

Article

## A Bayesian Algorithm for Functional Mapping of Dynamic Complex Traits

Tian Liu <sup>1,2</sup> and Rongling Wu <sup>1,3,4,\*</sup>

<sup>1</sup> Department of Statistics, University of Florida, Gainesville, FL 32611 USA; E-Mail: liut2@gis.a-star.edu.sg (T.L.)

<sup>2</sup> Genome Institute of Singapore, 60 Biopolis Street 02-01 Genome, Singapore

<sup>3</sup> Department of Public Health Sciences, Pennsylvania State College of Medicine, Hershey, PA 17033 USA

<sup>4</sup> Department of Statistics, Pennsylvania State University, University Park, PA 16802 USA

\* Author to whom correspondence should be addressed; E-Mail: rwu@hes.hmc.psu.edu

*Received: 8 January 2009; in revised form: 6 March 2009 / Accepted: 24 March 2009 /*

*Published: 21 April 2009*

---

**Abstract:** Functional mapping of dynamic traits measured in a longitudinal study was originally derived within the maximum likelihood (ML) context and implemented with the EM algorithm. Although ML-based functional mapping possesses many favorable statistical properties in parameter estimation, it may be computationally intractable for analyzing longitudinal data with high dimensions and high measurement errors. In this article, we derive a general functional mapping framework for quantitative trait locus mapping of dynamic traits within the Bayesian paradigm. Markov chain Monte Carlo techniques were implemented for functional mapping to estimate biologically and statistically sensible parameters that model the structures of time-dependent genetic effects and covariance matrix. The Bayesian approach is useful to handle difficulties in constructing confidence intervals as well as the identifiability problem, enhancing the statistical inference of functional mapping. We have undertaken simulation studies to investigate the statistical behavior of Bayesian-based functional mapping and used a real example with  $F_2$  mice to validate the utilization and usefulness of the model.

**Keywords:** Bayesian; functional mapping; MCMC; Quantitative trait loci.

---

## 1. Introduction

Because of polygenic control and environmental modification, many traits of agricultural, biological and biomedical importance vary in a quantitative way [1]. For this reason, the genetic analysis of these so-called quantitative traits has been one of the most difficult issues in genetic research for many decades. With the advent of powerful molecular biotechnologies and statistical methodologies, it has now been possible to precisely dissect the phenotypic variation of a quantitative trait into individual loci at the molecular level (known as quantitative trait loci or QTLs) and understand the effects of interactions between QTLs and environments on the trait phenotype [2–4]. Thanks to tremendous efforts by geneticists worldwide, hundreds of thousands of QTLs have now been identified for a variety of complex quantitative traits in plants, animals, and humans. Meanwhile, a vast body of literature has been available about the development of statistical methods for QTL mapping [5–7], although most of these methods ignore the developmental or dynamic features of a trait in time course.

More recently, a library of statistical models, called functional mapping, has been developed to map QTLs that control dynamic traits [8, 9]. Functional mapping incorporates fundamental biological principles behind trait growth and development into a mapping framework. It capitalizes on the mathematical aspect of a ubiquitous growth law that every biological trait experiences a growth and developmental change with time through altering the ratio of the anabolic or metabolic rate and the rate of catabolism [10]. The advantages of this approach lie in its increased biological relevance by embedding biological principles into the estimation process and its flexibility to generate a number of quantitative testable hypotheses about the developmental and genetic regulation of growth. From a statistical perspective, functional mapping estimates parameters that determine the shape of a genotype-specific growth curve and/or the covariance structure, thus strikingly increasing its statistical power for QTL detection.

Functional mapping was originally constructed within the maximum likelihood (ML) context, incorporated by a finite mixture model and implemented with the EM algorithm [11, 12]. Although ML-based approaches have many favorable statistical properties for parameter estimation, they may often face some significant drawbacks when applied to functional mapping. First, functional mapping usually uses nonlinear mathematical equations to model the mean and covariance structures, it is extremely difficult or eventually impossible to derive log-likelihood equations for nonlinear parameters in the maximization step. Second, functional mapping concerns genetic analyses of longitudinal data whose intrinsic high-dimensional complexity makes computation increasingly prohibitive, especially when permutation tests are used to determine critical thresholds. Third, but not specific to functional mapping, ML-based approaches do not automatically provide confidence interval estimates of the estimators, and thus affect the inference of parameter estimation.

Many of the problems described above for ML can be overcome by using a Bayesian method. In ML, the unknown parameters are treated as unknown variables (unobservables) and the likelihood function is maximized in these variables. In the Bayesian paradigm, each unobservable parameter is given a prior distribution, and we then infer the posterior distribution of each unobservable conditional on the data (the observables). The summary statistics of the posterior distribution, e.g., the mean, the mode or the median, can be regarded as Bayesian estimates of unobservables [13]. The interval estimate can be obtained simply by examining the posterior distribution. The mean of this marginal posterior distribution is

a candidate Bayesian estimator of an unknown parameter. Although this marginal distribution rarely has an explicit form, and numerical integration is often prohibited because of high dimensionality of parameters, a Markov chain Monte Carlo (MCMC) algorithm can be used to simulate the sample from the joint posterior distribution. The potential of the Bayesian approach implemented with the Gibbs sampler or Metropolis-Hastings algorithm for QTL mapping has been explored for various genetic designs [14–18]. Yang and Xu [19] incorporated a Bayesian approach to map QTLs involved in a dynamic trait based on nonparametric Legendre polynomials, although they did not take a full advantage of functional mapping in quantifying biologically meaningful hypothesis tests about the genetic control of development and modeling the structure of the covariance matrix. For such biological processes as growth curve, HIV dynamics and pharmacodynamics, in which explicit mathematical functions exist to specify their dynamic changes, functional mapping based on parametric modeling has proven to be powerful for asking and addressing the biological questions at the interplay between genetic actions and developmental patterns.

In longitudinal data analysis, parametric covariance modeling has several advantages compared to conventional multivariate approaches ignoring the covariance structure. The most significant advantage is that parametric modeling generally results in more efficient estimation of the covariance matrix (and therefore the mean structure). Also, it can deal more effectively with data when the number of measurement times is relatively large. Lastly, it can handle the data more effectively in the cases that the measurement times are not common across subjects. For those reasons, the development of explicit parametric models for the variance-covariance structure has been an important topic over the last two decades [20, 21].

In this article, we will develop a general Bayesian framework for functional mapping of complex dynamic traits based on parametric modeling of the mean-covariance structures. This framework is constructed by a mixture model in which multiple mixture components corresponding to the genotypes of the underlying QTLs are involved. We will implement the MCMC algorithm to estimate the posterior distribution of each parameter contained within the mixture model including those that define curve shapes and the covariance structure. A real example for the  $F_2$  mouse progeny is used to demonstrate the utilization of the model and validate its usefulness in a practical genomic project of dynamic QTLs. We perform simulation studies to investigate the statistical properties of the Bayesian functional mapping model in terms of its convergence rate, estimation precision and power for QTL detection.

## 2. Bayesian Functional Mapping

### 2.1. Linear Model

Suppose there is an  $F_2$  population of size  $n$ , initiated with two inbred lines. In this  $F_2$  population, many markers are genotyped to construct a genetic linkage map, aimed to identify the QTLs that affect growth curves. For each progeny, a particular growth trait, such as body weight, tail length or cell number, is measured at a series of time points  $(1, \dots, T)$ . Consider a putative QTL with genotypes  $qq$  (denoted as 0),  $Qq$  (denoted as 1) and  $QQ$  (denoted as 2) that affects the shape of growth curves. At a specific time point  $t$ , the phenotypic value of the growth trait for progeny  $i$  due to the QTL may be expressed by a

linear model, i.e.,

$$y_i(t) = \sum_{j=0}^2 \xi_{ij} u_j(t) + e_i(t), \quad (1)$$

where  $\xi_{ij}$  is an indicator variable for progeny  $i$  carrying a QTL genotype and defined as 1 if a particular QTL genotype  $j$  is indicated and 0 otherwise,  $u_j(t)$  is the expected phenotypic value for QTL genotype  $j$  at time  $t$ , and  $e_i(t)$  is independently and identically distributed as  $N(0, \sigma^2(t))$ . Note that measurements within an individual are likely to be correlated across times, with covariance  $\sigma(t_1, t_2)$  between times  $t_1$  and  $t_2$  ( $t_1, t_2 = 1, \dots, T$ ). These variances and covariances form a  $(T \times T)$  matrix  $\Sigma$ .

## 2.2. Modeling the Mean-Covariance Structures

Functional mapping models  $u_j(t)$  by a biologically meaningful mathematical equation and the time-dependent covariance matrix composed of  $\sigma^2(t)$  and  $\sigma(t_1, t_2)$  by statistically robust approaches. For example, the growth of a living entity can be defined as the irreversible increase of size with time. A series of mathematical models have been proposed to describe growth curves. Among these models, the sigmoidal (or logistic) growth function is regarded as being nearly universal in living systems to capture age-specific changes [10]. Specifically, this S-shaped curve for a QTL genotype  $j$  can be mathematically expressed as:

$$g_j(t | \Theta_j) = \frac{\alpha_j}{1 + \beta_j e^{-\gamma_j t}} \quad (2)$$

where  $\Theta_j = (\alpha_j, \beta_j, \gamma_j)$  is set of parameters that determine curve shape of genotype  $j$ . Parameter  $\alpha$  is the limiting value of growth as time  $t$  goes to infinity,  $\alpha/(1 + \beta)$  is the initial value of growth, and  $\gamma$  is the relative growth rate. Many biologically important features of growth curves, such as the time of maximal growth rate and the duration of maximal growth, can be described by these parameters or their combinations.

A number of statistical methods have been derived to model the structure of covariance between longitudinal measurements [21]. The first-order autoregressive (AR(1)) model has been applied to model the structure of the within-subject covariance matrix for functional mapping. This model uses two simplified assumptions, i.e., variance stationarity – the residual variance ( $\sigma^2$ ) is constant over time, and covariance stationarity – the correlation between different measurements decreases proportionally (in  $\rho$ ) with increased time interval. These two assumptions facilitate the computation of functional mapping. Wu *et al.* [22] embedded Carrol and Rupert's [23] transform-both-sides (TBS) model into the growth-incorporated finite mixture model, in order to reduce the heteroscedasticity of the residual variance and preserve original biological means of curve parameters after the data are transformed. Núñez-Antón and colleagues proposed a series of so-called structured antedependence (SAD) models to approximate the age-specific change of correlation in the analysis of longitudinal variables [20]. The SAD model has been employed in several studies and displays many favorable properties for genetic mapping of dynamic traits [24].

2.3. Likelihood

Although only phenotypic values and marker genotypes are observable, the probability distribution of the QTL genotypes in a mapping population can be expressed in terms of the location of the putative QTL, the marker genotypes, and the distance between the markers. Let  $y_i$  be the phenotypic value of progeny  $i$  at different time points,  $M_i = \{M_{ik}\}_{k=1}^m$  be the  $m$ -marker genotype of progeny  $i$ ,  $\lambda$  be the QTL position measured by the distance of the QTL from the first marker of an ordered linkage group, and  $D = \{D_k\}_{k=1}^m$  be the distances between markers 1 and  $k$ . Suppose this QTL is located between markers  $k$  and  $k + 1$ . Then, we use  $\omega_{j|i}$  to denote the distribution of the QTL genotype of progeny  $i$ , i.e.,

$$\omega_{j|i} = \text{Prob}(Q_i = j | \lambda, M_{ik}, M_{i(k+1)}, D_k, D_{k+1}), \tag{3}$$

which is the conditional probability of QTL genotype  $j$  given the marker genotype of progeny  $i$ .

The likelihood of parameters  $\lambda, \Theta$  and  $\Sigma$  given observations can be written as:

$$L(\lambda, \Theta, \Sigma | \mathbf{y}) = \prod_{i=1}^n \left[ \sum_{j=0}^2 \omega_{j|i} \cdot \pi(\mathbf{y}_i | Q_i = j, \Theta_j, \Sigma) \right], \tag{4}$$

where  $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^n$ ,  $\Theta = \{\Theta_j\}_{j=0}^2$  and  $\pi$  is assumed to follow a multivariate normal distribution with the genotype-specific mean vector expressed as:

$$\begin{aligned} \mathbf{u}_j &= g(\mathbf{t} | \Theta_j) \\ &= \{u_j(t)\}_{t=1}^T \\ &= \left\{ \frac{\alpha_j}{1 + \beta_j e^{\gamma_j t}} \right\}_{t=1}^T \end{aligned} \tag{5}$$

and covariance matrix  $\Sigma$ .

2.4. Parameter Estimation and Algorithm

Estimation theory

We will derive a Bayesian approach to estimate the unknown parameters. This approach needs specifying prior distributions for the unknowns. Given the data and the priors specified, the posterior distribution over all unknown parameters can be obtained by using Bayes' theorem. Let  $\mathbf{Q} = \{Q_i\}_{i=1}^n$  denote the QTL genotypes for all  $n$  progeny. Then, the posterior density of  $\lambda, \Theta, \Sigma$ , and  $\mathbf{Q}$  is given by:

$$\pi(\lambda, \Theta, \Sigma, \mathbf{Q} | \mathbf{y}) \propto \pi(\mathbf{y} | \mathbf{Q}, \Theta, \Sigma) \cdot \pi(\mathbf{Q} | \lambda) \cdot \pi(\lambda, \Theta, \Sigma), \tag{6}$$

where  $\pi(\mathbf{y} | \mathbf{Q}, \Theta, \Sigma) = \prod \pi(\mathbf{y}_i | Q_i = j, \Theta_j, \Sigma)$  denotes the probability mass of the observation  $\mathbf{y}$  given the QTL genotypes,  $\pi(\mathbf{Q} | \lambda) = \prod \pi(Q_i | \lambda)$  is the probability mass of the QTL genotypes of all  $n$  progeny given their marker genotypes ( $\mathbf{M}$ ) and the QTL position ( $\lambda$ ), and  $\pi(\lambda, \Theta, \Sigma)$  is the prior imposed on the genetic parameters. It is reasonable to assume the priors are independent for the parameters. Thus, we have:

$$\pi(\lambda, \Theta, \Sigma) = \pi(\lambda) \cdot \pi(\Sigma) \cdot \prod_{j=0}^2 \pi(\Theta_j). \tag{7}$$

The priors can be chosen from related studies. In principle, if there is reliable information for parameters, such as  $\Theta$ , priors with a small variance may be used. For a parameter without enough information used to determine the prior, such as  $\Sigma$  and  $\lambda$ , noninformative priors or priors with a large variance may be utilized. Here, we choose a uniform distribution on  $[0, \mathbf{D}_m]$  as a prior of  $\lambda$ . Multivariate normal priors with moderate variances are used for  $\Theta_j$  can be obtained. The standard prior distribution for the inverse of the covariance matrix  $\Sigma^{-1}$  is the Wishart ( $\mathbf{R}, \rho$ ) [25, 26], where the so-called scale matrix  $\mathbf{C} = \mathbf{R}^{-1}$  represents prior structural information about  $\Sigma$  and  $\rho$  is the degree of freedom, greater than  $T - 1$ . A small value of  $\rho$  gives a relative flat distribution. The Wishart prior with low degrees of freedom and a specified  $\mathbf{R}$  is regarded as a reference (or noninformative) proper prior. Despite less flexibility, this distribution offers the advantage of being a conjugate prior, leading to a relative simple form for the posterior.

Theoretically, the marginal posteriors of the unknown parameters can be obtained from the joint posterior (6) by integrating over the other unknowns. Unfortunately, in practice, the evaluation of such high-dimensional integrals in a closed form is not possible. However, it is straightforward to derive either full conditional posterior distributions for some parameters, or the explicit expressions that are proportional to the corresponding full conditional posterior distributions for other parameters, i.e.,

$$\begin{aligned} & \pi(\Theta_j | \mathbf{y}, \Theta_{-j}, \Sigma, \mathbf{Q}, \lambda) \\ \propto & \pi(\mathbf{y} | \Theta, \Sigma, \mathbf{Q}, \lambda) \cdot \pi(\Theta_j) \\ \propto & \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n_j} [(\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))' \Sigma^{-1} (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))] - \frac{1}{2} (\Theta_j - \eta)' \Lambda^{-1} (\Theta_j - \eta) \right\} \end{aligned} \quad (8)$$

and

$$\begin{aligned} & \pi(\Sigma^{-1} | \mathbf{y}, \Theta, \mathbf{Q}, \lambda) \\ = & \pi(\mathbf{y} | \Theta, \Sigma^{-1}, \mathbf{Q}, \lambda) \cdot \pi(\Sigma^{-1}) \\ \propto & |\Sigma^{-1}|^{\frac{n+\rho+T+1}{2}} \cdot \exp \left\{ -\frac{1}{2} \text{tr} \left[ \mathbf{R}^{-1} \Sigma^{-1} + \sum_{j=0}^2 \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j)) (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))' \right] \right\} \end{aligned} \quad (9)$$

$$\sim \text{Wi}(\mathbf{D}^{-1}, n + \rho) \quad (10)$$

where  $\Theta_{-j} = \{\Theta_{j'} : j' = 0, 1, 2, j' \neq j\}$ ,  $\mathbf{y}_{ij}$  contains the observations from those individuals with genotype  $j$ , and  $\mathbf{D} = \mathbf{R}^{-1} + \sum_{j=0}^2 \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j)) (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))'$ .

### Algorithm implementation

Within the Bayesian framework, with the explicit expressions described in the preceding section, a Markov chain Monte Carlo (MCMC) technique can be used to draw samples from the joint posterior distributions of unknown parameters (6). We will use a hybrid scheme of Gibbs sampler and Metropolis-Hastings (M-H) algorithm [27, 28]. In particular, Gibbs sampling steps update  $\Sigma^{-1}$ , while the M-H algorithm updates  $\Theta_j$  and the QTL position. After generating a random sequence of states  $(\lambda^0, \mathbf{Q}^0, \Theta^0, \Sigma^0)$ ,  $(\lambda^1, \mathbf{Q}^1, \Theta^1, \Sigma^1)$ ,  $\dots$ ,  $(\lambda^N, \mathbf{Q}^N, \Theta^N, \Sigma^N)$ , the final MCMC samples are collected, from which we are able to make the inference about the unknown parameters. The construction of the Markov chain process was described in Appendix A.

Estimation issues

For parameters  $\lambda$  and  $\Theta$ , the empirical means of their marginal posteriors are the Bayesian estimators. As justified by Tierney [28], for any unknown parametric family  $f$ , which is a square integrable with respect to the stationary distribution  $\pi$ , we have:

$$\bar{f}_N = \frac{1}{N} \sum_{k=1}^N f(\lambda^{(k)}, \Theta^{(k)}, \mathbf{Q}^{(k)}, \Sigma^{(k)}) \rightarrow E_{\pi}[f(\lambda, \Theta, \mathbf{Q}, \Sigma | \mathbf{y})], \tag{11}$$

under the assumption that  $(\lambda^{(k)}, \Theta^{(k)}, \mathbf{Q}^{(k)}, \Sigma^{(k)})$  are the samples from the Markov chain. In other words, the empirical averages of the corresponding MCMC samples may be regarded as the consistent estimators for the unknown parameters.

The marginal posterior densities of these parameters are determined from the kernel density estimator [29], since the closed form of their full conditional posteriors are not available. For example, the histogram kernel density estimator for  $\lambda$  is given by:

$$\hat{\pi}(\lambda | \mathbf{y}) = \frac{1}{Nh} \sum_{j=0}^{D_m/h} I(jh < \lambda \leq (j + 1)h) \sum_{k=0}^N I(0 < \frac{\lambda^{(k)}}{h} - j \leq 1). \tag{12}$$

A second important estimation issue is to obtain the confidence intervals for the unknowns. Box and Tao [30] suggested that highest posterior density (HPD) regions can be constructed to give the confidence intervals for the parameters of interest and the detailed method for developing an approximated HPD via MCMC samples can be seen in Ritter and Tanner [31]. Alternatively, we can obtain the approximate HPD for the parameters directly by from their corresponding smooth density estimators.

Finally, in order to estimate the Monte Carlo error via the central limit theorem, Geyer [32] suggested three types of consistent estimators, the window estimators, the method of standardized time series and the specialized Markov Chain estimators. Among them, the windows estimators probably provides the best estimates, although it requires stronger regularity conditions for consistency.

2.5. Structuring the Covariance Matrix

In longitudinal data analysis, statistical modeling of the covariance structure is generally desirable, given that time-dependent variances and correlations follow a certain pattern. Below, we incorporate two commonly used approaches for the covariance structure in functional mapping into the Bayesian framework.

Autoregressive model

When a stationary AR(1) model is used, we specify the structure of covariance matrix  $\Sigma$  by constant variance and covariance, i.e.,

$$\text{var}(y(t)) = \sigma^2, \quad \forall 1 \leq t \leq T \tag{13}$$

$$\text{cov}(y(t_1), y(t_2)) = \sigma^2 \rho^{|t_1 - t_2|}, \quad \forall 1 \leq t_1 \neq t_2 \leq T. \tag{14}$$

We use the inverse gamma prior for  $\sigma^2$  and an informative prior restricted on  $[-1, 1]$  for  $\rho$ . Assuming that the priors of these two parameters are independent, the posterior density of  $(\lambda, \Theta, \sigma^2, \rho)$  is given as:

$$\pi(\lambda, \Theta, \mathbf{Q}, \rho, \sigma^2 | \mathbf{y}) \propto \pi(\mathbf{y} | \Theta, \mathbf{Q}, \sigma^2, \rho) \cdot \pi(\mathbf{Q} | \lambda) \cdot \pi(\lambda) \cdot \pi(\sigma^2) \cdot \pi(\rho) \cdot \prod_{j=0}^2 \pi(\Theta_j), \tag{15}$$

where  $\pi(\sigma^2) = IG(\alpha, \beta)$  and  $\pi(\rho) = Uniform(-1, 1)$ .

The closed forms of the full conditional posterior distributions of  $\sigma^2$  and  $\rho$  are not available, but the explicit expressions that are proportional to these posteriors can be derived, which are:

$$\pi(\sigma^2 | \mathbf{y}, \lambda, \mathbf{Q}, \Theta, \rho) \propto \pi(\mathbf{y} | \lambda, \mathbf{Q}, \Theta, \sigma^2, \rho) \cdot \pi(\sigma^2) \tag{16}$$

$$\propto |\Sigma(\sigma^2, \rho)|^{-\frac{n}{2}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \sigma^{2(-\alpha-1)} \tag{17}$$

$$\cdot \exp\left\{-\frac{\beta}{\sigma^2} - \frac{1}{2} \sum_{j=0}^2 \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))' \Sigma^{-1}(\sigma^2, \rho) (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))\right\} \tag{18}$$

for  $\sigma^2$ , and:

$$\pi(\rho | \mathbf{y}, \lambda, \mathbf{Q}, \Theta, \sigma^2) \propto \pi(\mathbf{y} | \lambda, \mathbf{Q}, \Theta, \sigma^2, \rho) \cdot \pi(\rho) \tag{19}$$

$$= \frac{1}{2} \cdot |\Sigma(\sigma^2, \rho)|^{-\frac{n}{2}} \cdot \exp\left\{-\frac{1}{2} \sum_{j=0}^2 \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))' \Sigma^{-1}(\sigma^2, \rho) (\mathbf{y}_{ij} - g(\mathbf{t} | \Theta_j))\right\} \tag{20}$$

for  $\rho$ . Based on these expressions, the corresponding Metropolis-Hastings steps can be developed to update  $\sigma^2$  and  $\rho$  within the MCMC estimation scheme (Appendix B).

### Structured Antedependence Model

According to Gabriel [33], the error at the current time depends on those at the previous times and the innovative error generated at the current time. Such a structured antedependence (SAD) model can be described by antedependence coefficients and innovation variances. The first-order SAD or SAD(1) model only contains one antedependence coefficient ( $\phi$ ). Assuming a constant innovation variance ( $\nu^2$ ), Jaffrézic *et al.* [34] derived a SAD(1)-structured covariance as:

$$\begin{aligned} \text{var}(t) &= \frac{1 - \phi^{2t}}{1 - \phi^2} \nu^2, \quad \forall 1 \leq t \leq T \\ \text{corr}(t_1, t_2) &= \phi^{t_1 - t_2} \frac{1 - \phi^{2t_2}}{1 - \phi^2} \nu^2, \quad \forall 1 \leq t_1 \neq t_2 \leq T. \end{aligned} \tag{21}$$

It can be seen that neither the variance nor the correlation function is stationary for the SAD(1) model, i.e. the variance can change with time, and the correlation does not depend only on lag time  $|t_1 - t_2|$ .

We pose an inverse-gamma prior on innovation variance  $\nu^2$  and a normal prior on antedependence coefficient  $\phi$ . In a real data analysis, the priors are selected as  $\pi(\nu^2) = IG(\alpha, \beta) = IG(1, 1)$  and  $\pi(\phi) = N(\mu_\phi, \eta_\phi) = N(0, 10)$ . The explicit expressions that are proportional to the full conditional

posteriors of these two parameters can be derived as:

$$\pi(\nu^2 | \mathbf{y}, \lambda, \mathbf{Q}, \Theta, \phi) \propto \pi(\mathbf{y} | \lambda, \mathbf{Q}, \Theta, \nu^2, \phi) \cdot \pi(\nu^2) \tag{22}$$

$$\propto |\Sigma(\nu^2, \phi)|^{-\frac{n}{2}} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \nu^{2(-\alpha-1)} \tag{23}$$

$$\cdot \exp\left\{-\frac{\beta}{\nu^2} - \frac{1}{2} \sum_{j=0}^2 \sum_{\Omega_i=1}^{n_j} (\mathbf{y}_{ij} - \mathbf{f}(\Theta_j))' \Sigma^{-1}(\nu^2, \phi) (\mathbf{y}_{ij} - \mathbf{f}(\Theta_j))\right\} \tag{24}$$

for  $\nu^2$ , and:

$$\pi(\phi | \mathbf{y}, \lambda, \mathbf{Q}, \Theta, \nu^2) \propto \pi(\mathbf{y} | \lambda, \mathbf{Q}, \Theta, \nu^2, \phi) \cdot \pi(\phi) \tag{25}$$

$$= \frac{1}{2} \cdot |\Sigma(\nu^2, \phi)|^{-\frac{n}{2}} \tag{26}$$

$$\cdot \exp\left\{-\frac{1}{2\eta_\phi} (\phi - \mu_\phi)^2 - \frac{1}{2} \sum_{j=0}^2 \sum_{\Omega_i=1}^{n_j} (\mathbf{y}_{ij} - \mathbf{f}(\Theta_j))' \Sigma^{-1}(\nu^2, \phi) (\mathbf{y}_{ij} - \mathbf{f}(\Theta_j))\right\} \tag{27}$$

for  $\phi$ .

The detailed M-H steps for updating  $\nu^2$  and  $\phi$  are given in Appendix C.

### 2.6. Bayes Factor

The estimate of the number of QTLs is a very important but difficult issue for QTL mapping. In ML, the QTL number can be estimated on the basis of an assumed number by changing the dimensionality of the mode. The best QTL number is chosen using some model-selection criterion such as Akaike’s AIC or by hypothesis testing. However, the theory underlying AIC could not apply in the mixture mode context because of the absence of a single dominating local maximum in the likelihood. The unsuitability of the theory used to construct the null distribution of the likelihood ratio statistic for mixture model means that hypothesis testing is not straightforward in this context. Moreover, because the number of QTL ( $\kappa$ ) cannot be estimated as a parameter, it is not possible to estimate the sampling error and confidence interval of the estimate for QTL number. But in the Bayesian paradigm, inference is based on the posterior distribution of the parameters. Inference for  $\kappa$  is then based on the marginal posterior distribution,  $Pr(\kappa = l | \mathbf{y}), l = 1, 2, \dots, .$

Models  $\kappa = l_1$  and  $\kappa = l_2$  may be compared via the ratio of their posterior probabilities:

$$\frac{P(\kappa = l_1 | \mathbf{y})}{P(\kappa = l_2 | \mathbf{y})} = \frac{P(\mathbf{y} | \kappa = l_1) Pr(\kappa = l_1)}{P(\mathbf{y} | \kappa = l_2) Pr(\kappa = l_2)} = B_{l_1 l_2} \frac{Pr(\kappa = l_1)}{Pr(\kappa = l_2)}, \tag{28}$$

where the ratio of marginal probabilities of  $\mathbf{y}$ ,  $B_{l_1 l_2} = \frac{P(\mathbf{y} | \kappa = l_1)}{P(\mathbf{y} | \kappa = l_2)}$ , is known as the Bayes factor for comparing  $\kappa = l_1$  with  $\kappa = l_2$ . The Bayes factor does not depend on the prior distribution of  $\kappa$ . Recently, the Bayesian analysis of mixtures with an unknown number of components has received great attention [13, 35, 46]. In practice, a Bayes factor larger than 100 can often be regarded as an evidence for the preference of  $l_1$  QTLs over  $l_2$  QTLs.

### 3. A Worked Example

#### 3.1. Mapping Population

Cheverud *et al.* [36] constructed a linkage map with 76 microsatellite markers for 535 F<sub>2</sub> mice derived from two strains, the Large (LG/J) and Small (SM/J). The total length of this map is ~1,500 cM (in Haldane's units) and an average marker interval length is ~27.5 cM. The same experiment was repeated by Vaughn *et al.* [37] in which 502 F<sub>2</sub> mice were generated and a linkage map of 1,780 cM long was constructed. In both experiments, each F<sub>2</sub> progeny was measured for their body mass at 10 weekly intervals starting at age seven days. The raw weights were corrected for the effects of each covariate due to dam, litter size at birth, parity and sex. We combine these two F<sub>2</sub> populations for QTL mapping of growth trajectories. Overall, about 10% of the marker genotypes were randomly missing. The mice with missing data were excluded from the analyses.

#### 3.2. Results

Zhao *et al.* [23 38] first analyzed Vaughn *et al.*'s [37] data set by using functional mapping with maximum-likelihood based methods. They showed that body masses in the F<sub>2</sub> mice follow a logistic curve which can be described by Equation 2, but display substantial variation in the shape of curves. Bayesian-based functional mapping was used to genome-wide search for the QTLs that control mouse growth curves by estimating the QTL location ( $\lambda$ ), genotype-specific curve parameters ( $\Theta$ ), and the structure or unstructured covariance matrix ( $\Sigma$ ). The prior for the locations of  $s$  QTLs ( $\lambda_1, \dots, \lambda_s$ ) along each chromosome is simply assumed to be uniform over  $[0, D_m]$ . Maximum likelihood estimates of growth curve parameters by Zhao *et al.* [24,37] provide the information about priors of these parameters. The priors for  $\Theta_{j_1 j_2 \dots j_s}$  ( $j_1, \dots, j_s = 0, 1, 2$ ) are a multivariate normal, centered at  $(30, 10, 0.6)^T$  with a large dispersion  $\text{diag}(9, 4, 1)$ . The prior for the covariance matrix is set to be inverse-Wishart( $R^{-1}, \rho$ ), where  $R$  is given by the sample covariance matrix.

We used the distance of 40 cM from the first marker on each chromosome as the initial value of  $\lambda$ . The initial values of the other parameters were generated from their priors. For each analysis, the Markov chain was run for 60,000 cycles after discarding the first 2,000 cycles for the burn-in period. In order to reduce the serial correlation among samples, the chain was thinned by saving one iteration in every 60 cycles so that the total number of samples stored for calculating the targeted posterior distributions of the unknowns was 1,000 [26].

Bayesian functional mapping was used to genome-wide scan for the existence of QTLs for body mass growth trajectories in the F<sub>2</sub> mouse population. Estimated marginal posteriors of the QTL locations over all 19 mouse chromosomes are illustrated in Figure 1. from which the posterior peaks were observed on chromosome 6, 7, 10, 11, and 15. By calculating the Bayes factors of the model with equation 28 by assuming one QTL over the model assuming no QTL, chromosomes 6, 7 and 10 were each detected to harbor a QTL given that the logarithmic scaled Bayes factors are 12.91, 13.47 and 7.99 for the three chromosomes, respectively. The estimated locations of these QTLs are between between markers *D6nds* and *D6Mit58* on chromosome 6, marker *D7nds1* and *D7Mit17* on chromosome 7 and marker *D10Mit133* and *D10Mit14* on chromosome 10.

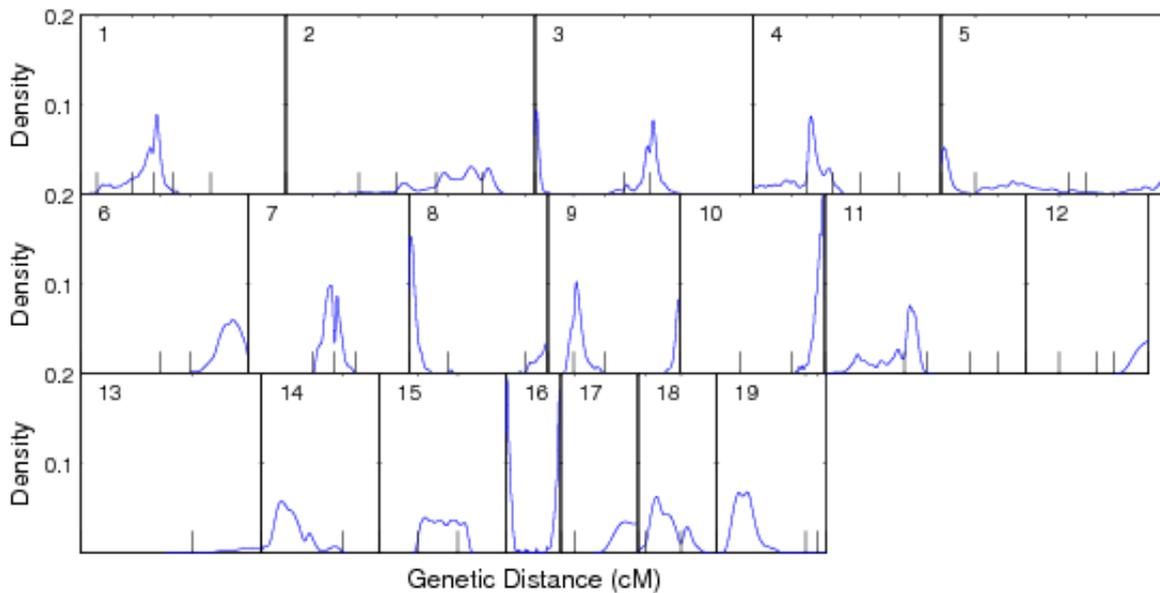
**Table 1.** Bayesian estimates of QTL locations and genotype-specific growth curves for the QTLs detected on mouse chromosome 6, 7 and 10. Numbers in parentheses are the 95% equal-tail confidence intervals.

| Parameter                          | <i>QQ</i>            |                | <i>Qq</i> |                | <i>qq</i> |                |
|------------------------------------|----------------------|----------------|-----------|----------------|-----------|----------------|
| <b>Chromosome 6</b>                |                      |                |           |                |           |                |
| Location, cM from the first marker | 82.68 (67.77, 92.96) |                |           |                |           |                |
| $\alpha$                           | 36.09                | (35.20,37.04)  | 34.94     | (34.36,35.52)  | 33.12     | (32.36,33.93)  |
| $\beta$                            | 11.93                | (11.44,12.45)  | 11.58     | (11.16,12.03)  | 11.07     | (10.65,11.51)  |
| $\gamma$                           | 0.65                 | (0.64, 0.66)   | 0.65      | (0.64, 0.66)   | 0.65      | (0.64, 0.67)   |
| <b>Chromosome 7</b>                |                      |                |           |                |           |                |
| Location, cM from the first marker | 46.84 (38.80,56.02)  |                |           |                |           |                |
| $\alpha$                           | 36.55                | (35.50, 37.73) | 35.61     | (34.56,36.50)  | 33.38     | (32.54,34.33)  |
| $\beta$                            | 11.83                | (11.43, 12.34) | 11.27     | (10.90, 11.73) | 11.25     | (10.76, 11.70) |
| $\gamma$                           | 0.65                 | (0.63, 0.66)   | 0.64      | (0.63, 0.65)   | 0.65      | (0.63, 0.66)   |
| <b>Chromosome 10</b>               |                      |                |           |                |           |                |
| Location, cM from the first marker | 77.78 (68.75,80.96)  |                |           |                |           |                |
| $\alpha$                           | 35.41                | (34.33, 36.52) | 34.71     | (33.67,35.70)  | 33.59     | (32.61,34.42)  |
| $\beta$                            | 11.67                | (11.23, 12.19) | 11.47     | (11.23, 11.81) | 11.01     | (10.61, 11.44) |
| $\gamma$                           | 0.65                 | (0.64, 0.66)   | 0.64      | (0.63, 0.66)   | 0.65      | (0.63, 0.66)   |

Table 1 tabulates Bayesian estimates for the locations of the three QTLs detected and their genotype-specific curve parameters from the marginal posterior means of these parameters. From the estimated 95% equal-tail confidence intervals as HPD regions, Bayesian estimates are considered to be reasonably precise (Table 1).

The estimated curve parameters were used to draw the growth curves of three different genotypes at each of the three detected QTLs (Figure 2). The three QTLs exert an effect on body mass growth curves in a similar pattern. The homozygote for the LG/J allele has the best growth during the time course of measurement, followed by the heterozygote for the LG/J and SM/J alleles and the homozygote for the SM/J allele. Based on quantitative genetic theory, we can partition time-dependent genotypic values into time-dependent additive ( $a(t)$ ) and time-dependent dominance effect components ( $d(t)$ ) for each QTL, illustrated in Figure 3. The additive effect due to the substitution of the SM/J allele with the LG/J allele is positive and increases with age for all the three detected QTLs. The dominant effect due to the interaction between the LG/J and LG/J alleles is consistently small in time course.

**Figure 1.** A profile of Estimated marginal posterior distribution of the QTL location by assuming that exactly one QTL is located on one of the chromosome respectively.



#### 4. Monte Carlo Simulation

Simulation studies were performed to study the statistical behavior of Bayesian-based functional mapping and demonstrated its applicability and utilization. The simulation design is based on an  $F_2$  population containing 450 individuals. A linkage group of length 100 cM was simulated with 11 equally-spaced, ordered codominant markers. A QTL that affects growth curves was assumed at 34 cM from the first marker. The true values of curve parameters for three QTL genotypes are given as:

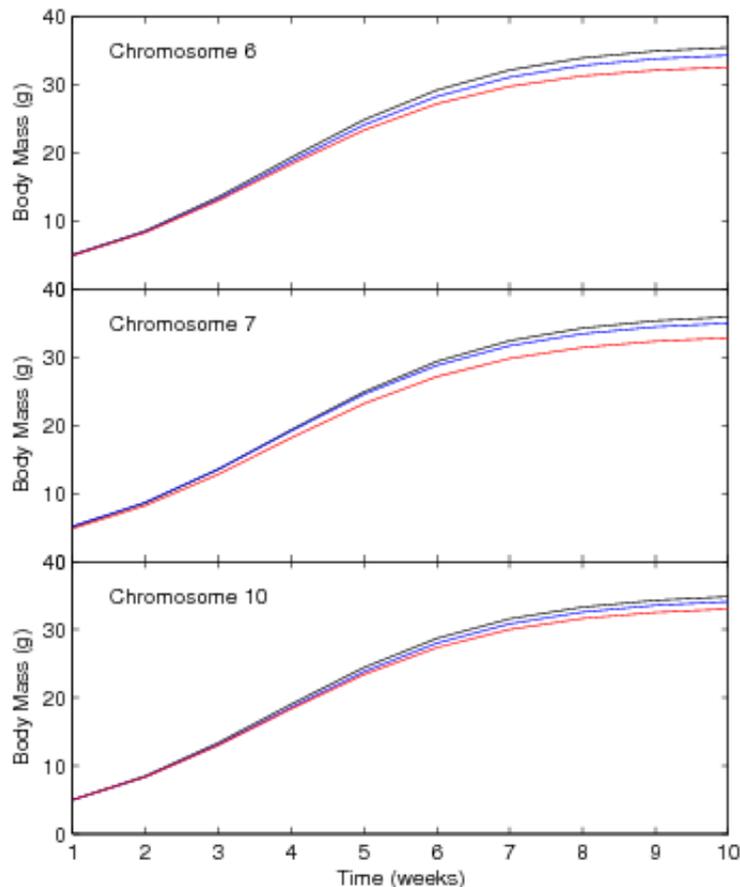
$$\Theta_0 = (33.4, 11.2, 0.65), \quad \Theta_2 = (36.7, 11.9, 0.65), \quad \Theta_1 = (35.6, 11.2, 0.64),$$

which are close to the estimates for body mass growth from the real mouse example used above. A total of 10 evenly spaced time points were assumed to measure growth curves. The residual covariance matrix among different time points was set to be the same as the sample covariance matrix estimated from the mouse example. Given this hypothesized covariance matrix and curves parameters for the three QTL genotypes, the heritability of the simulated growth trait at a middle time point is about 0.1. The time-dependent phenotypic values of the growth trait are assumed to follow a multivariate normal distribution.

The priors for curve parameters  $\Theta_j$  ( $j = 0, 1, 2$ ) are assumed to be a multivariate normal distribution centered at  $(30, 10, 0.7)$  with covariance matrix  $\Lambda = \text{diag}\{10, 5, 4\}$ . We used a uniform distribution as a prior for the QTL location over the simulated linkage group. Three different approaches were used to estimate the residual covariance matrix  $\Sigma$ : (1) estimating the unstructured covariance matrix based on an inverse Whishart prior with  $T = 10$  degrees of freedom, (2) estimating the structured covariance matrix based on the SAD(1) model by imposing  $IG(1, 1)$  as a prior for innovation variance  $\nu^2$  and

$Normal(0, 10)$  as a prior for antedependence parameter  $\phi$ , and (3) estimating the structured covariance matrix based on the AR(1) model by imposing  $IG(10, 1)$  as a prior for variance  $\sigma^2$  and  $Uniform(-1, 1)$  as a prior for correlation  $\rho$ .

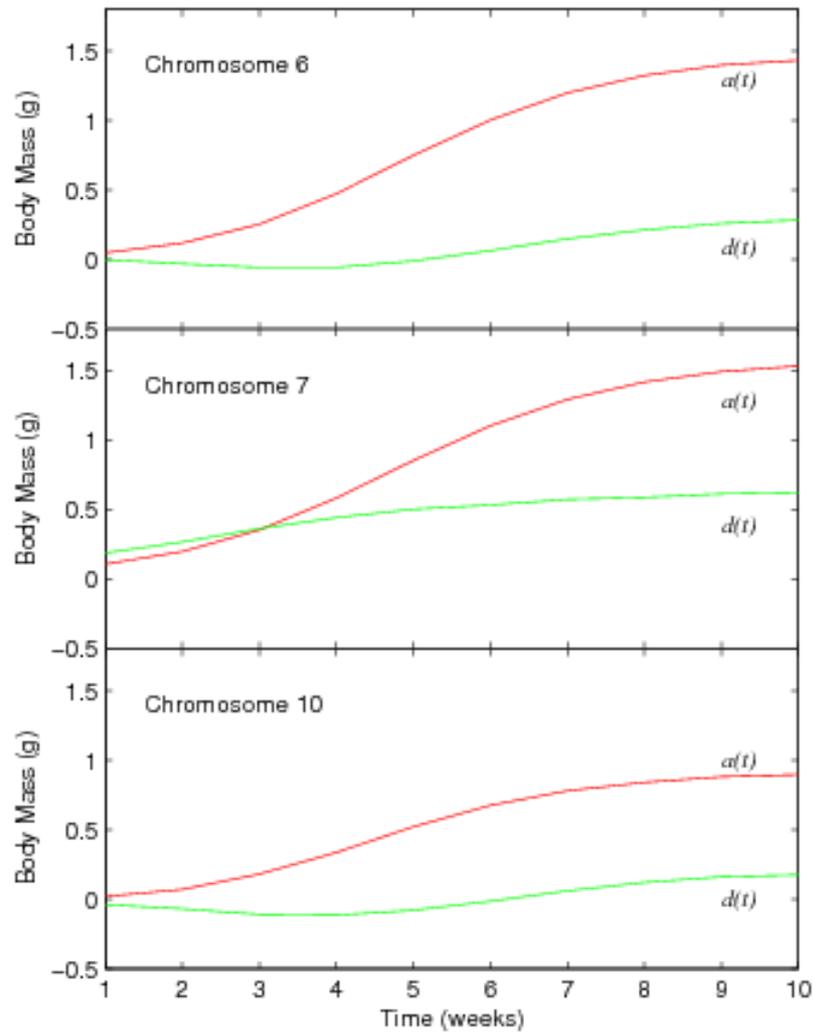
**Figure 2.** Fitted growth curves for the three QTL genotypes ( $qq$ , red;  $Qq$ , blue;  $QQ$ , black) assuming a single QTL is located on mouse chromosome 6, 7 and 10.



For each MCMC-implemented run, 10,000 initial "burn-in" iterations were discarded and, after then, samples are collected once for every 60 cycles so that a working set of 1,000 states were used to estimate the posteriors of parameters. The MCMC experiment was repeated several times with the same simulated data set to ensure the chain to converge to a stationary distribution. Table 2 gives the Bayesian estimates of the QTL location and growth curves for different QTL genotypes, along with the 95% empirical highest posterior density (HPD) confidence regions, under three different matrix-estimating approaches described above, respectively. It can be seen that our Bayesian-based functional mapping provides reasonably good estimates of the parameters in terms of accuracy and precision. The estimates of parameters, especially the QTL location (Figure 4), are affected by matrix-estimating approaches. Overall, the SAD(1)-structured approach is more precise in parameter estimation than the unstructured approach, whereas the unstructured approach is more precise than the AR(1)-structured approach. The SAD(1)-structured approach gives a narrowest confidence interval [33.01, 36.30] for the localization of

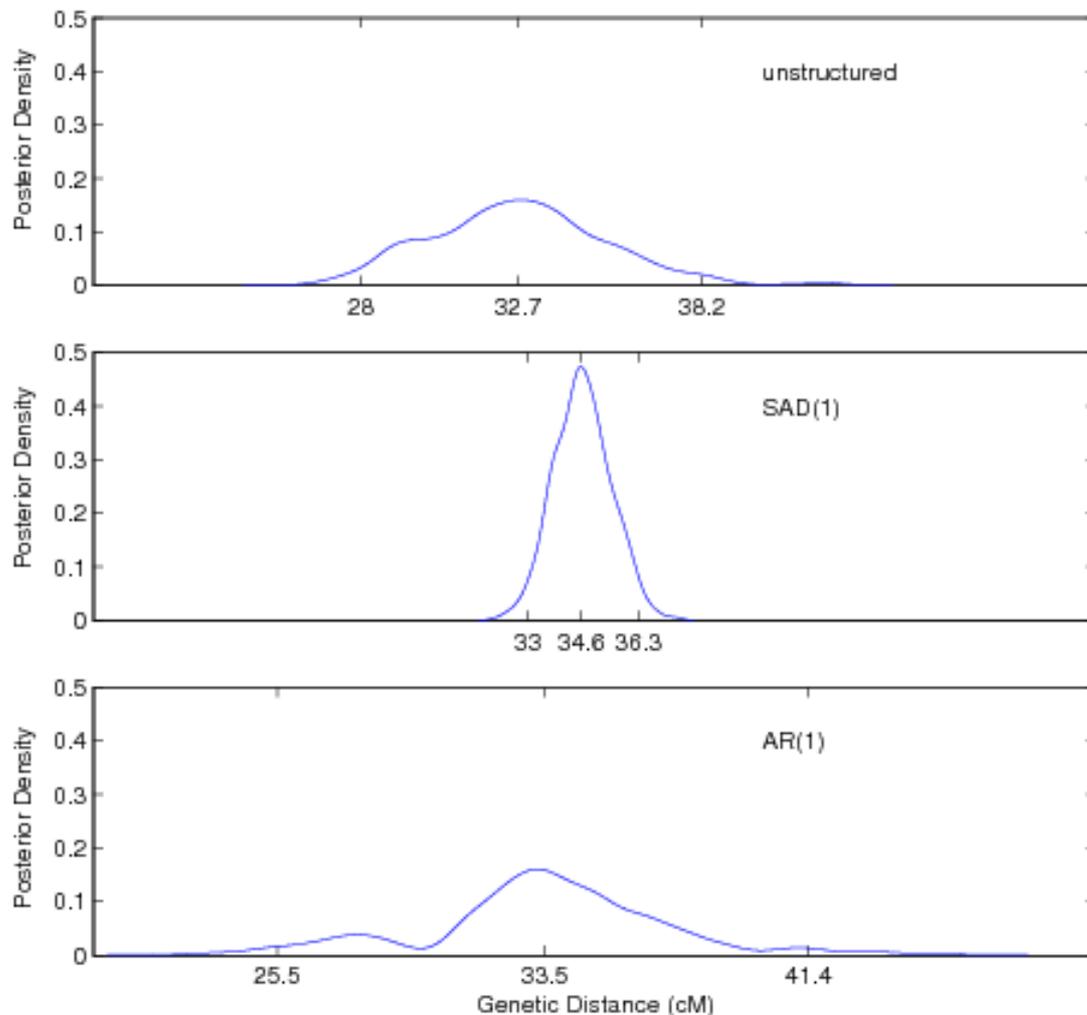
QTL, followed by the unstructured [28.02, 38.16] and AR(1)-structured approach [25.54, 41.39]. Also, the AR(1)-structured approach generates a small peak for the marginal posterior distribution at a wrong location (Figure 4), thus increasing a risk to detect a spurious QTL.

**Figure 3.** Dynamic changes of the additive and dominant effect due to the QTL located on mouse chromosome 6,7,and 10 respectively.



The simulated data set was further analyzed by ML-based functional mapping in which the likelihood ratio (LR) test statistic was computed and plotted across the linkage group (Figure 5). We found two drawbacks for the ML-based approach. First, the LR profile has two small peaks, giving a chance to claim a false positive QTL. On the other hand, none of these peaks is sharp over a flatted profile, suggesting less power and precision for QTL detection by ML-based functional mapping. Second, the significance test for the ML-based approach is based on computationally expensive permutation tests. ML-based functional mapping costs about 20 times in computing times (including 100 permutation tests) as much as does Bayesian-based functional mapping.

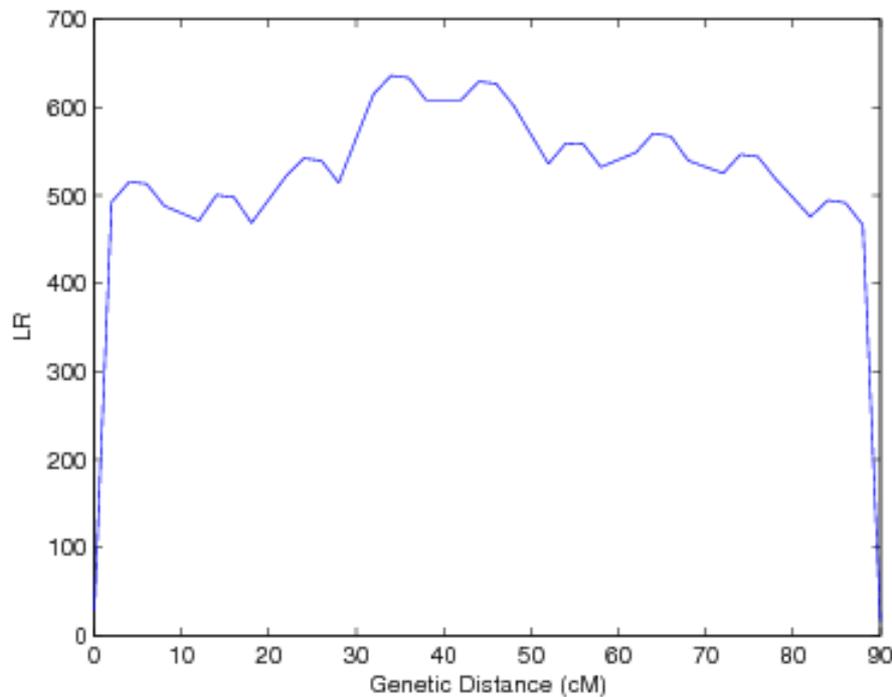
**Figure 4.** Estimated marginal posterior distributions of the QTL location over a simulated linkage group with different matrix-structuring approaches, unstructured (top), SAD(1)-structured (middle) and AR(1)-structured (bottom).



## 5. Discussion

The genetic architecture of complex traits can well be understood by incorporating their underlying developmental features described by mathematical functions. Functional mapping that integrates genetics, statistics and developmental biology as a whole can be useful for deciphering the ontogenetic development of the genetic control of a complex quantitative trait [8, 9]. Original models for functional mapping were derived within the maximum likelihood (ML) context and implemented with the EM algorithm. A similar approach for mapping logistic QTLs was used by Malosetti [8]. Although ML-based approaches possess many favorable statistical properties in parameter estimation, they may not be powerful enough to handle the complexity of high-dimensional QTL mapping models, as often seen in functional mapping. As an increasingly popular approach, Bayesian methods display remarkable capacity to estimate genetic parameters in QTL mapping [17–19, 39, 40].

**Figure 5.** The profile of the log-likelihood ratio (LR) test statistics between the full (there is a QTL) and reduced (there is no QTL) models across a simulated linkage group. The covariance matrix was structured by the SAD(1) model.



In this article, we derived a general Bayesian framework for functional QTL mapping of dynamic traits and implemented Markov chain Monte Carlo (MCMC) algorithms to locate genomic positions of QTLs, and estimate the mathematical parameters that define a biological process and the statistical parameters that model the covariance structure. The Bayesian-based model allows the estimation of these parameters and their confidence intervals based on posterior distributions, and has great power to handle complex estimation issues related to functional mapping in an effective way. Like original parametric functional mapping [9], the new model allows to approximate the ontogenetic changes of the genetic effects triggered by a QTL. Because many biological processes, such as growth, follow a particular pattern of development, the ontogenetic control of a QTL can be mathematically described and, thereby, tested by estimating the parameters that define a biological process. The new model also take an advantage of functional mapping to model the structure of covariance matrix by a stationary or nonstationary approach.

The application of Bayesian approaches to functional mapping was also considered by other authors. Yang and Xu [19] integrated Bayesian shrinkage approaches to map dynamic QTL, but their model was based on a nonparametric Legendre polynomial fitting. Although this treatment may be statistically flexible, its biological relevance may be limited because no biologically sensible functions are deployed. Also, Yang and Xu's [19] approach did not utilize the high-efficiency characteristic of functional mapping through modeling the structure of an autocorrelated covariance matrix.

**Table 2.** Bayesian estimates of QTL locations and genotype-specific growth curves for an assumed QTL from the simulated data set for an  $F_2$  population of 450 individuals based on different covariance-structuring approaches. Numbers in parentheses are the 95% equal-tail confidence intervals.

| Parameter                         | $QQ$                 | $Qq$                 | $qq$                 |
|-----------------------------------|----------------------|----------------------|----------------------|
| <b>Unstructured approach</b>      |                      |                      |                      |
| Location                          |                      | 32.74 (28.02, 38.16) |                      |
| $\alpha$                          | 36.67 (35.81, 37.46) | 36.03 (35.47, 36.63) | 33.67 (32.67, 34.31) |
| $\beta$                           | 11.83 (11.95, 12.64) | 11.22 (10.97, 11.50) | 11.30 (10.94, 11.60) |
| $\gamma$                          | 0.66 (0.64, 0.67)    | 0.64 (0.63, 0.65)    | 0.64 (0.63, 0.66)    |
| <b>SAD(1)-structured approach</b> |                      |                      |                      |
| Location                          |                      | 34.63 (33.01, 36.30) |                      |
| $\alpha$                          | 36.58 (35.85, 37.36) | 35.88 (35.37, 36.39) | 33.81 (33.09, 34.55) |
| $\beta$                           | 12.04 (11.65, 12.45) | 11.27 (11.01, 11.54) | 11.46 (11.09, 11.83) |
| $\gamma$                          | 0.66 (0.65, 0.67)    | 0.64 (0.63, 0.65)    | 0.65 (0.64, 0.65)    |
| <b>AR(1)-structured approach</b>  |                      |                      |                      |
| Location                          |                      | 33.54 (25.54, 41.39) |                      |
| $\alpha$                          | 36.60 (35.59, 37.66) | 35.57 (34.88, 36.35) | 33.61 (32.65, 34.54) |
| $\beta$                           | 12.04 (11.61, 12.48) | 11.23 (10.99, 11.56) | 11.42 (11.05, 11.83) |
| $\gamma$                          | 0.65 (0.64, 0.66)    | 0.63 (0.62, 0.65)    | 0.66 (0.64, 0.67)    |

Given a simulated time-dependent covariance matrix, Bayesian-based functional mapping obtains different results about the precision of parameter estimation, depending on whether and how such a covariance matrix is structured. This suggests the importance of estimating the residual covariance matrix effectively and efficiently. The unstructured approach captures the full information of time-dependent variances and covariances, which is effective but not efficient. The SAD(1)-structured model approximates the changes of variance and correlation over time [20], which is not only efficient (through estimating fewer parameters) but also effective (by reflecting the reality of this simulated data set to great extent). The AR(1)-structure model assuming the stationarity in both variance and correlation is efficient but not effective as much as the SAD(1) model. Overall, the SAD(1)-structured approach performs best for this particular simulated longitudinal data, followed by the unstructured and AR(1)-structured approach in an order. If the covariance does not have a structure, an unstructured approach, such as one based on the Wishart prior, should be used. In Appendix D, we describe a different approach for estimating the full matrix based on a reference prior.

The model was used to reanalyze a published data set on the growth of body mass in mice, confirming the discovery of a few QTL detected by conventional ML-based functional mapping. Yet, the new model provides estimates of confidence intervals of curve parameters, thus allowing better statistical inferences about the genetic control of dynamic QTLs. Simulation studies show that Bayesian-based functional mapping is more robust than ML-based functional mapping, in that the former provides more reasonable

estimates of QTL positions and dynamic QTL effects when the heritability of growth curves is modest (0.1) compared to the latter. However, a unique issue related to the Bayesian approach is about its stability or sensitivity in parameter estimation to different priors. One approach for testing the model’s stability is to initiate the chain with several different starting points and/or different priors under the same MCMC sampling scheme [41], and further examine how the estimates depend on the choices of initial values and priors.

The model proposed in this article can be modified by considering a network of genetic control. As a basic Bayesian framework, our single-QTL interval mapping model is not adequate to explore the effects of interactions on developmental variation for a growth trait between different QTLs from the same genome [43] or different genomes [44], and between QTLs and environments [24]. One of the significant advantages of Bayesian approaches lies in the estimation of the optimal number of QTLs involved. Variable selection via stepwise regression is used in ML mapping, but it is highly computationally expensive. Corresponding to this variable selection procedure, reversible jump MCMC is proposed in Bayesian analysis [45, 46], although it is subject to poor mixing and a slow convergence to the stationary distribution [47, 48]. More efficient methods based on Bayesian shrinkage analysis [16] and stochastic search variable selection [17] have now been proposed. These methods do not rely upon any explicit form of variable selection; rather they proceed implicitly by shrinking the effects of excessive QTLs to zero. Our Bayesian-based model modified by considering complicated features of growth and development will certainly prove its value in elucidating the genetic architecture of dynamic traits and will probably be the beginning of detecting the driving forces behind developmental genetics and its relationship to the organism as a whole.

**Appendix A: Estimating  $(\lambda, \mathbf{Q}, \Theta, \Sigma)$**

The Markov chain for Bayesian functional mapping described in the **Parameter Estimation and Algorithm** section is constructed as follows:

**Step 1.** Initialize the iteration at an arbitrary point  $(\lambda^0, \mathbf{Q}^0, \Theta^0, \Sigma^0)$ , which has a positive posterior density;

**Step 2.** Modify four blocks of the unknowns parameters and move to a new state from the previous step  $(\lambda^{k-1}, \mathbf{Q}^{k-1}, \Theta^{k-1}, \Sigma^{k-1})$  through a successive generation of new values  $\lambda^k, \mathbf{Q}^k, \Theta^k$ , and  $\Sigma^k$ . More specifically, given the values of the unknowns  $(\lambda, \mathbf{Q}, \Theta, \Sigma)$  from the current state, we proceed as follows:

*Updating  $\lambda$*  : In each step, following the idea of Satagopan *et al.* [14],  $\lambda$  is updated by using the Metropolis algorithm. A new value of  $\lambda^*$  is generated from  $Uniform(\max(0, \lambda - \delta), \min(\lambda + \delta, D_m))$  (where  $\delta > 0$  is the tuning parameter), and this proposed distribution is denoted by  $q(\lambda, \lambda^*)$ . This proposed  $\lambda$  is accepted with probability  $\min(\alpha_\lambda, 1)$ , and the state keeps the current value if the proposal is rejected, where  $\alpha_\lambda$  is given as:

$$\alpha_\lambda(\lambda, \lambda^*) = \frac{\pi(\lambda^* | \mathbf{y}, \mathbf{Q}, \Theta, \Sigma) \cdot \mathbf{q}(\lambda^*, \lambda)}{\pi(\lambda | \mathbf{y}, \mathbf{Q}, \Theta, \Sigma) \cdot \mathbf{q}(\lambda, \lambda^*)} \tag{A 1}$$

Note that,

$$\begin{aligned} \pi(\lambda^* | \mathbf{y}, \mathbf{Q}, \Theta, \Sigma) &= \pi(\lambda^* | \mathbf{y}, \mathbf{Q}, \Theta, \Sigma, \mathbf{M}) \\ &= \pi(\lambda^* | \mathbf{Q}, \mathbf{M}) \\ &\propto \pi(\mathbf{Q} | \lambda^*, \mathbf{M}) \cdot \pi(\lambda^*) \\ &= \prod_{i=1}^n \pi(Q_i | \lambda^*, M_i) \cdot \pi(\lambda^*) \end{aligned} \tag{A 2}$$

Similarly, we have:

$$\pi(\lambda | \mathbf{y}, \mathbf{Q}, \Theta, \Sigma) \propto \prod_{i=1}^n \pi(Q_i | \lambda, M_i) \cdot \pi(\lambda) \tag{A 3}$$

Hence, the acceptance probability (A 1) can be simplified as:

$$\alpha_\lambda(\lambda, \lambda^*) = \min \left( \frac{\prod_{i=1}^n \pi(Q_i | \lambda^*, M_i) \cdot q(\lambda^*, \lambda)}{\prod_{i=1}^n \pi(Q_i | \lambda, M_i) \cdot q(\lambda, \lambda^*)}, 1 \right). \tag{A 4}$$

*Updating Q* : Because of independence among  $n$  progeny,  $\mathbf{Q}$  is updated by separately updating each  $Q_i$ . For each progeny  $i$  and QTL genotype  $j$ , the full conditional density is in the form of a multinomial with cell probabilities:

$$p_{ij} = \pi(Q_i = j | \mathbf{y}, \Theta, \Sigma, \lambda) = \frac{\pi(Q_i = j | \lambda) \cdot \pi(\mathbf{y}_i | \Theta, \Sigma, Q_i = j)}{\sum_{q_i=0}^2 \pi(Q_i = q_i | \lambda) \cdot \pi(\mathbf{y}_i | \Theta, \Sigma, Q_i = q_i)}. \tag{A 5}$$

Hence, at each cycle, we can sample the QTL genotype  $Q_i$  directly from this full conditional density.

*Updating Θ* : We update each  $\Theta_j$  successively by a Metropolis-Hastings algorithm. For each QTL genotype, a new value  $\Theta_j^*$  is generated from a proposed density  $q(\Theta_j, \Theta_j^*)$ , given the current  $\Theta$ . Evaluate the acceptance probability of the move is  $\min(1, \alpha_{\Theta_j})$ . In general,  $\alpha_{\Theta_j}$  can be expressed as:

$$\alpha_{\Theta_j} = \frac{\pi(\Theta_j^* | \mathbf{y}, \Theta_{-j}, \Sigma, \mathbf{Q}, \lambda) \cdot q(\Theta_j^*, \Theta_j)}{\pi(\Theta_j | \mathbf{y}, \Theta_{-j}, \Sigma, \mathbf{Q}, \lambda) \cdot q(\Theta_j, \Theta_j^*)}. \tag{A 6}$$

Note that the choice of the Metropolis kernel  $q$  is essentially arbitrary, and a symmetric  $q$  in the sense that  $q(\Theta_j, \Theta_j^*) = q(\Theta_j^*, \Theta_j)$  is usually preferred. And in that case, the ratio  $q(\Theta_j^*, \Theta_j)/q(\Theta_j, \Theta_j^*)$  is canceled in the above expression (A 6). Here, for the proposed density, we use a multivariate normal distribution centered at the current  $\Theta$ , with variance-covariance matrix given by an information-type matrix [49] whose inverse has  $(u, v)$ th element:

$$\sum_{i=1}^{n_j} \left( \frac{\partial g(\mathbf{t} | \Theta_j)}{\partial \Theta_{j,u}} \right)' \left( \frac{\partial g(\mathbf{t} | \Theta_j)}{\partial \Theta_{j,v}} \right) + \frac{1}{2} \frac{\partial}{\partial \Theta_{j,u}} \frac{\partial}{\partial \Theta_{j,v}} (\Theta_j - \eta)' \Sigma^{-1} (\Theta_j - \eta). \tag{A 7}$$

This expression combines the expected information (the first term) with the prior information (the second term) and offers an advantage of avoiding singular information matrices. Unfortunately, a tedious initial analysis has to be conducted to obtain estimates of  $\Theta_j$  and  $\Sigma$  from which to evaluate (A 7). So, the Metropolis algorithm described above can be carried out by using an arbitrary variance-covariance matrix. The posterior means of  $\Theta_j$  and  $\Sigma$  are then plugged into (A 7) from the subsequent analysis.

*Updating Σ* : We generate a new value of  $\Sigma^{-1}$  directly from its full conditional posterior distribution. This is straightforward since it has an explicit expression for its full conditional posterior distribution.

**Appendix B. Estimating  $(\sigma^2, \rho)$**

Below, we describe the Metropolis-Hastings steps for updating  $\sigma^2$  and  $\rho$  within the MCMC estimation scheme when the residual covariance matrix is structured by the AR(1) model.

**Updating  $\sigma^2$**  : In each MCMC cycle, a candidate value of  $\sigma^2$  denoted by  $\sigma^{2*}$  is generated from its proposal distribution, which can be specified as:

$$q(\sigma^{2*} | \sigma^2) = IG\left(\frac{1}{\sigma^2} + 1, 1\right).$$

This proposal will be accepted with probability:

$$\min(\alpha_{\sigma^2}, 1),$$

where:

$$\alpha_{\sigma^2} = \frac{\pi(\sigma^{2*} | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \rho) \cdot q(\sigma^2 | \sigma^{2*})}{\pi(\sigma^2 | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \rho) \cdot q(\sigma^{2*} | \sigma^2)}. \tag{B1}$$

**Updating  $\rho$**  : The proposal distribution of  $\rho$  can be specified as a uniform with a moderate range around the current value of  $\rho$ . In other words,  $q(\rho^* | \rho) = Uniform(max(-1, \rho - \delta_\rho), min(\rho + \delta_\rho, 1))$ . A new value of  $\rho, \rho^*$ , is generated from this proposal distribution and accepted with probability:

$$\min(\alpha_\rho, 1),$$

with:

$$\alpha_\rho = \frac{\pi(\rho^* | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \sigma^2) \cdot q(\rho | \rho^*)}{\pi(\rho | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \sigma^2) \cdot q(\rho^* | \rho)}. \tag{B2}$$

**Appendix C. Estimating  $(\nu^2, \phi)$**

The Metropolis-Hastings steps for updating  $\nu^2$  and  $\phi$  within the MCMC estimation scheme for a SAD(1)-structured residual covariance matrix is described as follows:

**Updating  $\nu^2$** : In each MCMC cycle, a candidate value of  $\nu^2$  denoted by  $\nu^{2*}$  is generated from its proposal distribution, which can be specified as:

$$q(\nu^{2*} | \nu^2) = IG\left(\frac{1}{\nu^2} + 1, 1\right).$$

This proposal will be accepted with probability:

$$\min(\alpha_{\nu^2}, 1),$$

where:

$$\alpha_{\nu^2} = \frac{\pi(\nu^{2*} | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \phi) \cdot q(\nu^2 | \nu^{2*})}{\pi(\nu^2 | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \phi) \cdot q(\nu^{2*} | \nu^2)}. \tag{C1}$$

**Updating  $\phi$ :** The proposal distribution of  $\rho$  can be specified as a uniform with a moderate range around the current value of  $\rho$ , i.e.,  $q(\phi^* | \phi) = N(\phi, V_\phi)$ . A new value of  $\phi$ ,  $\phi^*$ , is then generated from this proposal distribution and accepted with probability:

$$\min(\alpha_\phi, 1),$$

with:

$$\alpha_\phi = \frac{\pi(\phi^* | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \nu^2) \cdot q(\phi | \phi^*)}{\pi(\phi | \mathbf{y}, \boldsymbol{\lambda}, \mathbf{Q}, \boldsymbol{\Theta}, \nu^2) \cdot q(\phi^* | \phi)}. \tag{C2}$$

**Appendix D. Alternative Approach for Modeling  $\Sigma$**

Although the Wishart is a standard prior for the covariance matrix and convenient to use, it has been criticized for being too restricted and the lack of flexibility. Also, as pointed out by Dempster *et al.* [52] and Stein [50], if the true covariance matrix  $\Sigma$  is close to  $I$ , the eigenstructure of  $\Sigma$  can be systematically distorted by the estimator, so this conventional prior can behave poorly, especially when the sample size is small or the data are sparse. To overcome these drawbacks, several other more flexible priors have been introduced, including a log-matrix prior [51], a reference noninformative prior [53], and a constrained Wishart prior [54]. The covariance matrix can also be modeled with a different parameterization [55]. This strategy is based on the key idea that a covariance matrix for longitudinal data can be diagonalized, i.e.,

$$\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} = \mathbf{S}, \tag{D1}$$

where  $\mathbf{S}$  is a diagonal matrix with positive entries and  $\mathbf{A}$  is a unique lower triangular matrix with 1's on the diagonal.

Without nice properties due to the conjugate priors, the resulting full conditional for  $\Sigma$  is no longer inverse Wishart. Often, one might have to generate  $\Sigma$  componentwise by using Gibbs sampling. Consequently, we will focus on computationally easy methods that can generate the entire  $\Sigma$  at a time, given the problem's complexity. Therefore, as an option to further improve the estimation for the covariance matrix, the reference (noninformative) prior, first introduced by Berger and Bernado [56] and thoroughly discussed by Yang and Berger [53], is investigated.

The Jeffreys prior,

$$\pi(\boldsymbol{\Sigma}) = |\boldsymbol{\Sigma}|^{-(p+1)/2} \tag{D2}$$

might be the most commonly used reference prior. However, care must be taken when using this prior, as it can lead to an improper posterior distribution, and it also fails to shrink the eigenvalues appropriately. However, the approach proposed by Yang and Berger [53] has proven to be able to overcome these inadequacies of the Jeffrey's prior remarkably. Note that  $\Sigma$  can be decomposed as  $\Sigma = \mathbf{O}\mathbf{D}\mathbf{O}'$ , where  $\mathbf{O}$  is orthogonal with positive entries in the first row, and  $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_p)$ , with  $d_1 \geq d_2 \geq \dots \geq d_p$ . Hence, providing the monotonically ordered  $\{d_i\}$ , the reference prior for  $(\mathbf{D}, \mathbf{O})$  is given by:

$$\pi(\mathbf{D}, \mathbf{O}) \propto \frac{1}{|\boldsymbol{\Sigma}| \prod_{i < j} (d_i - d_j)}, \tag{D3}$$

and the resulting posterior distribution is:

$$\pi(\Sigma|\mathbf{y}, \Theta, \mathbf{Q}, \lambda) \propto \frac{\exp[\text{tr}(-\frac{1}{2n}\mathbf{O}\mathbf{D}^{-1}\mathbf{O}'(\sum_{j=0}^2 \sum_{i=1}^{n_j}(\mathbf{y}_{ij} - g(\mathbf{t}|\Theta_j))(\mathbf{y}_{ij} - g(\mathbf{t}|\Theta_j))'))]}{|\Sigma|^{n/2+1}}. \quad (\text{D4})$$

Comparing the reference prior with the Jeffrey's prior, it is noted that the posterior given in equation (D3) is always proper. Also note that, since this reference prior put more mass near the region for equal eigenvalues, it can produce an estimator with better eigenstructure.

Again, it is very difficult to analytically evaluate the posterior (D3). Yang and Berger [53] suggested using a Metropolized hit-and-run sampler algorithm to obtain the integration. The detail sampling procedure at the  $k$ th iteration is given as follows:

- (1) Given the current positive-definitive matrix  $\Sigma_k$ , we set  $\mathbf{W}_k = \log \Sigma_k$ , in the sense that  $\mathbf{W}_k = \sum_{i=0}^{\infty} \frac{(\Sigma_k)^i}{i!}$ ;
- (2) Randomly generate a symmetric  $p$  by  $p$  matrix  $T$ , with elements  $t_{ij} = z_{ij}/(\sum_{l \leq m} z_{lm}^2)^{1/2}$ , where  $z_{ij} \sim i.i.d.N(0, 1)$ , for  $i \leq j$ ;
- (3) Set  $\mathbf{W}^* = \mathbf{W}_k + \nu T$  where  $\nu$  is generated from  $N(0, 1)$ ;
- (4) Update  $\mathbf{W}_k$  with an acceptance probability  $\min(1, \pi^*(\mathbf{W}^*|\mathbf{y}, \Theta, \mathbf{Q}, \lambda)/\pi^*(\Sigma_k|\mathbf{y}, \Theta, \mathbf{Q}, \lambda))$ .

## Acknowledgements

We thank four reviewers for their constructive comments on the manuscript and Dr. Jim Cheverud for providing his mouse data. This work is partially supported by Joint NSF/NIH grant DMS/NIGMS-0540745 and NNSFC grant 30671704.

## References and Notes

1. Lynch, M.; Walsh, B. *Genetics and Analysis of Quantitative Traits*; Sinauer Associates: Sunderland, MA, USA, 1998.
2. Paterson, A.H.; Lander, E.S.; Hewitt, J.D.; Peterson, S.; Lincoln, S.E.; Tanksley, S.D. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **1988**, *335*, 721-726.
3. Lander, E.S.; Botstein, D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **1989**, *121*, 185-199.
4. Zeng, Z.B. Precision mapping of quantitative trait loci. *Genetics* **1994**, *136*, 1457-1468.
5. Weller, J.I. *Quantitative Trait Loci Analysis in Animals*; CABI Publishing: London, UK, 2001.
6. Siegmund, D.; Yakir, B. *The Statistics of Gene Mapping*; Springer: New York, USA, 2007.
7. Lin, M.; Li, H.; Hou, W.; Johnson, J.; Wu, R.L. Modeling sequence-sequence interactions for drug response. *Bioinformatics* **2007**, *23*, 1251-1257.
8. Ma, C.X.; Casella, G.; Wu, R.L. Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics* **2002**, *161*, 1751-1762.

9. Wu, R.L.; Lin M. Functional mapping – how to map and study the genetic architecture of dynamic complex traits. *Nat. Rev. Genet.* **2006**, *7*, 229–237.
10. West, G.B.; Brown, J.H.; Enquist, B.J. A general model for ontogenetic growth. *Nature* **2001**, *413*, 628–631.
11. Dempster, A.P. *Elements of Continuous Multivariate Analysis*; Addison-Wesley: Reading, 1969.
12. Meng, X.L.; Rubin, D.B. Maximum likelihood via the ECM algorithm: A general framework. *Biometrika* **1993**, *80*, 267–278.
13. Carlin, B.P.; Louis, T.A., *Bayes and Empirical Bayes Methods for Data Analysis*; Chapman & Hall: New York, 1996.
14. Satagopan, J.M.; Yandell, B.S.; Newton, M.A.; Osborn, T.C. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **1996**, *144*, 805–816.
15. Sillanpää, M.J.; Arjas, E. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **1999**, *151*, 1605–1619.
16. Xu, S. Estimating polygenic effects using markers of the entire genome. *Genetics* **2003**, *163*, 789–801.
17. Yi, N.; Yandell, B.S.; Churchill, G.A.; Allison, D.B.; Eisen, E.J.; Pomp, D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **2005**, *170*, 1333–1344.
18. Zhang, M.; Montooth, K.L.; Wells, M.T.; Clark, A.G.; Zhang, D. Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* **2005**, *169*, 2305–2318.
19. Yang, R.Q.; Xu, S. Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics* **2007**, *176*, 1169–1185.
20. Zimmerman, D.L.; Núñez-Antón, V. Parametric modeling of growth curve data: An overview (with discussion). *Test* **2001**, *10*, 1–73.
21. Diggle, P.J.; Heagerty, P.; Liang, K.Y.; Zeger, S.L. *Analysis of Longitudinal Data*; Oxford University Press: Oxford, UK, 2002.
22. Wu, R.L.; Ma, C.M.; Lin, M.; Wang, Z.H.; Casella, G. Functional mapping of growth QTL using a transform-both-sides logistic model. *Biometrics* **2004**, *60*, 729–738.
23. Carrol, R.J.; Rupert, D. Power transformations when fitting theoretical models to data. *J. Am. Stat. Assoc.* **1984**, *79*, 321–328.
24. Zhao, W.; Ma, C.M.; Cheverud, J.M.; Wu, R.L. A unifying statistical model for QTL mapping of genotype-sex interaction for developmental trajectories. *Physiol. Genomics* **2004**, *19*, 218–227.
25. Evans, I.G. Bayesian estimation of parameters of multivariate normal distribution. *J. Roy. Statist. Soc. B* **1965**, *27*, 279–283.
26. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE* **1984**, *6*, 721–741.
27. Hastings, W.K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 397–409.
28. Tierney, L. Markov chain for exploring posterior distributions. *Ann. Stat.* **1994**, *22*, 1701–1762.
29. Fan, J.; Gilbels, I. *Local Polynomial Modelling and Its Applications*; Chapman & Hall: New York, NY, USA, 1996.

30. Box, G.; Tao, G. *Bayesian Inference in Statistical Analysis*; Wiley Interscience: New York, NY, USA, 1973.
31. Ritter, C.; Tanner, M.A. Facilitating the Gibbs sampler: The Gibbs stopper and Griddy-Gibbs sampler. *J. Am. Stat. Assoc.* **1992**, *87*, 861–868.
32. Geyer, C. Practical Markov chain Monte Carlo. *Stat. Sci.* **1992**, *7*, 473–483.
33. Gabriel, K.R. Ante-dependence analysis of an ordered set of variables. *Trans. Roy. Soc. Edinb-Earth Sci.* **1962**, *33*, 201–212.
34. Jaffrézic, F.; Thompson, R.; Hill, W.R. Structured antedependence models for genetic analysis of repeated measures on multiple quantitative traits. *Genet. Res.* **2003**, *82*, 55–65.
35. Stephens, M. Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods. *Ann. Stat.* **2000**, *28*, 40–74.
36. Cheverud, J.M.; Routman, E.J.; Duarte, F.A.M.; van Swinderen, B.; Cothran, K.; Perel, C. Quantitative trait loci for murine growth. *Genetics* **1996**, *142*, 1305–1319.
37. Vaughn, T.T.; Pletscher, L.S.; Peripato, A.; King-Ellison, K.; Adams, E.; Erikson, C.; Cheverud, J.M. Mapping quantitative trait loci for murine growth - A closer look at genetic architecture. *Genet. Res.* **1999**, *74*, 313–322.
38. Zhao, W.; Chen, Y.Q.; Casella, G.; Cheverud, J.M.; Wu, R.L. A non-stationary model for functional mapping of longitudinal quantitative traits. *Bioinformatics* **2005**, *21*, 2469–2477.
39. Satagopan, J.M.; Yandell, B.S.; Newton, M.A.; Osborn, T.C. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **1996**, *144*, 805–816.
40. Sillanpää, M.J.; Arjas, E. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **1999**, *151*, 1605–1619.
41. Raftery, A.E.; Lewis, S. *Bayesian Statistics, 4th ED*; Oxford University Press: Oxford, UK, 1992.
42. Malosetti Zunin, M.; Visser, R.G.F.; Celis Gamboa, B.C.; van Eeuwijk, F.A. QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. *Theor. Appl. Genet.* **2006**, *113*, 288–300.
43. Kao, C.H.; Zeng, Z.B. Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* **2002**, *160*, 1243–1261.
44. Cui, Y.H.; Wu, R.L. Mapping genome-genome epistasis: A multi-dimensional model. *Bioinformatics* **2005**, *21*, 2447–2455.
45. Green, P.J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **1995**, *82*, 711–732.
46. Green, P.J.; Hjort, L.; Richardson, S. *Trans-Dimensional Markov Chain Monte Carlo*; Oxford University Press: London, UK, 2003.
47. Brooks, S.P.; Giudici, P. Markov chain Monte Carlo convergence assessment via two-way analysis of variance. *J. Comput. Graph. Stat.* **2000**, *9*, 266–276.
48. Godsill, S.J. On the relationship between MCMC model uncertainty methods. *J. Comput. Graph. Stat.* **2002**, *10*, 230–248.
49. Dempster, A.P. *Elements of Continuous Multivariate Analysis*; Addison-Wesley: Reading, MA, USA, 1969.
50. Stein, C. *Estimation of a Covariance Matrix*; Rietz Lecture: Atlanta, Georgia, 1975.

51. Wakefield, J. The Bayesian analysis of population pharmacokinetics models. *J. Am. Stat. Assoc.* **1969**, *91*, 62–75.
52. Leonardo, T.; Hsu, J.S. Bayesian inference for a covariance matrix. *Ann. Stat.* **1993**, *21*, 1–25.
53. Yang, R.; Berger, J.O. Estimation of a covariance using the reference prior. *Ann. Stat.* **1994**, *22*, 1195–1211.
54. Everson, P.J.; Morris, C.N. Inference for multivariate normal hierarchical models. *J. Roy. Statist. Soc. B* **2000**, *62*, 399–412.
55. Pourahmadi, M. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika* **1999**, *86*, 677–690.
56. *In Bayesian Statistics* Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A. F.M., Eds.; Oxford University Press: Oxford, UK, 1992; pp. 3560.

© 2009 by the authors; licensee Molecular Diversity Preservation International, Basel, Switzerland. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).