayesian QTL mapping methods utilize pre-specified simple statistical models, which fit the effects of only one QTL whose putative position is scanned across the genome (for example, Lander and Botstein, 1989; Zeng, 1994; Jansen and Stam, 1994). Extensions of this approach can allow for main and epistatic effects at two or perhaps a few QTL at a time and employ a multidimensional scan to detect QTL. Rather than fitting pre-specified models to the observed data, model selection approaches proceed by identifying from a set of potential models the subset of models that are best supported by the data. Various model selection methods for multiple QTL mapping have been recently proposed from both non-Bayesian and Bayesian perspectives. Non-Bayesian approaches sequentially add or delete QTL using forward or stepwise selection procedures and apply criteria such as *P*-values or a modified Bayesian information criterion to identify the 'best multiple QTL model' (Kao *et al*., 1999; Carlborg *et al*., 2000; Reifsnyder *et al*., 2000; Zeng *et al*., 2000; Bogdan *et al*., 2004; Baierl *et al*., 2006).

Our emphasis in this review is on the application of Bayesian methodology and its related algorithms in multiple QTL mapping. Emergence of the Bayesian approach has been driven by not only the availability of new and powerful computational techniques but also the pragmatic advantages of the Bayesian framework. In Bayesian analysis, a comprehensive probabilistic model is employed to describe relationships among observed (data and knowledge) and unobserved (parameters and hypotheses) quantities (Carlin and Louis, 2000; Gelman *et al*., 2004). Inference is then based on the conditional distribution of the unknowns, given the observed data. The Bayesian paradigm has inherent flexibility and generality, which in principle allows it to cope with models with virtually arbitrary complexity. The Bayesian approach can fully take into account the uncertainties associated with all unknowns. Inferences about any particular parameter of interest can be obtained by

Correspondence: Professor N Yi, Department of Biostatistics, The University of Alabama at Birmingham, Birmingham, AL 35294-0022, USA.
Email: nyi@ms.soph.uab.edu
Received 14 May 2007; revised 30 August 2007; accepted 31 August 2007; published online 7 November 2007

averaging over possible models, rather than using a single selected model. Therefore, Bayesian methods can provide more robust inferences than non-Bayesian methods. Another attractive feature of Bayesian analysis is the ability to incorporate prior information into the specification of the model.

Applied to multiple QTL analysis, the Bayesian framework can treat the number of QTL (and thus the size or dimensionality of the model) as an unknown, and can simultaneously model main effects of QTL, environmental factors, gene–gene interactions (epistatic effects) and gene–environment interactions (G × E) (Yi *et al*., 2005, 2007b). Uncertainties about unobserved quantities are built directly into the Bayesian QTL model. That is, we can assign a reasonable prior to each unobserved quantity before observing any phenotypes. Assumptions and information from previous studies can be incorporated into model priors about the shape of the distribution of phenotypic values, the relative importance of different regions of the genome, and the likely number and pattern of genomic regions that might be detected.

Over the past decade, a variety of Bayesian methods have been developed to map multiple QTL for complex traits in experimental crosses (Satagopan and Yandell, 1996; Satagopan *et al*., 1996; Sillanpää and Arjas, 1998; Stephens and Fisch, 1998; Hoeschele, 2001; Sen and Churchill, 2001; Gaffney, 2001; Yi and Xu, 2002; Xu, 2003; Yi *et al*., 2003a, b, 2005; Narita and Sasaki, 2004; Wang *et al*., 2005; Zhang *et al*., 2005). In this review, we describe these recently developed and still developing Bayesian multiple QTL mapping methods. We compare and contrast these methods to clearly describe the relationships among different Bayesian methods. We illustrate improvements in QTL mapping using Bayesian vs frequentist methods using hypertension data from a murine backcross (Sugiyama *et al*., 2001). We conclude this review by highlighting some areas of future research.

QTL data structure and notation

Experimental crosses for QTL mapping are usually derived from two inbred parental lines. Parental lines are first crossed to produce a hybrid F_1 generation. Subsequent segregating generations are obtained by selfing, sib-mating or backcrossing to the parental lines. Observed data in QTL mapping consist of phenotypic values of a complex trait and molecular marker data. We denote the phenotypic values by the vector $y=(y_1, y_2, \ldots, y_n)^{\mathrm{T}}$, and the marker data by the $n \times k$ matrix $m=(m_{ij})$, where n and k represent the number of individuals and markers, respectively, in the mapping population. This review does not address the problem of building marker linkage maps and assumes that the marker linkage map has been built before QTL mapping. The observed marker data include not only the marker genotypes but also the genomic positions of the markers. Besides phenotypic values and marker data, most QTL studies usually measure some discrete or continuous environmental factors that may affect the phenotype. We use the term covariates synonymously with environmental factors, and denote their observed values by the matrix X_{E}.

The marker data provide information about the segregation of alleles at various genomic positions in a mapping population. When the markers are densely and regularly spaced, we restrict possible QTL to the marker positions; otherwise, we insert some loci (called pseudomarkers) between flanking markers separated by some minimal distance, say 1 cM, and assume that possible QTL occur at the genotyped markers and the un-genotyped pseudomarkers (Sen and Churchill, 2001; Wang *et al*., 2005; Yi *et al*., 2005). Inserting pseudomarkers into marker intervals enables us to detect potential QTL within the marker intervals, similar to the idea of traditional interval mapping and composite interval mapping methods (Lander and Botstein, 1989; Haley and Knott, 1992; Zeng, 1994). However, inserting pseudomarkers introduces a special statistical problem, that is, QTL genotypes are generally unobserved and thus are missing data. Unobserved QTL genotypes play a central role in modeling the observed phenotype data. We denote the QTL genotypes by the $n \times l$ matrix $g=(g_{iq})$, where n and l represent the number of individuals and QTL, respectively.

QTL mapping is the process of inferring the number of QTL, their genomic positions, and the activity and size of the associated genetic effects, given *the observed data* (y, m, X_{E}). Genetic effects include main effects of QTL as well as gene–gene and gene–environment interactions. The activity of a genetic effect refers to its inclusion or exclusion from the model and will be described in detail later. We organize the number of QTL, the positions of QTL and the activity of the genetic effects into the vector H. The vector H comprises a model index that identifies the genetic architecture of the trait. As can be seen in the next section, the statistical models and methods for Bayesian QTL mapping are largely influenced by the specification of H. We use the vector θ to include the corresponding environmental effects, genetic effects and other model parameters (for example, overall mean, residual variance, etc.). Therefore, the *unobserved quantities* in QTL mapping include θ, g and H.

Basic principles of Bayesian multiple QTL mapping

Bayesian analysis refers to statistical methods for making inferences from data using probability models for quantities we observe and for quantities about which we wish to learn. The full process of a typical Bayesian analysis can be idealized by division into the following main steps (Gelman *et al*., 2004): (1) setting up a full probability model, the *joint probability distribution*, that captures the relationships among all observable and unobservable quantities (modeling); (2) calculating the appropriate *posterior distribution*, the conditional distribution of the unobserved quantities of ultimate interest, given the observed data (computation) and (3) summarizing and interpreting the posterior distribution (posterior inference).

In Bayesian QTL mapping, the joint probability distribution of observed phenotypes y and unobserved quantities (θ, g, H) can be expressed as

$$p(y,\theta,g,H) = p(y|X_{\mathrm{E}},\theta,g,H)\cdot p(g|m,H)\cdot p(\theta,H) \qquad (1)$$

where $p(y \mid X_{\mathrm{E}}, \theta, g, H)$ is the *likelihood function* or the *sampling distribution* of phenotypes conditional on all the unknowns, $p(g \mid m, H)$ is the conditional probability of QTL genotypes, given the observed marker data and the QTL positions, and $p(\theta, H)$ is the prior distribution of the parameters.

Conditioning on the observed data (y, m, X_E), using the basic property of conditional probability known as Bayes' rule, yields the *joint posterior distribution*:

$$p(\theta, g, H|y, m, X_E) = \frac{p(y, \theta, g, H)}{p(y)} \propto p(y|X_E, \theta, g, H) \cdot p(g|m, H) \cdot p(\theta, H) \quad (2)$$

where the denominator $p(y)$ is the *marginal likelihood* of the model, which does not depend on the unknown quantities (θ, g, H) and thus can be omitted from the joint posterior distribution, yielding the *unnormalized posterior distribution* (the right side of (2)). This joint posterior distribution contains all information about the genetic architecture of the phenotype.

From Equations ((1) and (2), the first challenge of Bayesian QTL analysis is to develop the likelihood function $p(y \mid X_E, \theta, g, H)$, the conditional probability of QTL genotypes $p(g \mid m, H)$ and the prior distribution $p(\theta, H)$, which must effectively capture the key features of the underlying scientific problem. The second challenge is to develop efficient algorithms to calculate the posterior distribution $p(\theta, g, H \mid y, m, X_E)$. Great advances addressing these two major challenges have been made in the past decade, and many of these are reviewed throughout the article.

The likelihood function $p(y \mid X_E, \theta, g, H)$ specifies the distribution of phenotypes, given the QTL genotypes, the genetic effects, the covariates and other model parameters, involving the problems such as how many loci we should include in the model, and whether or not we simultaneously model main effects of QTL, gene–gene interactions (epistatic effects) and gene–environment interactions (G × E effects). The specification of the prior distribution $p(\theta, H)$ is an important part of Bayesian analysis. Indeed, through the prior distribution, we can incorporate prior knowledge and information about the unknown quantities. This is especially important in multiple QTL analysis, since geneticists often have substantial knowledge about the genetic architecture of the phenotype under study. However, formal incorporation of prior knowledge is not trivial.

Evaluating $p(g, \theta, H \mid y, m, X_E)$ is analytically infeasible in multiple QTL mapping and therefore requires computational methodology. Recent advances in computing technology coupled with developments in *Markov chain Monte Carlo* (MCMC) algorithms have opened up new and promising directions for addressing the challenge of sampling from a complicated posterior distribution (for example, George and McCulloch, 1997; Chipman *et al*., 2001; Godsill, 2001). MCMC methods simulate a Markov chain $\{(\theta, g, H)^{(t)}; t = 1, 2, \ldots, T\}$, which are random samples (called *posterior samples*) from $p(g, \theta, H \mid y, m, X_E)$. The posterior samples contain all of the information about the joint posterior distribution and thus can be used to infer the genetic architecture of the phenotype. Summarizing and interpreting the posterior samples pose another challenge, however, since the joint posterior distribution includes a huge number of parameters.

R/qtlbim: QTL Bayesian interval mapping

A variety of Bayesian methods for mapping multiple QTL are available. It is important that Bayesian methods be easily accessible to scientists through user-friendly software. Yandell *et al*. (2007) developed a comprehensive package, called R/qtlbim, implementing several Bayesian multiple QTL mapping methods in experimental crosses (www.qtlbim.org). R/qtlbim is implemented as an add-on package for the freely available and widely used statistical language/software R (R Development Core Team, 2006), and provides an extensible, interactive environment for Bayesian analysis of multiple QTL. It is built on the widely used R/qtl framework (Broman *et al*., 2003), and includes all its advantages for extensibility. Computationally intensive algorithms are written in C, with data manipulation and graphics in R. R/qtlbim is available across Window, Linux and MacOS platforms and accepts a variety of input formats via R/qtl.

R/qtlbim can simultaneously handle arbitrary covariates, gene–gene interactions and gene–environment interactions, and can analyze not only continuous traits but also binary and ordinal traits. It includes several efficient MCMC algorithms for generating posterior samples from the joint posterior distribution of unknown quantities, provides extensive informative graphical and numerical summaries, and provides model selection and convergence diagnostics of the posterior samples.

The implementation of R/qtlbim includes the full process of a typical Bayesian analysis, (a) setting up the conditional probability of QTL genotypes $p(g \mid m, H)$, the likelihood function of phenotypes $p(y \mid X_E, g, \theta, H)$ and the prior distribution of the parameters $p(\theta, H)$; (b) generating samples from the joint posterior distribution $p(g, \theta, H \mid y, m, X_E)$ and (c) graphically and numerically summarizing the posterior samples and inferring the genetic architecture of the trait.

In the following sections, we describe Bayesian multiple QTL mapping methods, focusing on the methods that have been implemented in R/qtlbim and discussing other approaches.

Bayesian modeling of multiple QTL

As shown in Equation (1), the joint probability distribution of observed phenotypes y and unobservable quantities (θ, g, H) can be divided into three components: the conditional probability of QTL genotypes $p(g \mid m, H)$, the likelihood function $p(y \mid \theta, g, H)$ and the prior distribution $p(\theta, H)$. Bayesian modeling of multiple QTL requires specification of these three components. In this section, we review recent developments of the Bayesian multiple QTL modeling and highlight connections and differences in terms of these three specifications.

The conditional probability of QTL genotypes $p(g|m, H)$

For regular experimental designs (for example, F_2, backcross (BC) and recombinant inbred lines (RILs), we can directly calculate the conditional probability distribution of genotypes for any locus, given the observed marker data using multipoint methods (Jiang and Zeng, 1997; Rao and Xu, 1998).

Assume that a locus q (pseudomarker or marker) on chromosome c is located between markers j and $j+1$, and there are k_c ordered markers on chromosome c. We denote the genotype of individual i at locus q by g_{iq}, and the marker data of individual i on chromosome c by m^c.

Then, the multipoint conditional probability of g_{iq} can be computed by

$$p(g_{iq}|m^c) = \frac{\mathbf{1}^{\mathrm{T}} D_1 T_{12} \cdots T_{jq} D_{g_{iq}} T_{q(j+1)} \cdots D_{k_c-1} T_{(k_c-1)k_c} \mathbf{1}}{\sum_{g_{iq}} \mathbf{1}^{\mathrm{T}} D_1 T_{12} \cdots T_{jq} D_{g_{iq}} T_{q(j+1)} \cdots D_{K_c-1} T_{(k_c-1)k_c} \mathbf{1}} \quad (3)$$

The terms in this equation depend on the experimental cross design. For example, for a backcross population, there are two genotypes for any locus. We denote the two genotypes by $\mathrm{b}_q\mathrm{b}_q$ and $\mathrm{B}_q\mathrm{b}_q$ for locus q. In Equation (3), g_{iq} takes $\mathrm{b}_q\mathrm{b}_q$ or $\mathrm{B}_q\mathrm{b}_q$, $\mathbf{1}=(1\ 1)^{\mathrm{T}}$, $D_k=\mathrm{diag}(p(\mathrm{b}_k\mathrm{b}_k)\ p(\mathrm{B}_k\mathrm{b}_k))$, $k=1, 2, \ldots, k_c-1$, $D_{g_{iq}}=\mathrm{diag}(1\,0)$ for $g_{iq}=\mathrm{b}_q\mathrm{b}_q$ or $\mathrm{diag}(0\ 1)$ for $g_{iq}=\mathrm{B}_q\mathrm{B}_q$, and T_{jq} is the genotype transition probability matrix from marker j to locus q, computed as

$$T_{jq} = \begin{pmatrix} 1-r_{jq} & r_{jq} \\ r_{jq} & 1-r_{jq} \end{pmatrix},$$

with r_{jq} being the recombination ratio between marker j and locus q. The transition probability matrix for two markers is similarly defined. Note that when a marker is fully informative, each genotype is uniquely identified and thus $p(\mathrm{b}_k\mathrm{b}_k)$ and $p(\mathrm{B}_k\mathrm{b}_k)$ equal 1 or 0. On the other hand, if a marker is non-informative or missing, $p(\mathrm{b}_k\mathrm{b}_k)$ and $p(\mathrm{B}_k\mathrm{b}_k)$ equal 0.5.

By using Equation (3), we can calculate the conditional probabilities of genotypes for all pseudomarkers and markers before QTL mapping (Broman *et al*., 2003; Yi *et al*., 2005). This probability distribution is used as the prior distribution of QTL genotypes.

The likelihood function $p(y|X_{\mathrm{E}}, \theta, g, H)$

The likelihood function $p(y \mid X_{\mathrm{E}}, \theta, g, H)$ specifies the distribution of phenotypes, given the QTL genotypes, the genetic effects, the covariates and other model parameters. For a continuous trait, we usually use a normal linear model to describe the likelihood function. For a binary or ordinal trait, a generalized linear model should be used (Yi and Xu, 2000; Yi *et al*., 2004, 2007a). In this review, we focus on continuous traits. The likelihood function $p(y \mid X_{\mathrm{E}}, \theta, g, H)$ depends on how many loci are included in the model, and whether or not we simultaneously model main effects of QTL, covariates, gene–gene interactions (epistatic effects) and gene–environment interactions ($\mathrm{G}\times\mathrm{E}$ effects). Most of earlier Bayesian multiple QTL mapping methods only considered main effects of multiple QTL (Satagopan and Yandell, 1996; Satagopan *et al*., 1996; Sillanpää and Arjas, 1998; Stephens and Fisch, 1998; Gaffney, 2001; Xu, 2003). Recently, Bayesian methods have been extended to simultaneously include main and epistatic effects of QTL (Yi and Xu, 2002; Yi *et al*., 2003a, b, 2005), and arbitrary covariates and $\mathrm{G}\times\mathrm{E}$ effects (Yi *et al*., 2007b).

Assume that L loci are included in the model. For a continuous trait, the phenotype can be expressed as a linear model

$$y = \mu + X_{\mathrm{G}}\beta_{\mathrm{G}} + X_{\mathrm{GG}}\beta_{\mathrm{GG}} + X_{\mathrm{E}}\beta_{\mathrm{E}} + X_{\mathrm{GE}}\beta_{\mathrm{GE}} + e \triangleq X\beta + e \quad (4)$$

where μ is the overall mean; β_{G} and β_{GG} represent the vectors of all main and epistatic effects associated with L loci, respectively; β_{E} and β_{GE} represent the vectors of environmental effects and gene–environment interactions, respectively; X_{G}, X_{GG}, X_{E} and X_{GE} are the design matrices of effect predictors and e is the vector of independent normal errors with mean zero and variance σ^2_{e}. Model (4) can be equivalently expressed as

$$y|X_{\mathrm{E}}, \theta, g, H \sim N_n(X\beta, \sigma^2_{\mathrm{e}} I) \quad (5)$$

with I being the $n \times n$ identity matrix.

The number of genetic effects (and effect predictors) depends on the experimental design. For a mapping population with $(K+1)$ genotypes per locus, there are K main effects for each locus and K^2 epistatic effects for any two loci. The ith row of $X_{\mathrm{G}}\beta_{\mathrm{G}}$ and $X_{\mathrm{GG}}\beta_{\mathrm{GG}}$ can be expressed as

$$(X_{\mathrm{G}}\beta_{\mathrm{G}})_i = \sum_{q=1}^{L}\sum_{k=1}^{K} x_{ik}^{(q)}\beta_k^{(q)}, \quad \text{and}$$
$$(X_{\mathrm{GG}}\beta_{\mathrm{GG}})_i = \sum_{q<q'}^{L}\sum_{k<k'}^{K} x_{ik}^{(q)} x_{ik'}^{(q')} \beta_{kk'}^{(qq')} \quad (6)$$

where $x_{ik}^{(q)}$ and $\beta_k^{(q)}$ are the main-effect predictors and the main effects of locus q, respectively, and $x_{ik}^{(q)}x_{ik'}^{(q')}$ and $\beta_{kk'}^{(qq')}$ are the epistasis predictors and the epistatic effects between loci q and q', respectively. Effect predictors are determined from the genotypes of locus q by using a particular transformation called a genetic model. A commonly used genetic model is the Cockerham genetic model (Kao and Zeng, 2002; Yi *et al*., 2005; Zeng *et al*., 2005). For a backcross design with two segregating genotypes denoted by $\mathrm{b}_q\mathrm{b}_q$ and $\mathrm{B}_q\mathrm{b}_q$ at locus q, the Cockerham model defines $x_{i1}^{(q)}=z_{iq}-0.5$, where z_{iq} denotes the number of allele B_q. For an intercross (F_2) design with three segregating genotypes denoted by $\mathrm{b}_q\mathrm{b}_q$, $\mathrm{B}_q\mathrm{b}_q$ and $\mathrm{B}_q\mathrm{B}_q$ at locus q, the Cockerham model defines $x_{i1}^{(q)}=z_{iq}-1$ and $x_{i2}^{(q)}=z_{iq}\,(2-z_{iq})-0.5$, respectively. $\beta_k^{(q)}$, for $k=1, 2$, represent additive and dominance effects of locus q, respectively, and $\beta_{kk'}^{(qq')}$, for $k, k'=1, 2$, are called additive–additive, additive–dominance, dominance–additive and dominance–dominance interactions, between loci q and q', respectively.

The environmental term $X_{\mathrm{E}}\beta_{\mathrm{E}}$ is defined as in convenient hierarchical linear models and quantitative genetics models (for example, Gelman *et al*., 2004; Lynch and Walsh, 1998). The gene–environment interaction predictors X_{GE} are formed by multiplying two corresponding predictors X_{G} and X_{E}. In Model (4), we only include those (continuous or discrete) covariates that may be important in understanding the effect of genotype on phenotype in the model (for example, gender, family indicators, locations and some other traits correlated to the phenotype under study). Including relevant covariates can make data collection approximately ignorable or help identify alternate sets of QTL involved in different pathways. We only consider gene–environment interaction terms that are highly probable (for example, gene–sex interactions).

Three ways to deal with unobserved effect predictors

The above phenotype model reveals two special statistical problems in multiple QTL mapping. First, the effect predictors include many missing values because genotypes at all pseudomarkers and at markers with missing values are unobserved. Second, we need to define the number of loci included in the model.

There are three approaches to deal with the problem of unobserved genotypes. All three approaches need the

conditional probability of QTL genotypes $p(g \mid m, H)$. The first approach takes uncertainty of the missing genotypes into account by treating QTL genotypes as unknowns and sampling them in the MCMC update procedure. The second approach, a Bayesian analog to Haley and Knott (1992), replaces all missing genotypes by their expected values conditioning on the observed marker data, and thus essentially removes QTL genotypes $g = (g_{il})$ from the list of unknowns. Although this second method ignores the uncertainty of missing genotypes, which is unwise when the rate of missing genotypes is high (say 20%) or there is selective genotyping (Lander and Botstein 1989), it has a big computational advantage over the first method. These two methods are available in the package R/qtlbim (Yandell *et al*., 2007). The third method, known as multiple imputation, is to sample genotypes from the conditional probability $p(g \mid m, H)$ multiple times for each locus. These multiple imputations are then averaged in a careful way (Sen and Churchill, 2001).

Four ways to specify *L*, the number of loci included in the model

Bayesian QTL mapping methods are largely determined by the specification of L, the number of loci included in the model. There are four ways to specify L, leading to four types of Bayesian QTL mapping methods. As discussed below, these ways affect the definition of the model index H.

Setting up *L* as a small number: Earlier Bayesian QTL mapping approaches were developed to estimate the positions and the effect parameters of multiple QTL based on models with a small number of included loci, say $L = 1$, 2 or 3 (Stephens and Smith, 1993; Uimari *et al*., 1996). With a few included loci, it is possible to evaluate all loci or pairwise combinations across the genome. The advantages of such approaches are simplicity and similarity to traditional QTL mapping methods. Although successful in many applications, such approaches ignore the nature of complex traits in statistical modeling and require prohibitive corrections for multiple testing.

Treating *L* as the number of QTL—the variable dimension model space approach: In QTL mapping studies, the number of QTL is in fact unknown. A natural choice is to directly treat L as the number of QTL (that is, $L = l$), an unknown random variable. In this setting, the model index H includes the number of QTL L and the positions of L QTL denoted by λ. Even with a moderate number of L, the multiple interacting QTL model (4) includes many close-to-zero genetic effects that can be removed from the model. The unknown vector H also includes an additional vector γ of binary indicator variables, indicating inclusion ($\gamma_j = 1$) or exclusion ($\gamma_j = 0$) of each genetic effect associated with the L QTL (Yi *et al*., 2003a, b).

This choice results in an unknown dimension of the parameter space in Model (4), and thus requires MCMC algorithms to sample from the joint posterior distribution of parameters with variable dimension. Green (1995) developed the reversible jump-MCMC (RJ-MCMC) algorithm that can move between spaces of differing dimensions. The RJ-MCMC technique has become a widely used tool in Bayesian multiple QTL mapping (Hoeschele, 2001). Over the past decade, a variety of RJ-MCMC algorithms have been proposed to map multiple non-epistatic QTL (Satagopan and Yandell, 1996; Sillanpää and Arjas, 1998; Stephens and Fisch, 1998; Yi and Xu, 2000; Gaffney, 2001), and epistatic QTL in experimental crosses (Yi and Xu, 2002; Yi *et al*., 2003a, b).

Setting up *L* as a large number and including all possible effects in the model—shrinkage and stochastic search variable selection methods: Xu (2003) proposed a Bayesian hierarchical model in inbred line crosses that simultaneously fits a large number of fixed loci (for example, all observed markers) and always includes all possible main effects, similar to the work of Meuwissen *et al*. (2001) for outbred populations where each locus may have multiple alleles. This approach removes the model index H from the list of unknowns. Wang *et al*. (2005) extended this method to fit a fixed number of loci for each chromosome in a hierarchical model and include the position of each locus as an unknown, thus allowing the possibility of detecting QTL within marker intervals. The key to the success of the above methods is Bayesian hierarchical modeling, that is, each effect is assumed to have its own variance parameter that is estimated from the data. The hierarchical model approach shrinks negligible effects close to zero and is thus able to handle a large number of loci. The key advantage of this shrinkage method is that it is easy to implement MCMC algorithms and it avoids complicated model selection procedures.

An alternative method that always includes all possible effects in the model was proposed by Yi *et al*. (2003b). This method is based on a variable selection method, called stochastic search variable selection (SSVS), developed by George and McCulloch (1993). The difference between SSVS and other variable selection approaches is that the dimensionality is kept constant across all possible models by limiting the posterior distribution of genetic effects for nonsignificant terms in a small neighborhood near zero instead of removing them from the model as is usually done. Due to this unique property, SSVS is able to be easily implemented via MCMC algorithms and can evaluate each effect on the dependent response.

Setting up *L* as the upper bound of detectable QTL and removing small effects from the model—the composite model space approach: Yi (2004) and Yi *et al*. (2005) developed a unified Bayesian model selection framework to identify multiple QTL for complex traits in experimental designs, based upon a composite model space approach (Godsill, 2001). The composite model space approach deals with the number of included loci L as an upper bound on the number of detectable QTL across the entire genome. The upper bound L is treated as a fixed constant and is chosen to be larger than the number of detectable QTL for a given data set. Even with a moderate value for the upper bound, there are many possible genetic effects, especially when considering interactions, but most are negligible (that is, close to zero) and can be excluded from the model. The composite model space approach thus uses an unobserved vector γ of binary indicator variables to indicate which genetic effects (main effects, epistatic effects and gene–environment interactions) are included in ($\gamma_j = 1$) or excluded from ($\gamma_j = 0$) the model. In this

setting, the actual number of QTL l is not treated as an explicit parameter but can be determined by γ and L (Yi *et al*., 2005). Thus, we have $H=(\gamma, \lambda)$, where the vector λ represents the genomic positions of l QTL.

The key advantages of the composite model space approach are that it provides a convenient way to reasonably reduce the model space and to construct efficient MCMC algorithms, especially for simultaneously mapping main effects, epistatic effects and gene–environment interactions (Yi *et al*., 2005, 2007). The composite model space approach has been implemented in the package R/qtlbim.

The prior distribution $p(\theta, H)$

A Bayesian QTL analysis proceeds by placing prior distributions on the unknowns (θ, H). We outline in detail the composite model space approach that has been implemented in the package R/qtlbim (Yi, 2004; Yi *et al*., 2005, 2007b) and discuss other methods described in the last section.

The prior for the overall mean μ is chosen to be normally distributed with mean η_0 and variance τ_0^2. We choose $\eta_0=1/n\sum_{i=1}^{n}y_i$, and $\tau_0^2=(1/n-1)\sum_{i=1}^{n}(y_i-\bar{y})^2$. We choose an inverse-Gamma(a, b) as the prior of σ_e^2. Gaffney (2001) suggested $a=3$ and $b=s_y^2$, which has prior mean and variance equal to $s_y^2/2$. In the package R/qtlbim, we take the non-informative prior $p(\sigma_e^2)\propto 1/\sigma_e^2$.

For the positions of QTL, the simplest and most widely used prior assumption is that the positions are independently and uniformly distributed over the pre-set loci across the genome. The basic framework of the composite model space approach provides flexible ways to reduce the model space by putting some constraints on models. We have incorporated two global constraints on models into our algorithms and software R/qtlbim as options (Yi *et al*., 2007b). These constraints dramatically reduce the model space and may be useful for efficiently detecting multiple interacting QTL. The first constraint restricts the spacing among multiple linked QTL. On chromosome c, forcing QTL to be at least d_c cM apart excludes the possibility of fitting closely linked QTL if d_c is large. The distance d_c should depend on the density of markers on chromosome c and on the sample size n. We suggest setting it to the average length of marker intervals on chromosome c. The second constraint restricts the number of detectable QTL on each chromosome to L_c with $L\leqslant\sum L_c$ and $L_c\leqslant D_c/d_c$, where D_c is the length of chromosome c. End users can use these global constraints to rule out many unrealistic or undistinguishable models from consideration.

A variety of prior distributions for genetic effects β have been proposed. It is desirable that effect priors be invariant to the scales of the phenotype and the effect predictors and model complexity. This can be accomplished by hierarchical models in which the priors have empirical hyper-priors that depend on the total phenotypic variance and the sample variances of the predictors. Following Yi *et al*. (2007b), we partition the genetic effects into batches, corresponding to different types of effects, for example, additive, dominance, additive–additive, additive–environment interactions, etc. Effects in the same batch k follow the same prior,

$$\beta_{kj}|\gamma_{kj}\sim(1-\gamma_{kj})I_0+\gamma_{kj}N(0,\sigma_k^2) \tag{7}$$

where γ_{kj} is the indicator variable for β_{kj}, and I_0 is a point mass at 0. Under this prior, when $\gamma_{kj}=0$, β_{kj} is assigned to be 0 and thus is actually removed from the model; when $\gamma_{kj}=1$, β_{kj} follows a normal distribution $N(0, \sigma_k^2)$. We treat the variance σ_k^2 as a random variable with an inverse χ^2 hyper-prior distribution:

$$\sigma_k^2\sim \mathrm{Inv}-\chi^2(v_k,s_k^2) \tag{8}$$

The prior degrees of freedom v_k and scale parameters s_k^2 are chosen to control the prior expected mean and the prior confidence region of the proportion of the phenotypic variance explained by β_{kj}. One attractive feature of this strategy of specifying the hyperparameters is that it causes the above priors to be invariant to the scales of the phenotype and the effect predictors in Model (4). Under the prior (8), σ_k^2 has expected value $\mathrm{E}(\sigma_k^2)=v_ks_k^2/(v_k-2)$. The degrees of freedom v_k control the skew of the prior for σ_k^2, with larger values recommended (here $v_k=6$) to tightly center the prior around s_k^2 (see Chipman, 2004). The scale s_k^2 controls the prior heritability per effect (also see Gaffney, 2001). The proportion of phenotypic variance explained by β_{kj} is $h_{kj}=V_{kj}\beta_{kj}^2/V_p$, with V_{kj} the sample variance for the column of X associated with effect β_{kj}. Setting $s_k^2=(v_k-2)\mathrm{E}(h_{kj})V_p/(v_k\ \ V_{kj})$ yields $\mathrm{E}(h_{kj})=V_{kj}\mathrm{E}((\sigma_k^2)/V_p$. Expected effect heritabilities, $\mathrm{E}(h_{kj})$, can be set small (say 0.05–0.2) to reflect prior knowledge about genetic architecture.

Priors on environmental effects in β_E can be assigned uniform distributions or normal distributions with mean 0 and unknown variances, labeled fixed or random effects from the non-Bayesian tradition, respectively (Gelman *et al*., 2004). For the unknown variances, conjugate prior distributions are scaled inverse χ^2 distributions with prior degrees of freedom and scale parameters specified as they were for genetic effects.

For the vector of genetic-effect indicators γ, we could use an independence prior of the form

$$p(\gamma)=\prod w_j^{\gamma_j}(1-w_j)^{1-\gamma_j} \tag{9}$$

where $w_j=p(\gamma_j=1)$ is the prior inclusion probability for the jth effect and equals the predetermined hyperparameter w_m or w_e, depending on the jth effect being a main effect or an epistatic effect, respectively. Under this prior, the importance of any effect is independent of the importance of any other effect and the prior inclusion probability of a main effect is different from that of an epistatic effect. The hyperparameters w_m and w_e control the expected numbers of active main and epistatic effects, respectively, and thus the expected number of QTL; small w_m and w_e would concentrate the priors on parsimonious models with few main effects and epistatic effects. Instead of directly specifying w_m and w_e, it would be better to first determine the prior expected numbers of main-effect QTL, l_m, and all QTL, $l_0\geqslant l_m$ (that is, main-effect and epistatic QTL) and then solve for w_m and w_e from the expressions of the prior expected numbers (Yi *et al*., 2005). The prior expected number of main-effect QTL, l_m, could be set to the number of QTL detected by traditional non-epistatic mapping methods, for example, interval mapping or composite interval mapping (Lander and Botstein, 1989; Zeng, 1994). The prior expected number of all QTL, l_0, should be chosen to be at least l_m. The number of QTL detected by traditional epistatic

mapping methods, for example, a two-dimensional genome scan from R/qtl, could provide a rough guide for choosing l_0.

Independence priors for γ work well for many situations (Yi *et al*., 2005, 2006), but may not be appropriate when either (1) loci with large main effects are more likely to have large interactions or (2) many loci have detectable main effects and thus the probability of detecting additional QTL with weak main effects but strong interactions is low. Yi *et al*. (2007b) proposed dependence priors capturing relations between interaction and main effect terms (see Chipman, 1996, 2004; Chipman *et al*., 2001). Consider two QTL indexed by j and k, with main effect and epistasis indicators γ_j, γ_k and γ_{jk}. Setting a common inclusion probability for main effects, $P(\gamma_j=1)=P(\gamma_k=1)p_m$ (Yi *et al*., 2005), we construct conditional inclusion probabilities for epistasis as

$$P(\gamma_{jk}=1|\gamma_j,\gamma_k) = \begin{cases} c_0 p_m & \text{if } (\gamma_j,\gamma_k)=(0,0) \\ c_1 p_m & \text{if } (\gamma_j,\gamma_k)=(1,0) \text{ or } (0,1) \\ c_2 p_m & \text{if } (\gamma_j,\gamma_k)=(1,1) \end{cases} \quad (10)$$

Typically, $0\leqslant c_0\leqslant c_1\leqslant c_2\leqslant 1$, implying that main effects are more likely to be detected than epistasis, and that the importance of an interaction depends on the importance of its 'parent' terms. Setting some c_i to zero rules out certain interactions: $c_0=c_1=0$ and $c_2>0$ allows interactions only if both main effects are included. These values establish a principle of variable selection, modifying prior mass across possible genetic architectures and greatly reducing the model space.

For the varying dimensional model space approach, the prior distribution of the number of QTL l may be a truncated Poisson distribution with mean l_0 and maximum integer L, or a uniform distribution between 0 and L. The choice of l_0 influences the posterior of l but Bayes factors for l are relatively insensitive to the choice of prior distribution for this hyperparameter (Satagopan and Yandell, 1996; Gaffney, 2001; Yi *et al*., 2003a).

MCMC algorithms

MCMC is a class of algorithms for drawing values of unknown parameters θ from the target posterior distribution $p(\theta|y)$. The keys to MCMC algorithms are to design and simulate a Markov chain (that is, the distribution of the sampled draws depending on the last value drawn) whose stationary distribution is the target distribution $p(\theta|y)$ and to run the simulation long enough that the distribution of the current samples is close enough to this stationary distribution. MCMC is used when it is not possible (or not computationally efficient) to analytically calculate $p(\theta|y)$ or directly sample θ from $p(\theta|y)$. A major advantage of MCMC algorithms is their ability to deal with high-dimensional and complex problems. These algorithms serve our purpose ideally because in Bayesian multiple QTL analysis we want to evaluate the joint posterior distribution $p(\theta, g, H|y, m, X_E)$, which includes a large number of parameters.

For high-dimensional models, MCMC algorithms usually proceed by partitioning the set of parameters into components or subvectors $\theta=(\theta_1,\ldots,\theta_d)$ and then drawing each subset from the conditional distribution $p(\theta_j|\theta_{-j}, y)$, $j=1,\ldots,d$, given the latest values of all other parameters θ_{-j} and the data y. Each iteration of the MCMC algorithm cycles through the subvectors of θ. This process continues for a large number of iterations to obtain a random sample from the joint posterior distribution $p(\theta|y)$. Various methods have been devised for constructing and sampling from the conditional distribution $p(\theta_j|\theta_{-j}, y)$. The Metropolis–Hastings (M–H) algorithm is a general term for a family of Markov chain simulation methods that are useful for drawing samples from many distributions (Metropolis *et al*., 1953; Hastings, 1970). The Gibbs sampler and the Metropolis algorithm are two commonly used special cases of the M–H algorithm. These algorithms can be used as building blocks for sampling from complicated distributions. If the conditional distribution $p(\theta_j|\theta_{-j}, y)$ has a standard form, the Gibbs sampler can be used to directly sample from it; otherwise, we have to use the M–H algorithm (Gelman *et al*., 2004).

We now describe the MCMC algorithms for sampling from the joint posterior distribution $p(\theta, g, H|y, m, X_E)$, focusing on those implemented in R/qtlbim and discussing others. In our notation, we have partitioned the set of unknown quantities into three subvectors θ, g and H, where θ includes all model parameters, that is, $\theta=(\beta, \sigma_e, \sigma_\beta) \triangleq (\beta, \sigma)$, $g=(g_{iq})$, is the $n\times L$ matrix of genotypes, and the model index H includes the indicator variables of genetic effects γ and the QTL positions λ. For the varying dimensional model space approach, H also includes the number of QTL l. The posterior distribution for the unknown quantities (θ, g, H) can be simulated using MCMC, alternately updating the model parameters θ given (g, H), the genotypes g given (θ, H) and the model index H given (θ, g).

Updating θ

A notable feature of the multiple QTL model is that, given (g, H), Model (4) is a conventional hierarchical normal model. Therefore, given (g, H), θ can be drawn using standard Gibbs algorithms for hierarchical linear models (Gelman *et al*., 2004). The variance parameters are sampled one at a time from their conditional posterior distributions; for each j, $p(\sigma_j^2|y, X_E, \beta, g, H, \sigma_{-j})$, is a scaled inverse χ^2 distribution and can be directly sampled, where σ_{-j} is all elements of σ except σ_j. For hierarchical normal models, there are two Gibbs sampler algorithms to update β. In one version, the vector β is drawn all at once from the conditional posterior distribution $p(\beta|y, X_E, g, H, \sigma)$; this algorithm requires large matrix operations at each simulation iteration. In the other version, the components of β are drawn one at a time; for each j, β_j is sampled from $p(\beta_j|y, X_E, g, H, \sigma, \beta_{-j})$, where β_{-j} represents all of β except β_j, so that β_j is sampled from a simple univariate normal distribution. In the package R/qtlbim, we use the second algorithm. This one-at-a-time algorithm has the advantage of never requiring matrix operations; if set up carefully, with the appropriate intermediate results held in storage, this algorithm can be very efficient in terms of computation time (Gelman *et al*., 2004; Yi *et al*., 2005, 2007b). There is also the potential for extending the model to include additional genetic or non-genetic factors by simply adding additional steps in the Gibbs algorithm. It is worth noting that at each iteration the

composite model space approach drops many possible genetic effects from the model and hence significantly reduces computation (Yi, 2004; Yi *et al*., 2005, 2007b). In contrast, the shrinkage and the SSVS methods always fit and update all effects (Xu, 2003; Yi *et al*., 2003b; Wang *et al*., 2005).

Updating g

The matrix of genotypes g is updated one at a time from the conditional posterior distributions. If locus q is included in the model and the genotype g_{iq} is not observed, then g_{iq} is sampled from the conditional posterior distribution

$$p(g_{iq}=k|y,X_{\mathrm{E}},\theta,g_{-iq},H) = \frac{p(y_i|X_{\mathrm{E}},\theta,g_{-iq},g_{iq}=k,H)p(g_{iq}=k|m,\lambda_q)}{\sum_{g_{iq}}p(y_i|X_{\mathrm{E}},\theta,g_{-iq},g_{iq},H)p(g_{iq}|m,\lambda_q)} \quad (11)$$

where g_{iq} is the genotype of individual i at locus q, g_{-iq} represents all elements of g except g_{iq}, $p(y_i \mid X_{\mathrm{E}}, \theta, g, H)$ is the likelihood for individual i, and $p(g_{ij}=k \mid m, \lambda_q)$ is the prior probability of $g_{iq}=k$ that has been calculated by Equation (3). This posterior is a simple multinomial distribution, and thus can be sampled directly. If g_{iq} is observed (for example, for fully observed markers), we do not need to sample g_{iq}.

When the pre-set upper bound L is large, the composite model space approach usually excludes many of L loci from the model and thus the genotypes at these excluded loci do not need to be updated (Yi, 2004; Yi *et al*., 2005, 2007b). However, the shrinkage and the SSVS methods update genotypes of all L loci because they always include L loci in the model (Xu, 2003; Yi *et al*., 2003b; Wang *et al*., 2005).

Updating λ

The vector of QTL positions λ is updated one at a time using the Metropolis algorithm. For QTL q, the joint conditional posterior distribution of the position λ_q and the genotypes $g_q=(g_{1q}, \ldots, g_{nq})$ is

$$p(\lambda_q,g_q|y,m,X_{\mathrm{E}},\theta,g_{-q},H_{-\lambda_q}) \propto p(y|X_{\mathrm{E}},\theta,g_{-q},H_{-\lambda_q},\lambda_q,g_q)p(\lambda_q|\lambda_{-q})p(g_q|\lambda_q,m) \quad (12)$$

where $g_{-q}\,(H_{-\lambda_q})$ represents all elements of $g(H)$ except $g_q(\lambda_q)$, $p(y|X_{\mathrm{E}},\theta,g_{-q},H_{-\lambda_q},\lambda_q,g_q)$ is the likelihood calculated by Model (4), $p(\lambda_q \mid \lambda_{-q})$ is the conditional prior of λ_q given all other elements of λ, and $p(g_q \mid \lambda_q, m)$ is the prior probability of g_{iq} that has been calculated by Equation (3).

This posterior is not a standard distribution, and thus an M–H algorithm is needed to update λ_q and g_q jointly. We first propose a new position λ_q^* from a proposal distribution $q(\lambda_q^*;\lambda_q)$, and then generate new genotypes, g_q^*, at this new position for all individuals from the conditional posterior $q(g_q^*)=\Pi_i p(g_{iq} \mid y, X_{\mathrm{E}}, \theta, g_{-iq}, H_{-\lambda_q}, \lambda_q^*)$. The proposals for λ_q^* and g_q^* are then accepted simultaneously with probability (Yi and Xu, 2002)

$$\alpha=\min\left(1,\frac{p(\lambda_q^*,g_q^*|y,m,X_{\mathrm{E}},\theta,g_{-q},H_{-\lambda_q})q(\lambda_q;\lambda_q^*)q(g_q)}{p(\lambda_q,g_q|y,m,X_{\mathrm{E}},\theta,g_{-q},H_{-\lambda_q})q(\lambda_q^*;\lambda_q)q(g_q^*)}\right) \quad (13)$$

The proposal distribution for the new position $q(\lambda_q^*;\lambda_q)$ is usually constructed as uniformly distributed over $2d$ most flanking loci of λ_q, with d being a predetermined tuning integer. This local proposal never allows the QTL to move to different chromosomes. An alternative scheme—which allows long-distance moves—has been proposed by Gaffney (2001). In R/qtlbim, we use the local move scheme and take $d=2$, incorporating the previously described pre-set constraints on QTL positions into our algorithm.

Proposing the new genotypes from the conditional posterior $q(g_q^*)$ is equivalent to integrating over the genotypes at QTL q, that is, the acceptance probability equals

$$\alpha=\min\left(1,\frac{p(\lambda_q^*|y,m,X_{\mathrm{E}},\theta,g_{-q},H_{-\lambda_q})q(\lambda_q;\lambda_q^*)}{p(\lambda_q|y,m,X_{\mathrm{E}},\theta,g_{-q},H_{-\lambda_q})q(\lambda_q^*;\lambda_q)}\right) \quad (14)$$

In principle, the genetic effects associated with QTL q can also be integrated out, allowing further improvement of the algorithm, especially for long-range moves, as observed by Gaffney (2001). However, for multiple interacting QTL models, many genetic effects are associated with a QTL, and thus integrating out these effects involves large matrix operations.

Updating γ

We here describe two algorithms to update the indicator vector γ: The first one is a Gibbs sampler similar to that of Yi *et al*. (2005), modified by incorporating the new priors and the constraints, and the second is a novel M–H scheme developed by Yi *et al*. (2007b). The M–H algorithm offers significant computational savings over the Gibbs sampler, especially when the number of effects is large (Yi *et al*., 2007b). Both of these algorithms are available in the package R/qtlbim.

At each iteration of the MCMC simulation, the full Gibbs sampler generates each of the indicator variables, $\gamma B_{j,}$ from its conditional posterior distribution

$$p(\gamma_j=1|y,m,X_{\mathrm{E}},\theta_{-j},g,H_{-j})=\frac{wL_1}{(1-w)L_0+wL_1} \quad (15)$$

where θ_{-j} is all elements of θ except β_j, H_{-j} is all elements of H except γ_j, $w=p(\gamma_j=1 \mid \gamma_{-j})$ is the prior inclusion probability of the effect β_j, and $L_k=p(y \mid X_{\mathrm{E}}, \theta_{-j}, g, H_{-j}, \gamma_j=k)$ for $k=0, 1$. Note that β_{j} is integrated out from L_1. L_1 can be calculated using the identity of simple conditional probability

$$L_1=\frac{p(y|X_{\mathrm{E}},\theta_{-j},g,H_{-j},\gamma_j=1,\beta_j)p(\beta_j)}{p(\beta_j|y,X_{\mathrm{E}},g,H,\sigma,\beta_{-j})} \quad (16)$$

where $p(y \mid X_{\mathrm{E}}, \theta_{-j}, g, H_{-j}, \gamma_j=1, \beta_j)$ is the phenotype likelihood, $p(\beta_{\mathrm{j}})$ is the prior distribution of β_j, and $p(\beta_j \mid y, X_{\mathrm{E}}, g, H, \sigma, \beta_{-j})$ is the conditional posterior distribution of β_j. Notationally, the right side of (16) depends on β_j, but from the definition of L_1, we know it cannot depend on β_j in a real sense. That is, the factors that depend on β_j in the numerator and denominator must cancel. Thus, we can compute (16) by inserting any value of β_{j} into the expression. A convenient, stable choice for β_{j} is the conditional posterior mean of β_{j} (Gelman *et al*., 2004).

The full Gibbs sampling scheme works reliably (Yi *et al*., 2005, 2006). However, when the number of possible genetic effects (that is, the number of indicator variables) is large, most of the genetic effects are near zero and thus γ_j is zero for most j. If the current value of γ_j is 0, γ_j is likely to be regenerated as zero because the prior

probability $w = p(\gamma_j = 1 \mid \gamma_{-j})$ in (15) is very small. In the Gibbs sampler, it is always necessary to calculate the conditional posterior probability (15) when γ_j is currently 0. Such computation may be wasteful.

As with the Gibbs sampler, at each iteration of the MCMC simulation, the M–H scheme of Yi *et al*. (2007b) proceeds to update all indicator variables. Denote the current value of γ_j by C ($=0$ or 1). The M–H algorithm proposes a new value P ($=0$ or 1) for γ_j from the conditional prior probability $p(\gamma_j = C \mid \gamma_{-j})$. If $P = C$, the M–H acceptance probability is 1, and thus γ_j remains at C and there is no need to compute any values. Otherwise, we update γ_j from the current value C to the proposal $1-C$ with acceptance probability

$$\alpha = \min\left(1, \frac{p(\gamma_j = 1-C|y, m, X_E, \theta_{-j}, g, H_{-j}) \quad p(\gamma_j = C|\gamma_{-j})}{p(\gamma_j = C|y, m, X_E, \theta_{-j}, g, H_{-j}) \quad p(\gamma_j = 1-C|\gamma_{-j})}\right) = \min\left(1, \frac{L_{1-C}}{L_C}\right) \tag{17}$$

in which all terms are defined in (15). If γ_j is currently 1 (that is, β_j is currently included in the model), we can calculate the two values L_0 and L_1 using the prior variance of β_j and the column of $\mathbf{X}$ corresponding to the effect β_j. If γ_j is currently 0 (that is, β_j is currently excluded from the model) and the involved QTL(s) is (are) not currently in the model, we first expand $\mathbf{X}$; sample from the corresponding priors one or two new QTL position(s) as needed, new genotypes for all individuals, and the prior variance of β_j if this parameter is currently out of the model; and then calculate the acceptance probability to update γ_j. This procedure is also needed for the full Gibbs sampler (Yi *et al*., 2005).

In this M–H algorithm, the proposal probability to generate $\gamma_j = 1$ when it is currently 0 is $p(\gamma_j = 1 \mid \boldsymbol{\gamma}_{-j})$, which is very small when the number of possible genetic effects is large and most of them are near 0, and thus γ_j is likely to be proposed as 0. Therefore, it is unnecessary to compute any values for most γ_j, and hence this new algorithm is much faster than the full Gibbs sampler.

We can illustrate the relative advantages of the Gibbs sampler to the M–H algorithm in terms of statistical efficiency. The transition probability for γ_j from C to P, $Q(C \to P)$, for the Gibbs sampler and the M–H algorithm is

$$Q_G(0 \to 1) = \frac{wL_1}{(1-w)L_0 + wL_1},$$

$$Q_G(1 \to 0) = \frac{(1-w)L_0}{(1-w)L_0 + wL_1}$$

and

$$Q_{MH}(0 \to 1) = w \cdot \min(1, \frac{L_1}{L_0}),$$

$$Q_{MH}(1 \to 0) = (1-w) \cdot \min(1, \frac{L_0}{L1})$$

respectively, with $w = p(\gamma_j = 1 \mid \gamma_{-j})$. Following Kohn *et al*. (2001), $Q_G(C \to 1-C) > Q_{MH}(C \to 1-C)$. Thus, the Gibbs sampler is statistically more efficient per scan than the M–H algorithm in terms of transition probabilities. When the upper bound of QTL is large and w is small, the new faster algorithm does not sacrifice much statistical efficiency, since it can be easily shown that $Q_{MH}(C \to 1-C) \approx Q_G(C \to 1-C)$.

The above M–H algorithm is derived using the conventional M–H technique based on the composite model space. However, it is similar to a RJ-MCMC algorithm, which cycles through each indicator variable and, using the prior probability as the proposal, generates one or two new QTL position(s), new genotypes for all individuals and the prior variance of β_j from the corresponding priors and the associated effect β_j from the full conditional posterior. This RJ-MCMC algorithm can be derived from our composite model space approach. For non-epistatic models, Yi (2004) showed that the composite model space approach includes many RJ-MCMC algorithms as special cases.

Updating *l*

The traditional M–H algorithm can only be used to generate samples from the posterior distributions with fixed dimension, and thus cannot be applied to the variable dimensional model space approach. Green (1995) introduced a generalization of M–H algorithms for sampling from models with variable dimensionality, called RJ or trans-dimensional MCMC. This method is extremely flexible and can jump from one model to another, provided that we carefully select appropriate proposal densities. The RJ-MCMC sampler has been successfully applied to mapping multiple non-epistatic QTL (Satagopan and Yandell, 1996; Stephens and Fisch, 1998; Sillanpää and Arjas, 1998; Yi and Xu, 2000; Gaffney, 2001). Recently, we have extended RJ-MCMC algorithms to map epistatic QTL (Yi and Xu, 2002; Yi *et al*., 2003a).

The algorithm of Yi *et al*. (2003a) includes two steps: (a) adding one new QTL with main effects or epistatic effects with some of the existing QTL, or deleting a QTL from all existing QTL and (b) adding two QTL with main effects or epistatic effects among themselves or with some other existing QTL, or deleting two QTL from all existing QTL. Here, we use step (b) as an example to show how to perform the RJ-MCMC. For step (b), we first randomly decide to propose adding two new QTL with probability $j(l+2;\ l)$, or deleting two existing QTL with $j(l;\ l+2) = 1 - j(l+2;\ l)$. To add two QTL, we need to generate additional parameters associated with the new QTL, that is, two new positions λ_1^* and λ_2^*, genotypes g_1^* and g_2^*, effect indicators γ^* associated with these two QTL, and new main and epistatic effects β^*. New positions, genotypes and indicators are sampled from their priors. β^* are sampled from the conditional posterior distribution, which is a multivariate normal distribution. The change in the number of QTL from l to $l+2$, together with the proposed parameters, is accepted or rejected according to the RJ algorithm. Deleting two QTL is simply the reverse process. Two QTL are randomly chosen among the existing QTL. The chosen QTL, together with all corresponding parameters, are then proposed to be deleted. In most of Bayesian mapping, the proposal probabilities for birth and death, $j(l+2;\ l)$ and $j(l;\ l+2)$, have been chosen to be constants, for example, $j(l+2;\ l) = j(l;\ l+2) = 0.5$ (for example, Yi *et al*., 2003a). Alternatively, these proposal probabilities can be chosen so that $\frac{p(l+2)j(l;l+2)}{p(l)} \frac{1}{j(l+2;l)(l+2)(l+1)}$ is unity (Satagopan and Yandell, 1996; Gaffney, 2001).

Summarizing and interpreting the posterior samples

The MCMC algorithms described above are used to simulate a Markov chain $\{(\theta, g, H)^{(t)}; t = 1, 2, \ldots, T\}$ from the joint posterior distribution $p(\theta, g, H \mid y, m, X_E)$ to generate posterior samples. If enough iterates have been run, the posterior samples contain all the information about the posterior distribution. Inference using the posterior samples requires some care, however (Gelman *et al*., 2004). First, if insufficient iterates have been run, the simulation may not have converged and thus may not be representative of the target distribution. Even when the simulations have reached convergence, early iterates may still be influenced by initial values. A second problem is within-sequence correlation; inference from correlated draws is generally less precise than from the same number of independent draws.

We handle these special problems in different ways. To diminish the dependence on initial values, we generally discard (thousands of) early iterates, referred to as 'burn-in.' To reduce sequential correlation, the subsequent sample is thinned by keeping every kth simulation draw and discarding the rest (for example, $k = 40$). For a high-dimensional problem, the mixing behavior and convergence rates of MCMC algorithms are critical issues. It is very difficult to say conclusively that a chain has converged, only to diagnose when it definitely has not. The package R/qtlbim provides tools to monitor mixing behavior and convergence of the simulated Markov chain, either by examining trace plots of the sample values of scalar quantities of interest, such as the numbers of QTL and epistatic effects or by using formal diagnostic methods provided in the package R/coda (Plummer *et al*., 2007).

For all of the Bayesian multiple QTL mapping methods we have described, the basic principle of posterior inference is to use all of the saved iterates of the Markov chain, corresponding to model averaging, which assesses characteristics of the genetic architecture by averaging over possible models weighted by their posterior probability. Model averaging accounts for model uncertainty and hence provides more robust inference compared to a single 'best' model approach (Raftery *et al*., 1997; Ball, 2001; Sillanpää and Corander, 2002). For Bayesian methods involving model selection,

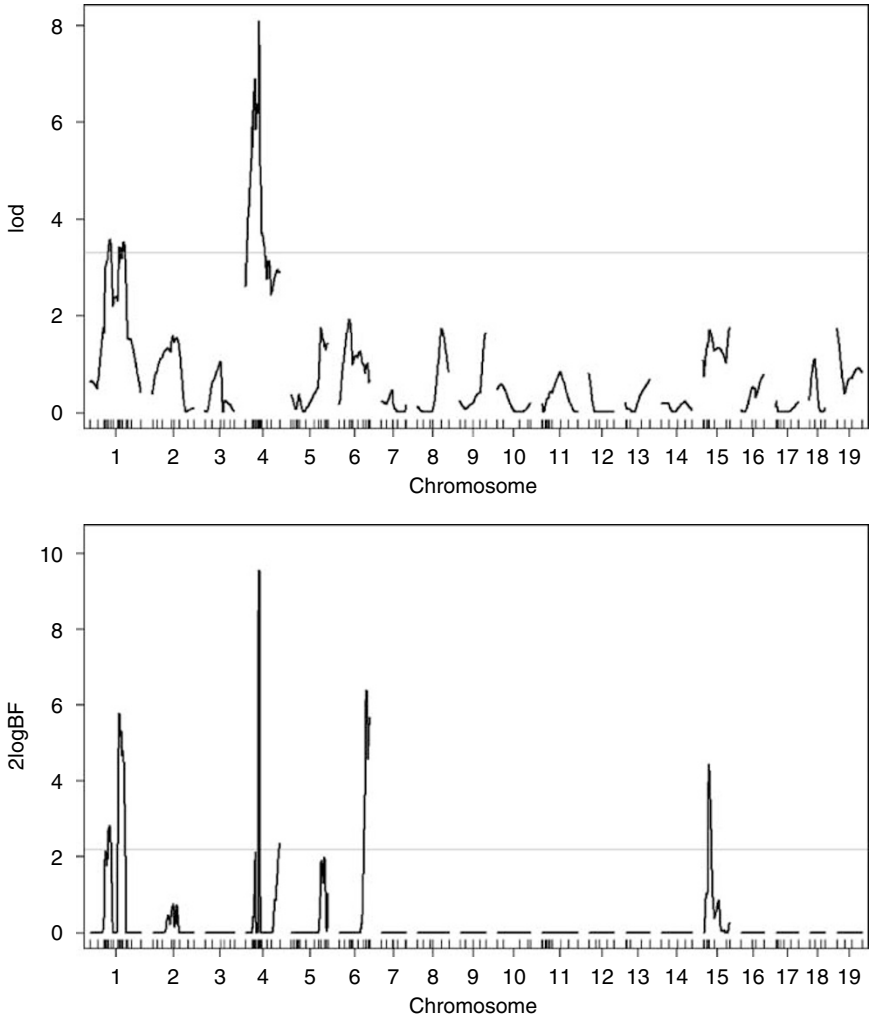

Figure 1 Genome-wide scan for main effects using R/qtl (the top panel) and R/qtlbim (the bottom panel). The gray lines indicate the significance threshold for lod scores of 3.3 (the top) and Bayes factors of 3.0 (the bottom). Inner ticks on the x-axis depict the locations on observed markers.

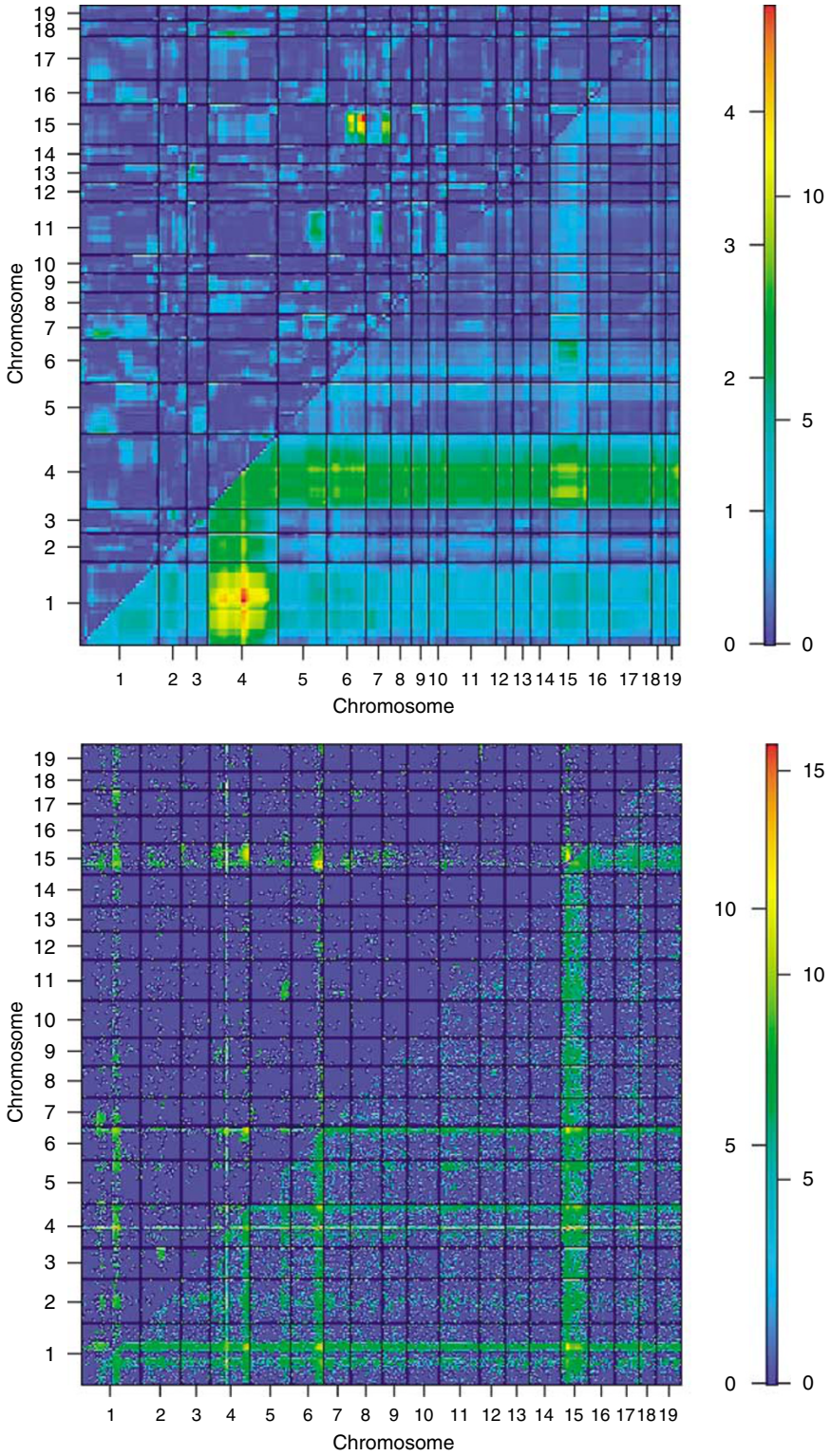

Figure 2 Genome-wide scan for epistatic effects using R/qtl (the top panel) and R/qtlbim (the bottom panel). In the top panel, epistatic lod scores are plotted in the upper left triangle using the left scale, and joint lod scores are plotted in the lower right triangle using the right scale. In the bottom panel, epistatic Bayes factors are plotted in the upper left triangle using the left scale, and joint Bayes factors scores are plotted in the lower right triangle using the right scale.

the posterior samples can be used to search for models with high posterior probability. The idea here is that larger effects should tend to appear more often and early in the posterior sample, making them easier to identify.

A key advantage of the Bayesian approach, as implemented by MCMC simulation, is the flexibility with which posterior inferences can be summarized. The package R/qtlbim provides various graphical and

tabular summaries that assess the contribution of individual loci and pairs of loci while adjusting for effects of all other possible loci and covariates via model averaging (Yandell *et al*., 2007). One such summary is the posterior inclusion probability for each locus or each pair of loci, estimated as its frequency in the posterior samples. Taking prior probabilities into consideration, we can then use Bayes factors to compare models with and without the locus or loci (Kass and Raftery, 1995). Because each locus may be included in the model through its main effects and/or interactions with other loci (epistasis) or environmental effects, we can separately estimate the posterior inclusion probabilities and corresponding Bayes factors of main effects, epistasis and gene–environment interactions per locus. These estimates can be further divided into Cockerham effects (additive and dominance for main effects or the four types of epistatic interactions), if desired. In addition to posterior inclusion probabilities and Bayes factors, the package R/qtlbim provides tools to estimate marginal heritabilities, genetic effects, genotypic means, Bayesian log posterior densities and other features (Yandell *et al*., 2007).

Example analysis

To illustrate these above methods, we compare and contrast results from R/qtl and R/qtlbim using murine backcross data. Briefly, Sugiyama *et al*. (2001) described a backcross of salt-sensitive C57BL/6J and non-salt-sensitive A/J mice. They measured blood pressure for 250 male mice. A total of 170 markers were genotyped at approximately 15 cM intervals over the 19 autosomes.

We first performed a one-dimensional scan using R/qtl (the top panel of Figure 1). This analysis suggested the presence of three QTL, two on chromosome 1 and one on chromosome 4. We also scanned the genome for main effects using R/qtlbim (the bottom panel of Figure 1). The Bayesian analysis revealed, in addition to the same three QTL, evidence supporting another QTL on chromosome 4 and QTL on chromosomes 6 and 15. Note the improved separation between noise and signal the Bayesian method provides over the frequentist method.

We then scanned the genome for epistatic interactions using a two-dimensional scan in R/qtl (the top panel of Figure 2) and in R/qtlbim (the bottom panel of Figure 2). The frequentist analysis yielded evidence for an epistatic interaction between chromosomes 6 and 15. In addition to that interaction, the Bayesian analysis also yielded evidence for epistatic interactions between chromosomes 1 and 4, 1 and 6, 1 and 15, 4 and 6, 4 and 15, and 15 and 15. As with the scan for main effects, the Bayesian method of scanning for epistatic effects yields improved separation between noise and signal.

Conclusions

Bayesian modeling of multiple QTL, coupled with advances in posterior search and computation, has led to an explosion of research in mapping multiple QTL for complex traits. To illustrate the rapid evolution of these methods, we have highlighted some of these developments. We have a clear sense of the potential gains to be achieved using the Bayesian approach to mapping multiple interacting QTL. Bayesian methods and associated computer software provide us with tools to comprehensively unravel the genetic basis and architecture of complex trait variation. What is standard in complex trait analysis has changed much in the past years, and with the continuing development of sophisticated statistical mapping methods, further dramatic improvement may be possible. Future research directions include extensions to joint analysis of multiple traits, and experimental crosses derived from multiple inbred lines and outbred populations. Computationally efficient algorithms are an essential feature for the practical analysis of complex genetic architectures in these more complicated cases.

Acknowledgements

We thank two anonymous reviewers and the associated editor for constructive comments on an earlier version of the manuscript. This research was supported in part by National Institutes of Health Grant GM069430 to NY.

References

Baierl A, Bogdan M, Frommlet F, Futschik A (2006). On locating multiple interacting quantitative trait loci in intercross designs. *Genetics* **173**: 1693–1703.

Ball RD (2001). Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian information criterion. *Genetics* **159**: 1351–1364.

Bogdan M, Ghosh JK, Doerge RW (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics* **167**: 989–999.

Broman KW, Speed TP (2002). A model selection approach for identification of quantitative trait loci in experimental crosses. *J R Stat Soc B* **64**: 641–656.

Broman KW, Wu H, Sen Ś, Churchill GA (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**: 889–890.

Carlin BP, Louis TA (2000). *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd edn. Chapman & Hall: London, UK.

Carlborg Ö, Andersson L, Kinghorn B (2000). The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics* **155**: 2003–2010.

Carlborg Ö, Haley C (2004). Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* **5**: 618–625.

Chipman H (1996). Bayesian variable selection with related predictions. *Can J Stat* **24**: 17–36.

Chipman H (2004). Prior distributions for Bayesian analysis of screening experiments. In: Dean A, Lewis SM (eds). *Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics*. Springer: New York. pp 235–267.

Chipman H, Edwards EI, McCulloch RE (2001). The practical implementation of Bayesian model selection. In: Lahiri P (ed). *Model Selection*. Institute of Mathematical Statistics: Beachwood, Ohio. pp 65–116.

Gaffney PJ (2001). An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. PhD dissertation, Department of Statistics, University of Wisconsin—Madison, Madison, WI, USA.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis*, 2nd edn. Chapman & Hall: London, UK.

George EI, McCulloch RE (1993). Variable selection via Gibbs sampling. *J Am Stat Assoc* **88**: 881–889.

George EI, McCulloch RE (1997). Approaches for Bayesian variable selection. *Stat Sinica* **7**: 339–373.

Godsill SJ (2001). On the relationship between MCMC model uncertainty methods. *J Comput Graph Stat* **10**: 230–248.

Green PJ (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.

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

Hastings WK (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.

Hoeschele I (2001). Mapping quantitative trait loci in outbred pedigrees. In: Balding DJ, Bishop M, Cannings C (eds). *Handbook of Statistical Genetics*. Wiley: New York. pp 599–644.

Jansen RC, Stam P (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**: 1447–1455.

Jiang C, Zeng ZB (1997). Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. *Genetica* **101**: 47–58.

Kao CH, Zeng ZB (2002). odeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* **160**: 1243–1261.

Kao CH, Zeng ZB, Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.

Kass RE, Raftery AE (1995). Bayes factors. *J Am Stat Assoc* **90**: 773–795.

Kohn R, Smith M, Chen D (2001). Nonparametric regression using linear combinations of basis functions. *Stat Comput* **11**: 313–322.

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

Lynch M, Walsh B (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates Inc.: Sunderland, MA.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953). Equation of state calculations by fast computing machines. *J Chem Phys* **21**: 1087–1092.

Meuwissen THE, Hayes BJ, Goddard ME (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819–1829.

Moore JH (2005). A global view of epistasis. *Nat Genet* **37**: 13–14.

Narita A, Sasaki Y (2004). Detection of multiple QTL with epistatic effects under a mixed inheritance model in an outbred population. *Genet Sel Evol* **36**: 415–433.

Plummer M, Best N, Cowles K, Vines K (2007). coda: output analysis and diagnostics for MCMC, R package version 0.11-2. Institute of Mathematics Statistics, Beachwood, OH (http://www-fis.iarc.fr/coda).

R Development Core Team (2006). *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing: Vienna, Austria, ISBN 3-900051-07-0, www.R-project.org.

Raftery AE, Madigan D, Hoeting JA (1997). Bayesian model averaging for linear regression models. *J Am Stat Assoc* **92**: 179–191.

Rao S, Xu S (1998). Mapping quantitative trait loci for ordered categorical traits in four-way crosses. *Heredity* **81**: 214–224.

Reifsnyder PR, Churchill G, Leiter EH (2000). Maternal environment and genotype interact to establish diabesity in mice. *Genome Res* **10**: 1568–1578.

Satagopan JM, Yandell BS (1996). Estimating the number of quantitative trait loci via Bayesian model determination. Special contributed paper session on Genetic Analysis of Quantitative Traits and Complex Disease, Biometric Section. *Joint Statistical Meetings*, August 5, Chicago.

Satagopan JM, Yandell BS, Newton MA, Osborn TC (1996). Markov chain Monte Carlo approach to detect polygene loci for complex traits. *Genetics* **144**: 805–816.

Sen Ś, Churchill G (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.

Sillanpää MJ, Arjas E (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.

Sillanpää MJ, Corander J (2002). Model choice in gene mapping: what and why. *Trends Genet* **18**: 301–307.

Stephens D, Smith A (1993). Bayesian inference in multipoint gene mapping. *Ann Hum Genet* **57**: 65–82.

Stephens DA, Fisch RD (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics* **54**: 1334–1347.

Stylianou IM, Korstanje R, Li R, Sheehan S, Paigen B, Churchill GA (2006). Quantitative trait locus analysis for obesity reveals multiple networks of interacting loci. *Mamm Genome* **17**: 22–36.

Sugiyama F, Churchill GA, Higgins DC, Johns C, Makaritsis KP, Gavras H *et al.* (2001). Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics* **71**: 70–77.

Uimari P, Thaller G, Hoeschele I (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* **143**: 1831–1842.

Valdar W, Solberg LC, Gauguier D, Cookson WO, Rawlins JNP, Mott R *et al.* (2006). Genetic and environmental effects on complex traits in mice. *Genetics* **174**: 959–984.

Wang H, Zhang YM, Li X, Masinde GL, Mohan S, Baylink DJ *et al.* (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**: 465–480.

Wang S, Yehya N, Schadt EE, Wang H, Drake TA, Lusis AJ (2006). Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet* **2**: 0148–0159.

Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.

Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY *et al.* (2007). R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. *Bioinformatics* **23**: 641–643.

Yi N (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* **167**: 967–975.

Yi N, Allison DB, Xu S (2003a). Bayesian model choice and search strategies for mapping multiple epistatic quantitative trait loci. *Genetics* **165**: 867–883.

Yi N, Banerjee S, Pomp D, Yandell BS (2007a). Bayesian mapping of genome-wide interacting QTL for ordinal traits. *Genetics* **176**: 1855–1864.

Yi N, George V, Allison DB (2003b). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**: 1129–1138.

Yi N, Shriner D, Banerjee S, Mehta T, Pomp D, Yandell BS (2007b). An efficient Bayesian model selection approach for interacting QTL models with many effects. *Genetics* **176**: 1865–1877.

Yi N, Xu S (2000). Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* **155**: 1391–1403.

Yi N, Xu S (2002). Mapping quantitative trait loci with epistatic effects. *Genet Res* **79**: 185–198.

Yi N, Xu S, George V, Allison DB (2004). Mapping multiple quantitative trait loci for complex ordinal traits. *Behav Genet* **34**: 3–15.

Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D (2005). Bayesian model selection for genome-wide QTL analysis. *Genetics* **170**: 1333–1344.

Yi N, Zinniel DK, Kim K, Eisen EJ, Bartolucci A, Allison DB *et al.* (2006). Bayesian analysis of multiple epistatic QTL models for body weight and body composition in mice. *Genet Res* **87**: 45–60.

Zeng ZB (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.

Zeng ZB, Kao C, Basten CJ (2000). Estimating the genetic architecture of quantitative traits. *Genet Res* **74**: 279–289.

Zeng ZB, Wang T, Zou W (2005). Modeling quantitative trait loci and interpretation of models. *Genetics* **169**: 1711–1725.

Zhang M, Montooth KL, Wells MT, Clark AG, Zhang D (2005). Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* **169**: 2305–2318.