Biometrics 000, 000–000 DOI: 000
000 0000

Gaussian Process Based Bayesian Semiparametric Quantitative Trait Loci Interval Mapping

Hanwen Huang[1], Haibo Zhou[1], Fuxia Cheng[2], Ina Hoeschele[3], Fei Zou[1]

[1] *Department of Biostatistics*

University of North Carolina at Chapel Hill

Chapel Hill, North Carolina 27599, U.S.A.

[2] *Department of Mathematics, Illinois State University*

[3] *Virginia Bioinformatics Institute and Department of Statistics, Virginia Tech*

email: fzou@bios.unc.edu

Summary: In linkage analysis, it is often necessary to include covariates such as age or weight to increase power or avoid spurious false positive findings. However, if a covariate term in the model is specified incorrectly (e.g., a quadratic term misspecified as a linear term), then the inclusion of the covariate may adversely affect power and accuracy of the identification of Quantitative Trait Loci (QTL). Furthermore, some covariates may interact with each other in a complicated fashion. We implement semiparametric models for single and multiple QTL mapping. Both mapping methods include an unspecified function of any covariate found or suspected to have a more complex than linear but unknown relationship with the response variable. They also allow for interactions among different covariates. This analysis is performed in a Bayesian inference framework using Markov chain Monte Carlo. The advantages of our methods are demonstrated via extensive simulations and real data analysis.

Key words: Higher order interaction; MCMC; Multiple QTL; Non-linear; Variable selection.

1. Introduction

Unlike monogenic traits where success in associating genotype to phenotype is assured, complex traits pose significantly greater challenges. Parametric genetic mapping using experimental populations, such as backcrosses, F2 intercross or Recombinant Inbred Lines (RIL) have been well developed during the past 15 years (see Doerge et al., 1997 for an introduction to QTL mapping in inbred line crosses). Many excellent open source software packages, such as QTLCart (Basten et al., 1999), MapManager (Manly and Olson, 1999), MapMaker (Lincoln et al., 1993), and R/qtl (Broman et al., 2003) are freely available on-line. These QTL mapping packages often model the effects of non-genetic covariates linearly, for example,

$$y = \mu + \beta x + \zeta t + e, \tag{1}$$

where y is the measured quantitative trait; t is the non-genetic covariate; x is the genetic factor, and e is the random error term. However, in practice, the QTL position is unknown, resulting in missing x (missing for all individuals).

Although most available QTL mapping methods only map one or a few QTL at a time and are not efficient for complex trait mapping, recently multiple QTL have been mapped simultaneously by treating QTL mapping as a large-scale variable selection problem: for example, for a backcross population and with q potential QTL positions (selected grid of positions across the genome), where q is in the hundreds or thousands and typically (much) larger than the sample size, there are 2^q possible main effect models. Variable selection methods are needed that are capable of selecting variables that are not necessarily all individually important but rather together important. By treating multiple quantitative trait locus (QTL) mapping as a model/variable selection problem (Broman and Speed, 2002), forward and step-wise selection procedures have been proposed to search for multiple QTL. Although simple, these methods have their limitations, such as uncertainty about the number

of QTL, the sequential model building that makes it unclear how to assess the significance of the associated tests, etc. Bayesian QTL mapping methods (Satagopan, 1996; Sillanpää and Arjas, 1998; Stephens and Fisch, 1998; Yi and Xu, 2000, 2001; Hoeschele, 2007) have been developed, in particular, for the detection of multiple QTL by treating the number of QTL as a random variable and by specifically modeling it using reversible jump Markov chain Monte Carlo (MCMC) (Green, 1995). Due to the variable dimensionality of the parameter spaces associated with different models (different numbers of QTL), care must be taken in determining the acceptance probability for such changes in dimension, which in practice may not be handled correctly (Ven, 2004). To avoid this problem, another leading approach to variable selection in QTL analysis implemented by MCMC is based on the composite model space framework (Godsill, 2001, 2003) and has been introduced to genetic mapping by Yi (2004). Bayesian variable selection methods such as reversible jump MCMC (Green, 1995) and stochastic search variable selection (SSVS) (George and McCulloch, 1993) are special cases of this framework. A modification that treats (variance) hyperparameters as unknown was recently found to produce a better mixing MCMC sampler for multiple QTL mapping (Yi et al., 2007). Recently, Yi and Xu (2008) have developed a Bayesian LASSO (Tibshirani, 1996) for QTL mapping.

In some studies, however, the relationship between y and t may not be linear. In their study of the metabolic syndrome, McQueen et al (2003) have found a nonlinear effect of alcohol consumption on the quantitative traits they investigated. Incorrect modeling of the covariate effect may adversely affect power and accuracy of QTL identification. Semiparametric regression modeling, where ζt in (1) is replaced by an unspecified function $\eta(t)$, has attracted considerable attention in the statistical literature. When x is observed, model (1) reduces to the semiparametric regression model, which is well investigated in the spline literature as well as in the kernel regression literature. Examples for spline regression

include Wahba (1984), Heckman (1986), Chen (1988), Speckman (1988), Cuzick (1992), Hastie and Loader (1993) and Mays (1995) while examples for kernel regression include Härdle (1990), Wand and Jones (1995), and Fan (1992). Spline regression requires a penalty weight to balance between goodness-of-fit and complexity. To account for the non-linear effect of the alcohol consumption, McQueen et al. (2003) categorized the alcohol consumption into five non-overlapping groups in their linear regression analysis, which is essentially a special form of spline regression, so-called local polynomial regression. Kernel regression, on the other hand, needs a bandwidth to determine the degree of localness and smoothness of η. The choice of a bandwidth, and not the choice of a kernel function, is critical for the performance of the nonparametric fit (Härdle, 1990). Bayesian approaches to semiparametric regression have also been developed. Bayesian nonparametric methods achieve flexibility by putting priors on distribution spaces, corresponding to infinitely dimensional parameterizations. The leading Bayesian methods include Dirichlet process (Müller, et al. 1996), splines (Smith and Kohn, 1996; Denison et al., 1998; DiMatteo et al., 2001), wavelets (Abramovich et al., 1998) and Gaussian process (Neal, 1997, 1996). Gaussian process priors date back to at least O′Hagan (1978) and have a large support in the space of all smooth functions through an appropriate choice for the covariance kernel. Gaussian process is particularly flexible for curve estimation because of their flexible sample path shapes. Wahba (1978) has shown that for an appropriate choice of the covariance kernel of the Gaussian process, the Bayesian estimator is a smoothing spline. However, Gaussian process better suited for modeling with multiple (even many) covariates than the smoothing spline approach.

Applying spline regression and kernel regression techniques to semiparametric interval QTL mapping is challenging, especially when mapping multiple QTL, due to the missing QTL genotypes. The Bayesian approach, however, is very flexible in handling missing data. In this paper, we propose novel Bayesian methods for interval QTL mapping of a single QTL and

of multiple QTL which incorporate an unspecified function of a single covariate or multiple covariates using a Gaussian process prior. The rest of the paper is organized as follows. We first introduce the underlying semiparametric QTL model for single QTL mapping and the general development of the Markov Chain Monte Carlo (MCMC) sampler in Section 2. We then develop semiparametric multiple QTL mapping in Section 3. Simulation results are presented in Sections 4.1 amd 4.2, and the analysis of a real data set is presented in Section 4.3. We end the paper with comments and conclusions in Section 5.

2. Single QTL Mapping

Quantitative traits are usually controlled by both genetic and environmental factors, such as diet and exposure to chemical toxicities. When studying natural populations, like humans, it is difficult to separate environmental and genetic effects. With experimental organisms, uniform genetic backgrounds and controlled breeding schemes can avoid environmental variability which may obscure genetic effects. It is considerably easier to map quantitative traits with experimental populations than with natural populations. For this reason, crosses between completely inbred lines are often used for detecting QTL. QTL line-cross analysis has been widely applied in the plant sciences. It has also been used successfully in a number of animal species, such as mice and rats (Stoehr et al., 2000; Lan et al., 2001). Because of the homology between humans and rodents, animal models are extremely useful in helping us to understand human diseases.

The backcross (BC) and F2 intercross are two of the most popular mapping populations in QTL studies. Suppose two inbred parents (P1 and P2) differ in some quantitative traits. At each locus, we label the allele of parent P1 as a and the allele of P2 as A. An F1 generation is completely heterozygous with genotype Aa at all loci, receiving one allele from each parent. Thus, there is no segregation in F1 individuals. A BC population is generated when F1 is crossed back with one of its parents, for example, P2. At each locus, every BC individual

has equal probability of 1/2 to be Aa or AA, respectively. Thus there is segregation in BC since BC individuals are no longer genetically identical at each locus. Similarly, crossing F1 individuals generates an F2 population in which each individual has probability 1/4, 1/2 and 1/4 of being aa, Aa and AA, respectively. The combination of the two alleles in an individual is a genotype, and the genotypes can be determined at a number of loci, called marker loci (or genetic markers), throughout the genome. These loci are often not the functional loci (QTL) that we wish to identify and whose genotypes can generally not be measured but rather inferred (with some uncertainty) from the genotypes at nearby markers.

The QTL data available include the trait values or phenotypes (the dependent variable) y_i $(i = 1, \cdots, n)$, the discrete marker genotypes m_{ij} $(i = 1, \cdots, n,\ j = 1, \cdots, m)$, and an additional covariate t_i $(i = 1, \cdots, n)$, where n is sample size (the number of individuals) and m is the number of genetic markers. The genotypes at a putative QTL may be denoted by $\{bb, Bb, bb\}$ to distinguish the QTL genotypes from the marker genotypes $\{aa, Aa, AA\}$.

The single QTL model assumes that there is only one QTL affecting the trait according to the linear regression model in Equation (1). If the QTL genotypes are observed, QTL mapping is a simple linear regression problem. However, in practice, the QTL position is rarely known and the genotypes are typically unobserved, resulting in all missing x_is. The idea behind interval mapping (Lander and Botstein, 1989) is that at any putative QTL position located in an interval between two marker loci (typically on an evenly spaced, genome-wide grid), we can compute the probabilities of the unobserved QTL genotypes for each individual given its genotypes at the pair of closest flanking marker loci (see chapter 15 of Lynch and Walsh, 1998). The distribution of the quantitative trait given the marker genotypes thus follows a finite mixture model. Under this model, the null (alternative) hypothesis for a putative QTL position is that its genotypes do not (do) affect the phenotype of interest. The null hypothesis is evaluated via the likelihood ratio, and a plot of likelihood ratios versus

putative QTL positions is called a log-likelihood (ratio) profile. In any region where the profile exceeds a (genome-wide) significance threshold, a QTL is declared at the position with the highest log-likelihood ratio.

When a parametric model is specified incorrectly, estimation bias can result, and the incorrect or inefficient modeling of the covariate effect may adversely affect the power and precision of the QTL identification. To alleviate this problem, we propose a semiparametric model of the form

$$y_i = \beta x_i + \eta(t_i) + e_i,\ i = 1, \cdots, n, \tag{2}$$

where x_i is the indicator of the QTL genotype (e.g., with values -1 and 1 depending on whether the QTL genotype is Aa or AA in a backcross population); β is the effect of the QTL; $\eta(t)$ is an arbitrary function with no particular parametric form specified, and e_i is the error term with independent distribution $N(0, \sigma_e^2)$. Note that here t_i can be a scalar corresponding to a single covariate or a vector representing multiple covariates.

As all our inference is performed conditional on the marker genotypes, i.e. on $\boldsymbol{M} = \{m_{ij}\}$, we suppress this conditioning notation for the remainder of the paper. In Bayesian analysis, a prior distribution of the unobserved variables is combined with the likelihood of the observed data to obtain a posterior distribution of the unknown variables. Since the QTL position, λ, and therefore, the QTL genotypes are unknown, the unobserved variables of model (2) are η, λ, $\boldsymbol{x} = \{x_i\}_{i=1}^n$, β and σ_e^2. Below we describe in detail some of the prior and conditional posterior distributions.

2.1 Prior Specifications.

2.1.1 Gaussian process prior for covariate function. A prior probability density $p(\eta|t)$ is induced by using a Gaussian process. A Gaussian process is a stochastic process such that each finite dimensional distribution is multivariate normal. Thus any Gaussian

process is specified by its mean function and covariance kernel. A wide array of functions can be derived as sample paths of a Gaussian process.

Let $t_1, t_2, \cdots, t_k$ be the set of distinct covariate values and d_j the number of occurrences of $t_j, j = 1, \cdots, k$. A Gaussian process on domain $\mathcal{T}$ is a random, real valued function $\eta(t)$ such that all possible finite dimensional distributions $(\eta(t_1), \cdots, \eta(t_k))^T$ are multivariate normal, with $E(\eta(t_j)) = \mu(t_j)$ and $\text{cov}(\eta(t_{j1}), \eta(t_{j2})) = \sigma(t_{j1}, t_{j2})$ for $j1, j2 = 1, \cdots, k$. The fixed real valued function $\mu(t)$ is known as the mean function, and the function $\sigma(t, t')$ is known as the covariance kernel, which must satisfy the condition that the resulting $k \times k$ matrix $\mathbf{\Sigma}$ with (i, j)th element equal to $\sigma(t_i, t_j)$ is positive definite. Smoothness of the covariance kernel essentially controls the smoothness of the sample paths of η. For an appropriate choice of the covariate kernel, a Gaussian process has a large support in the space of all smooth functions (Abrahamsen, 1997; Mackay, 1998).

To specify a Gaussian process prior with $\boldsymbol{\mu} = \{\mu(t_j)\}_{j=1}^k$ and $\mathbf{\Sigma}$, one may consider some parametric forms for the functional parameters while putting priors on the hyper-parameters. To build the prior around a parametric family, we consider

$$\mu(t; \boldsymbol{\alpha}) = \alpha_1 f_1(t) + \cdots + \alpha_l f_l(t), \tag{3}$$

where l is a fixed integer, and $\{f_1, \cdots, f_l\}$ are known (for example polynomial) functions in t with the scaled hyper-parameters $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^l$ unknown. Such a class of parametric families covers a wide variety of functions with appropriate choice of $f_j(t)$. For the covariance kernel, consider the simplest parametric form $\sigma(t, t') = \sigma_0(t, t')/\tau$ for unknown hyper-parameters $\tau > 0$, and known kernel σ_0, such as, $\sigma_0(t, t') = \exp\{-(t - t')^2\}$. There are many other types of kernel functions that can be applied (see Mackay, 1998 for details). Note that the posterior mean of η almost interpolates the data as $\tau \to 0$, while the posterior distribution is concentrated near the prior mean function as $\tau \to \infty$. Clearly τ controls the smoothness of η.

In practice, reasonable choices of conjugate prior distributions for τ and $\boldsymbol{\alpha}$ are an inverse Gamma distribution on τ and an independent l-variate normal distribution on $\boldsymbol{\alpha}$. Specifically, we consider the following hierarchical model

$$\begin{aligned} \tau &\sim \textit{inverse Gamma}(a, b), \\ \boldsymbol{\alpha} &\sim N(\boldsymbol{\alpha}_0, \Gamma), \\ \eta|\boldsymbol{\alpha}, \tau &\sim \textit{Gaussian process}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_0/\tau) \end{aligned} \tag{4}$$

where $\boldsymbol{\Sigma}_0$ is the $k \times k$ matrix with element $(j1, j2)$ equal to $\sigma_0(t_{j1}; t_{j2})$.

2.1.2 Priors for the remaining parameters. The prior on the QTL position, λ, is non-informative with a uniform distribution over the entire genome. Given λ, for each individual i, the probability of its QTL genotype is a function of the genotypes of the two markers flanking λ and of the locations of the two flanking markers, and it is denoted by $p(x_i|m_{iL}(\lambda), m_{iR}(\lambda), d_L(\lambda), d_R(\lambda), \lambda)$, where $m_{iL}(\lambda)$ and $m_{iR}(\lambda)$ are the genotypes of the markers to the left and right of locus λ, respectively; $d_L(\lambda)$ and $d_R(\lambda)$ are the locations of the flanking markers (see chapter 15 of Lynch and Walsh, 1998 for detailed calculations). QTL effect and error variance are assumed to be independent a priori with noninformative priors $p(\beta) \propto 1$ and $p(\sigma_e^2) \propto 1/\sigma_e^2$. The choice of these priors is due to their computational simplicity and our lack of knowledge on these parameters. If prior data is available, more informative priors can be employed.

2.2 MCMC algorithm for posterior computation. Note that given the nonparametric component η, (2) becomes a linear model with y_i replaced by $y_i - \eta(t_i)$. Given the parameters of the genetic component (β, $\boldsymbol{x}$), model (2) reduces to a traditional nonparametric model if y_i is replaced by $y_i - \beta x_i$. Below we present the MCMC algorithm for the posterior computation, which is largely based on the Gibbs sampling approach.

First we initialize parameters and hyperparameters σ_e^2, τ, β, λ, a and b. Then we perform the following updating steps many times:

Step 1. Sample η: Conditional on η, the y_i's are independent normal random variables with variance σ_e^2 and mean $\beta x_i + \eta(t_i)$. Let U_j be the average of all $y_i - \beta x_i$ for those individuals whose corresponding covariate value is $t_j, j = 1 \cdots, k$ and $\boldsymbol{U} = \{U_j\}_{j=1}^k$. Further, let $\mathbf{D}$ be a diagonal matrix with diagonal element i equal to d_i/σ_e^2. Then the conditional distribution of η is k-variate normal $\eta|\mathbf{y}, \boldsymbol{\alpha}, \beta, \tau \sim N_k(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$, where $\boldsymbol{\Sigma}^\star = (\mathbf{D} + \boldsymbol{\Sigma}^{-1})^{-1}$ and mean vector $\boldsymbol{\mu}^\star = \boldsymbol{\Sigma}^\star \mathbf{D}(\boldsymbol{U} - \boldsymbol{\mu}) + \boldsymbol{\mu}$.

Step 2. Sample $\boldsymbol{\alpha}$ and τ: Let $\boldsymbol{F}$ be the $k \times l$ matrix with the (j, r)th element equal to $f_r(t_j)$. Since $\mathbf{y}$ affects the distributions of $\boldsymbol{\alpha}$ and τ only through η, we have the following conditional posterior distributions:

$$\begin{aligned} \boldsymbol{\alpha}|\tau, \eta, \mathbf{y} &= \boldsymbol{\alpha}|\tau, \eta \sim N(\boldsymbol{\alpha}_0^\star, \Gamma^\star), \\ \tau|\boldsymbol{\alpha}, \eta, \mathbf{y} &= \tau|\boldsymbol{\alpha}, \eta \sim \textit{inverse Gamma}(a^*, b^*), \end{aligned}$$

where $\Gamma^* = (\tau \boldsymbol{F}^T \boldsymbol{\Sigma}_0^{-1} \boldsymbol{F} + \Gamma^{-1})^{-1}$, $\boldsymbol{\alpha}_0^* = \tau \Gamma^* \boldsymbol{F}^T \boldsymbol{\Sigma}_0^{-1}(\eta - \boldsymbol{F}\boldsymbol{\alpha}_0) + \boldsymbol{\alpha}_0, a^* = a + k/2$ and $b^* = b + (\eta - \boldsymbol{F}\boldsymbol{\alpha})^T \boldsymbol{\Sigma}_0^{-1}(\eta - \boldsymbol{F}\boldsymbol{\alpha})/2$.

Step 3. Sample β from its conditional posterior distribution, which is normal $N\left(\frac{\sum_{i=1}^n x_i\{y_i - \eta(t_i;\boldsymbol{\alpha})\}}{\sum_{i=1}^n x_i^2}, \frac{\sigma_e^2}{\sum_{i=1}^n x_i^2}\right)$.

Step 4. Sample σ_e^2 from the inverse Gamma distribution with parameters $n/2$ and $\sum_{i=1}^n \{y_i - \beta x_i - \eta(t_i; \boldsymbol{\alpha})\}^2/2$.

Step 5. Sample QTL position λ and QTL genotypes $\boldsymbol{x}$ jointly in a Metropolis-Hastings step detailed below. A new QTL position λ^* and new QTL genotypes $x_i^*, i = 1, \ldots, n$ are proposed jointly by first sampling λ^* from a uniform proposal distribution centered on the current λ, $U[\lambda - \delta, \lambda + \delta)$, where δ is prespecified to yield a desirable acceptance rate (say 20-40%), and $q(\lambda^*|\lambda)$ is the density of this distribution (it needs a slight modification when λ is located near the end of a chromosome). Given the new position λ^*, the QTL genotypes

x_i^* are sampled directly from their fully conditional posterior distributions

$$q(x_i^* \mid \lambda^*) \stackrel{def}{=} \frac{p(y_i|x_i^*,\beta,\eta(t_i)) \cdot p(x_i^*|m_{iL}(\lambda^*),m_{iR}(\lambda^*),d_L(\lambda^*),d_R(\lambda^*),\lambda^*)}{\sum_{h\in\{-1,+1\}} p(y_i|h,\beta,\eta(t_i)) \cdot p(h|m_{iL}(\lambda^*),m_{iR}(\lambda^*),d_L(\lambda^*),d_R(\lambda^*),\lambda^*)},$$

where $p(y_i|x_i,\beta,\eta(t_i))$ follows from equation (2), i.e., $p(y_i|x_i,\beta,\eta(t_i)) \propto \exp\{-(y_i-\beta x_i-\eta(t_i))^2/(2\sigma_e^2)\}$. The sampled new position and QTL genotypes are accepted jointly with a probability equal to min(1,γ), where

$$\gamma = \frac{p(\lambda^*,\boldsymbol{x}^*|\mathbf{y},\beta,\eta)}{p(\lambda,\boldsymbol{x}|\mathbf{y},\beta,\eta)} \frac{q(\lambda)\prod_{i=1}^n q(x_i \mid \lambda)}{q(\lambda^*)\prod_{i=1}^n q(x_i^* \mid \lambda^*)}$$

The ratio of the joint posteriors of QTL position and genotypes evaluated at the proposed and current values is

$$\frac{p(\lambda^*,\boldsymbol{x}^*|\mathbf{y},\beta,\eta)}{p(\lambda,\boldsymbol{x}|y,\beta,\eta)} = \prod_{i=1}^n \frac{p(y_i|x_i^*,\beta,\eta(t_i)) \cdot p(x_i^*|m_{iL}(\lambda^*),m_{iR}(\lambda^*),d_L(\lambda^*),d_R(\lambda^*),\lambda^*)}{p(y_i|x_i,\beta,\eta(t_i)) \cdot p(x_i|m_{iL}(\lambda),m_{iR}(\lambda),d_L(\lambda),d_R(\lambda),\lambda)}.$$

Consequently, γ is simplified as

$$\gamma = \prod_{i=1}^n \frac{\sum_{h\in\{-1,+1\}} p(y_i|h,\beta,\eta(t_i)) \cdot p(h|m_{iL}(\lambda^*),m_{iR}(\lambda^*),d_L(\lambda^*),d_R(\lambda^*),\lambda^*)}{\sum_{h\in\{-1,+1\}} p(y_i|h,\beta,\eta(t_i)) \cdot p(h|m_{iL}(\lambda),m_{iR}(\lambda),d_L(\lambda),d_R(\lambda),\lambda)} \frac{q(\lambda)}{q(\lambda^*)}.$$

One iteration or cycle of our MCMC sampler consists of steps 1 to 5. 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.

3. Multiple QTL Mapping

We now extend the single QTL model to a multiple QTL model, or

$$y_i = \sum_{j=1}^q \beta_j x_{ij} + \eta(t_i) + e_i,\ i = 1,\cdots,n, \tag{5}$$

where x_{ij} is the indicator of the genotype of the ith individual at the jth putative QTL; q is the total number of QTL; β_j is the effect of the jth QTL; $\eta(t)$ and e_i are defined as for model (2) in the previous section.

For multiple Bayesian QTL interval mapping, we follow Wang et al. (2005) and assume (at most) one QTL within each marker interval. That is, in this paper, we assume q in model

(5) equals the number of marker intervals. When marker intervals are short, it is reasonable to make this assumption. For intervals that are not short enough to meet this assumption, we can further divide them into sub-intervals with pseudo-markers. Similar to the single QTL mapping, here the QTL positions, $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^q$, and therefore, the QTL genotypes $\boldsymbol{X} = \{\boldsymbol{x}_j\}_{j=1}^q$ (with $\boldsymbol{x}_j = \{x_{ij}\}_{i=1}^n$) are unknown. Therefore, the unobserved variables of model (5) are function η, error variance σ_e^2, QTL positions $\boldsymbol{\lambda}$, QTL genotypes $\boldsymbol{X}$, and QTL effects, $\boldsymbol{\beta} = \{\beta_j\}_{j=1}^q$.

3.1 *Prior Specifications*

The priors of η and σ_e^2 are unchanged from the single QTL mapping.

The prior specifications of $\boldsymbol{\lambda}$ and $\boldsymbol{X}$ are: For each $j \in \{1, \cdots, q\}$, jth QTL position λ_j has a uniform prior distribution over the entire jth interval, and the λ_js are independent of each other. Conditional on λ_j, QTL genotype x_{ij} of the ith individual at the jth QTL has conditional probability $p(x_{ij}|m_{iL}(\lambda_j), m_{iR}(\lambda_j), d_L(\lambda_j), d_R(\lambda_j), \lambda_j)$.

For QTL selection, we employ the SSVS method (George and McCulloch 1993). The SSVS approach imposes a normal mixture prior on the regression parameters $\boldsymbol{\beta}$ and uses latent variables to identify subset choices. Specifically, the latent variables γ_j ($\gamma_j = 0$ or 1), are introduced and the normal mixture prior are represented by

$$\beta_j|\gamma_j \sim (1-\gamma_j)N(0, \sigma^2) + \gamma_j N(0, c_j^2\sigma^2)$$

with $P(\gamma_j = 1) = 1 - P(\gamma_j = 0) = p_j$. The promising subsets of predictors can be identified by applying Bayesian multiple comparison rules that use the posterior probabilities $p(\gamma_j = 1|\mathbf{y})$ (e.g., Müller et al., 1996). Setting $\sigma^2 (> 0)$ small ensures that if $\gamma_j = 0$, β_j could be "safely" claimed to be 0. While setting c_j large ($c_j > 1$ always) makes sure that if $\gamma_j = 1$, a non-zero estimate of β_j will be included in the model. Such mixture model results in a normal posterior distribution of β_j, allowing the use of the Gibbs sampler.

3.2 *Posterior computation.*

The MCMC steps for sampling η, $\boldsymbol{\alpha}$, τ and σ_e^2 are identical to those in Section 3 if we replace $y_i - \beta x_i$ $(i = 1, \cdots, n)$ by $y_i - \sum_j \beta_j x_{ij}$ in all corresponding posterior updates. Therefore, below we concentrate on the updates for the remaining parameters. Let θ be the vector containing all the unknown parameters and let θ_{-z} be vector of θ after removing parameter z.

Sample β_j and $\gamma_j (j = 1, \cdots, q)$: Sample β_j from its conditional posterior distribution, which is normal

$$\beta_j | \mathbf{y}, \theta_{-\beta_j} \sim N\left(\hat{\beta}_j, \frac{\sigma_e^2}{\boldsymbol{x}_j^T \boldsymbol{x}_j + \sigma_e^2/\sigma_j^2}\right), j = 1, \cdots, q,$$

where $\hat{\beta}_j = \boldsymbol{x}_j^T \boldsymbol{w} / (\boldsymbol{x}_j^T \boldsymbol{x}_j + \sigma_e^2/\sigma_j^2)$, $\boldsymbol{w} = \{w_i\}_{i=1}^n$, with $w_i = y_i - \eta(t_i) - \Sigma_{j' \neq j} \beta_{j'} x_{ij'}$, and $\sigma_j = \sigma$ or $c_j \sigma$ depending on whether $\gamma_j = 0$ or $\gamma_j = 1$. The conditional distribution of γ_j does not depend on $\mathbf{y}$ and is of the form

$$p(\gamma_j = 1 | \theta_{-\gamma_j}) = \frac{c_j}{c_j + \frac{p_j}{1-p_j} \exp\left\{\frac{\beta_j^2}{2\sigma^2}\left(1 - \frac{1}{c_j^2}\right)\right\}}.$$

Sample QTL positions: λ_j is sampled via Metropolis-Hastings approach since there is no closed form for the conditional posterior probability density of a QTL position. We first sample a new position $\lambda_j^\star$ uniformly from the neighborhood of λ_j, $[max\{\lambda_j - \delta, d_L(\lambda_j)\}, min\{\lambda_j + \delta, d_R(\lambda_j)\}]$ with δ being a tuning parameter (set to 2 cM in subsequent simulations). Then, λ_j^* will be accepted or rejected according to the probability $min(1, \alpha)$, where

$$\alpha = \prod_{i=1}^{n} \frac{p(x_{ij} | m_{iL}(\lambda_j)^*, m_{iR}(\lambda_j^*), d_L(\lambda_j^*), d_R(\lambda_j^*), \lambda_j^*)}{p(x_{ij} | m_{iL}(\lambda_j), m_{iR}(\lambda_j), d_L(\lambda_j), d_R(\lambda_j), \lambda_j)} \frac{q(\lambda_j)}{q(\lambda_j^*)}.$$

If neither λ_j nor $\lambda_j^\star$ is within δ distance away from the ends of the interval, $q(\lambda_j^\star) = q(\lambda_j) = 1/(2\delta)$. However, if λ_j or/and $\lambda_j^\star$ are within δ distance from one end of the interval , then

$$q(\lambda_j^\star) = 1/[min\{\lambda_j^* + \delta, d_R(\lambda_j^*)\} - max\{\lambda_j^* - \delta, d_L(\lambda_j^*)\}], \text{ or/and}$$

$$q(\lambda_j) = 1/[min\{\lambda_j + \delta, d_R(\lambda_j)\} - max\{\lambda_j - \delta, d_L(\lambda_j)\}].$$

Sample QTL genotypes: The QTL genotype $x_{ij} (i = 1, 2, \cdots, n; j = 1, 2, \cdots, q)$ is updated

one individual and one locus at a time, based on flanking marker information. Specifically, x_{ij} is sampled from the conditional probability distribution $p_{ij}(x)$, for $x = \pm 1$, which equals

$$\frac{p(y_i|x, \beta, \boldsymbol{x}_{i(-j)}, \eta(t_i)) \cdot p(x|m_{iL}(\lambda_j), m_{iR}(\lambda_j), d_L(\lambda_j), d_R(\lambda_j), \lambda_j)}{\sum_{h\in\{-1,+1\}} p(y_i|h, \beta, \boldsymbol{x}_{i(-j)}, \eta(t_i)) \cdot p(x|m_{iL}(\lambda_j), m_{iR}(\lambda_j), d_L(\lambda_j), d_R(\lambda_j), \lambda_j)},$$

where

$$\boldsymbol{x}_{i(-j)} = (x_{i1}, \cdots, x_{i(j-1)}, x_{i(j+1)}, \cdots, x_{iq})$$

and

$$p(y_i|x, \beta, \boldsymbol{x}_{i(-j)}, \eta(t_i)) \propto \exp\left[-\frac{1}{2\sigma_e^2}\{y_i - \eta(t_i) - \beta_j x - \sum_{j'\neq j}^{q} \beta_{j'} x_{ij'}\}^2\right].$$

One iteration or cycle of our MCMC sampler consists of steps 1 to 7 and produces a sample from the joint posterior after completing the burn-in period.

4. Simulations and Real Data Analysis

4.1 *Simulation I*

We perform simulations to evaluate the small sample performance of the proposed semiparametric Bayesian interval QTL mapping method. Backcross populations of size 100, 300, 500 and 1000 individuals were simulated, with a single large chromosome of genetic length 10 Morgan. This genome was covered by 101 evenly spaced markers (100 marker intervals, each 10 centi-Morgan (cM) long). A single QTL was located at position $\lambda = 155$ cM with an effect of $\beta = 1$. The residual variance was $\sigma_e^2 = 1$. The covariate values were sampled from the uniform distribution $[0, 10)$, and the unknown function was $\eta(t) = sin(t)$, $\eta(t) = 3sin(t/3)$, or $\eta(t) = 5sin(t/5)$.

For the analysis, we chose the following mean and covariance function for the Gaussian process prior of η:

$$\mu(t; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 t, \text{ and } \sigma_0(t, t') = \exp\{-(t - t')^2\}.$$

with the priors on α_is being improper uniform. That is, $p(\alpha_i) \propto 1$. For initializations, we set

$\sigma_e^{(0)}$ to some value equal to or less than the overall variance of the trait (response variable), we set $\tau^{(0)} = 1$, $\beta^{(0)} = 0$, and $\lambda^{(0)}$ was chosen by randomly selecting a position in the genome. Conditional on $\lambda^{(0)}$, the initial genotype of each individual was generated conditional on the genotypes at the two markers flanking the QTL position. Lastly, we set $a = 0.1$ and $b = 0.1$ to specify a proper but vague prior for τ. We investigated and ensured that our posterior distribution is proper given the improper priors employed here, in particular for $\boldsymbol{\alpha}$ and σ_e^2.

For the analysis of the simulated and the real (see below) data, the MCMC sampler was run for a total of 25,000 cycles. The first 5000 cycles were discarded as burn-in, and the remainder of the chain was thinned by keeping one out of every ten samples, resulting in a total of 2000 samples for post-MCMC analysis. Several convergence checks were performed, including the Convergence Diagnosis and Output Analysis (CODA) by Best, Cowles and Vines (1995) and two convergence diagnostic tests by Geweke (1992) and Gelman and Rubin (1992). All analyses indicated convergence of our MCMC samplers.

The posterior mean estimates of QTL position and effect are summarized in Table 1 for the backcross of size 500 and $\eta(t) = sin(t)$, along with the true parameter values, showing that estimates and true values were in very good agreement. Figure 1 compares the true and the estimated unknown functions of the covariate, where the estimate is evaluated at grid points equally spaced between [0,10] with the increment of 0.05, and the posterior mean is given at each point, for sample sizes n of 100, 300, 500 and 1000 and for $\eta(t) = sin(t)$. The figure supports the result in Neal (1996) about the consistency of the semiparametric estimator in the context of neural networks. As we increased the sample size, the estimated function became closer to the true function and is expected to converge to the true function as the sample size approaches infinity.

[Table 1 about here.]

[Figure 1 about here.]

Table 2 provides comparisons between the linear regression model and our semiparametric model for the three different η functions and for sample size n=300. The estimates from the semiparametric analysis agree well with the true values for all three functions, while the results from the linear parametric model are quite different. The estimated residual variances from linear model are much larger than the true values, while the QTL effect is underestimated and its estimated position is rather inaccurate. In contrast, our semiparametric model provides the flexibility to fit the data well.

[Table 2 about here.]

For further comparison, we simulated two additional cow weight datasets based on a simple linear growth model and a well-known generalized logistic function for modeling the growth curve (Richards, 1959). The linear growth model is $\eta(t) = 187.459 + 2.682t$, while the generalized logistic function is

$$\eta(t) = -243 + \frac{968}{\{1 + 0.15e^{-0.01955(t-20.1)}\}^{1/0.15}}.$$

We set QTL effect $\beta = 3$ and $\sigma_e^2 = 9$ so that the phenotypic variations due to the QTL, the covariate and the random error are roughly balanced. All other simulation parameters (including the marker data and QTL position) were the same as before (sample size $n = 500$).

The simulated responses are plotted against covariate time in Figure 2. Table 3 presents the posterior summaries of the analysis of the two data sets. For the data set simulated with the linear function, the estimates from the semiparametric and linear regression models are very similar. But for the data set simulated with the generalized logistic function, the semiparametric method produces again more accurate results.

[Figure 2 about here.]

[Table 3 about here.]

4.2 *Simulation II*

We simulated a backcross population with 10 chromosomes, each containing 12 evenly spaced markers (11 intervals of length 20 centiMorgan). Four QTLs were located at the center of chromosomes 1, 3, 5, and 7 with effects 0.5, -0.5, 0.5, and -0.5 respectively. A total of 500 individuals were sampled. The covariate function was of the form $\eta(\mathbf{t}) = \mathbf{5}(\mathbf{1} - \mathbf{2t_2})\sin(\mathbf{2t_1})$, therefore two (interacting) variables t_1 and t_2 were simulated, where t_1 was continuous and sampled from the uniform distribution $\mathcal{U}[0, 10)$, while t_2 was binary 0/1 and sampled with equal probability 0.5.

For the prior on η, we have mean $\mu(\mathbf{t}; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 t_1 + \alpha_3 t_2 + \alpha_4 t_1 t_2$, and covariance kernel $\sigma_0(\mathbf{t}; \mathbf{t}') = \exp[-\{(t_1 - t_1')^2 + (t_2 - t_2')^2\}]$. As before, the prior distribution for the αs is improper uniform. The prior for τ is again an inverse gamma distribution and is independent of α's. The lower panel of Figure 3 shows the posterior mean for the probability of each interval containing a QTL. For comparison, we also performed linear regression analysis by setting the covariance kernel of η to 0. The linear model results are given in the upper panel of Figure 3. Clearly the semiparametric model performed better than the parametric model.

[Figure 3 about here.]

4.3 *Real data analysis*

In addition to the simulations, we tested our method on a real mouse study of 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, NZO (New Zealand Obese)/HlLt strain to the relatively lean NON (Nonobese Nondiabetic)/Lt strain, and then reciprocally backcrossing obese F1 mice to the lean NON/Lt parental strain. They measured the body weights of 203 backcross males. The total number of markers is 85 with average distance between adjacent markers of 20.5 cM. In addition, inguinal, gonadal, retroperitoneal and mesenteric fat pad weights have also been measured (the data set can be

downloaded at http://cgd.jax.org/nav/qtlarchive1.htm). 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. The lean body weight (LBwt) is defined as the difference between the total body weight and the total weight of the fat pads. In their QTL analysis of the mesenteric fat pad weight (MFPwt), Stylianou et al. (2006) adjusted for the effect of LBwt. We applied our semiparametric single and multiple Bayesian QTL mapping methods to the MFPwt using LBwt as a covariate. Both analyses strongly suggest a single QTL on chromosome 4 (Figure 4, panels (b) and (d)). The estimated function of LBwt on mesenteric fat pad weight is presented in Figure 4(a) and is nearly linear. For comparison, we performed parametric multiple Bayesian QTL mapping, which includes a linear effect of LBwt. This analysis also identified a single QTL on chromosome 4 (Figure 4(c)). Hence, as in our simulations, the semiparametric and parametric mapping methods yield very similar results when the relationship between covariate and trait is (nearly) linear, although the posterior probability for QTL presence was reduced for the semiparametric method, relative to the parametric analysis (Figure 4 (c) and (d)).

[Figure 4 about here.]

5. Discussion

We have proposed efficient and robust Bayesian semiparametric, single and multiple interval QTL mapping, which allows for an unknown function of non-genetic covariates. The covariates may have a non-linear relationship with the response and/or interact with each other in a complex way. A Gaussian process is used as the prior for the unknown function. This prior does not require any assumptions like monotonicity or additivity. A hierarchical scheme to construct a prior around a parametric family has been described. We specified the Gaussian process prior via the simple, hierarchical model in Equation (4), and we used approximately

uninformative default prior for the hyperparameters: Flat priors for the location parameters (α_is), and an inverse Gamma(0.1,0.1) prior for the precision parameter τ, which controls the smoothness of the sample paths of the unknown function. The method performed well on simulated and real datasets. The software and simulated data are available at the website:

`http://www.bios.unc.edu/~fzou/semiparam/index.html.`

One of our reasons for choosing the Gaussian process is its ability to deal with multiple covariates. With the Gaussian process, we can specify a simple function as in Equation (3), and use the extra layer of randomness represented by Equation (4) to fine tune the curve to fit the data. Alternatively fitting a spline function to $\eta(t)$ will not require the extra randomness of Equation (4). However, the drawback of the spline approach is its inflexibility in handling high-dimensional covariates. This is important for a substantial extension of our method. Essentially all parametric multiple QTL mapping method assume additive QTL action, or incorporate at most two-locus interactions, but no higher-order interactions. How to accommodate a large number of potential QTL with possibly higher-order interactions and interactions with environmental covariates in multiple QTL mapping remains a challenging problem. Our current work focuses on extending our semiparametric Gaussian process model to unknown functions including not only non-genetic covariates but also multiple putative QTL, with variable selection.

6. Acknowledgements

The authors wish to thank the Editor, Associate editor and two Referees for their helpful comments and suggestions, which have led to a great improvement of this article. This research was partially supported by NIH grants GM074175 (F.Z and H. H.) and CA79949 (H. Z.).

References

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

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B* **60**, 725-749.

Basten, C. J., Weir, B. S., and Zeng, Z. B. (1999). QTL Cartographer: a reference manual and tutorial for QTL mapping. Department of Statistics, North Carolina State University.

Best, N. C. , Cowles, M. K. , and Vines, S. K. (1995). CODA manual version 0.30. MRC Biostatistics Unit, Cambridge, UK.

Broman, K. W., Speed, T. P. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses (with discussion). *Journal of the Royal Statistical Society, Series B* **64**, 641-656, 731-775.

Broman, K. W., Wu, H., Sen, S., and Churchill, G. A. (2003). R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889-890.

Chen, H. (1988). Convergence rates for parametric components in a partly linear model. *Ann. Statist.* **16**, 136-146.

Cuzick, J. (1992). Semiparametric additive regression. *Journal of the Royal Statistical Society, Series B* **54**, 831-843.

Denison, D. G. T., Mallick, B. K. and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society Series B* **60**, 333-350.

DiMatteo, I., Genovese, C. R. and Kass, R. (2001). Bayesian curve fitting with free-knot splines. *Biometrika* **88**, 1055-1071.

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

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical*

Association **87**, 998-1004.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* **7**, 457-72.

George, E. I., and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* **88**: 881-889.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. Bayesian Statistics, 4 (ed. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith), Clarendon Press, Oxford, UK.

Godsill, S. J. (2001). On the relationship between Markov chain Monte Carlo methods for model uncertainty. *Journal of Computational and Graphical Statistics* **10**, 230-248.

Godsill, S. J. (2003). Proposal densities, and product space methods, in Highly Structured Stochastic Systems. Oxford University Press.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Härdle, W. (1990). Applied nonparametric regression. Cambridge University Press.

Hastie, T. J. and Loader, C. (1993). Local regression: Automatic kernel carpentry (with discussion). *Statistical Science* **8**, 120-143.

Heckman, N. (1986). Spline smoothing in a partly linear model. *Journal of the Royal Statistical Society, Series B* **48**, 244-248.

Hoeschele, I. (2007). Mapping quantitative trait loci in outbred pedigrees, pp. 477-525 in Handbook of Statistical Genetics, edited by D. J. Balding, M. Bishop and C. Cannings. Wiley, New York.

Lan, H., Kendziorski, C. M., Haag, J. D., Shepel, L. A., Newton, M. A., and Gould, M. N. (2001). Genetic loci controlling breast cancer susceptibility in the Wistar-Kyoto rat. *Genetics* **157**, 331-339.

Lander, E. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185-199.

Lincoln, S. E., Daly, M. J., and Lander, E. S. (1993). A tutorial and reference manual for MAPMAKER/ QTL. Whitehead Institute.

Lynch, M. and Walsh, B. (1998). Genetics and analysis of quantitative traits. Sunderland, Mass., Sinauer.

MacKay, D. J. (1998). Introduction to Gaussian processes. In C. M. Bishop (Ed.), Neural networks and machine learning. Berlin. Springer.

Manly, K.F., and Olson, J.M. (1999). Overview of QTL mapping software and introduction to Map Manager QT. *Mammalian Genome* **10**, 327-334.

Mays, J. E. (1995). Model robust regression: combining parametric, nonparametric, and semiparametric methods. Unpublished Ph.D. dissertation. Virginia Polytechnic Institute and State University. Blacksburg, VA. 200p.

McQueen, M. B., Bertram, L., Rimm, E. B., Blacker, D. and Santangelo, S. L. (2003). A QTL genome scan of the metabolic syndrome and its component traits. *BMC Genetics* **4**, S96.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 67-79.

Neal, R. M. (1996). Bayesion learning for neural networks. New York: Springer-Verlag.

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

O′Hagan, A. (1978). On curve fitting and optimal design for regression. *Journal of the Royal Statistical, Society B* **40**, 1-42.

Reifsnyder, P. C., Churchill, G. A., and Leiter, E. H. (2000). Maternal environment and

genotype interact to establish diabesity in mice. *Genome Research* **10**, 1568-1578.

Richards, P. J. (1959). A flexible growth function for empirical use. *Journal of Experimental Botany* **10**, 290-300.

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

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

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75**, 317-343.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical, Society B* **50**, 413-436.

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

Stoehr, J. P., Nadler, S. T., Schueler, K. L., Rabaglia, M. E., Yandell, B. S., Metz, S. A., and Attie, A. D. (2000). Genetic obesity unmasks nonlinear interactions between murine type 2 diabetes susceptibility loci. *Diabetes* **49**, 1946-1954.

Stylianou, I. M., Korstanje, R. Li, R., Sheehan, S., Paigen, B., Churchill, G. A. (2006). Quantitative trait locus analysis for obesity reveals multiple networks of interacting loci. *Mammalian Genome* **17**, 22-36.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical, Society B* **58**, 267-288.

Ven, R. V. (2004). Reversible-Jump Markov chain Monte Carlo for quantitative trait loci mapping. *Genetics* **167**, 1033-1035.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against

model errors in regression, *Journal of the Royal Statistical, Society B* **40**, 364-372.

Wahba, G. (1984). Cross validated spline methods for the estimation of multivariate functions from data on functionals. In Statistics: An appraisal, proceedings 50th anniversary conference Iowa state statistical laboratory (H. A. David and H. T. David, 4s.) 205-235. Iowa State Univ. Press, Ames, Ia.

Wand, M. P. and Jones, M. C. (1995). Kernel smoothing. London: Chapman and Hall.

Wang, H., Zhang, Y. M., Li, X., Masinde, G. L., Mohan, S., Baylink, D. J. and Xu, S. (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**, 465-480.

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

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

Yi, N. and Xu, S. (2001). Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics* **157**, 1759-1771.

Yi, N. and Xu, S. (2008). Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics* **179**, 1045-1055.

Yi, N., Shriner, D., Banerjee, S., Mehta, T., Pomp, D. and Yandell, B. S. (2007). An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. *Genetics* **176**, 1865-1877.

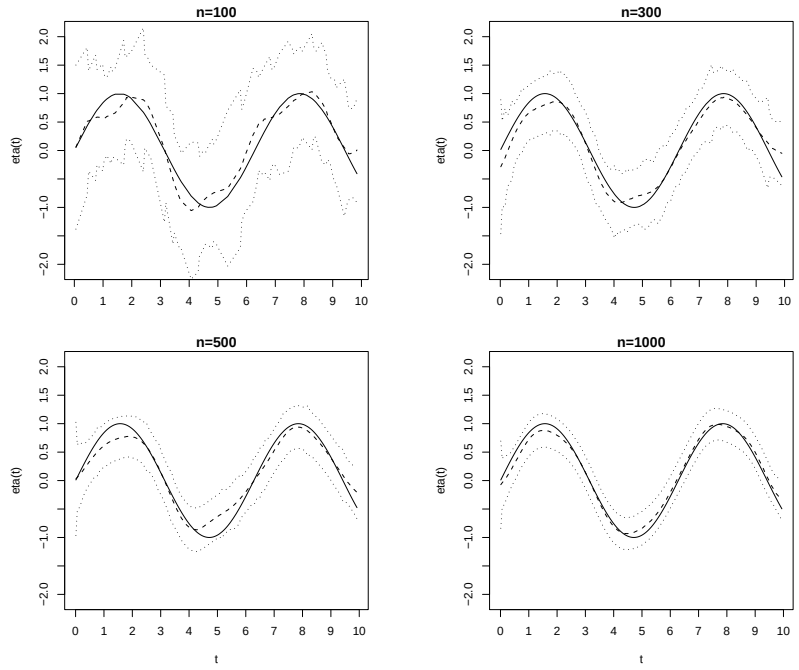

Figure 1. Convergence of estimator to the true function ($\eta(t) = \sin(t)$) with increasing sample size (n = 100, 300, 500 and 1000). True functions are in bold. The dashed line is the posterior mean of the unknown function. The dotted lines represent the 95% confidence band of the estimator.

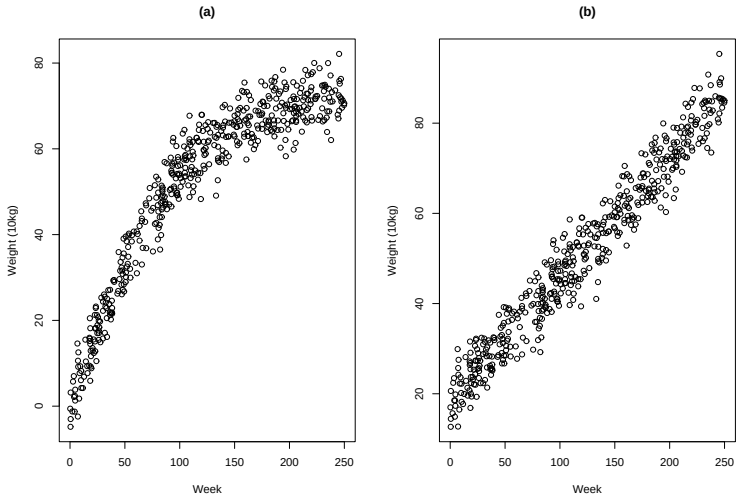

Figure 2. (a) Simulated cow weight against weeks based on the generalized logistic growth curve. (b) Simulated cow weight against weeks based on the linear growth curve.

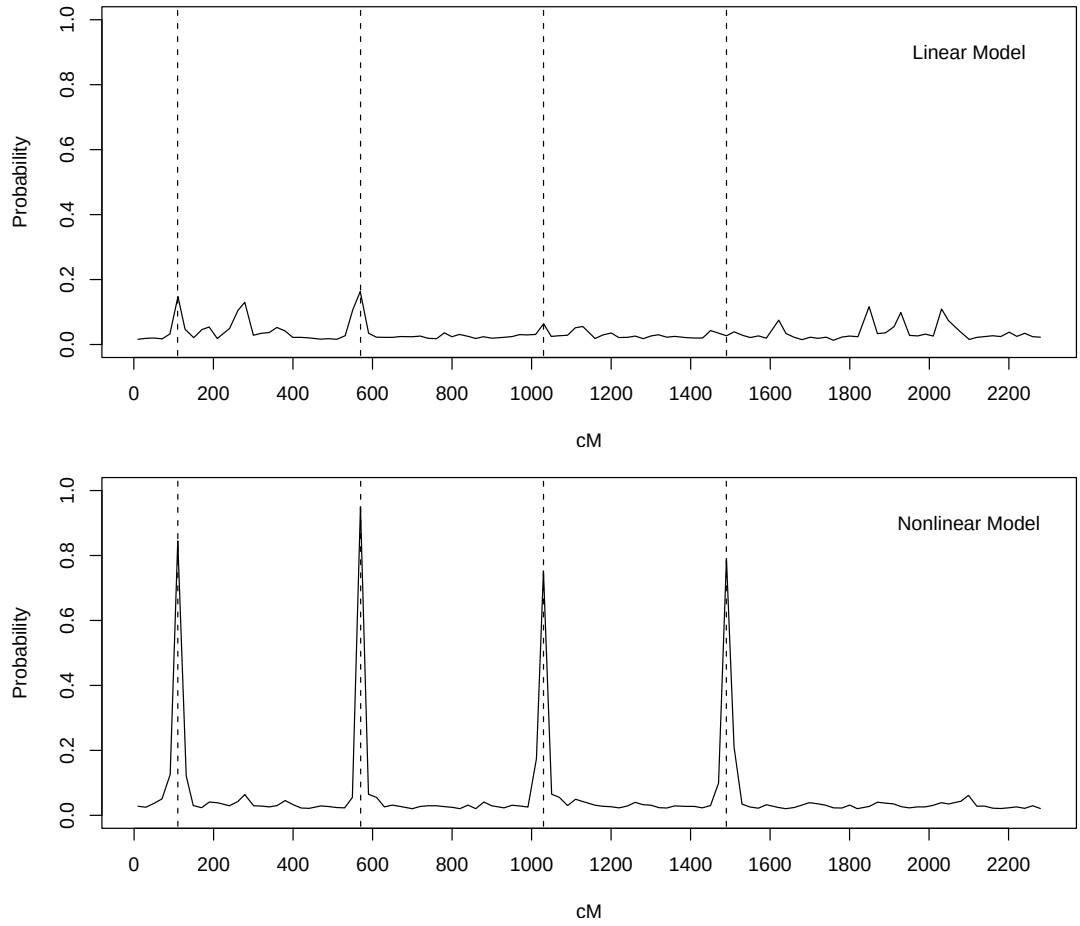

Figure 3. Posterior mean for the probability of each interval containing genes. Dashed lines represent true gene positions.

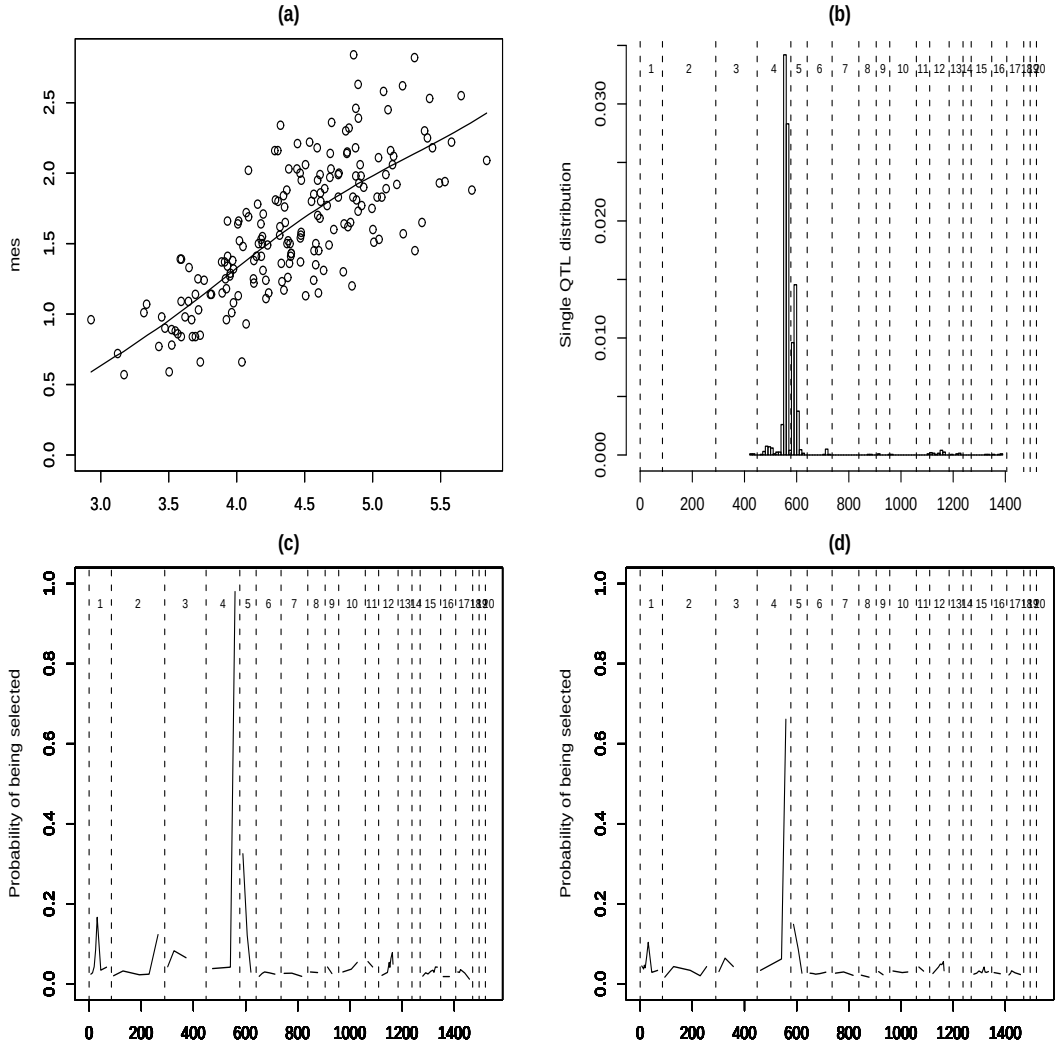

Figure 4. Bayesian analysis of the backcross mesenteric fat pad data. (a) Scatter plot of LBwt vs MFPwt. The solid line represents the estimated curve from the semiparametric single QTL mapping method. (b) Marginal posterior distribution of the QTL position based on the single QTL mapping model described in Section 2. (c) and (d) Posterior probability of QTL presence in each marker interval for the parametric and semiparametric multiple QTL mapping methods, respectively. Dashed lines separate the chromosomes.

Table 1
Posterior mean parameter estimates for the data simulated Data with sample size $n = 500$ and true covariate function $\eta(t) = \sin(t)$

	True Value	Posterior Mean	SD	P5	P95
QTL Position (cM)	155	155.52	0.90	153.97	156.99
QTL Effect	1	0.98	0.05	0.90	1.06
Residual Variance	1	0.99	0.07	0.88	1.12

SD, standard deviation of the estimated effect; P5 and P95, the fifth and ninety-fifth percentiles of the posterior distribution, respectively.

Table 2
Comparison between parameter estimates obtained with the linear parametric analysis and the semiparametric analysis (values in parentheses are for the linear parametric model)

$\eta(t) = \sin(t)$					
	True Value	Posterior Mean	SD	P5	P95
QTL Position	155 cM	151.86(153.18)	4.40(9.10)	143.40(144.65)	157.22(158.53)
QTL Effect	1	0.90(0.85)	0.13(0.15)	0.68(0.61)	1.12(1.09)
Residual Variance	1	1.07(1.47)	0.22(0.28)	0.77(1.09)	1.46(1.98)
$\eta(t) = 3\sin(t/3)$					
	True Value	Posterior Mean	SD	P5	P95
QTL Position	155 cM	151.80(299.51)	4.44(245.13)	143.27(71.80)	157.23(848.23)
QTL Effect	1	0.89(0.44)	0.13(0.41)	0.68(-0.30)	1.12(1.06)
Residual Variance	1	1.09(5.52)	0.23(0.98)	0.78(4.17)	1.49(7.26)
$\eta(t) = 5\sin(t/5)$					
	True Value	Posterior Mean	SD	P5	P95
QTL Position	155 cM	151.64(448.29)	4.33(284.40)	143.50(61.65)	157.13(937.22)
QTL Effect	1	0.90(0.21)	0.13(0.52)	0.67(-0.64)	1.12(1.05)
Residual Variance	1	1.10(13.11)	0.23(2.27)	0.77(9.99)	1.50(17.16)

Table 3
Comparison between the linear parametric model and the semiparametric model (values in parentheses are from the linear parametric model)

$\eta(t) = generalized\ logistic(t)$					
	True Value	Posterior Mean	SD	P5	P95
QTL Position	155 cM	155.03(153.13)	1.69(3.87)	152.98(145.77)	156.82(157.78)
QTL Effect	3	2.97(2.70)	0.16(0.43)	2.70(1.94)	3.22(3.37)
Residual Variance	9	8.40(68.25)	0.75(5.08)	7.34(60.73)	9.72(77.42)
$\eta(t) = linear(t)$					
	True Value	Posterior Mean	SD	P5	P95
QTL Position	155 cM	155.41(155.41)	0.87(0.87)	153.94(153.96)	156.86(156.81)
QTL Effect	3	2.98(2.94)	0.15(0.15)	2.73(2.69)	3.21(3.17)
Residual Variance	9	9.29(9.80)	0.63(1.09)	8.30(8.47)	10.36(11.88)