DOI: 10.1534/genetics.104.040386

Bayesian Model Selection for Genome-Wide Epistatic Quantitative Trait Loci Analysis

Nengjun Yi,*,†,1 Brian S. Yandell,‡ Gary A. Churchill,§ David B. Allison,*,† Eugene J. Eisen and Daniel Pomp††**

**Department of Biostatistics, Section on Statistical Genetics, †Clinical Nutrition Research Center, University of Alabama, Birmingham, Alabama 35294, ‡Departments of Statistics and Horticulture, University of Wisconsin, Madison, Wisconsin 53706, §The Jackson Laboratory, Bar Harbor, Maine 04609, **Department of Animal Science, North Carolina State University, Raleigh, North Carolina 27695 and ††Department of Animal Science, University of Nebraska, Lincoln, Nebraska 68583*

Manuscript received December 29, 2004
Accepted for publication April 4, 2005

ABSTRACT

The problem of identifying complex epistatic quantitative trait loci (QTL) across the entire genome continues to be a formidable challenge for geneticists. The complexity of genome-wide epistatic analysis results mainly from the number of QTL being unknown and the number of possible epistatic effects being huge. In this article, we use a composite model space approach to develop a Bayesian model selection framework for identifying epistatic QTL for complex traits in experimental crosses from two inbred lines. By placing a liberal constraint on the upper bound of the number of detectable QTL we restrict attention to models of fixed dimension, greatly simplifying calculations. Indicators specify which main and epistatic effects of putative QTL are included. We detail how to use prior knowledge to bound the number of detectable QTL and to specify prior distributions for indicators of genetic effects. We develop a computationally efficient Markov chain Monte Carlo (MCMC) algorithm using the Gibbs sampler and Metropolis-Hastings algorithm to explore the posterior distribution. We illustrate the proposed method by detecting new epistatic QTL for obesity in a backcross of CAST/Ei mice onto M16i.

MANY complex human diseases and traits of biological and/or economic importance are determined by multiple genetic and environmental influences (Lynch and Walsh 1998). Mounting evidence suggests that interactions among genes (epistasis) play an important role in the genetic control and evolution of complex traits (Cheverud 2000; Carlborg and Haley 2004). Mapping quantitative trait loci (QTL) is a process of inferring the number of QTL, their genomic positions, and genetic effects given observed phenotype and marker genotype data. From a statistical perspective, two key problems in QTL mapping are model search and selection (*e.g.*, Broman and Speed 2002; Sillanpää and Corander 2002; Yi 2004). Traditional QTL mapping methods utilize a statistical model, which estimates the effects of only one QTL whose putative position is scanned across the genome (*e.g.*, Lander and Botstein 1989; Jansen and Stam 1994; Zeng 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. However, such an approach neglects potential confounding effects from additional QTL and requires prohibitive corrections for multiple testing. Non-Bayesian model selection methods combine simultaneous search with a sequential procedure such as forward or stepwise selection and apply criteria such as *P*-values or modified Bayesian information criterion (BIC) to identify well-fitting multiple-QTL models (Kao *et al.* 1999; Carlborg *et al.* 2000; Reifsnyder *et al.* 2000; Bogdan *et al.* 2004). These methods, although appealing in their simplicity and popularity, have several drawbacks, including: (1) the uncertainty about the model itself is ignored in the final inference, (2) they involve a complex sequential testing strategy that includes a dynamically changing null hypothesis, and (3) the selection procedure is heavily influenced by the quantity of data (Raftery *et al.* 1997; George 2000; Gelman *et al.* 2004; Kadane and Lazar 2004).

Bayesian model selection methods provide a powerful and conceptually simple approach to mapping multiple QTL (Satagopan *et al.* 1996; Hoeschele 2001; Sen and Churchill 2001). The Bayesian approach proceeds by setting up a likelihood function for the phenotype and assigning prior distributions to all unknowns in the problem. These induce a posterior distribution on the unknown quantities that contains all of the available information for inference of the genetic architecture of the trait. Bayesian mapping methods can treat the unknown number of QTL as a random variable,

[1] *Corresponding author:* Department of Biostatistics, University of Alabama, Ryals Public Health Bldg., 1665 University Blvd., Birmingham, AL 35294-0022. E-mail: nyi@ms.soph.uab.edu

which has several advantages but results in the complication of varying the dimension of the model space. The reversible jump Markov chain Monte Carlo (MCMC) algorithm, introduced by Green (1995), offers a powerful and general approach to exploring posterior distributions in this setting. However, the ability to "move" between models of different dimension requires a careful construction of proposal distributions. Despite the challenges of implementation of reversible jump algorithms, effective approaches for mapping multiple noninteracting QTL have been developed (Satagopan and Yandell 1996; Heath 1997; Thomas *et al.* 1997; Uimari and Hoeschele 1997; Sillanpää and Arjas 1998; Stephens and Fisch 1998; Yi and Xu 2000; Gaffney 2001). Bayesian model selection methods employing the reversible jump MCMC algorithm have been proposed to map epistatic QTL in inbred line crosses and outbred populations (Yi and Xu 2002; Yi *et al.* 2003, 2004a,b; Narita and Sasaki 2004). However, the complexity of the reversible jump steps increases computational demand and may prohibit improvements of the algorithms.

Recently, Yi (2004) proposed a unified Bayesian model selection framework to identify multiple nonepistatic QTL for complex traits in experimental designs, based upon a composite space representation of the problem. The composite space approach, which is a modification of the product space concept developed by Carlin and Chib (1995), provides an interesting viewpoint on a wide variety of model selection problems (Godsill 2001). The key feature of the composite model space is that the dimension remains fixed, allowing for MCMC simulation to be performed on a space of fixed dimension, thus avoiding the complexities of reversible jump. In Yi (2004), the varying dimensional space is augmented to a fixed dimensional space (the composite model space) by placing an upper bound on the number of detectable QTL. In the composite model space, latent binary variables indicate whether each putative QTL has a nonzero effect. The resulting hierarchical model can vastly simplify the MCMC search strategy.

In this work we extend the composite model space approach to include epistatic effects. We develop a framework of Bayesian model selection for mapping epistatic QTL in experimental crosses from two inbred lines. We show how to incorporate prior knowledge to select an upper bound on the number of detectable QTL and prior distributions for indicator variables of genetic effects and other parameters. A computationally efficient MCMC algorithm using a Gibbs sampler or Metropolis-Hastings (M-H) algorithm is developed to explore the posterior distribution on the parameters. The proposed algorithm is easy to implement and allows more complete and rapid exploration of the model space. We first describe the implementation of this algorithm and then illustrate the method by analyzing a mouse backcross population.

A BAYESIAN MODEL SELECTION FRAMEWORK FOR QTL MAPPING

We consider experimental crosses derived from two inbred lines. In QTL studies, the observed data consist of phenotypic trait values, $\mathbf{y}$, and marker genotypes, $\mathbf{m}$, for individuals in a mapping population. We assume that markers are organized into a linkage map and restrict attention to models with, at most, pairwise interactions. We partition the entire genome into H loci, $\boldsymbol{\zeta} = \{\zeta_1, \ldots, \zeta_H\}$, and assume that the possible QTL occur at these fixed positions. This introduces only a minor bias in estimating the position of QTL when H is large. When the markers are densely and regularly spaced, we set $\boldsymbol{\zeta}$ to the marker positions; otherwise, $\boldsymbol{\zeta}$ includes not only the marker positions but also points between markers. In general, the genotypes, $\mathbf{g}$, at loci $\boldsymbol{\zeta}$ are unobservable except at completely informative markers, but their probability distribution, $p(\mathbf{g}|\boldsymbol{\zeta}, \mathbf{m})$, can be inferred from the observed marker data using the multipoint method (Jiang and Zeng 1997). This probability distribution is used as the prior distribution of QTL genotypes in our Bayesian framework.

The problem of inferring the number and locations of multiple QTL is equivalent to the problem of selecting a subset of $\boldsymbol{\zeta}$ that fully explains the phenotypic variation. Although a complex trait may be influenced by multitudes of loci, our emphasis is on a set of at most L QTL with detectable effects. Typically L will be much smaller than H. Let $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_L\}$ $(\in\{\zeta_1, \ldots, \zeta_H\})$ be the current positions of L putative QTL. Each locus may affect the trait through its marginal (main) effects and/or interactions with other loci (epistasis). The phenotype distribution is assumed to follow a linear model,

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1}$$

where $\boldsymbol{\mu}$ is the overall mean, $\boldsymbol{\beta}$ denotes the vector of all possible main effects and pairwise interactions of L potential QTL, $\mathbf{X}$ is the design matrix, and $\mathbf{e}$ is the vector of independent normal errors with mean zero and variance σ^2. The number of genetic effects depends on the experimental design, and the design matrix $\mathbf{X}$ is determined from those genotypes $\mathbf{g}$ at the current loci $\boldsymbol{\lambda}$ by using a particular genetic model (see appendix a for details of the Cockerham genetic model used here).

There is prior uncertainty about which genetic effects should be included in the model. As in Bayesian variable selection for linear regression (*e.g.*, George and McCulloch 1997; Kuo and Mallick 1998; Chipman *et al.* 2001), we introduce a binary variable γ for each effect, indicating that the corresponding effect is included ($\gamma = 1$) or excluded ($\gamma = 0$) from a model. Letting $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$, the model becomes

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta} + \mathbf{e}. \tag{2}$$

This linear model defines the likelihood, $p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}, \boldsymbol{\theta})$, with $\boldsymbol{\theta} = (\mu, \boldsymbol{\beta}, \sigma^2)$, and the full posterior can be written as

$$p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{m}) \propto p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}, \boldsymbol{\theta})\ p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta}|\mathbf{m}). \qquad (3)$$

Specifications of priors $p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta}|\mathbf{m})$ and posterior calculation are given in subsequent sections.

The vector $\boldsymbol{\gamma}$ determines the number of QTL (see APPENDIX B). Hereafter, we denote the included positions of QTL by $\boldsymbol{\lambda}_\gamma$. The vector $(\boldsymbol{\gamma}, \boldsymbol{\lambda}_\gamma)$ comprises a model index that identifies the genetic architecture of the trait. A natural model selection strategy is to choose the most probable model $(\boldsymbol{\gamma}, \boldsymbol{\lambda}_\gamma)$ on the basis of its marginal posterior, $p(\boldsymbol{\gamma}, \boldsymbol{\lambda}_\gamma|\mathbf{y}, \mathbf{m})$ (George and Foster 2000). For genome-wide epistatic analysis, however, no single model may stand out, and thus we average over possible models when assessing characteristics of genetic architecture, with the various models weighted by their posterior probability (Raftery *et al.* 1997; Ball 2001; Sillanpää and Corander 2002).

PRIOR DISTRIBUTIONS

The above Bayesian model selection framework provides a conceptually simple and general method to identify complex epistatic QTL across the entire genome. However, its practical implementation entails two challenges: prior specification and posterior calculation. In this section, we first propose a method to choose an upper bound for the number of QTL and then describe the prior specifications for the model index and other unknowns.

Choice of the upper bound *L*: We suggest first specifying the prior expected number of QTL, l_0, on the basis of initial investigations with traditional methods, and then determining a reasonably large upper bound, L. We assign the prior probability distribution for the number of QTL, l, to be a Poisson distribution with mean l_0. The value of L can be selected to be large enough that the probability $\Pr(l > L)$ is very small. On the basis of a normal approximation to the Poisson distribution, we could take L as $l_0 + 3\sqrt{l_0}$.

Prior on $\boldsymbol{\gamma}$: For the indicator vector $\boldsymbol{\gamma}$, we use an independence prior of the form

$$p(\boldsymbol{\gamma}) = \prod w_j^{\gamma_j}(1 - w_j)^{1-\gamma_j}, \qquad (4)$$

where $w_j = p(\gamma_j = 1)$ is the prior inclusion probability for the jth effect. We assume that w_j equals the predetermined hyperparameter w_{m} or w_{e}, depending on the jth effect being main effect or 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 main effect is different from that of epistatic effect.

The hyperparameters w_{m} and w_{e} control the expected numbers of main and epistatic effects included in the model, respectively; 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 may be better to first determine the prior expected numbers of main-effect QTL, l_{m}, and all QTL, $l_0 \geq l_{\mathrm{m}}$ (*i.e.*, main-effect and epistatic QTL), and then solve for w_{m} and w_{e} from the expressions of the prior expected numbers. It is reasonable to require that $w_{\mathrm{m}} \geq w_{\mathrm{e}}$, which requires some adjustment below when $l_{\mathrm{m}} = 0$.

As shown in APPENDIX B, the prior expected number of main-effect QTL can be expressed as

$$l_{\mathrm{m}} = L[1 - (1 - w_{\mathrm{m}})^K], \qquad (5)$$

and the prior expected number of all QTL as

$$l_0 = L[1 - (1 - w_{\mathrm{m}})^K(1 - w_{\mathrm{e}})^{K^2(L-1)}], \qquad (6)$$

where K is the number of possible main effects for each QTL and K^2 is the number of possible epistatic effects for any two QTL.

The prior expected number of main-effect QTL, l_{m}, could be set to the number of QTL detected by traditional nonepistatic mapping methods, *e.g.*, 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, *e.g.*, two-dimensional genome scan, could provide a rough guide for choosing l_0. From Equations 5 and 6, we obtain

$$w_{\mathrm{m}} = 1 - \left[1 - \frac{l_{\mathrm{m}}}{L}\right]^{1/K} \qquad (7)$$

and

$$w_{\mathrm{e}} = 1 - \left[\frac{1 - (l_0/L)}{(1 - w_{\mathrm{m}})^K}\right]^{1/K^2(L-1)}. \qquad (8)$$

We note above that if no main-effect QTL is detected by traditional nonepistatic mapping methods and $l_{\mathrm{m}} = 0$, then $w_{\mathrm{m}} = 0$. In this case, we suggest making all weights equal, $w_{\mathrm{m}} = w_{\mathrm{e}} \triangleq w$, and using (6) to obtain

$$w = 1 - \left(1 - \frac{l_0}{L}\right)^{1/(K+K^2(L-1))}. \qquad (9)$$

Prior on $\boldsymbol{\lambda}$: When there is no prior information concerning QTL locations, these could be assumed to be independent and uniformly distributed over the H possible loci. Thus, given l_0 the prior probability that any locus is included becomes l_0/H. In practice, it may be reasonable to assume that any intervals of a given length (*e.g.*, 10 cM) contain at most one QTL. Although this assumption is not necessary, it can substantially reduce the model space and thus accelerate the search procedure.

Prior on $\boldsymbol{\beta}$: We propose the following hierarchical mixture prior for each genetic effect,

$$\beta_j|(\gamma_j, \sigma^2, \mathbf{x}_{\cdot j}) \sim N(0, \gamma_j c\sigma^2(\mathbf{x}_{\cdot j}^T\mathbf{x}_{\cdot j})^{-1}), \qquad (10)$$

where $\mathbf{x}_{\cdot j} = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ is the vector of the coefficients of β_j, and c is a positive scale factor. Many suggestions have been proposed for choice of c for variable selection problems of linear regression (*e.g.*, Chipman *et al.* 2001; Fernandez *et al.* 2001). In this study, we take $c = n$, which is a popular choice and yields the BIC if the prior inclusion probability for each effect equals 0.5 (*e.g.*, George and Foster 2000; Chipman *et al.* 2001).

In this prior setup, a point mass prior at 0 is used for the genetic effect β_j when $\gamma_j = 0$, effectively removing β_j from the model. If $\gamma_j = 1$, the prior variances reflect the precision of each β_j and are invariant to scales changes in the phenotype and the coefficients. The value $(\mathbf{x}_{\cdot j}^{\mathrm{T}}\mathbf{x}_{\cdot j})^{-1}$ varies for different types of genetic effects. For a large backcross population with no segregation distortion, for example, $(\mathbf{x}_{\cdot j}^{\mathrm{T}}\mathbf{x}_{\cdot j})^{-1}/n \approx \frac{1}{4}$ for marginal effects and $[1 - (1 - 2r)^2]/16$ for epistatic effects, with r the recombination fraction between two QTL, under Cockerham's model (Zeng *et al.* 2000).

Priors on μ and σ^2: The prior for the overall mean μ is $N(\eta_0, \tau_0^2)$. We could empirically set

$$\eta_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \tau_0^2 = s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

We take the noninformative prior for the residual variance, $p(\sigma^2) \propto 1/\sigma^2$ (Gelman *et al.* 2004). Although this prior is improper, it yields a proper posterior distribution for the unknowns and so can be used formally (Chipman *et al.* 2001).

MARKOV CHAIN MONTE CARLO ALGORITHM

To develop our MCMC algorithm, we first partition the vector of unknowns $(\boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta})$ into $(\boldsymbol{\lambda}_\gamma, \mathbf{g}_\gamma, \boldsymbol{\theta}_\gamma)$ and $(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma})$, representing the unknowns included or excluded from the model, respectively, where $\boldsymbol{\lambda}_\gamma$ and $\mathbf{g}_\gamma$ ($\boldsymbol{\lambda}_{-\gamma}$ and $\mathbf{g}_{-\gamma}$) are the positions and the genotypes of QTL included (excluded), respectively, $\boldsymbol{\beta}_\gamma$ ($\boldsymbol{\beta}_{-\gamma}$) represent the genetic effects included (excluded), $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu, \sigma^2)$, $\boldsymbol{\theta}_\gamma = (\boldsymbol{\beta}_\gamma, \mu, \sigma^2)$, and $\boldsymbol{\theta}_{-\gamma} = \boldsymbol{\beta}_{-\gamma}$. Similarly, $\mathbf{X}_\gamma$ ($\mathbf{X}_{-\gamma}$) represent the model coefficients included (excluded), which are determined by $\mathbf{g}$ and $\boldsymbol{\gamma}$.

We suppress the dependence on the observed marker data below. For a particular $\boldsymbol{\gamma}$ the likelihood function depends only upon the parameters $(\mathbf{X}_\gamma, \boldsymbol{\theta}_\gamma)$ used by that model, *i.e.*,

$$p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}_\gamma, \boldsymbol{\theta}_\gamma). \tag{11}$$

The prior distribution of $(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{g}, \boldsymbol{\theta})$ can be partitioned as

$$p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta}) = p(\boldsymbol{\gamma})p(\boldsymbol{\lambda}_\gamma, \mathbf{g}_\gamma, \boldsymbol{\theta}_\gamma|\boldsymbol{\gamma})p(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma}|\boldsymbol{\gamma}). \tag{12}$$

The full posterior distribution for $(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta})$ can now be expressed as

$$p(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}_\gamma, \boldsymbol{\theta}_\gamma)p(\boldsymbol{\gamma})p(\boldsymbol{\lambda}_\gamma, \mathbf{g}_\gamma, \boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}) \times p(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma}|\boldsymbol{\gamma}). \tag{13}$$

From (13), we can derive the conditional posterior distributions

$$p(\boldsymbol{\lambda}_\gamma, \mathbf{g}_\gamma, \boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}_\gamma, \boldsymbol{\theta}_\gamma)p(\boldsymbol{\lambda}_\gamma, \mathbf{g}_\gamma, \boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}), \tag{14}$$

$$p(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma}|\boldsymbol{\gamma}, \mathbf{y}) \propto p(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma}|\boldsymbol{\gamma}), \tag{15}$$

and

$$p(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathbf{g}, \boldsymbol{\theta}, \mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}_\gamma, \boldsymbol{\theta}_\gamma)p(\boldsymbol{\gamma})p(\boldsymbol{\lambda}_\gamma, \mathbf{g}_\gamma, \boldsymbol{\theta}_\gamma|\boldsymbol{\gamma}) \times p(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma}|\boldsymbol{\gamma}). \tag{16}$$

It can be seen that the unused parameters do not affect the conditional posterior of $(\boldsymbol{\lambda}_\gamma, \mathbf{g}_\gamma, \boldsymbol{\theta}_\gamma)$ and thus do not need to be updated conditional on $\boldsymbol{\gamma}$. Since the unused parameters do not contribute to the likelihood, the posterior of $(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma})$ is identical to its prior. From (16), the conditional posterior of $\boldsymbol{\gamma}$ depends on $(\boldsymbol{\lambda}_{-\gamma}, \mathbf{g}_{-\gamma}, \boldsymbol{\theta}_{-\gamma})$ and thus the update of $\boldsymbol{\gamma}$ requires generation of the corresponding unused parameters in the current model. These properties lead us to develop MCMC algorithms as described below. We first briefly describe the algorithms for updating $\boldsymbol{\theta}_\gamma$, $\mathbf{g}_\gamma$, and $\boldsymbol{\lambda}_\gamma$ and then develop a novel Gibbs sampler and Metropolis-Hastings algorithm to update the indicator variables for main and epistatic effects, respectively.

Conditional on $\boldsymbol{\gamma}$, $\mathbf{X}_\gamma$, and $\boldsymbol{\lambda}_\gamma$, the parameters μ, σ^2, and $\boldsymbol{\beta}_\gamma$ can be sampled directly from their posterior distributions, which have standard form (Gelman *et al.* 2004). Conditional on $\boldsymbol{\gamma}$, $\boldsymbol{\lambda}_\gamma$, and $\boldsymbol{\theta}_\gamma$, the posterior distribution of each element of $\mathbf{g}_\gamma$ is multinomial and thus can be sampled directly as well (Yi and Xu 2002). We adapt the algorithm of Yi *et al.* (2003) to our model to update locations $\boldsymbol{\lambda}_\gamma$: (1) $\boldsymbol{\lambda}$ is restricted to the discrete space $\boldsymbol{\zeta} = \{\zeta_1, \ldots, \zeta_H\}$, and (2) any intervals of some length δ include at most one QTL. To update λ_q, therefore, we propose a new location λ_q^* for the qth QTL uniformly from $2d$ most flanking loci of λ_q, where d is a predetermined integer (*e.g.*, $d = 2$), and then generate genotypes at the new location for all individuals. The proposals for the new location and the genotypes are then jointly accepted or rejected using the Metropolis-Hastings algorithm.

At each iteration of the MCMC simulation, we update all elements of $\boldsymbol{\gamma}$ in some fixed or random order. For the indicator variable of a main effect, we need to consider two different cases: a QTL is currently (1) in or (2) out of the model. For (1), the QTL position and genotypes were generated at the preceding iteration. For (2), we sample a new QTL position from its prior distribution and generate its genotypes for all individuals. An epistatic effect involves two QTL, hence three different cases: (1) both QTL are in, (2) only one QTL is in, and (3) both QTL are out of the model. Again, the new QTL position(s) and genotypes are sampled as needed.

We update γ_j, the indicator variable for an effect, using its conditional posterior distribution of γ_j, which is Bernoulli,

$$p(\gamma_j = 1|\boldsymbol{\gamma}_{-\gamma_j}, \mathbf{X}, \boldsymbol{\theta}_{-\beta_j}, \mathbf{y}) = 1 - p(\gamma_j = 0|\boldsymbol{\gamma}_{-\gamma_j}, \mathbf{X}, \boldsymbol{\theta}_{-\beta_j}, \mathbf{y})$$

$$= \frac{wR}{(1 - w) + wR}, \qquad (17)$$

where

$$R = \frac{p(\mathbf{y}|\gamma_j = 1, \boldsymbol{\gamma}_{-\gamma_j}, \mathbf{X}, \boldsymbol{\theta}_{-\beta_j})}{p(\mathbf{y}|\gamma_j = 0, \boldsymbol{\gamma}_{-\gamma_j}, \mathbf{X}, \boldsymbol{\theta}_{-\beta_j})} = \left(\frac{\sigma_{\beta_j}^{-2} + \sigma^{-2}\sum_{i=1}^{n} x_{ij}^2}{\sigma_{\beta_j}^{-2}}\right)^{-0.5}$$

$$\times \exp\left(\frac{1}{2}\frac{(\sum_{i=1}^{n} x_{ij}(y_i - \mu - \mathbf{x}_{i\cdot}\boldsymbol{\beta} + x_{ij}\beta_j)\sigma^{-2})^2}{\sigma_{\beta_j}^{-2} + \sigma^{-2}\sum_{i=1}^{n} x_{ij}^2}\right),$$

$\mathbf{x}_{i\cdot}$ is the vector of the coefficients of $\boldsymbol{\beta}$ for the ith individual, $w = \mathrm{pr}(\gamma_j = 1)$ is the prior probability that β_j appears in the model, $\sigma_{\beta_j}^2$ is the prior variance of β_j (see Equation 10), $\boldsymbol{\gamma}_{-\gamma_j}$ means all the elements of $\boldsymbol{\gamma}$ except for γ_j, and $\boldsymbol{\theta}_{-\beta_j}$ represents all the elements of $\boldsymbol{\theta}$ except for β_j. We can sample γ_j directly from (17) or update γ_j with probability $\min(1, r)$, where $r = ((w/1 - w)R)^{1-2\gamma_j}$.

The effect β_j was integrated from (17). We can generate β_j as follows. If γ_j is sampled to be zero, $\beta_j = 0$. Otherwise, β_j is generated from its conditional posterior

$$p(\beta_j|\gamma_j = 1, \boldsymbol{\gamma}_{-\gamma_j}, \mathbf{X}, \boldsymbol{\theta}_{-\beta_j}, \mathbf{y}) = N(\tilde{\mu}_j, \tilde{\sigma}_j^2), \qquad (18)$$

where

$$\tilde{\mu}_j = (\sigma^2\sigma_{\beta_j}^{-2} + \sum_{i=1}^{n} x_{ij}^2)^{-1}\sum_{i=1}^{n} x_{ij}(y_i - \mu - \mathbf{x}_{i\cdot}\boldsymbol{\beta} + x_{ij}\beta_j)$$

and

$$\tilde{\sigma}_j^{-2} = \sigma_{\beta_j}^{-2} + \sigma^{-2}\sum_{i=1}^{n} x_{ij}^2.$$

POSTERIOR ANALYSIS

The MCMC algorithm described above starts from initial values and updates each group of unknowns in turn. Initial iterations are discarded as "burn-in." To reduce serial correlation, we thin the subsequent samples by keeping every kth simulation draw and discarding the rest, where k is an integer. The MCMC sampler sequence $\{(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\lambda}_{\gamma}^{(t)}, \mathbf{g}_{\gamma}^{(t)}, \boldsymbol{\theta}_{\gamma}^{(t)});\ t = 1, \ldots, N\}$ is a random draw from the joint posterior distribution $p(\boldsymbol{\gamma}, \boldsymbol{\lambda}_{\gamma}, \mathbf{g}_{\gamma}, \boldsymbol{\theta}_{\gamma}|\mathbf{y})$, and thus the embedded subsequence $\{(\boldsymbol{\gamma}^{(t)}, \boldsymbol{\lambda}_{\gamma}^{(t)});\ t = 1, \ldots, N\}$ is a random sample from its marginal posterior distribution $p(\boldsymbol{\gamma}, \boldsymbol{\lambda}_{\gamma}|\mathbf{y})$, which is used to infer the genetic architecture of the complex trait. For genome-wide epistatic analysis, no single model may stand out, and we may average over all possible models to assess genetic architecture. Bayesian model averaging provides more robust inferences about quantities of interest than any single model since it incorporates model uncertainty (Raftery *et al.* 1997; Ball 2001; Sillanpää and Corander 2002).

The most important characteristic may be the posterior inclusion probability of each possible locus ζ_h, estimated as

$$p(\zeta_h|\mathbf{y}) = \frac{1}{N}\sum_{t=1}^{N}\sum_{q=1}^{L} 1(\lambda_q^{(t)} = \zeta_h, \xi_q^{(t)} = 1), \quad h = 0, 1, \ldots, H, \qquad (19)$$

where ξ_q is the binary indicator that QTL q is included or excluded from the model. Thus, we can obtain the cumulative distribution function per chromosome, defined as $F_c(x|\mathbf{y}) = \sum_{\zeta_h=0}^{x} p(\zeta_h|\mathbf{y})$ for any position x on chromosome c. It is worth noting that the cumulative distribution function defined here can be >1 if the corresponding chromosome contains more than one QTL. Both $p(\zeta_h|\mathbf{y})$ and $F_c(x|\mathbf{y})$ can be graphically displayed and show evidence of QTL activity across the whole genome. Commonly used summaries include the posterior probability that a chromosomal region contains QTL, the most likely position of QTL (the mode of QTL positions), and the region of highest posterior density (HPD) (*e.g.*, Gelman *et al.* 2004). To take the prior specifications, $p(\zeta_h)$, into consideration, we can use the Bayes factor to show evidence for inclusion of ζ_h against exclusion of ζ_h (Kass and Raftery 1995),

$$\mathrm{BF}(\zeta_h) = \frac{p(\zeta_h|\mathbf{y})}{1 - p(\zeta_h|\mathbf{y})} \cdot \frac{1 - p(\zeta_h)}{p(\zeta_h)}. \qquad (20)$$

In a similar fashion, we can compute the Bayes factor comparing a chromosomal region containing QTL to that excluding QTL.

We can estimate the main effects at any locus or chromosomal intervals Δ,

$$\beta_k(\Delta) = \frac{1}{N}\sum_{t=1}^{N}\sum_{q=1}^{L} 1(\lambda_q^{(t)} \in \Delta, \xi_q^{(t)} = 1)\beta_{qk}^{(t)}, \quad k = 1, 2, \ldots, K. \qquad (21)$$

The heritabilities explained by the main effects can also be estimated. In epistatic analysis, we need to estimate two types of additional parameters, the posterior inclusion probability and the size of epistatic effects, both involving pairs of loci. These two types of unknowns can be estimated with natural extensions of (19) and (21), respectively.

EXAMPLE

We illustrate the application of our Bayesian model selection approach by an analysis of a mouse cross produced from two highly divergent strains: M16i, consisting of large and moderately obese mice, and CAST/Ei, a wild strain of small mice with lean bodies (Leamy *et al.* 2002). CAST/Ei males were mated to M16i females, and F_1 males were backcrossed to M16i females, resulting in 54 litters and 421 mice (213 males, 208 females) reaching 12 weeks of age. All mice were genotyped for 92 microsatellite markers located on 19 autosomal chromosomes. The marker linkage map covered 1214 cM with average spacing of 13 cM. In this study, we analyze FAT, the sum of right gonadal and hindlimb subcutaneous fat pads. Prior to QTL analysis, the phenotypic data were

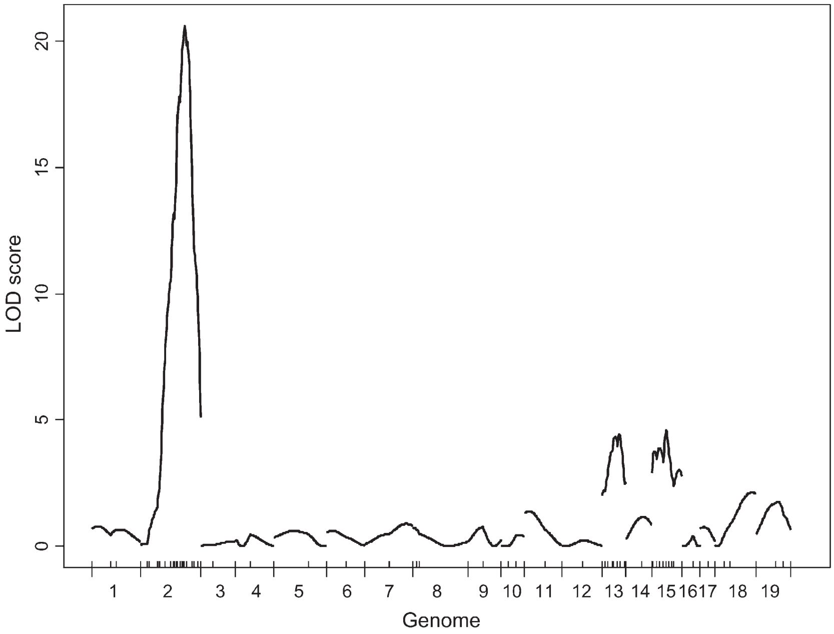

FIGURE 1.—Profiles of LOD scores from maximum-likelihood interval mapping. On the x-axis, large tick marks represent chromosomes and small tick marks represent markers.

linearly adjusted by sex and dam and standardized to mean 0 and variance 1, although this transformation may result in the possibility of destroying true biological interaction (JANSEN 2003). We used the Cockerham genetic model (APPENDIX A), in which the coefficients of main effects are defined as 0.5 and −0.5 for the two genotypes, CM and MM, where C and M represent the CAST/Ei and M16i alleles, respectively.

We partitioned each chromosome with a 1-cM grid, resulting in 1214 possible loci across the genome. A nonepistatic and an epistatic QTL model were evaluated. For all analyses, the MCMC started with no QTL and ran for 4×10^5 cycles after discarding the first 2000 burn-ins. The chain was thinned by one in $k = 20$, yielding 2×10^4 samples for posterior Bayesian analysis.

An initial interval map scan revealed three significant QTL (LOD > 3.2) on chromosomes 2, 13, and 15 (Figure 1), explaining 20.7, 4.9, and 5.1% of the phenotypic variance, respectively.

Under the nonepistatic analysis, epistatic effects are always excluded from the model and thus putative QTL are chosen only on the basis of their main effects. As

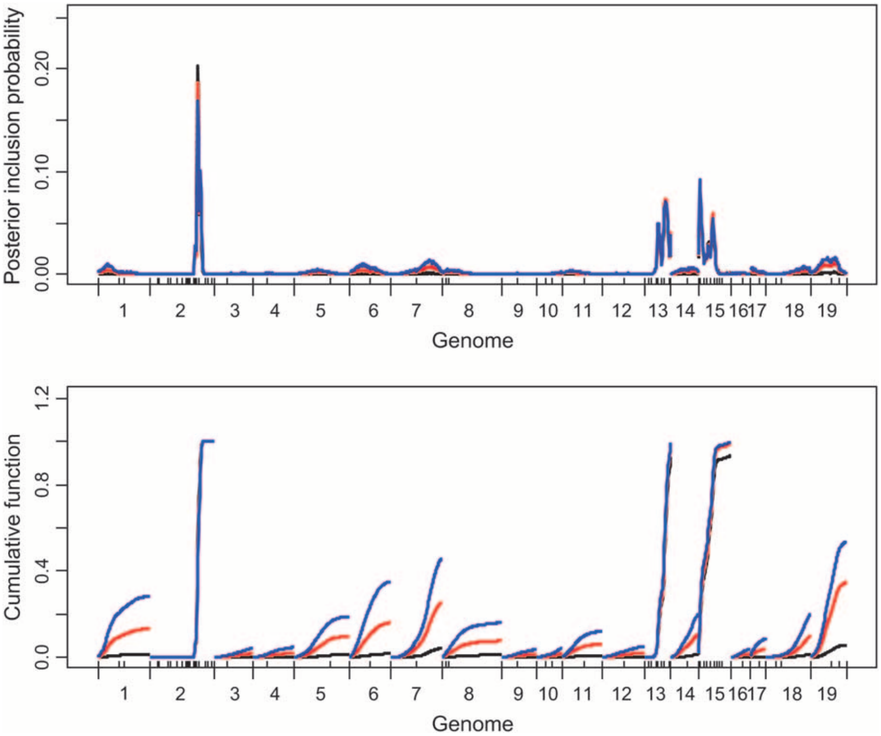

FIGURE 2.—Bayesian nonepistatic analysis: profiles of posterior inclusion probability and cumulative probability function. Black line, $l_m = 1$; red line, $l_m = 3$; blue line, $l_m = 6$. On the x-axis, large tick marks represent chromosomes and small tick marks represent markers.

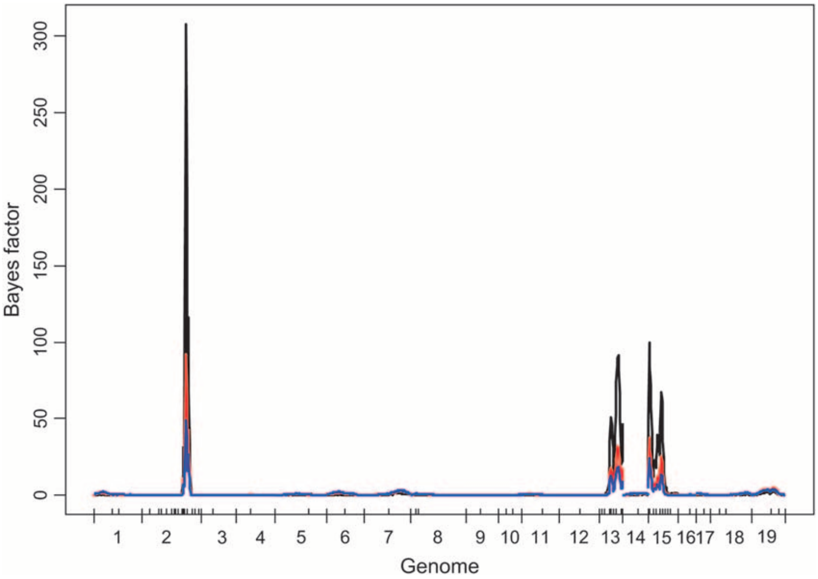

FIGURE 3.—Bayesian nonepistatic analysis: profiles of Bayes factor. Black line, $l_{\mathrm{m}} = 1$; red line, $l_{\mathrm{m}} = 3$; blue line, $l_{\mathrm{m}} = 6$. On the *x*-axis, large tick marks represent chromosomes and small tick marks represent markers.

described earlier, we took the number of significant QTL detected in the interval mapping as the prior number of main-effect QTL (l_{m}). To check prior sensitivity, we reran the algorithm for $l_{\mathrm{m}} = 1, 6$. The upper bound of the number of QTL was calculated as $L \approx l_{\mathrm{m}} + 3\sqrt{l_{\mathrm{m}}}$, or $L = 9$, 4, and 14 for $l_{\mathrm{m}} = 3$, 1, and 6, respectively. Therefore, the prior probabilities of inclusion for each main effect were $w_{\mathrm{m}} = 1 - [1 - (l_{\mathrm{m}}/L)]^{1/K} = \frac{1}{3}, \frac{1}{4}$, and $\frac{3}{7}$, respectively. Figure 2, top, displays the posterior probability of inclusion for each locus across the genome. Note the similarity to Figure 1, with clear evidence of QTL and flat profiles on other chromosomes. The peaks

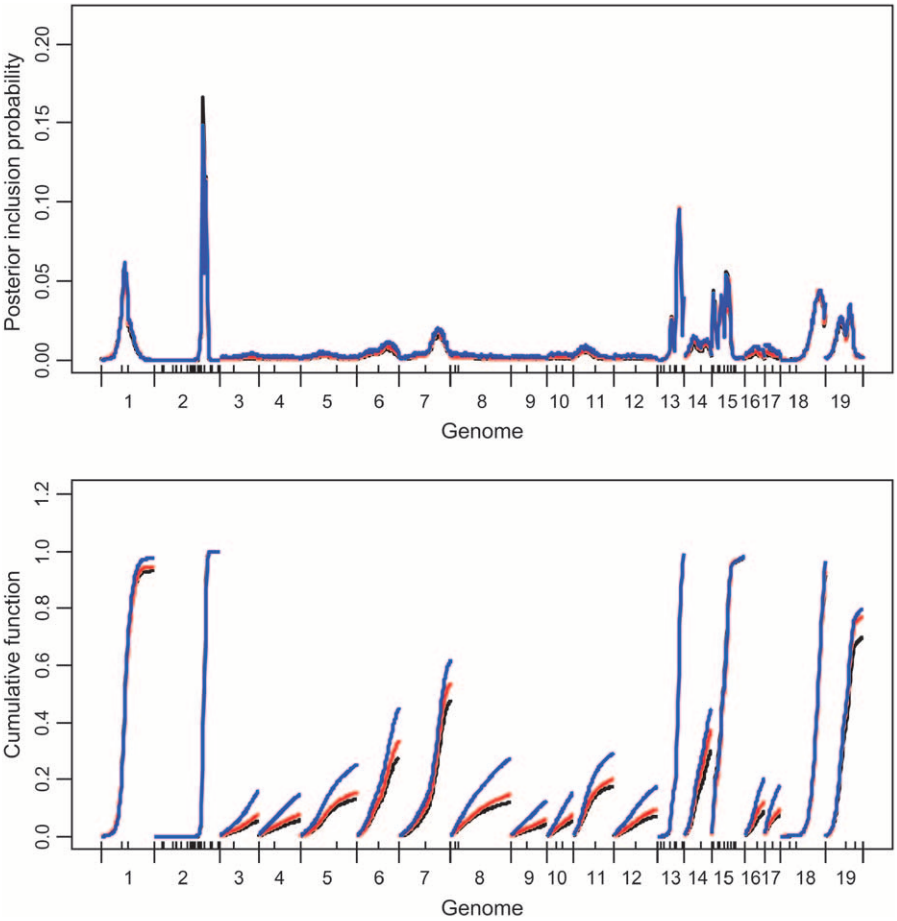

FIGURE 4.—Bayesian epistatic analysis: profiles of posterior inclusion probability and cumulative probability function. Black line, $l_0 = 4$; red line, $l_0 = 6$; blue line, $l_0 = 8$. On the *x*-axis, large tick marks represent chromosomes and small tick marks represent markers.

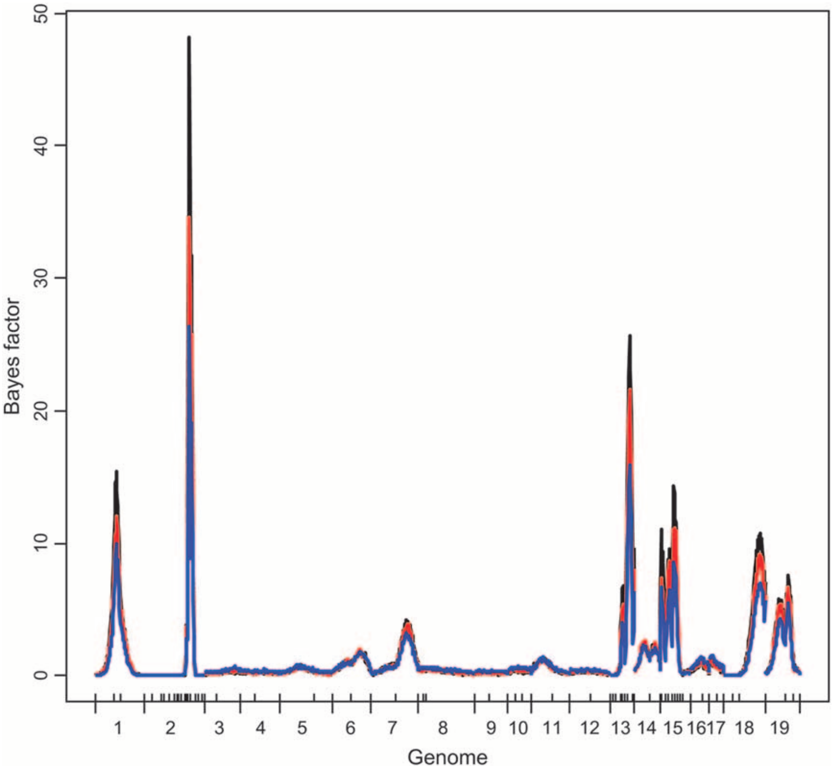

FIGURE 5.—Bayesian epistatic analysis: profiles of Bayes factor. Black line, $l_0 = 4$; red line, $l_0 = 6$; blue line, $l_0 = 8$. On the *x*-axis, large tick marks represent chromosomes and small tick marks represent markers.

on chromosomes 2, 13, and 15 overlap those identified by interval mapping. The graphs of the cumulative distribution function, displayed in Figure 2, bottom, show that the posterior inclusion probability of each chromosome is close to 1 for chromosomes 2, 13, and 15. The results show that, at least in this data set, detection of large-effect QTL is not sensitive to the choice of l_{m}. However, larger l_{m} tend to pick up more small-effect QTL as expected. The profiles of the Bayes factor are depicted in Figure 3. For the three choices of l_{m}, the regions on chromosomes 2, 13, and 15 show strong evidence for being selected, and other regions show a very low Bayes factor.

The epistatic analysis took $l_{\mathrm{m}} = 3$, the number of QTL detected in the nonepistatic analyses, as the prior expected number of main-effect QTL. Three values, $l_0 = 4$, 6, and 8, were chosen as the prior expected number of all QTL under the epistatic model. The upper bound of the number of QTL, L, was thus $L = 10$, 14, and 17, respectively. From Equations 7 and 8, the prior inclusion probabilities were 0.30, 0.21, and 0.18 for main effects and 0.017, 0.025, and 0.027 for epistatic effects, for the three values of (l_0, L), respectively. The profiles of the posterior inclusion probability for each locus across the genome and the cumulative posterior probability for each chromosome are depicted in Figure 4, top and bottom, respectively. It can be seen that the three different prior specifications of (l_0, L) provided fairly similar profiles of the posteriors, indicating that the posterior inference may be not very sensitive toward the small or mediate change of l_0. As expected, the choice of a smaller prior expected number of QTL tended to provide smaller posteriors, especially for infrequently arising loci. However, the identification of frequent arising loci remained the same. The profiles of the Bayes factor are depicted in Figure 5. The three choices of l_{m} provided similar profiles of the Bayes factor, especially for infrequently arising loci.

As shown in Figures 4 and 5, the epistatic analyses detected the same regions on chromosomes 2, 13, and 15 as the nonepistatic analyses. In addition to those on chromosomes 2, 13, and 15, our epistatic analyses found strong evidence of QTL on chromosomes 1, 18, and 19 with high cumulative probabilities (close to 1) and suggestive evidence of QTL on chromosomes 7 and 14. In the nonepistatic analyses, these chromosomes were found to have weak main effects and hence were detected in the epistatic model mainly due to epistatic interactions.

The profiles of the location-wise main effects and the variances explained by the main effects are depicted in Figure 6. For the three prior specifications, the posterior inferences were essentially identical. Therefore, we reported only the summary statistics for $l_0 = 6$ (see Tables 1 and 2). For the HPD regions on chromosomes 2, 13, and 15, the posterior inclusion probabilities are close to 1, and the corresponding Bayes factors are high. The estimated main effects were −0.856, 0.371, and −0.342 and explained 18.4, 3.5, and 3.1% of the phenotypic variance, respectively. For the HPD regions on chromosomes 1, 18, and 19, the posterior inclusion probabilities were ∼82, 88, and 70%, and the corresponding Bayes factors were ∼28, 47, and 12, respectively. In these HPD regions, the average main effects were weak and ex-

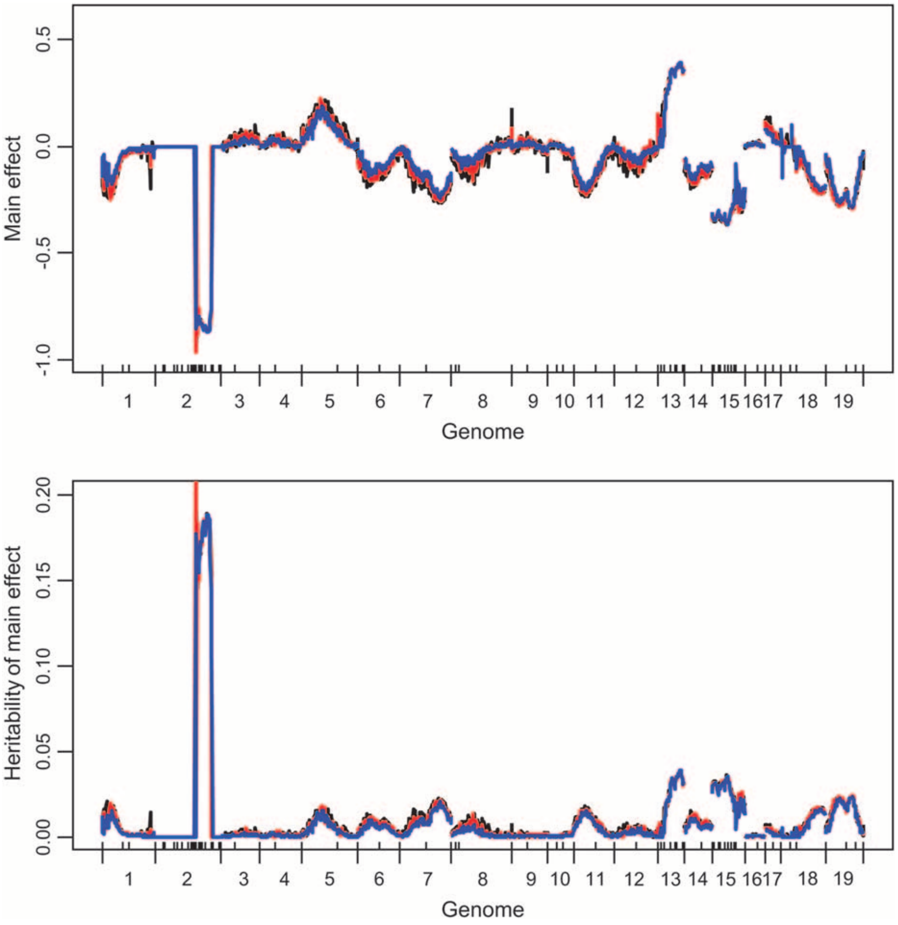

FIGURE 6.—Bayesian epistatic analysis: profiles of main effect and heritability explained by main effect. Black line, $l_0 = 4$; red line, $l_0 = 6$; blue line, $l_0 = 8$. On the *x*-axis, large tick marks represent chromosomes and small tick marks represent markers.

plained low proportions of the phenotypic variance. However, our epistatic analyses detected strong epistatic interactions associated with the HPD regions on chromosomes 1, 18, and 19. As shown in Table 2, the strongest epistasis is the interaction between chromosomes 1 and 18. This epistatic effect was estimated to be 0.936 and explained 5.6% of the phenotypic variance. The posterior inclusion probability of this epistasis was 81.9%. The region of chromosome 19 was found to interact with chromosomes 15 and 7. The interaction between the regions of chromosomes 19 and 15 was 0.604 and explained 2.5% of the phenotypic variance. The epistatic analyses also revealed interactions among chromosomes 2, 13, and 15. For example, the interaction between the HPD regions on chromosomes 2 and 13 was included in the model with probability of ~60% and explained ~2.5% of the phenotypic variance.

DISCUSSION

The Bayesian model selection approach provides a comprehensive solution to mapping multiple epistatic QTL across the entire genome using the posterior distribution as a selection criterion. MCMC algorithms based on the composite model space representation mix rapidly, thus ensuring that high-probability models are visited frequently and quickly, resulting in good inference from relatively short runs. The Bayesian framework provides a robust inference of genetic architecture that incorporates model uncertainty by averaging over all possible models (RAFTERY *et al.* 1997; BALL 2001; SILLANPÄÄ and CORANDER 2002).

One of the most challenging statistical problems presented by QTL mapping is that the number of QTL is unknown. Most previous Bayesian mapping methods treat QTL models as models of varying dimension and employ the reversible jump MCMC algorithm to explore the posterior. Although such a framework is very general and powerful (GREEN 1995), it is difficult to implement efficient search strategies. The key idea of the proposed Bayesian approach is to turn varying dimensional space of multiple-QTL models into fixed dimensional model space by using a fixed but large set of known loci, $\boldsymbol{\zeta}$, and putting a constraint on the upper bound of the number of detectable QTL. In this setting, posterior simulation then can be achieved with a relatively simple Gibbs sampler or M-H algorithm (GODSILL 2001; YI 2004). The algorithm proposed herein is easier to implement than the reversible jump method and it reduces the computational time of model search, an essential feature for the practical analysis of complex genetic architectures.

A prerequisite of the proposed method is a reasonable choice of the upper bound of the number of detectable

TABLE 1

Summary statistics for epistatic analysis: high posterior density (HPD) regions of QTL locations, posterior inclusion probabilities of main effects, Bayes factors, estimated main effects, and heritabilities explained by main effects in the HPD regions

	Chromosome							
	2	13	15	1	18	19	7	14
HPD region (cM)	[72, 85]	[20, 42]	[1, 29]	[26, 54]	[43, 71]	[15, 45]	[50, 75]	[12, 41]
Posterior probability (%)	98.3	97.2	93.5	81.9	88.4	70.6	36.7	30.1
Bayes factor	821.4	291.3	92.2	28.1	47.3	12.2	4.1	2.7
Main effect	−0.856	0.371	−0.342	−0.037	0.103	−0.167	−0.137	−0.147
Heritability	0.184	0.035	0.031	0.002	0.015	0.020	0.019	0.009

QTL. A minimal requirement is that the predetermined upper bound is greater than the true number of QTL with high probability. As an extreme case, we could take the total number of loci (H) as the upper bound. Since the number of detectable QTL is usually much less than H, such a choice is unlikely to be optimal. The suggestion made here utilizes the expected number of QTL and the prior probability distribution of the number of QTL to determine the upper bound. The expected number of QTL could be roughly estimated using standard genome scans. In practice, one could experiment with several values of the expected number of QTL and investigate their impact on the posterior inference. In high-dimensional problems, specifying the prior distributions on both the model space and parameters is perhaps the most difficult aspect of Bayesian model selection. We propose a novel method for elicitation of prior distribution on the indicator variables. Instead of directly specifying the prior inclusion probabilities w_m and w_e, the expected numbers of main-effect QTL and all QTL can first be given incorporating previous results and then are used to determine w_m and w_e. Here we have fixed w_m and w_e but we could relax this by treating w_m and w_e as unknown model parameters and assigning priors (Kohn *et al.* 2001).

A major difficulty of genome-wide epistatic analysis is created by the huge size of the model space. Strategies to reasonably reduce the model space, such as our proposed composite model space approach, can improve the performance of the MCMC algorithms and enhance our ability to detect complex epistatic QTL. We partition the entire genome into intervals by a number of points and restrict putative QTL to these fixed points, reducing loci to a discrete space. Additional speedup is achieved by computing the conditional probability of the genotypes given the marker data on this fixed (but dense) grid of possible locations before the MCMC procedure starts.

Several other strategies of reducing the model space could be incorporated into the proposed approach to improve the procedure. We could adopt a two-stage search method, first searching for main-effect QTL and second searching for epistatic effects of these and additional epistatic QTL given the already detected main-effect QTL. The positions and main effects of the QTL detected in the first stage should be updated in the second stage since inclusion of epistatic effects may yield more accurate estimation of the positions and the effects. Alternatively, we could selectively ignore some genetic effects. Even with a moderate number of detectable QTL, the epistatic models must accommodate many potential genetic effects. In a backcross population, for example, there are a total of $L(L + 1)/2$ (= 210, if $L =$ 20, say) possible effects, but many may be negligible.

TABLE 2

Summary statistics for epistatic analysis: posterior inclusion probabilities of epistatic effects, estimated epistatic effects, and heritability explained by each epistatic effect

	Posterior probability (%)	Epistatic effect	Heritability
Chr 1 [26, 54] × Chr 18 [43, 71]	81.9	0.936	0.056
Chr 2 [72, 85] × Chr 13 [20, 42]	59.5	−0.575	0.025
Chr 15 [1, 29] × Chr 19 [15, 45]	43.2	0.606	0.024
Chr 2 [72, 85] × Chr 14 [12, 41]	18.4	0.567	0.022
Chr 7 [50, 75] × Chr 19 [15, 45]	17.2	0.552	0.021
Chr 13 [20, 42] × Chr 15 [1, 29]	13.6	−0.501	0.018

Chr, chromosome.

To see this, categorize putative QTL into three types: (1) QTL with main effects (main-effect QTL), (2) QTL with weak main effects but epistatic effects with other main-effect QTL, and (3) QTL with weak main effects but epistatic effects among themselves. Letting the numbers of these three types of QTL be L_1, L_2, and L_3 ($L = L_1 + L_2 + L_3$), respectively, and ignoring the main effects of (2) and (3) QTL, the number of possible effects reduces to $L_1(L_1 + 1)/2 + L_1L_2 + L_3(L_3 - 1)/2$ ($= 115$, if $L_1 = 10$, $L_2 = 5$, and $L_3 = 5$). These three types of QTL can be detected either simultaneously or conditionally with a three-stage approach.

A number of extensions of the basic model are possible within this framework. The simplicity of the MCMC search enhances the overall flexibility of this approach and enables one to consider analysis in more complex settings. Extensions to binary or ordinal traits, inclusion of fixed- or random-effect covariates, and gene-by-environment interactions are feasible. In principle, the composite space method can be directly applied to identify higher-order interactions. However, the dramatic increase in the size of model space is likely to limit the performance of the MCMC algorithm. We regard the methods proposed here as a step toward achieving more efficient and comprehensive analysis of complex genetic architectures. There are many opportunities to extend and improve upon this general approach.

N.Y. and D.B.A. were supported by National Institutes of Health (NIH) grants NIH RO1GM069430, NIH RO1ES09912, NIH RO1 DK056366, NIH P30DK056336, and an obesity-related pilot/feasibility studies grant at the University of Alabama (Birmingham) (528176). G.A.C. was supported by NIH GM070683. B.S.Y. was supported by NIH/National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) 5803701, NIH/NIDDK 66369-01, American Diabetes Association 7-03-IG-01, and U.S. Department of Agriculture Cooperative State Research, Education and Extension Service grants to the University of Wisconsin (Madison) (C.J. and B.S.Y.). This research is a contribution of the University of Nebraska Agricultural Research Division (Lincoln, NE; journal series no. 14858) and the North Carolina Agricultural Research Service and was supported in part by funds provided through the Hatch Act.

LITERATURE CITED

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

Bogdan, M., J. K. Ghosh and R. W. Doerge, 2004 Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics **167:** 989–999.

Broman, K. W., and T. P. Speed, 2002 A model selection approach for identification of quantitative trait loci in experimental crosses. J. R. Stat. Soc. B **64:** 641–656.

Carlborg, O., and C. Haley, 2004 Epistasis: Too often neglected in complex trait studies? Nat. Rev. Genet. **5:** 618–625.

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

Carlin, B. P., and S. Chib, 1995 Bayesian model choice via Markov chain Monte Carlo. J. Am. Stat. Assoc. **88:** 881–889.

Cheverud, J. M., 2000 Detecting epistasis among quantitative trait loci, pp. 58–81 in *Epistasis and the Evolutionary Process*, edited by M. Wade, B. Brodie and J. Wolf. Oxford University Press, New York.

Chipman, H., E. I. Edwards and R. E. McCulloch, 2001 The practical implementation of Bayesian model selection, pp. 65–116 in *Model Selection*, edited by P. Lahiri. Institute of Mathematical Statistics, Beachwood, OH.

Fernandez, C., E. Ley and M. F. J. Steel, 2001 Benchmark priors for Bayesian model averaging. J. Econom. **100:** 381–427.

Gaffney, P. J., 2001 An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. Ph.D. Dissertation, Department of Statistics, University of Wisconsin, Madison, WI.

Gelman, A., J. Carlin, H. Stern and D. Rubin, 2004 *Bayesian Data Analysis*. Chapman & Hall, London.

George, E. I., 2000 The variable selection problem. J. Am. Stat. Assoc. **95:** 1304–1308.

George, E. I., and D. P. Foster, 2000 Calibration and empirical Bayes variable selection. Biometrika **87:** 731–747.

George, E. I., and R. E. McCulloch, 1997 Approaches for Bayesian variable selection. Stat. Sin. **7:** 339–373.

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

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

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Hoeschele, I., 2001 Mapping quantitative trait loci in outbred pedigrees, pp. 599–644 in *Handbook of Statistical Genetics*, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, New York.

Jansen, R. C., 2003 Studying complex biological systems using multifactorial perturbation. Nat. Rev. Genet. **4:** 145–151.

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

Kadane, J. B., and N. A. Lazar, 2004 Methods and criteria for model selection. J. Am. Stat. Assoc. **99:** 279–290.

Kao, C. H., and Z-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham's model. Genetics **160:** 1243–1261.

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

Kass, R. E., and A. E. Raftery, 1995 Bayes factors. J. Am. Stat. Assoc. **90:** 773–795.

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

Kuo, L., and B. Mallick, 1998 Variable selection for regression models. Sankhya Ser. B **60:** 65–81.

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

Leamy, L. J., D. Pomp, E. J. Eisen and J. M. Cheverud, 2002 Pleiotropy of quantitative trait loci for organ weights and limb bone lengths in mice. Physiol. Genomics **10:** 21–29.

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

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

Raftery, A. E., D. Madigan and J. A. Hoeting, 1997 Bayesian model averaging for linear regression models. J. Am. Stat. Assoc. **92:** 179–191.

Reifsnyder, P. R., G. Churchill and E. H. Leiter, 2000 Maternal environment and genotype interact to establish diabesity in mice. Genome Res. **10:** 1568–1578.

Satagopan, J. M., and B. S. Yandell, 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 Meeting, Chicago.

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

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

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

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

Stephens, D. A., and R. D. Fisch, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics **54:** 1334–1347.

Thomas, D. C., S. Richardson, J. Gauderman and J. Pitkaniemi, 1997 A Bayesian approach to multipoint mapping in nuclear families. Genet. Epidemiol. **14:** 903–908.

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

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

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

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

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

Yi, N., A. Diament, S. Chiu, J. Fisler and C. Warden, 2004a Characterization of epistasis influencing complex spontaneous obesity in the BSB model. Genetics **167:** 399–409.

Yi, N., A. Diament, S. Chiu, J. Fisler and C. Warden, 2004b Epistatic interaction between chromosomes 7 and 3 influences hepatic lipase activity in BSB mice. J. Lipid Res. **45:** 2063–2070.

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

Zeng, Z-B., C. Kao and C. J. Basten, 2000 Estimating the genetic architecture of quantitative traits. Genet. Res. **74:** 279–289.

Communicating editor: J. B. Walsh

APPENDIX A: THE MODIFIED COCKERHAM EPISTATIC MODEL FOR BACKCROSS AND INTERCROSS POPULATIONS

For a mapping population with $K + 1$ genotypes per locus, there are K marginal effect degrees of freedom (d.f.) for each locus and K^2 interaction-effect d.f. for any two loci. The design matrix $\mathbf{X}$ for model (1) has KL main-effect coefficients, x_{iqk}, and $K^2L(L-1)/2$ epistatic effect coefficients, $x_{iqq'k}$, obtained from the genotypes at the corresponding loci by using a particular epistatic model. The main and epistatic effects are denoted by β_{qk} and $\beta_{qq'k}$, respectively.

For a backcross population, there are two segregating genotypes denoted by b_qb_q, B_qb_q at locus q. For the commonly used Cockerham epistatic model (Kao and Zeng 2002), the coefficients are defined as

$$x_{iq1} = z_{iq} - 0.5 \quad \text{and} \quad x_{iqq'1} = x_{iq1}x_{iq'1},$$

where z_{iq} denotes the number of alleles B_q. For an intercross derived from two inbred lines, there are three segregating genotypes denoted by b_qb_q, B_qb_q, and B_qB_q at locus q. For the commonly used Cockerham epistatic model, the coefficients are defined as

$$x_{iq1} = z_{iq} - 1,$$

$$x_{iq2} = (1 + x_{iq1})(1 - x_{iq1}) - 0.5,$$

$$x_{iqq'k} = \begin{cases} x_{iq1}x_{iq'1}, & k = 1 \\ x_{iq1}x_{iq'2}, & k = 2 \\ x_{iq2}x_{iq'1}, & k = 3 \\ x_{iq2}x_{iq'2}, & k = 4. \end{cases}$$

For the Cockerham model, β_{q1} and β_{q2} correspond to additive and dominance effects of QTL q, respectively; and $\beta_{qq'1}$, $\beta_{qq'2}$, $\beta_{qq'3}$, and $\beta_{qq'4}$ are the epistatic effects between loci q and q', called additive-by-additive, additive-by-dominance, dominance-by-additive, and dominance-by-dominance effects, respectively. The Cockerham model keeps the same interpretation of main effects with or without epistatic effects. However, main effects should always be interpreted with caution in the presence of epistatic interactions.

APPENDIX B: THE PRIOR EXPECTED NUMBER OF QTL INCLUDED IN THE MODEL

We define ξ_q as the binary variable to indicate inclusion ($\xi_q = 1$) or exclusion ($\xi_q = 0$) of QTL q. QTL q is included into the model when and only when at least one of the genetic effects associated with QTL q is included. Therefore, we have

$$\xi_q = 1 - \prod_{k=1}^{K}(1 - \gamma_{qk})\prod_{k=1}^{K^2}\left[\prod_{q'>q}^{L}(1 - \gamma_{qq'k})\prod_{q'<q}^{L}(1 - \gamma_{q'qk})\right],$$

where K is the number of possible main effects for each locus, K^2 is the number of possible epistatic effects for any two loci, and γ_{qk} and $\gamma_{qq'k}$ are the indicators of main and epistatic effects, respectively. The actual number of QTL then equals $\Sigma_{q=1}^{L}\xi_q$. The prior expected number of all QTL is the expectation of the actual number of QTL and thus can be derived as

$$\begin{aligned} l_0 &= \sum_{q=1}^{L}\mathrm{pr}(\xi_q = 1) \\ &= L - \sum_{q=1}^{L}\left\{\prod_{k=1}^{K}\mathrm{pr}(\gamma_{qk} = 0)\prod_{k=1}^{K^2}\left[\prod_{q'>q}^{L}\mathrm{pr}(\gamma_{qq'k} = 0)\prod_{q'<q}^{L}\mathrm{pr}(\gamma_{q'qk} = 0)\right]\right\} \\ &= L[1 - (1 - w_{\mathrm{m}})^{K}(1 - w_{\mathrm{e}})^{K^2(L-1)}]. \end{aligned}$$

If we consider only main effects, then QTL q is included into the model when at least one of the main effects of QTL q is included. The binary indicator variable of QTL q then becomes $\xi_q = 1 - \Pi_{k=1}^{K}(1 - \gamma_{qk})$. Therefore, the prior expected number of main-effect QTL is

$$l_{\mathrm{m}} = L - \sum_{q=1}^{L}\left[\prod_{k=1}^{K}\mathrm{pr}(\gamma_{qk} = 0)\right] = L[1 - (1 - w_{\mathrm{m}})^{K}].$$