DOI: 10.1534/genetics.109.113688

Nonparametric Bayesian Variable Selection With Applications to Multiple Quantitative Trait Loci Mapping With Epistasis and Gene–Environment Interaction

Fei Zou,*[,1] Hanwen Huang,* Seunggeun Lee* and Ina Hoeschele†

**Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27599 and †Virginia Bioinformatics Institute and Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061*

Manuscript received December 29, 2009
Accepted for publication May 28, 2010

ABSTRACT

The joint action of multiple genes is an important source of variation for complex traits and human diseases. However, mapping genes with epistatic effects and gene–environment interactions is a difficult problem because of relatively small sample sizes and very large parameter spaces for quantitative trait locus models that include such interactions. Here we present a nonparametric Bayesian method to map multiple quantitative trait loci (QTL) by considering epistatic and gene–environment interactions. The proposed method is not restricted to pairwise interactions among genes, as is typically done in parametric QTL analysis. Rather than modeling each main and interaction term explicitly, our nonparametric Bayesian method measures the importance of each QTL, irrespective of whether it is mostly due to a main effect or due to some interaction effect(s), via an unspecified function of the genotypes at all candidate QTL. A Gaussian process prior is assigned to this unknown function. In addition to the candidate QTL, nongenetic factors and covariates, such as age, gender, and environmental conditions, can also be included in the unspecified function. The importance of each genetic factor (QTL) and each nongenetic factor/covariate included in the function is estimated by a single hyperparameter, which enters the covariance function and captures any main or interaction effect associated with a given factor/covariate. An initial evaluation of the performance of the proposed method is obtained via analysis of simulated and real data.

TRAITS showing continuous variation are called quantitative traits and are typically controlled by multiple genetic and nongenetic factors, which tend to have relatively small effects individually. Crosses between inbred lines produce suitable populations for quantitative trait locus (QTL) mapping and are available for agricultural plants and for animal (*e.g.*, mouse) models of human diseases. Such crosses are often used to detect QTL. For these inbred line crosses, uniform genetic backgrounds, controlled breeding schemes, and controlled environment ensure that there is little or no confounding of uncontrolled sources of variability with genetic effects. The potential for such confounding complicates and limits the analysis and interpretation of human data. Because of the homology between humans and rodents, rodent models can be extremely useful in advancing our understanding of certain human diseases. In the past 2 decades, various statistical approaches have been developed to identify QTL in inbred line crosses (see, for example, Doerge *et al.* 1997 for review). To perform QTL mapping (identification), a large number of candidate positions (candidate QTL) along the genome are selected. These candidate QTL may all be located at genetic markers (positions of sequence variants in the genome where the genotypes of all individuals in a mapping population can be measured) or also in between markers if the marker density is not high. QTL mapping may then be performed by considering one candidate QTL at a time or multiple candidate QTL simultaneously. For inbred line crosses with low marker density and considering a single candidate QTL at a time, the interval-mapping method was proposed by Lander and Botstein (1989). However, these authors showed that interval mapping tends to identify a "ghost" QTL located in between two actual linked QTL if two or more closely linked QTL exist. This problem can be reduced or eliminated in two ways: (1) by using composite-interval mapping (Jansen and Stam 1994; Zeng 1994) which still performs a one-dimensional QTL search but conditional on the genotypes at a pair of markers flanking the marker interval containing the current QTL, to absorb the effects of background (nontarget QTL) outside of the target interval; or (2) by performing multiple QTL mapping, where two or more QTL are mapped simultaneously. Furthermore, if several QTL affect a quantitative trait mostly through

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.109.113688/DC1.

[1] *Corresponding author:* Department of Biostatistics, University of North Carolina, 4115D McGavran–Greenberg Hall, CB 7420, Chapel Hill, NC 27599. E-mail: fzou@bios.unc.edu

Genetics **186:** 385–394 (September 2010)

their interactions (epistasis) while having nonexistent or weak main effects, then interval mapping or single-marker analysis will fail to detect such QTL. QTL interactions may not be limited to pairwise interactions. Marchini *et al.* (2005) have shown by simulation that searching for three loci jointly in the presence of a three-way interaction is more powerful than searching for a single or a pair of QTL. There are various different implementations of multiple QTL mapping. Most methods still perform only pairwise searches, with and without epistasis. The most recent methods are based on Bayesian variable selection and consider a group of candidate QTL or all candidate QTL in the genome simultaneously (*e.g.*, Yi *et al.* 2007). These methods are typically still limited to pairwise interactions among QTL and do not consider gene–environment interactions.

The identification of QTL can be viewed as a very large variable selection problem: for p candidate QTL, with p typically in the hundreds or thousands and sample size in the low hundreds, there are 2^p possible main-effect models, $2^{\binom{p}{2}}$ possible two-way interactions, and $2^{\binom{p}{k}}$ possible higher-order ($k > 2$) interactions. For inbred line crosses, where multiple-QTL mapping models can be represented as multiple linear regression models, classical variable selection methods such as forward and stepwise selection (Broman and Speed 2002) have been used in searching for main and two-way interaction effects. Bayesian analysis implemented by Markov chain Monte Carlo (MCMC) and based on the composite model space framework (Godsill 2001, 2003) has been introduced to genetic mapping (Yi 2004). Well-known Bayesian variable selection methods such as reversible jump MCMC (Green 1995) and stochastic search variable selection (SSVS) (George and McCulloch 1993) are special cases. SSVS and similar methods employ mixture priors for the regression coefficients, which specify different distributions for the coefficients under the null (effect negligible) and alternative (effect nonnegligible) hypotheses. The marginal posterior probabilities of the alternative hypotheses can be used to identify a subset of important parameters on the basis of Bayesian multiple comparison rules, including the median probability model (with a threshold of 0.5) and Bayesian false discovery rate control (*e.g.*, Müller *et al.* 2006).

An alternative to variable selection with mixture priors is classical and Bayesian shrinkage- or penalty-based inference. For the classical approach of penalized regression, while an L_2-based shrinkage method (ridge regression) cannot perform variable selection, other methods, in particular the L_1-based lasso of Tibshirani (1996) and later lasso extensions, are capable of performing variable selection by reducing the effects of unimportant variables effectively to zero. The lasso has been applied to parametric, regression-based QTL mapping (Yi and Xu 2008). The penalized regression methods can be interpreted as Bayesian regression models with particular sparsity priors imposed on the regression coefficients (Park and Casella 2008).

Regression methods are also used for association mapping in human populations. Recently, Kwee *et al.* (2008) proposed a semiparametric regression-based approach for candidate regions in human association mapping, where a quantitative trait is regressed on a nonparametric function of the tagSNP genotypes within a region. They analyzed a (small) subset of the genome and tested for the joint significance of the subset. Their method potentially can be used to model interactions among SNPs and covariates. However, Kwee *et al.* (2008) fit their model using least-squares kernel machines, a dimension-reducing technique that is identical to an analysis based on a specific linear mixed model. Model selection for different types of kernels and different sets of variables is performed using criteria such as Akaike's information criteria (Akaike 1974) and Bayesian information criteria (Schwarz 1978), which may not be appropriate or feasible in large-scale, sparse variable selection situations.

We (Huang *et al.* 2010) recently developed a Bayesian semiparametric QTL mapping method, where nongenetic covariate effects are modeled nonparametrically. This method was implemented via MCMC, and a Gaussian process prior (O'Hagan 1978; Neal 1996, 1997) was placed on the unknown covariate function. The Gaussian process is particularly well suited for curve estimation due to its flexible sample path shapes. This method allows one or more nongenetic covariates to have an arbitrary (nonlinear) relationship with the phenotype. Another strong advantage of the Gaussian process is its ability to deal with high-dimensional data compared to other nonparametric techniques such as spline regression (Wahba 1984; Heckman 1986; Chen 1988; Speckman 1988; Cuzick 1992; Hastie and Loader 1993). There has been a growing interest in using Gaussian processes as a unifying framework for studying multivariate regression (Rasmussen 1996), pattern classification (Williams and Barber 1998), and hierarchical modeling (Menzefricke 2000). In this article, we build on this work and propose a nonparametric Bayesian method for multiple QTL mapping by including not only nongenetic covariates but also all candidate QTL in the unknown function. A Gaussian process prior (GPP) is again placed on the unknown function, and a variable selection approach is implemented for the hyperparameters of the GPP (one for each QTL and nongenetic covariate). Here, we rely on mixture priors and MCMC implementation, and we focus on linkage mapping in inbred line crosses, while in ongoing and future work we are considering shrinkage priors, deterministic algorithms, and association mapping. Our application of the GPP differs from "standard" applications in that the QTL covariates included in the unknown function are discrete, not continuous, with a small number (two or three) of possible values (the genotype codes). The goal of using a GPP here is not curve or response surface modeling but rather high-dimensional variable selection (QTL and nongenetic covariates) with a method requiring only a single parameter for each variable while accounting for any multiway interactions among the candidate variables.

To improve current methods for linkage mapping in inbred line crosses and for association analysis of human populations, we need to be able to detect QTL irrespective of whether they act mostly through main effects, interactions with other QTL, or interactions with environment. Fitting a parametric model including all these potential effects for a genome-wide search would substantially increase the multiple-testing problem, in addition to being computationally extremely demanding. Here we offer an alternative. We show that our nonparametric Bayesian method can identify QTL irrespective of whether they act through main effects, through interactions with other QTL, or with environmental factors. This method cannot identify the source(s) of a QTL's importance (main or interaction effects involving this QTL). Therefore, once a small number of important QTL have been identified in a genome-wide scan, then these QTL can be further analyzed with detailed parametric models to determine the source(s) of their importance.

The remainder of the article is organized as follows. We first present the nonparametric multiple-QTL model and outline the MCMC sampler in the next section. Simulation results and the analysis of a real data set are presented in the section following that. And we end the article with a discussion and conclusions.

NONPARAMETRIC REGRESSION WITH GAUSSIAN PROCESS

Model and prior: For the *i*th individual, we observe (i) the genotype codes $\boldsymbol{x}_i = \{x_{ik}\}_{k=1}^{p}$ at p markers, where x_{ik} is the genotype code at the kth marker; (ii) q nongenetic covariates or factors $\mathbf{t}_i = \{t_{ij}\}_{j=1}^{q}$, where t_{ij} is the value of nongenetic covariate j; and (iii) the phenotype or trait value y_i. The primary goal of the analysis is to map QTL (also loosely referred to as "genes") associated with the phenotype. Assuming for simplicity of presentation that the set of candidate QTL is the set of markers, the problem reduces to identifying which markers influence the phenotype through their genotypes. We considered the following semiparametric QTL mapping model in Huang *et al.* (2010),

$$y_i = \sum_{k=1}^{p} x_{ik}\beta_k + \eta(t_{i1}, \ldots, t_{iq}) + e_i, i = 1, \ldots, n, \qquad (1)$$

where η is an unknown function of nongenetic covariates that we modeled via a Gaussian process prior; β_k is the partial regression coefficient associated with the kth marker; and e_i is a random error with distribution $N(0, \sigma_e^2)$. Model (1) considers only main QTL effects, which of course can be extended to pairwise interactions by including the terms $\sum_{j \neq k} x_{ij}x_{ik}\beta_{jk}$, into (1) and similarly to higher-order interactions. The explicit modeling of interactions among genes causes an increase in the number of parameters in (1) which is exponential in the order of the interactions considered. Consequences are computational difficulties and poor inferences due to small sample sizes. To overcome this problem, we can alternatively consider the following model:

$$y_i = \eta(x_{i1}, \ldots, x_{ip}, t_{i1}, \ldots, t_{iq}) + e_i,\ i = 1, \ldots, n. \qquad (2)$$

This model is flexible and considers all interactions among genes and gene–environment interactions nonexplicitly. For example, if the unknown function $\eta(x_{i1}, \ldots, x_{ip}, t_{i1}, \ldots, t_{iq}) = x_{i1}x_{i2} + x_{i3}t_{i1}$, then Equation 2 nonexplicitly models the two-way interaction between genes 1 and 2 and the gene–environmental interaction between gene 3 and environmental covariate 1. Let $\eta_i = \eta(x_{i1}, \ldots, x_{ip}, t_{i1}, \ldots, t_{iq})$ and define $\boldsymbol{\eta} = \{\eta_i\}_{i=1}^{n}$.

To estimate $\boldsymbol{\eta}$, we again assume that $\boldsymbol{\eta}$ has a Gaussian process prior (as in Huang *et al.* 2010) with mean 0 and with a covariance matrix $\boldsymbol{\Sigma}$ whose element ii' $(i \neq i')$ associated with individuals i and i' is

$$\Sigma_{ii'} = \mathrm{Cov}[\eta_i, \eta_{i'}] = \xi \exp\left(-\sum_{k=1}^{p} \rho_{xk}^2 (x_{ik} - x_{i'k})^2 - \sum_{j=1}^{q} \rho_{tj}^2 (t_{ij} - t_{i'j})^2\right), \qquad (3)$$

where ξ, the ρ_{xk}^2's, and the ρ_{tj}^2's are hyperparameters. This is the most commonly employed stationary covariance function for a Gaussian process (a detailed presentation of Gaussian processes with many valid covariance functions is in Abrahamsen 1997; see also MacKay 1998). Hyperparameter ξ defines the vertical scale of variations, *i.e.*, controls the magnitude of the exponential part. Hyperparameters ρ_{xk}^2 and ρ_{tj}^2 are related to length scales that characterize the distance in that particular direction over which y is expected to vary significantly. When for example, $\rho_{xk}^2 = 0$, then η is expected to be essentially a constant function of variable (gene) x_k, which is therefore deemed irrelevant (MacKay 1998). When ρ_{xk}^2 is large, then the resulting function has a short characteristic length and will vary rapidly along the corresponding axis of x_k, indicating that variable x_k is of high importance. Similarly, ρ_{tj}^2 indicates the importance of nongenetic covariate j in combination with the genetic factors and other nongenetic covariates.

The original articles on the Gaussian process (Neal 1997; MacKay 1998) did not view this method as an approach for variable selection and imposed an inverse Gamma prior on the ρ^2 parameters. Though ρ_{xk}^2 does provide information about the relevance of any QTL k with values near zero indicating an irrelevant QTL [similar to the parameters β_k in the parametric linear QTL model (1)], determining which ρ^2_{xk}'s are significantly nonzero is challenging. It is convenient to represent the hyperparameters ρ^2_{xk} in terms of their reciprocals, defined to be $\tau_{xk} = 1/\rho^2_{xk}$ and $\tau_{tj} = 1/\rho^2_{tj}$. We perform Bayesian variable selection by imposing Gamma mixture priors on the parameters τ_{xk} and τ_{tj}. We introduce the latent variables γ_{xk} ($\gamma_{xk} = 0$ or 1) and γ_{tj} ($\gamma_{tj} = 0$ or 1). Then the Gamma mixture priors for the QTL associated parameters are represented as

$$\begin{aligned} p_{xk} &\sim \mathrm{Be}(p_{xk} \mid a_{x\gamma}, b_{x\gamma}), \\ P(\gamma_{xk} = 1) &= 1 - P(\gamma_{xk} = 0) = p_{xk}, \\ \tau_{xk} &\sim (1-\gamma_{xk})\mathrm{Ga}\left(\tau_{xk}\middle|\frac{\alpha_{x0}}{2}, \frac{\alpha_{x0}}{2\mu_{x0}}\right) + \gamma_{xk}\mathrm{Ga}\left(\tau_{xk}\middle|\frac{\alpha_{x1}}{2}, \frac{\alpha_{x1}}{2\mu_{x1}}\right). \end{aligned} \tag{4}$$

Here $\mathrm{Be}(p \mid a, b)$ represents the Beta density $p^{a-1}(1-p)^{b-1}/B(a, b)$, $\mathrm{Ga}(\tau \mid a, b)$ represents the Gamma density $\tau^{a-1}\exp(-b\tau)b^a/\Gamma(a)$, with $E(\tau) = \mu$ and $\beta = \alpha/2\mu$, and (α_{x0}, μ_{x0}, α_{x1}, μ_{x1}, $a_{x\gamma}$, $\mathrm{b}_{x\gamma}$) are hyperparameters to be specified or inferred. Similarly, for the nongenetic covariate associated parameters, we assume the mixture priors

$$\begin{aligned} p_{tj} &\sim \mathrm{Be}(p_{tj} | a_{t\gamma}, b_{t\gamma}), \\ P(\gamma_{tj} = 1) &= 1 - P(\gamma_{tj} = 0) = p_{tj}, \\ \tau_{tj} &\sim (1-\gamma_{tj})\mathrm{Ga}\left(\tau_{tj}\middle|\frac{\alpha_{t0}}{2}, \frac{\alpha_{t0}}{2\mu_{t0}}\right) + \gamma_{tj}\mathrm{Ga}\left(\tau_{tj}\middle|\frac{\alpha_{t1}}{2}, \frac{\alpha_{t1}}{2\mu_{t1}}\right). \end{aligned} \tag{5}$$

Note that here μ_{x0}, μ_{x1}, μ_{t0}, and μ_{t1} are the means of the two Gamma distributions in (4) and in (5), respectively. Setting μ_{x0} (as well as μ_{t0}) to a large value ensures that if $\gamma_{xk} = 0$, then ρ_{xk} will take on very small values, and thus the corresponding variable is irrelevant. In contrast, setting μ_{x1} (as well as μ_{t1}) to a smaller value ensures that if $\gamma_{xk} = 1$, then a nonzero value of ρ_{xk} will be included in the model.

Define $\tau_\xi = 1/\xi^2$, $\tau_e = 1/\sigma^2_e$ and let the prior distributions of the two parameters be Gamma and given by

$$p(\tau_\xi) = \frac{(\alpha_\xi/2\mu_\xi)^{\alpha_\xi/2}}{\Gamma(\alpha_\xi/2)}(\tau_\xi)^{\alpha_\xi/2-1}\exp\left(\frac{-\tau_\xi\alpha_\xi}{2\mu_\xi}\right) \tag{6}$$

$$p(\tau_e) = \frac{(\alpha_e/2\mu_e)^{\alpha_e/2}}{\Gamma(\alpha_e/2)}(\tau_e)^{\alpha_e/2-1}\exp\left(\frac{-\tau_e\alpha_e}{2\mu_e}\right). \tag{7}$$

Values for the parameters (α_ξ, μ_ξ, α_e, μ_e) are chosen prior to analysis.

The Gaussian process was originally proposed for modeling curves with continuous covariates, where the smoothness assumption on Gaussian process guarantees the smoothness of the estimated curves. However, in QTL mapping and other similar genetics analysis (Kwee *et al.* 2008), the primary goal is to map genes. The violation of the continuity assumption may be highly influential on the QTL effect estimation, but since the QTL effect estimation is only a secondary task in QTL mapping, the discreteness of genetic variables is less of a concern. As one extreme example, when only one gene is included, the Gaussian process model is equivalent to the random effect model where the genetic effect is treated as random with a normal distribution.

MCMC algorithm for posterior computation: Inference is based on the joint posterior distribution of the unknown parameters ($\tau_{x1}, \ldots, \tau_{xp}, \tau_{t1}, \ldots, \tau_{tq}, \tau_\xi, \tau_e$) and the unknown function vector $\boldsymbol{\eta}$, conditional on the phenotype ($\mathbf{y}$), covariate ($\mathbf{t}$), and marker ($\mathbf{x}$) data. One potential problem in working with this joint posterior arises due to the discrete nature of the marker data: When the number of significant markers (*i.e.*, markers with distinctly nonzero ρ_{xk}) is small, then the covariance matrix of $\boldsymbol{\eta}$, $\boldsymbol{\Sigma}$, may become (nearly) singular, because multiple individuals will share the same genotype configuration at these few markers. In this case the performance of the method deteriorates, and we therefore prefer to work with the joint posterior of the unknown parameters, or the joint posterior marginalized with respect to $\boldsymbol{\eta}$. Because of the normal prior on $\boldsymbol{\eta}$, this marginalization is equivalent to substituting the likelihood function of $\mathbf{y}$ conditional on $\boldsymbol{\eta}$ by the unconditional likelihood of $\mathbf{y}$, or

$$\mathbf{y} \sim N(0, \boldsymbol{\Sigma}_y), \text{ where } \boldsymbol{\Sigma}_y = \boldsymbol{\Sigma} + \sigma^2_e\mathbf{I}, \tag{8}$$

where $\boldsymbol{\Sigma}_y$ is nonsingular even if $\boldsymbol{\Sigma}$ is singular. We compared inferences based on the joint posterior of the unknown parameters and $\boldsymbol{\eta}$ *vs.* the joint posterior of the parameters (using the same simulated data as presented below) and found the latter to provide clearly superior results. Therefore, from here on we consider only the marginalized posterior.

Most of the posterior computation is quite straightforward, and details can be found in supporting information, File S1. Below we describe an efficient sampling scheme, hybrid MCMC, that is essential for dealing with the large-scale QTL mapping data.

Let $\boldsymbol{\upsilon} = (\log(\tau_{x1}), \ldots, \log(\tau_{xp}), \log(\tau_{t1}), \ldots, \log(\tau_{tq}), \log(\tau_\xi), \log(\tau_e))$. Due to the complexity of the covariance form (3), one cannot sample from the full conditional posterior distributions of $\boldsymbol{\upsilon}$ directly. The Metropolis–Hastings algorithm could be used with some proposal distribution, but it would explore the region of high probability by an inefficient random walk. To overcome this problem, the hybrid Monte Carlo method was proposed for sampling the hyperparameters in Gaussian process regression (Neal 1993, 1996; Rasmussen 1996; Barber and Williams 1997), and we adopt this approach here. The hybrid Monte Carlo method merges the Metropolis–Hastings algorithm with sampling techniques based on dynamics simulation.

To sample the $p + q + 2$ elements of vector $\boldsymbol{\upsilon}$ from their posterior distribution $p(\boldsymbol{\upsilon} \mid \mathbf{y}, \boldsymbol{\theta}_{-\boldsymbol{\upsilon}})$, we consider a physical system including $p + q + 2$ particles with the coordinate of the ith particle being υ_i. The potential energy of this system is defined in such a way that $\varepsilon(\boldsymbol{\upsilon}) = -\log p(\boldsymbol{\upsilon} \mid \mathbf{y}, \boldsymbol{\theta}_{-\boldsymbol{\upsilon}})$. To allow the use of the dynamic method, we introduce a "momentum" variable, $\boldsymbol{\phi}$, which has $p + q + 2$ real-valued components, $\boldsymbol{\phi}_i$, in one-to-one correspondence with the components of $\boldsymbol{\upsilon}$. The kinetic energy of this system is defined as $\mathcal{K}(\boldsymbol{\phi}) = \frac{1}{2}\sum_{i=1}^{p+q+2}\phi_i^2$. Therefore, sampling $\boldsymbol{\upsilon}$ from $p(\boldsymbol{\upsilon}) = e^{-\varepsilon(\boldsymbol{\upsilon})}$ is equivalent to sampling $(\boldsymbol{\upsilon}, \boldsymbol{\phi})$ from $p(\boldsymbol{\upsilon}, \boldsymbol{\phi}) = e^{-\varepsilon(\boldsymbol{\upsilon}) - \mathcal{K}(\boldsymbol{\phi})}$ by simply ignoring the momentum $\boldsymbol{\phi}$. The canonical distribution over $(\boldsymbol{\upsilon}, \boldsymbol{\phi})$ is defined to be $p(\boldsymbol{\upsilon}, \boldsymbol{\phi}) = e^{-\mathcal{H}(\boldsymbol{\upsilon}, \boldsymbol{\phi})}$, where $\mathcal{H}(\boldsymbol{\upsilon}, \boldsymbol{\phi}) = \varepsilon(\boldsymbol{\upsilon}) + \mathcal{K}(\boldsymbol{\phi})$ is the "Hamiltonian" function, which gives the total energy of the system. It is well known in physics that the evolutions of such a canonical dynamical system through fictitious time s are governed by the following differential equations:

$$\frac{d\upsilon_i}{ds} = \frac{\partial\mathcal{H}}{\partial\phi_i} = \phi_i, \; \frac{d\phi_i}{ds} = -\frac{\partial\mathcal{H}}{\partial\upsilon_i} = -\frac{\partial\varepsilon}{\partial\upsilon_i}. \qquad (9)$$

By simulating this dynamical system, the transitions of the Markov chain in the hybrid Monte Carlo method take place as follows:

a. Starting from the current state $(\boldsymbol{\upsilon}(s), \boldsymbol{\phi}(s))$, perform L steps on the basis of the discretized Equation 9 with step size ε, resulting in the state $(\boldsymbol{\upsilon}^*, \boldsymbol{\phi}^*) = (\boldsymbol{\upsilon}(s + L\varepsilon), \boldsymbol{\phi}(s + L\varepsilon))$. A single step from s to $s + \varepsilon$ can be explicitly written as

$$\phi_i\left(s + \frac{\varepsilon}{2}\right) = \phi_i(s) - \frac{\varepsilon}{2}\frac{\partial\varepsilon}{\partial\upsilon_i}(\boldsymbol{\upsilon}(s)), \qquad (10)$$

$$\upsilon_i(s + \varepsilon) = \upsilon_i(s) + \varepsilon\phi_i\left(s + \frac{\varepsilon}{2}\right), \qquad (11)$$

$$\phi_i(s + \varepsilon) = \phi_i\left(s + \frac{\varepsilon}{2}\right) - \frac{\varepsilon}{2}\frac{\partial\varepsilon}{\partial\upsilon_i}(\boldsymbol{\upsilon}(s + \varepsilon)). \qquad (12)$$

b. With probability $\min(1, \exp[\mathcal{H}(\boldsymbol{\upsilon}, \boldsymbol{\phi}) - \mathcal{H}(\boldsymbol{\upsilon}^*, \boldsymbol{\phi}^*)])$, accept the new state $(\boldsymbol{\upsilon}, \boldsymbol{\phi}) = (\boldsymbol{\upsilon}^*, \boldsymbol{\phi}^*)$; otherwise reject the new state and retain the old state with negated momentum $(\boldsymbol{\upsilon}, \boldsymbol{\phi}) = (\boldsymbol{\upsilon}, -\boldsymbol{\phi})$.

c. Update the total energy of the system by perturbing the momenta according to $\phi_i = c\phi_i + z_i\sqrt{1 - c^2}$ for all i, where z_i is drawn randomly from the standard normal distribution. The momentum causes the particle to continue in a consistent direction until a region of high energy (low probability) is encountered. Following Rasmussen (1996), we set $\varepsilon = 0.5n^{-1/2}$ and $c = 0.95$.

SIMULATION AND REAL DATA ANALYSIS

Simulation of multiple-QTL models with or without higher-order interactions: We simulated a backcross population with 200 individuals and a single chromosome with 151 evenly spaced markers at 5-cM intervals. To investigate the ability of the nonparametric Bayesian multiple-QTL analysis based on (2) to map higher-order interacting QTL that have no main effects, we simulated four interacting QTL without main effects and without lower-order interactions. The four simulated QTL are located at markers 9, 39, 69, and 99, respectively. The simulated function $\eta = \eta(x_{i1}, \ldots, x_{i151}) = x_{i9}x_{i39}x_{i69}x_{i99}$, where the x_{ik}, $k = 9, 39, 69, 99$, are the genotype codes (1 and -1) of the four simulated QTL of individual i and $\sigma_e^2 = 1$. The total heritability of this model is 50%.

For the analysis, we set $\alpha_{x0} = \alpha_{t0} = \alpha_{x1} = \alpha_{t1} = 1$, $\alpha_\xi = \alpha_e = 0.5$, $C = 100$, and $\mu_\xi = \mu_e = 400$. We also set $a_{x\gamma} = a_{t\gamma} = 0.95$ and $b_{x\gamma} = b_{t\gamma} = 0.05$, so that the prior probabilities that each variable (QTL or nongenetic covariate) is relevant or irrelevant for the phenotype are 0.05 and 0.95, respectively. Figure 1a provides a plot of the posterior mean estimate of the hyperparameter γ_{xk} for each marker k *vs.* the marker position from the general model (2). As we hoped, the estimates of the hyperparameters associated with the true QTL markers are much larger than the estimates of the hyperparameters associated with the irrelevant markers, and all four, purely interacting QTL were identified on the basis of the marginal posterior probability of inclusion >0.5. Selecting all variables with marginal posterior probability of inclusion >0.5 produces the median probability model that is known to frequently correspond to the optimal predictive model while often differing from the highest probability model.

For comparison, we also ran R/qtlbim (www.qtlbim.org/), a popular software for Bayesian multiple-QTL mapping developed by Yandell *et al.* (2007). R/qtlbim is an extensible, interactive environment for parametric Bayesian analysis of multiple interacting QTL models for experimental crosses (limited to two-way interactions). The results are summarized in Figure 1b. In the R/qtlbim manual (Banerjee *et al.* 2008), the following criteria are suggested for judging the significance of QTL: weak support if the Bayes factor (BF) falls between 3 and 10, moderate support if the BF falls between 10 and 30, strong support if the BF > 30, and no support if BF < 3. According to these criteria, R/qtlbim fails to detect any QTL simulated.

To further test the method, we then simulated data sets containing QTL that have only main effects $(\eta(x_{i1}, \ldots, x_{i151}) = 0.25(x_{i21} + x_{i51} + x_{i81} + x_{i111}))$ or main and two-way interaction effects $(\eta(x_{i1}, \ldots, x_{i151}) = 0.25 \cdot (\mathrm{x}_{i21} + x_{i81} + x_{i21}x_{i51} + x_{i81}x_{i111}))$. These are situations that R/qtlbim was specifically designed for. All other simulation parameters remained the same as in the previous simulation. Both models have a total

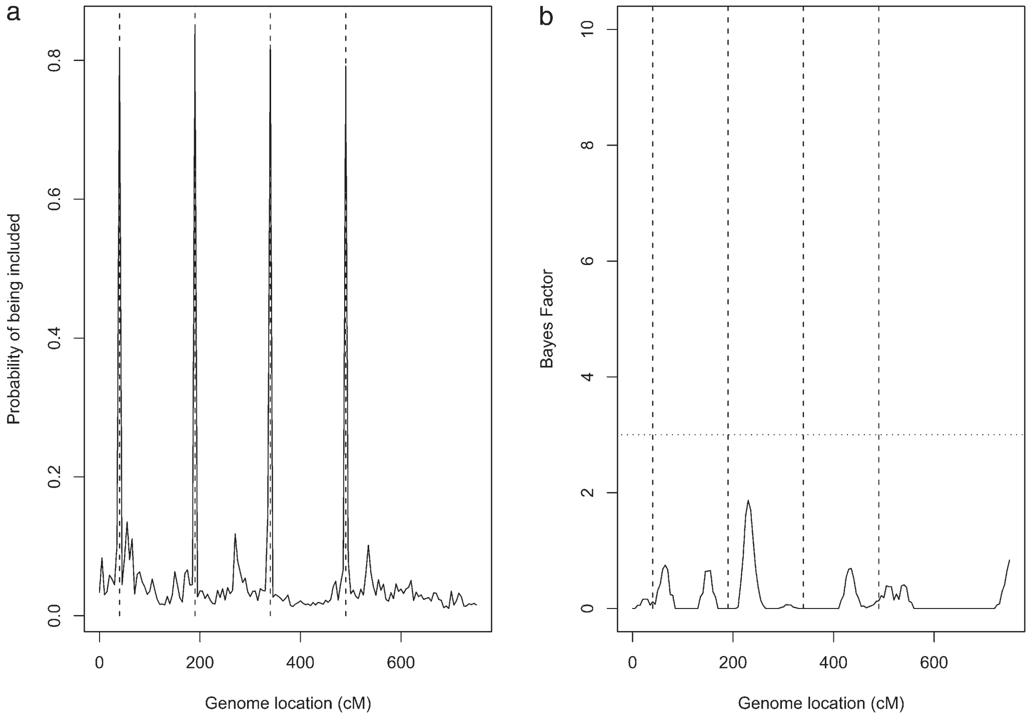

FIGURE 1.—Four-way interaction model. (a) Average marginal posterior probabilities of being included in the model for each marker. (b) Estimated marginal Bayes factor for each marker from R/qtlbim. The dashed vertical lines indicate the true QTL positions. The dotted horizontal line is the threshold suggested by R/qtlbim for weak support of significance.

heritability of 20% (∼5% heritability for each QTL). We used the same priors as in the previous simulation for the nonparametric method, and as before we used the default priors of R/qtlbim. Figure 2 summarizes the results for the additive model. Our nonparametric method detects three of the four QTL on the basis of the marginal posterior probability of inclusion (>0.5) and misses one QTL. Similarly, R/qtlbim detects the same three QTL with weak support ($3 < \mathrm{BF} < 10$). For the model with the two-way interactions, results were very similar and are therefore not shown.

Our method and R/qtlbim use different criteria (median inclusion probability *vs.* BF-based selection) for the selection of a relevant subset of QTL. This difference is confounded with the comparison between the nonparametric and the linear parametric method in terms of their ability to detect existing QTL correctly. To overcome this problem, we varied the cutoffs imposed on the inclusion probability and BF, respectively, for declaring the significance of QTL, and we generated receiver operating characteristic (ROC) curves. For each scenario (four-way interaction, additive, and additive plus two-way interaction models as above), we ran 100 simulations. Instead of fixing the positions of the four simulated QTL, we uniformly generated their positions subject to the restriction that any pair of QTL had to be at least 10 cM apart. We divided the whole genome into nonoverlapping 10-cM-wide intervals. For a given cutoff (on inclusion probability or BF), a significant interval was defined as an interval that contains at least one marker whose significance measure exceeds the cutoff. A significant interval is defined as a true positive if it includes one of the simulated QTL. Otherwise, it is called a false positive. We defined true positive rate = (no. of significant, true intervals)/(no. of significant intervals) and false positive rate = (no. of significant, false intervals)/(no. of significant intervals). The ROC curves up to a false positive rate of 0.1 are given in Figure 3 for all three models simulated. For the four-way interaction model, our nonparametric method performed much better than R/qtlbim, which essentially failed to detect any QTL. It is interesting to see that our method appears to perform essentially as well as R/qtlbim for the model with both main and two-way interactions. It is even more interesting to find that our method is superior to R/qtlbim for the main effects model. This is because we ran R/qtlbim by searching for both main effects and two-way interactions simultaneously, even when analyzing the data generated under the pure main effects (additive) model.

Real data analysis: In addition to the simulation, we tested our method on a real mouse study on obesity, a major risk factor for type II diabetes. To genetically dissect a polygenic mouse model of obesity-driven type II diabetes, REIFSNYDER *et al.* (2000) outcrossed the obese, diabetes-prone, New Zealand obese (NZO)/HlLt strain to the relatively lean nonobese nondiabetic (NON)/Lt strain and then reciprocally backcrossed obese F_1 mice to the lean NON/Lt parental strain. They measured the body weights of 187 backcross males. In addition, inguinal, gonadal, retroperitoneal, and mesenteric fat pad weights were also measured. STYLIANOU *et al.* (2006) studied the fat pad weights using F_2 progeny between the SM/J and NZB/BINJ inbred mouse strains. They identified several QTL associated with the gonadal fat pad weight after adjusting for the total lean body

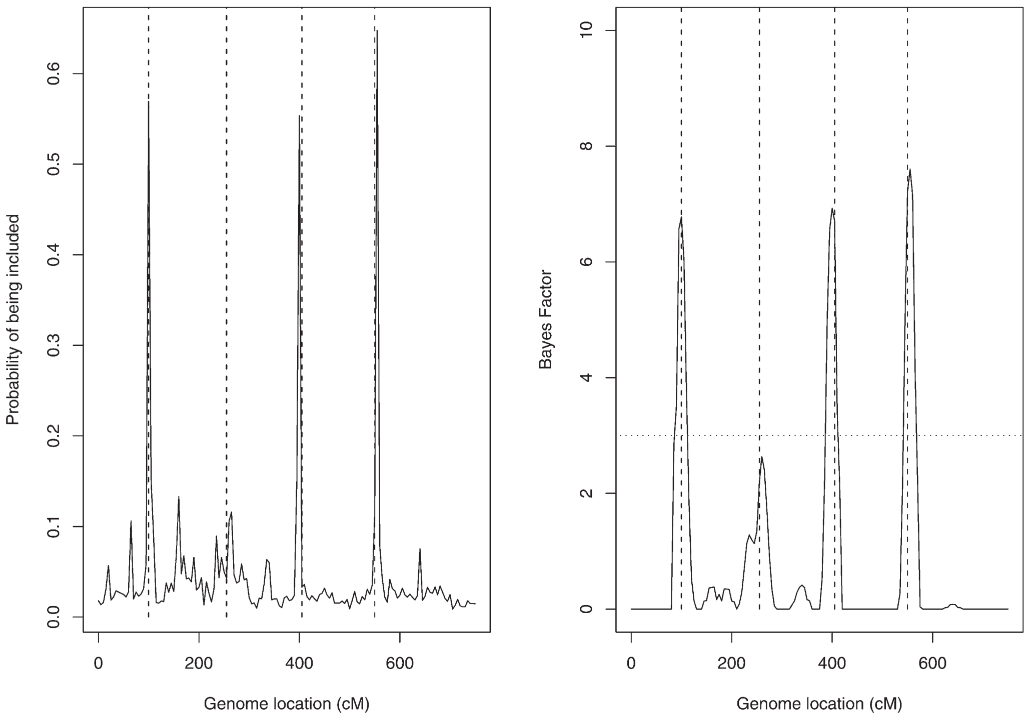

FIGURE 2.—Additive (main effects only) model. See Figure 1 for details.

weight (LBWT). Following STYLIANOU *et al.* (2006), we first calculated the total fat pad weight as the mesenteric fat pad weight plus twice the sum of the inguinal, gonadal, and retroperitoneal fat pad weights. Then the LBWT was obtained as the difference between the total body weight and the total weight of the fat pads. We applied our nonparametric Bayesian variable selection method to the REIFSNYDER *et al.* (2000) data. The results are presented in Figure 4. Clearly, 2 among the 86 predictors (85 markers plus the continuous covariate LBWT) are selected. The first ranked predictor is the covariate LBWT, and the second ranked predictor is marker D4Mit311 located on chromosome 4. Figure 4 strongly indicates a QTL on chromosome 4 while other QTL in the genome are (much) less likely. Further studies based on this observation can be done by investigating the relationship between the phenotype and these two variables in more detail. For each genotype of the QTL identified on chromosome 4, we estimated the weight curve function on LBWT, and the results are reported in Figure 5. From the two estimated curves, there is no clear evidence for an interaction between the QTL and LBWT.

DISCUSSION

In this article, we have proposed a novel nonparametric QTL mapping method where the genetic as well as nongenetic factors are modeled via a function η, whose form is unspecified. The advantage of our approach is that it models all potential genetic and nongenetic effects, including main effects and all interaction effects of any order, nonexplicitly. It determines only which of the genetic and nongenetic factors are important, on their own through main effects and/or in combination with other factors. This was achieved by combining the Gaussian process prior for the unknown function with variable selection. Although in this article we assumed that all putative QTL are located at the marker positions, it is straightforward to extend the method to consider any candidate QTL in between marker positions as in WANG *et al.* (2005) and HUANG *et al.* (2010). A similar nonparametric variable selection procedure has been proposed for computer experiments by LINKLETTER *et al.* (2006). These authors mainly focused on identifying active factors having nonlinear relationships with the response variable. However, mapping multiple interacting QTL is our main purpose, and our article appears to be the first one to propose modeling the joint action of multiple QTL with an unknown function having a Gaussian process prior, which accommodates any multiway interactions. Moreover, LINKLETTER *et al.* (2006) consider only a relatively small (<50) number of continuous covariates while in our article and in QTL linkage and association mapping in general, there are a large number of discrete marker covariates (hundreds or thousands) in addition to a small number of environmental, continuous covariates or discrete factors. Therefore, an efficient sampling scheme, such as the hybrid MCMC described in this article, is essential for dealing with these large-scale data sets.

While the linear parametric method in R/qtlbim may have little or no power to detect QTL acting through higher-order interactions, computationally it is fast, and it can handle large numbers of individuals and markers. We do not recommend replacing the linear parametric

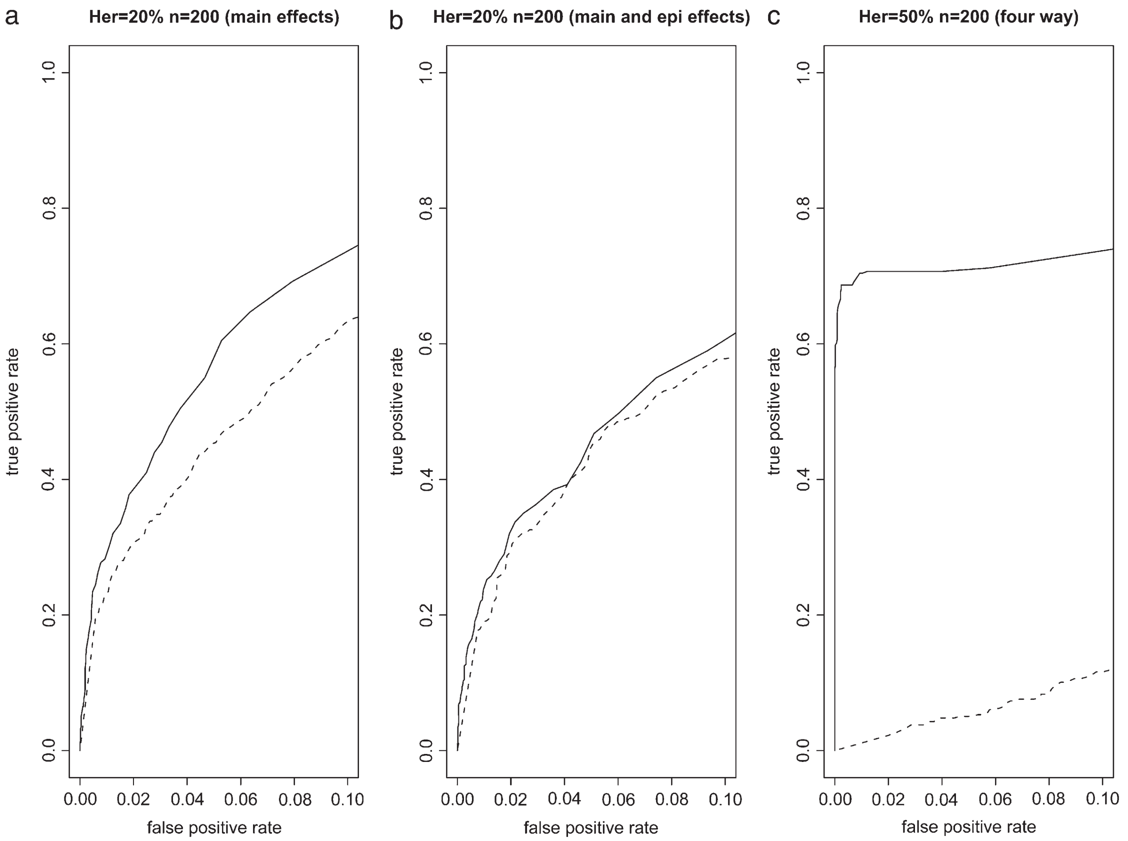

FIGURE 3.—ROC curves estimated for the three models simulated. The solid lines represent the nonparametric method and the dotted lines the linear parametric method in R/qtlbim. (a) Additive (main effects) model; (b) main plus two-way interaction effects model; (c) four-way interaction model.

analysis with the nonparametric method, but rather using it as an additional or preliminary tool to screen the genome for QTL acting through higher-order interactions, which existing QTL mapping methods fail to detect. Once important factors have been identified with the nonparametric method, they then can and should be further analyzed with a detailed parametric model to elucidate the mode of action of the identified QTL (and environmental factors). Application of a detailed parametric method on a genome-wide scale to search for all possible main and interaction effects would dramatically increase the multiple-testing problem, in addition to the computational burden, while the nonparametric method can identify all these effects with a single parameter per candidate QTL (and environmental factor).

Our current research focuses on further improving the computational feasibility of our nonparametric method. Our current implementation of the Bayesian Gaussian process prior method, with the mixture priors on the variable selection parameters (ρ's) and using the hybrid Monte Carlo method, allows us to analyze data sets with up to several hundred individuals and several hundred markers, in hours rather than in minutes as with R/qtlbim. A major reason for this increase in computing time is the need to compute the inverse of an $n \times n$ matrix in each MCMC cycle to sample $\boldsymbol{\upsilon}$. This is particularly a problem for genome-wide association studies (GWAS), for which our nonparametric method is also potentially useful. GWAS typically require a larger sample size than linkage studies (in the order of thousands or tens of thousands) and several hundred thousand markers (tag SNPs). Further, in this article, we propose a simple Gibbs sampler for the latent binary variables that code for inclusion of a marker in the covariance function. For QTL mapping with only hundreds of markers, this algorithm works well. For very large p, it is likely that the algorithm may not properly mix over the huge sample space, a legitimate concern when we apply the method directly to GWAS data where hundreds of thousands of SNPs are available. BERGER and MOLINA (2005) propose an approach to search for important models through large model space without visiting every model. Their approach provides a nice alternative, which deserves further investigation for GWAS data. Besides alternative sampling schemes we are currently investigating shrinkage priors to replace the mixture priors and increase computational efficiency, and we are exploring deterministic algorithms to replace MCMC sampling, in particular a conjugate gradient optimization technique to compute the maximum

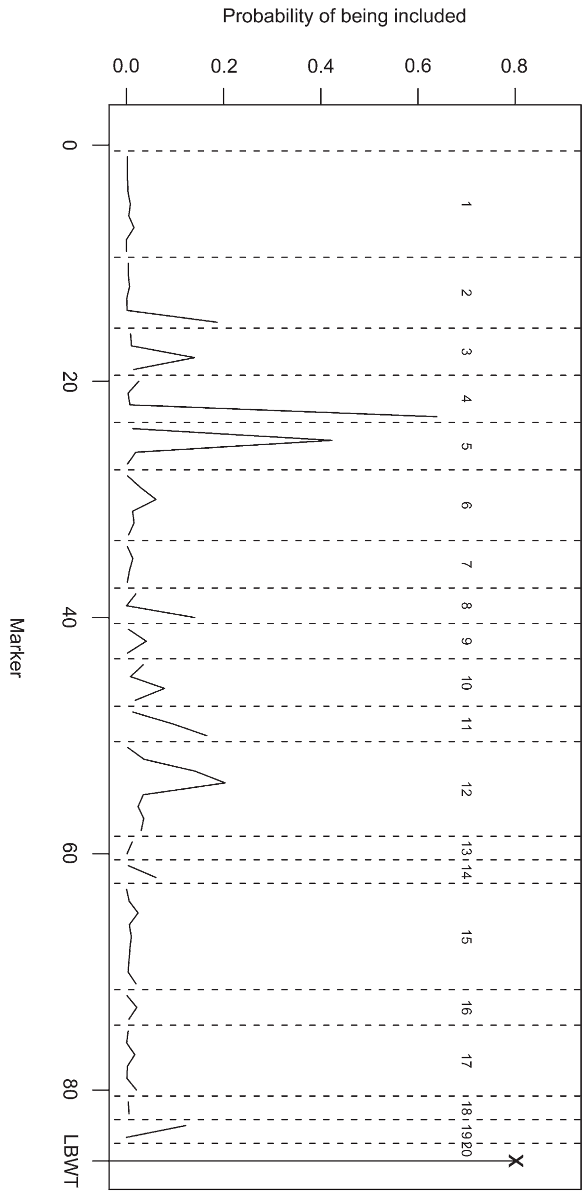

Figure 4.—Nonparametric Bayesian analysis of the backcross mesenteric fat pad data.

a posteriori estimates of the parameters (Rasmussen 1996). A genome-wide data set may first be analyzed with the deterministic implementation to screen out many variables (predictors) that are clearly irrelevant. Then the selected, promising subset of predictors (markers, genomic regions) may be reanalyzed by full MCMC, which provides much more information than a deterministic mode-finding algorithm. With an initial implementation of shrinkage priors and the conjugate gradient optimization technique we have been able to analyze a data set in a candidate gene association study with ~900 participants and 2500 tag SNPs.

Selection of a subset of QTL can be performed on the basis of the estimated marginal posterior probabilities of inclusion with cutoff determined using the median probability model or Bayesian false discovery rate. Alternatively, we may add pseudonull variable(s) into the model and use the posterior distribution of their γ's to guide the variable selection. Linkletter *et al.* (2006) suggested adding a single pseudonull variable but running the analysis many times (say 100). For computational reasons, this approach works for their smaller size problems but is computationally very demanding or infeasible in the QTL mapping context. Furthermore, adding a single pseudonull variable would not work (well) for QTL mapping because marker (null) variables are correlated due to linkage. Wu *et al.* (2007) proposed a similar idea for variable selection in linear regression models using a set of pseudonull variables. Their method requires no additional repeated analysis as in Linkletter *et al.* (2006) and can also incorporate the linkage structure of the observed markers into the generation of the pseudonull variables. We are planning to extend the method of Wu *et al.* (2007) to our Gaussian process-based QTL selection methodology.

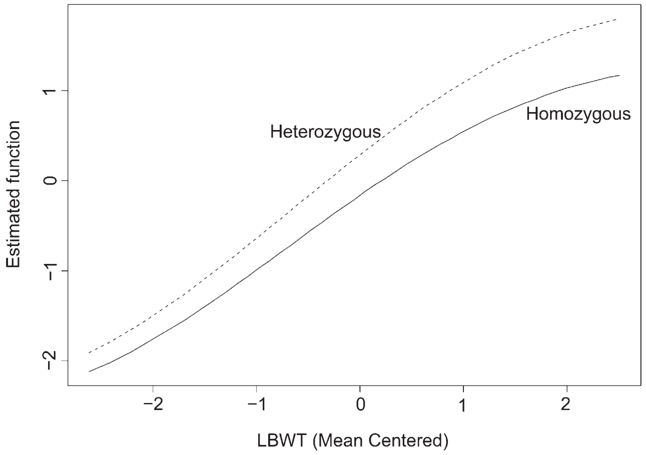

Figure 5.—Estimated η function separated for the QTL genotypes of marker D4Mit311 on chromosome 4.

Much work has been done recently on sparse signal detection in (generalized) linear regression models, where there are two groups of sparsity priors, shrinkage or one-group priors, and mixture or two (multiple) group priors. Here we have employed a mixture prior for the parameters related to variable selection. Our current and future work focuses on further studies and modifications of this mixture prior and of alternative shrinkage priors. The goal of our present article was to convincingly demonstrate that the nonparametric Bayesian analysis based on the Gaussian process prior is indeed able to detect QTL irrespectively of whether they act on the trait of interest through main effects, any order of interaction among QTL, or interactions of QTL with environmental factors.

The authors thank the editor, the associate editor, and referees for their helpful comments and suggestions, which led to a great improvement of this article. This research was partially supported by National Institutes of Health grant GM074175.

LITERATURE CITED

Abrahamsen, P., 1997 A review of Gaussian random fields and correlation functions. Technical Report 917. Norwegian Computing Center, Oslo.

Akaike, H., 1974 A new look at the statistical model identification. IEEE Trans. Automat. Contr. **19:** 716–723.

Banerjee, S., B. S. Yandell, W. W. Neely and N. Yi, 2008 *QTL Analysis Using Bayesian Interval Mapping*. University of Birmingham, Birmingham, AL. http://www.ssg.uab.edu/qtlbim/assets/docs/qtlbim.overview.pdf

Barber, D., and C. K. I. Williams, 1997 Gaussian processes for Bayesian classification via hybrid Monte Carlo, pp. 340–346 in *Advances in Neural Information Processing Systems 9*, edited by M. C. Mozer, M. I. Jordan and T. Petsche. MIT Press, Cambridge, MA.

Berger, J. O., and G. Molina, 2005 Posterior model probabilities via path-based pairwise priors. Stat. Neerl. **59:** 3–15.

Broman, K. W., and T. P. Speed, 2002 A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). J. R. Stat. Soc. Ser. B **64:** 641–656, 737–775.

Chen, H., 1988 Convergence rates for parametric components in a partly linear model. Ann. Stat. **16:** 136–146.

Cuzick, J., 1992 Semiparametric additive regression. J. R. Stat. Soc. Ser. B **54:** 831–843.

Doerge, R. W., Z.-B. Zeng and B. S. Weir, 1997 Statistical issues in the search for genes affecting quantitative traits in experimental populations. Stat. Sci. **12:** 195–219.

George, E. I., and R. E. McCulloch, 1993 Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88:** 881–889.

Godsill, S. J., 2001 On the relationship between Markov chain Monte Carlo methods for model uncertainty. J. Comput. Graph. Stat. **10:** 230–248.

Godsill, S. J., 2003 *Proposal Densities, and Product Space Methods, in Highly Structured Stochastic Systems*. Oxford University Press, London/New York/Oxford.

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

Hastie, T. J., and C. Loader, 1993 Local regression: automatic kernel carpentry (with discussion). Stat. Sci. **8:** 120–143.

Heckman, N., 1986 Spline smoothing in a partly linear model. J. R. Stat. Soc. Ser. B **48:** 244–248.

Huang, H., H. Zhou, F. Cheng, I. Hoeschele and F. Zou, 2010 Gaussian process based Bayesian semiparametric quantitative trait loci interval mapping. Biometrics **66:** 222–232.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple quantitative trait in line crosses using flanking markers. Heredity **69:** 315–324.

Kwee, L. C., D. Liu, X. Lin, D. Ghosh and M. P. Epstein, 2008 A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet. **82:** 386–397.

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

Linkletter, C., D. Bingham, N. Hengartner, D. Higdon and K. Q. Ye, 2006 Variable selection for Gaussian process models in computer experiments. Technometrics **48:** 478–490.

MacKay, D. J., 1998 Introduction to Gaussian processes, pp. 133–166 in *Neural Networks and Machine Learning* (NATO Asi Series, Vol. 168. Series F, Computer and Systems Sciences), edited by C. M. Bishop. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Marchini, J., P. Donnelly and L. R. Cardon, 2005 Genome-wide strategies for detecting multiple loci influencing complex diseases. Nat. Genet. **37:** 413–417.

Menzefricke, U., 2000 Hierarchical modeling with Gaussian processes. Commun. Stat. **29:** 1089–1108.

Müller, P., G. Parmigiani and K. Rice, 2006 *FDR and Bayesian Multiple Comparisons Rules* (Working Paper 115). Department of Biostatistics Working Papers, Johns Hopkins University, Baltimore. http://www.bepress.com/jhubiostat/paper115

Neal, R. M., 1993 Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93–1. Department of Computer Science, University of Toronto, Toronto.

Neal, R. M., 1996 *Bayesian Learning for Neural Networks*. Springer-Verlag, New York.

Neal, R. M., 1997 Monte Carlo implementation of Gaussian process models for Bayesian regression and classification. Technical Report No. 9702. Department of Statistics, University of Toronto, Toronto.

O'hagan, A. 1978 On curve fitting and optimal design for regression. J. R. Stat. Soc. B **40:** 1–42.

Park, T., and G. Casella, 2008 The Bayesian lasso. J. Am. Stat. Assoc. **103:** 681–686.

Rasmussen, C. E., 1996 Evaluation of Gaussian processes and other methods for non-linear regression. Ph.D. Thesis, University of Toronto, Toronto.

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

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

Speckman, P., 1988 Kernel smoothing in partial linear models. J. R. Stat. Soc. B **50:** 413–436.

Stylianou, I. M., R. Korstanje, R. Li, S. Sheehan, B. Paigen *et al.*, 2006 Quantitative trait locus analysis for obesity reveals multiple networks of interacting loci. Mamm. Genome **17:** 22–36.

Tibshirani, R., 1996 Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58:** 267–288.

Wahba, G., 1984 Cross validated spline methods for the estimation of multivariate functions from data on functionals, pp. 205–235 in *Statistics, an Appraisal, Proceedings of the Iowa State University Statistical Laboratory 50th Anniversary Conference*, edited by H. A. David and H. T. David. Iowa State University Press, Ames, IA.

Wang, H., Y. M. Zhang, X. Li, G. L. Masinde, S. Mohan *et al.*, 2005 Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics **170:** 465–480.

Williams, C. K. I., and D. Barber, 1998 Bayesian classification with Gaussian processes. IEEE Trans. Patt. Anal. Mach. Intell. **20:** 1342–1351.

Wu, Y., D. D. Boos and L. A. Stefanski, 2007 Controlling variable selection by the addition of pseudovariables. J. Am. Stat. Assoc. **102:** 235–243.

Yandell, B. S., T. Mehta, S. Banerjee, D. Shriner, R. Venkataraman *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., and S. Xu, 2008 Bayesian LASSO for QTL mapping. Genetics **179:** 1045–1055.

Yi, N., D. Shriner, S. Banerjee, T. Mehta, D. Pomp *et al.*, 2007 An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. Genetics **176:** 1865–1877.

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

Communicating editor: J. Wakefield

GENETICS

Supporting Information
http://www.genetics.org/cgi/content/full/genetics.109.113688/DC1

Nonparametric Bayesian Variable Selection With Applications to Multiple Quantitative Trait Loci Mapping With Epistasis and Gene–Environment Interaction

Fei Zou, Hanwen Huang, Seunggeun Lee and Ina Hoeschele

DOI: 10.1534/genetics.109.113688

File S1

MCMC algorithm for posterior computation. Let $\boldsymbol{\theta}$ be the vector of all unknown quantities in the model, including the τ_{xk}s, the τ_{tj}s, the γ_{xk}s, the γ_{tj}s, τ_ξ, τ_e and the latent variables, the η_is. Furthermore, we define $\boldsymbol{\theta}_{-z}$ as the remaining sub-vector of $\boldsymbol{\theta}$ after removing a parameter or parameter subset z from $\boldsymbol{\theta}$. Below we present the MCMC algorithm for the posterior computation.

For given values of the hyperparameters, which include $\alpha_{x0}, \alpha_{t0}, \alpha_{x1}, \alpha_{t1}, \alpha_\xi, \alpha_e$, C, μ_ξ, μ_e, $a_{x\gamma}, a_{t\gamma}$, $b_{x\gamma}$ and $b_{t\gamma}$, we first sample the τ_{xk}s, the τ_{tj}s and τ_e from their prior distributions. Then we perform the following updating steps many times:

Step 1. Sample $\boldsymbol{\eta}$ directly from its conditional distribution, which is the multivariate normal

$$\boldsymbol{\eta}|\mathbf{y}, \boldsymbol{\theta}_{-\boldsymbol{\eta}} \sim N_n(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*), \tag{1}$$

with the covariance matrix $\boldsymbol{\Sigma}^* = \left(\frac{1}{\sigma_e^2}\mathbf{I}_n + \boldsymbol{\Sigma}^{-1}\right)^{-1}$ and mean vector $\boldsymbol{\mu}^* = \frac{1}{\sigma_e^2}\boldsymbol{\Sigma}^*\mathbf{y}$. For the posterior distribution marginalized with respect to $\boldsymbol{\eta}$ we skip this step and directly go to Step 2.

Step 2: Sample the γ_{xk}s and the γ_{tj}s directly from their conditional posterior distributions. The conditional distribution of γ_{xk} does not depend on $\mathbf{y}$ but rather on ρ_{xk} and several hyper-parameters and is of the form

$$p(\gamma_{xk} = 1|\boldsymbol{\theta}_{-\gamma_{xk}}) \;=\; \frac{b_{x\gamma}\left(\frac{\alpha_{x1}}{2\mu_{x1}}\right)^{\frac{\alpha_{x1}}{2}} \rho_{xk}^{-\alpha_{x1}} e^{-\frac{\alpha_{x1}}{2\mu_{x1}\rho_{xk}^2}}}{a_{x\gamma}\left(\frac{\alpha_{x0}}{2\mu_{x0}}\right)^{\frac{\alpha_{x0}}{2}} \rho_{xk}^{-\alpha_{x0}} e^{-\frac{\alpha_{x0}}{2\mu_{x0}\rho_{xk}^2}} + b_{x\gamma}\left(\frac{\alpha_{x1}}{2\mu_{x1}}\right)^{\frac{\alpha_{x1}}{2}} \rho_{xk}^{-\alpha_{x1}} e^{-\frac{\alpha_{x1}}{2\mu_{x1}\rho_{xk}^2}}}, \tag{2}$$

Similarly, we have

$$p(\gamma_{tj} = 1|\boldsymbol{\theta}_{-\gamma_{tj}}) \;=\; \frac{b_{t\gamma}\left(\frac{\alpha_{t1}}{2\mu_{t1}}\right)^{\frac{\alpha_{t1}}{2}} \rho_{tj}^{-\alpha_{t1}} e^{-\frac{\alpha_{t1}}{2\mu_{t1}\rho_{tj}^2}}}{a_{t\gamma}\left(\frac{\alpha_{t0}}{2\mu_{t0}}\right)^{\frac{\alpha_{t0}}{2}} \rho_{tj}^{-\alpha_{t0}} e^{-\frac{\alpha_{t0}}{2\mu_{t0}\rho_{tj}^2}} + b_{t\gamma}\left(\frac{\alpha_{t1}}{2\mu_{t1}}\right)^{\frac{\alpha_{t1}}{2}} \rho_{tj}^{-\alpha_{t1}} e^{-\frac{\alpha_{t1}}{2\mu_{t1}\rho_{tj}^2}}}. \tag{3}$$

Step 3: Sample the τ_{xk}s, the τ_{tj}s, τ_ξ and τ_e from their posterior distribution, which is proportional to

$$\frac{1}{\mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\{-\frac{1}{2}\boldsymbol{\eta}'\boldsymbol{\Sigma}\boldsymbol{\eta}\} \prod_k p(\tau_{xk}|\gamma_{xk}) \prod_j p(\tau_{tj}|\gamma_{tj}) p(\tau_\xi)[\tau_e^{n/2} \exp\{-\frac{1}{2}\tau_e(\mathbf{y}-\boldsymbol{\eta})'(\mathbf{y}-\boldsymbol{\eta})\} p(\tau_e)]. \tag{4}$$

See Section 2.2 for the implementation details.

For the posterior distribution marginalized with respect to $\boldsymbol{\eta}$, Equation 4 is changed to

$$\frac{1}{\mid \boldsymbol{\Sigma}^* \mid^{1/2}} \exp\{-\frac{1}{2}\mathbf{y}'\boldsymbol{\Sigma}^{*-1}\mathbf{y}\} \prod_k p(\tau_{xk}|\gamma_{xk}) \prod_j p(\tau_{tj}|\gamma_{tj}) p(\tau_\xi) p(\tau_e). \tag{5}$$

One iteration or cycle of our MCMC sampler consists of steps 1 to 3 for the joint posterior of the parameters and $\boldsymbol{\eta}$, or steps 2 and 3 for the marginal posterior of the parameters. When the chain converges to its stationary distribution, the sampled values of all parameters are from their joint posterior distribution. Likewise, the samples of any single parameter represent the marginal posterior distribution of this parameter.