G&I
GENOMICS & INFORMATICS

Original article

eISSN 2234-0742
Genomics Inform 2022; 20(1):e8
https://doi.org/10.5808/gi.21080

Received: December 13, 2021
Revised: February 13, 2022
Accepted: March 3, 2022

*Corresponding author:
E-mail: wchung@ssu.ac.kr

Bayesian mixed models for longitudinal genetic data: theory, concepts, and simulation studies

Wonil Chung[1,2]*, Youngkwang Cho[1]

[1]Department of Statistics and Actuarial Science, Soongsil University, Seoul 06978, Korea
[2]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA

Despite the success of recent genome-wide association studies investigating longitudinal traits, a large fraction of overall heritability remains unexplained. This suggests that some of the missing heritability may be accounted for by gene-gene and gene-time/environment interactions. In this paper, we develop a Bayesian variable selection method for longitudinal genetic data based on mixed models. The method jointly models the main effects and interactions of all candidate genetic variants and non-genetic factors and has higher statistical power than previous approaches. To account for the within-subject dependence structure, we propose a grid-based approach that models only one fixed-dimensional covariance matrix, which is thus applicable to data where subjects have different numbers of time points. We provide the theoretical basis of our Bayesian method and then illustrate its performance using data from the 1000 Genome Project with various simulation settings. Several simulation studies show that our multivariate method increases the statistical power compared to the corresponding univariate method and can detect gene-time/ environment interactions well. We further evaluate our method with different numbers of individuals, variants, and causal variants, as well as different trait-heritability, and conclude that our method performs reasonably well with various simulation settings.

Keywords: Bayesian mixed model, gene-time interaction, grid-based model, longitudinal data

Introduction

In recent years, genome-wide association studies (GWAS) for longitudinal traits (e.g., body weight or cholesterol levels) have been carried out in cohorts, where multiple measurements have been collected from each individual [1-7]. Although GWAS have successfully discovered a large number of novel genetic variants associated with these traits, the identified variants typically account for only a small proportion of overall heritability [8-10]. A presumed explanation for the "missing heritability" is that existing methods have low power to identify gene-gene and gene-time/environment interactions [11]. Since traditional methodologies are limited to the identification of variants with marginal effects using a single measurement per individual, a large amount of useful information in longitudinal data is lost and variants that interact with other variants or have time-varying effects may not be detected [12]. It is more appropriate to analyze multiple variants simultaneously, using all available measurements, for longitudinal genetic studies.

There are methodological challenges associated with the genetic analysis of longitudinal traits for multiple variants. Most complex traits are typically controlled by multiple variants that interact with each other or environmental factors. It may be exceedingly difficult to model all candidate variants with epistatic effect and gene-time/environment interactions for longitudinal traits because genetic data are generally high-dimensional relative to the number of samples. Bayesian multiple quantitative trait loci (QTL) mapping methods [13-16] have been proposed for modeling epistatic effects. Multiple QTL can be simultaneously detected by treating the number of QTL as a random variable using the reversible jump Markov-chain Monte Carlo (MCMC) method [13,14]. Alternatively, multiple QTL can be viewed as a variable selection problem [15,16]. Bayesian model selection approaches are used for identifying QTL with main and epistatic effects [17], as well as QTL that interact with other covariates [18] based on the composite model space framework. These approaches use a fixed-dimensional parameter space by setting an upper bound on the number of detectable QTL and introduce latent binary variables for deciding which variables will be included in the model. This technique reasonably reduces the model space using efficient MCMC algorithms. For multiple QTL mapping with multivariate traits, Banerjee et al. [19] extended the Bayesian variable selection method of Yi [16] via a model that allows different genetic models for different traits. This method provides a multiple QTL mapping strategy for correlated traits, but it does not account for the dependence structure among repeated measurements from each subject.

Several statistical methods have been proposed for dealing with within-subject variation. For data collected at the same time points across all individuals, the measured values at each time point can be treated as one variable. The data can then be treated as multivariate outcomes and jointly analyzed [20-23]. For data collected at different time points across some or all individuals, the measured values cannot be effectively grouped; thus, standard multivariate analysis is no longer applicable. Alternatively, mixed models are used for longitudinal data to map QTL [24]. Mixed models are flexible in modeling such unbalanced data because they allow non-constant correlations among observations. Chung and Zou [25] developed a Bayesian multiple association-mapping algorithm based on a mixed model with a built-in variable selection feature. It models multiple genes simultaneously and allows gene-gene and gene-time/environment interactions for repeatedly measured phenotypes. However, in that model, we made the strong assumption that the covariance matrix is known up to a constant. We plan to relax that assumption here.

In this paper, we develop a Bayesian variable selection method for longitudinal data where phenotypes are not measured at a fixed set of time points for all samples. It jointly models the main and pairwise interactions of all candidate genetic variants. We propose a novel grid-based approach to parsimoniously model each subject's covariance matrix as a function of a covariance matrix defined on a set of pre-selected time points where each observed time point is mapped to its two adjacent grid time points via linear interpolation. This approach thus deals only with a covariance matrix of a fixed dimension. The covariance matrix is then modeled nonparametrically using the modified Cholesky decomposition of Chen and Dunson [26], which facilitates the use of normal conjugate priors. The deviance information criterion (DIC) and the Bayesian predictive information criterion (BPIC) are proposed for the selection of an optimal number of grid points. The paper is organized as follows. In the Methods section, we introduce a novel grid-based Bayesian method for longitudinal genetic data and provide its theoretical basis. In the Results section, we show numerous simulation results using whole-genome sequencing data from the 1000 Genome Project to evaluate the performance of the proposed methods and assess the effects of sample size, number of variants, causal variants, and heritability. We conclude the paper with some discussions on the proposed methods and future research.

Methods

Genotype data

For our simulation studies, we utilized the whole-genome sequencing data from the 1000 Genome Project, which created a catalogue of common human variations using samples from people who provided open consent who declared themselves healthy. It ran between 2008 and 2015, generating a large public catalogue of human variations and genotype data. We randomly selected 400 out of 504 individuals of East Asian (EAS) ancestry from the 1000 Genome Project data (phase 3 version 5) and then removed single-nucleotide polymorphisms (SNPs) with a minor allele frequency < 5% and p(Hardy-Weinberg equilibrium) < 10^{-6}, which resulted in 6, 247, 288 SNPs.

Bayesian mixed models

For a given trait, suppose we have n individuals where individual i has phenotypes measured at n_i time points ($i = 1, ..., n$) and p SNPs. Let $N = \Sigma_{i=1}^{n} n_i$. We set the number of main effect terms equal to p, the number of SNP-SNP interaction term s to $\frac{p(p-1)}{2}$, and the number of SNP-covariate interaction terms to pq, where q is the number of covariates in the model, including time. We define $\lambda = (\lambda_1, ..., \lambda_d)^T$ as the SNP positions associated with the above genetic effects, where $d = p + \frac{p(p-1)}{2} + pq$. Each SNP can be associated with the trait through its main effect or interactions with other SNPs (epistatic effects) or covariates. We introduce latent binary variables $\gamma = (\gamma_1, ..., \gamma_d)^T$ for the selection of genetic effects to be included in $(\gamma_i = 1)$ or excluded from $(\gamma_i = 0)$ the model. The vector (γ, λ) determines the number and positions of SNPs. For the ith individual, x_{ti} denotes the $n_i \times q$ design matrix of time/environmental covariates, x_{gi} denotes the $n_i \times p$ design matrix of the p SNPs, x_{ggi} denotes the $n_i \times \frac{p(p-1)}{2}$ design matrix of the epistatic effects, and x_{gti} denotes the $n_i \times pq$ design matrix of the SNP-time/SNP-environment interactions. We define the final design matrix as $x_i = (x_{ti}, x_{gi}, x_{ggi}, x_{gti})$.

Given γ, λ, and x_i, we consider the following mixed model:

$$y_i = \mu_i + x_i \Gamma\beta + p_i v_i + e_i (i = 1, ..., n), \qquad (1)$$

where $y_i = (y_{i1}, ..., y_{in_i})^T$ is an $n_i \times 1$ phenotype vector of individual i; $\mu_i = \mu 1_{n_i}$ is an $n_i \times 1$ overall mean vector; Γ is a diagonal matrix with upper diagonal elements 1_q (i.e., the model always contains all non-genetic covariates) and lower diagonal elements Υ;$\beta = (\beta_t^T, \beta_g^T, \beta_{gg}^T, \beta_{gt}^T)^T$ is a vector of genetic effects, time/environmental effects, epistatic effects, and SNP-time/environment interactions; and e_i is an $n_i \times 1$ vector of random errors with $e_i \sim N(0, \sigma^2 I_{n_i})$. To model the correlation among repeated measurements of the same individual, we partition the observed time interval by k pre-specified grid points, $t = (t_1, ... , t_k)^T$, and define v_i as a $k \times 1$ vector of random effects at the grid time points with $v_i \sim N(0, D)$ where D is a $k \times k$ covariance matrix. Let $p_\mathrm{i} = (p_{i1}^T, ... , p_{ini}^T)^T$ and $P = diag(p_1, ... , p_n)$ where p_i is defined as follows. If all subjects have k observations measured exactly on the k grid time points, then p_i becomes an identity matrix. We apply an interpolation procedure (e.g., linear, polynomial, or spline) to any observation that does not fall on any of the k grid time points. For simplicity, we choose a linear interpolation here. When the jth measurement of individual i falls at time t, which is in between the grid points t_r and t_{r+1} $(t_r \leq t \leq t_{r+1})$, we set $p_\mathrm{ij} = (0_{(r-1)}^T, \frac{t_{r+1}-t}{t_{r+1}-t_r}, \frac{t-t_r}{t_{r+1}-t_r}, 0_{(k-r-1)}^T)$. When $t = t_r$, we get $p_\mathrm{ij} = (0_{(r-1)}^T, 1, 0_{(k-r)}^T)$. We can re-express p_{ij} as $p_{ij} = a_{ij1}e_1 + ... + a_{ijk}e_k$, where a_{ijr} is the rth element of p_{ij} and e_r $(1 \leq r \leq k)$ is a $1 \times k$ vector whose elements are all zero except the rth component, which equals 1. Note that $\Sigma_{r=1}^{k} a_{ijr} = 1$, $0 \leq a_{ij1}, ... , a_{ijk} \leq 1$ and at most two adjacent a_{ijr} values can be non-zero due to the linear interpolation we employ here.

Re-parameterized model

For Bayesian estimation of the mixed model (1), we factor D, the covariance matrix of the random effects, by employing the modified Cholesky decomposition of Chen and Dunson [26]. Let L denote a $k \times k$ lower triangular Cholesky decomposition matrix that has nonnegative diagonal elements, such that $D = LL^T$. Let $L = \Delta\Psi$, where $\Delta = \mathrm{diag}(\delta_1, ..., \delta_k)$ and Ψ is a $k \times k$ matrix with the (l, m)th element denoted by ψ_{lm}. To make Δ and Ψ identifiable, we make the following assumptions: $\delta_l \geq 0$, $\psi_{ll} = 1$ and $\psi_{lm} = 0$, for $l = 1, ..., k$; $m = l+1, ..., k$. These conditions make Δ a nonnegative $k \times k$ diagonal matrix and Ψ a lower triangular matrix with 1's in the diagonal elements. This leads to the decomposition $D = \Delta\Psi\Psi^T\Delta$, and thus we reparametrize model (1) as

$$y_i = \mu_i + x_i \Gamma\beta + p_i \Delta\Psi b_i + e_i (i = 1, ..., n), \qquad (2)$$

where $b_i = (b_{i1}, ...b_{ik})^T$ such that $\mathrm{b}_{ij} \sim N(0, 1)$ and $b_{ij} \perp b'_{ij}$ $(j \neq j')$, $j = 1, ..., k$. For later use, we define $v_i = p_i \Delta\Psi = (\mathrm{v}_{i1}, ..., \mathrm{v}_{in_i})^T$ and $\mathrm{v} = \mathrm{diag}(v_1, ..., v_n)$.

Model identifiability

Model identifiability is a property that a model must satisfy for accurate inference to be possible. A model is identifiable if it is theoretically possible to estimate the true values of the underlying parameters of the model, while a model is non-identifiable or unidentifiable if two or more parametrizations are observationally equivalent [27]. The proposed Bayesian model has an identifiability issue associated with the covariance matrix of $y = (y_1^T, ... , y_n^T)^T$, which equals $PDP^T + \sigma^2 I_N$ where $D = I_n \otimes D$. The condition is that $PDP^T + \sigma^2 I_N = P\widehat{D}P^T + \hat{\sigma}^2 I_N$ if and only if $\widehat{D} = D$ and $\hat{\sigma}^2 = \sigma^2$. This is equivalent to the system of equations $P\widetilde{D}P^T + \tilde{\sigma}^2 I_N = 0$ having no non-zero solutions for $\widetilde{D}$ and $\tilde{\sigma}^2$ when $\widetilde{D} = D - \widehat{D}$ and $\tilde{\sigma}^2 = \sigma^2 - \hat{\sigma}^2$. Let the (r, s)th element of $\widetilde{D}$ be $\tilde{d}_{r,s}$. The system of equations $P\widetilde{D}P^T + \tilde{\sigma}^2 I_N = 0$ is equivalent to the system of equations $AX = 0$, where $A = (A_1^T, ... , A_n^T)^T$ is a $[2^{-1} \Sigma_{i=1}^{n} n_i(n_i+1)] \times [2^{-1} k(k+1)+1]$ matrix whose elements are functions of the a_{ijr}s and $\mathrm{X} = (\tilde{d}_{1,1}, \tilde{d}_{1,2}, ... , \tilde{d}_{1,k}, \tilde{d}_{2,2}, ... , \tilde{d}_{k,k}, \tilde{\sigma}^2)^T$ which contains all elements of the matrix $\widetilde{D}$ and $\tilde{\sigma}^2$ (see proof of Lemma 1 in Supplementary Data 1).Therefore, the proposed Bayesian model (2) is identifiable if and only if rank $(A) = 2^{-1} k(k+1)+1$ (see proof of Theorem 1 in Supplementary Data 1).

Lemma 1 and Theorem 1 enable us to check whether a given model is identifiable. A toy example is provided below. Suppose there are 3 grid points that produce 2 time intervals. According to the theorem, the rank of A must be $\frac{1}{2}\,3(3+1)+1 = 7$ for the model to be identifiable. Suppose the phenotypes of all individuals are observed exactly on the 3 grid points. Then

$$p_i = I_3, A_1 = \dots = A_n = \begin{pmatrix} 1&0&0&0&0&0&1 \\ 0&1&0&0&0&0&0 \\ 0&0&1&0&0&0&0 \\ 0&0&0&1&0&0&0 \\ 0&0&0&0&1&0&0 \\ 0&0&0&0&0&1&1 \end{pmatrix} \text{ and } X = \begin{pmatrix} \tilde{d}_{1,1} \\ \tilde{d}_{1,2} \\ \tilde{d}_{1,3} \\ \tilde{d}_{2,2} \\ \tilde{d}_{2,3} \\ \tilde{d}_{3,3} \\ \tilde{\sigma}^2 \end{pmatrix}$$

The rank of A is $\frac{1}{2}\,3(3+1)=6$. Therefore, $PDP^T+\sigma^2 I_N$ is non-identifiable. If we have one additional individual who has one phenotype measured not on any of the grid points, the model becomes identifiable since the rank of A now increases to 7. If we do not have any additional individuals, we can avoid the identifiability issue simply by setting $\sigma^2=0$ and modeling D directly.

Prior specifications

For the random effects of the proposed Bayesian model, we employ the priors presented by Chen and Dunson [26]. Specifically, independent half normal priors are imposed on the diagonal elements of Δ and normal priors on the lower triangular elements of Ψ. For the fixed effects, we straightforwardly extend the priors presented in Yi et al. [17, 18].

Priors on γ and λ

Let $w_a = P(\gamma_a = 1)$ be the inclusion probability of the ath genetic effect. We assume that all inclusion probabilities are independent of each other and thus the prior of γ is $\Pi_{a=1}^{r} w_{\mathrm{a}}^{\gamma_a}(1-w_{\mathrm{a}})^{1-\gamma_a}$. The inclusion probability w_a is pre-determined and can vary according to whether it corresponds to a main genetic effect, SNP-SNP interaction, or SNP-covariate interaction [17]. To specify a prior on λ, we assume that the locations are again independent and uniformly distributed over all SNPs. For the number of SNPs (i.e., p), the prior distribution of genetic variant location λ is therefore given by $P(\lambda) = \Pi_{a=1}^{r} P(\lambda_{\mathrm{a}})$.

Priors on b, Δ, and Ψ

In model (2), we let the distribution of each b_{ij} independently follow a standard normal distribution. Thus, the joint prior distribution of $b = (b_1^T, \dots, b_n^T)^T$ is $P(b) \overset{d}{=} N(0, I_{nk})$. As priors for Δ and Ψ, we define two vectors $\delta = (\delta_l : l = 1, \dots, k)^T$ and $\psi = (\psi_{ml} : m = 2, \dots, k; l = 1, \dots, m\text{-}1)^T$. The prior distribution for δ is $P(\delta) = \Pi_{l=1}^{k} P(\delta_l)$ $\Pi_{l=1}^{k} N^{+}(\delta_1 | m_{l0}, s_{l0}^2)$, where $N^{+}(\delta_1 | m_{l0}, s_{l0}^2)$ is the density of a half normal distribution that is a $N(\delta_1 | m_{l0}, s_{l0}^2)$ density truncated below by zero. The prior distribution for ψ is $P(\Psi) \overset{d}{=} N(\Psi_0, R_0)$, where ψ_0 and R_0 are pre-specified hyperparameters.

Priors on β, μ, and σ^2

The prior for the ath genetic effect is a normal distribution, $P(\beta_a | \gamma_a, \sigma_\beta^2) \overset{d}{=} N(0, \gamma_a \sigma_\beta^2)$ and the prior for the variance σ_β^2 is a scaled inverse χ^2 distribution, $P(\sigma_\beta^2) \overset{d}{=} inv\text{–}\chi^2(\nu_\beta, s_\beta^2)$ whose expectation is $E(\sigma_\beta^2) = \dfrac{\nu_\beta s_\beta^2}{\nu_\beta - 2}$. The degree of freedom ν_β controls the skewness of the prior for σ_β^2 (we set $\nu_\beta = 6$) and the scale parameter s_β^2 controls the prior confidence region for the heritability of the associated genetic factor. Let V be the total phenotypic variance and V_a be the sample variance of the column of x_i associated with β_a. The heritability of the ath genetic factor, h_a, is therefore $V_{\mathrm{a}}\beta_a^2/V$. Setting $E(\sigma_\beta^2) = E(\beta_a^2)$, we have $s_\beta^2 = (\nu_\beta - 2)E(\beta_a^2)/V_\beta = (V_\beta - 2)E(h_a)V/(\nu_\beta V_a)$, with $\mathrm{E}(h_a) = 0.1$. The prior for the overall mean μ is given by $P(\mu) \overset{d}{=} N(\eta_0, \tau_0^2)$. We empirically set $\eta_0 = \bar{y} = (\frac{1}{N})\Sigma_{i=1}^{n} \Sigma_{j=1}^{n_i} y_{ij}$ and $\tau_0^2 = s_y^2 = (\frac{1}{N-1})\ \Sigma_{i=1}^{n} \Sigma_{j=1}^{n_i}(y_{ij} - \bar{y})^2$. The prior for the residual variance σ^2 is chosen as an scaled inverse χ^2 distribution, $P(\sigma^2) \overset{d}{=} inv - \chi^2(\nu_\sigma, \mathrm{s}_\sigma^2)$.

Posterior calculation and MCMC algorithm

The joint posterior distribution is proportional to the product of the likelihood and the prior distributions of all unknown parameters, which can be expressed as

$$P(\gamma, \theta | y) \propto P(y | \gamma, \theta) P(\gamma) P(\lambda) P(\beta | \gamma) P(b) P(\delta) P(\psi) P(\mu) P(\sigma^2), \tag{3}$$

where $\theta = (\lambda, \beta, b, \delta, \psi, \mu, \sigma^2)^T$. To obtain MCMC samples of all parameters, we utilize the Metropolis-Hastings and Gibbs sampling algorithms, and alternately update each unknown parameter or set of unknown parameters conditional on all the other parameters and the observed data.

For γ and λ, we use the Metropolis-Hastings algorithm within Gibbs sampler since their conditional distributions have no known distributional forms. To update those parameters, we straightforwardly extend the Metropolis-Hastings algorithm proposed by Yi et al. [18] for our Bayesian model. These algorithms are described in the Supplementary Data 1. For the other parameters, we applied the Gibbs sampling algorithm. Specifically, since b, δ, and ψ have multivariate normal or half normal priors, the full conditional distributions are easy to derive by their conjugacy properties. The full conditional posterior distributions of b, δ and ψ are $P(b | y, \gamma, \theta_{-b}) \overset{d}{=} N(b^*, \Sigma_b^*)$, $P(\delta_l | y, \gamma, \theta_{-\delta_l}) \overset{d}{=} N^{+}(\delta_l^*, {}^3\sigma_l^{*2})$, and $P(\psi | y, \gamma, \theta_{-\psi}) \overset{d}{=} N(\psi^*, \Sigma_\psi^*)$, respectively, where $\theta_{\text{-}f}$ represents all the elements of θ except f. The expressions for b^*, Σ_b^*, δ_l^*, δ_l^{*2}, ψ^*, and Σ_ψ^* are again given in Supplementary Data 1. The full conditional distributions of β, σ_β^2, μ

and σ^2 are $P(\beta_a|\gamma_a = 1, \gamma_{-a}, \theta_{-\beta a}, y) \overset{d}{=} N(\tilde{\mu}_a, \tilde{\sigma}^2_\beta)$, $P(\sigma^2_\beta|\beta_a) \overset{d}{=} Inv-\chi^2(v_\beta+1, (\beta^2_a+v_\beta s^2_\beta)/(v_\beta+1))$, $P(\mu|\gamma, \theta_{-\mu}, y) \overset{d}{=} (\mu^*, \sigma^{2}_\mu{}^*)$, and $P(\sigma^2|\gamma, \theta_{-\sigma^2}, y) \overset{d}{=} Inv-\chi^2(v_\sigma + N, \frac{v_\sigma s^2_\sigma+N\hat{\sigma}^2}{v_\sigma+N})$, respectively, where $\tilde{\mu}_a$, $\tilde{\sigma}^2_\beta$, μ^*, $\sigma^{2}_\mu{}^*$, and $\hat{\sigma}^2$ are given in Supplementary Data 1 as well.

Posterior analysis

The posterior samples can be used to approximate the posterior distribution of the parameters. MCMC samples from the initial iterations are discarded as "burn-in" and the subsequent samples are thinned by keeping every *c*th MCMC sample, where *c* is an integer, and discarding the rest. The posterior inclusion probability of each SNP can be calculated using its inclusion proportion in the MCMC samples as $P(\kappa_l|y) = \frac{1}{T}\sum_{t=1}^{T} \sum_{w=1}^{r} 1(\lambda^{(t)}_w = \kappa_l, \gamma^{(t)}_w = 1)$ where κ_l is *l*th SNP position ($l = 1, ..., h$) and T is the total number of MCMC samples. With the prior $P(\kappa_l) = \frac{p}{h}$, the Bayes factor can be calculated to quantify the evidence for inclusion of the *l*th SNP (κ_l) against exclusion of the *l*th SNP as

$$BF(\kappa_l) = \frac{P(\kappa_l | y)/P(\kappa_l)}{(1-P(\kappa_l | y))/(1-P(\kappa_l))} = \frac{(P(\kappa_l | y)}{(1-P(\kappa_l | y)} \frac{1-P(\kappa_l)}{P(\kappa_l)}. \quad (4)$$

The Bayes factor $BF(\kappa_l)$ reflects how our belief in the importance of the lth SNP changes as we move from prior knowledge to posterior information. Jeffreys [28] and Yandell et al. [29] suggest the following criteria for judging the significance of each SNP: weak support if $BF(\kappa_l)$ falls between 3 and 10; moderate support if $BF(\kappa_l)$falls between 10 and 30; and strong support if $BF(\kappa_l)$ is larger than 30.

Choice of the number of grid points

A critical issue with the proposed Bayesian model is how to choose an optimal number of grid points, k. We achieve this goal by evaluating the goodness of the predictive distributions of our Bayesian models. Spiegelhalter et al. [30] proposed the DIC as $DIC = -2E_{\gamma,\theta|y}[logP(y|\gamma, \theta)]+P_D$. The second term of the DIC, P_D, is the effective number of parameters, which is defined as $P_D = -2E_{\gamma,\theta|y}[logP(y|\gamma, \theta)]+2logP(y|\bar{\gamma}, \bar{\theta})$, where $\bar{\gamma}$ and $\bar{\theta}$ are the posterior means of γ and θ. Since $P(y_i|\gamma, \theta) \overset{d}{=} N(\mu_i + x_i\Gamma\beta, p_iDp_i^T + \sigma^2 I_{n_i})$ in model (1), the DIC is easy to compute with the MCMC samples. However, as stated by Robert and Titterington [31], the observed data are used twice to calculate P_D, and thus the predictive distribution from the DIC tends to overfit the data. To overcome the overfitting problem, Ando [32] developed the BPIC, which is defined as $BPIC = -2E_{\gamma,\theta|y}[logP(y|\gamma, \theta)]+2n\hat{b}$ where $\hat{b}$ is the asymptotic bias in the posterior mean of the expected log-likelihood. Under a certain mild regularity condition, the bias term can be approximated by $n\hat{b} \approx P_D$, resulting in the simplified $BPIC = 2E_{\gamma,\theta|y}[logP(y|\gamma, \theta)]+2P_D$. It should be noted that the penalty term of the simplified BPIC is twice that of the original DIC. We select the optimal number of grid points for our model by minimizing DIC or simplified BPIC scores.

Implementation in gridbayes

The proposed grid-based Bayesian mixed models have been implemented in an R package named gridbayes [33], which is built on top of the R packages, qtl [34] and qtlbim [29]. The MCMC algorithm in C and the data manipulation procedure in R were modified for longitudinal analysis. The gridbayes package employs both DIC and simplified BPIC scores to select the optimal number of grid points. The software package and the source code are available for download at https://github.com/wonilchung/GridBayes.

Results

Simulation I

To evaluate the performance of the proposed method, we conducted the following simulations. We first used 400 individuals and 1, 000 SNPs from the 1000 Genome Project data. The number of measurements for each individual ranged from 3 to 7 and the total number of observations was set to 2, 000. Six different setups (Setups 1–6) were considered. We simulated the datasets containing 10 causal SNPs, which are randomly selected (i.e., the proportion of causal SNPs = 1%) with only main effects (Setup 1). For individual *i*, the phenotype values were generated from the model: $y_i = c_{g1} \cdot (\sum_{a=1}^{10} x_{ia} + t_i) + p_i v_i + e_i$, where x_{ia} (a = 1, ... , 10) were genotype values of the causal SNPs, c_{g1} is used to set trait-heritability to 40%, $t_i = (t_{i1}, ... , t_{in_i})^T$ were the time covariates generated from the uniform distribution $U[0, 1]$ and then standardized to have mean 0 and variance 1, and $e_i \sim N(0, \sigma^2 I_{ni})$. We set $\sigma^2 = 1$. The true number of grid points was set to 3 (i.e., true $k = 3$), and p_i was calculated from t_i by the linear interpolation as we described in the Methods section. We set $\delta = (\delta_1, \delta_2, \delta_3) = (1, 1.2, 0.8)$ and $\psi = (\psi_{21}, \psi_{31}, \psi_{32}) = (0.6, 0.4, 0.6)$. That is, $v_i \sim N(0, D)$ with $diag(D) = (1, 1.96, 0.97)$ and the lower triangle elements $(d_{21}, d_{31}, d_{32}) = (0.72, 0.32, 0.81)$. The prior distributions for the elements in δ were independent $N^+(0, 30)$ and the prior distributions for the elements in ψ were independent $N(0, 0.5)$. For each simulated dataset, the MCMC algorithm ran for 4 × 10^5 iterations after discarding the first 1, 000 burn-in iterations. The remaining samples were further thinned for every 40 iterations, yielding 10^4 MCMC samples for the posterior analysis.

To further investigate the Bayesian mixed model, we analyzed additional datasets containing two SNP-SNP interactions (Setup 2), five SNP-SNP interactions (Setup 3), two SNP-time interactions (Setup 4), five SNP-time interactions (Setup 5), or ten SNP-time interactions (Setup 6). Specifically, we simulated data according to the following models:
$y_i = c_{g2}\cdot(\sum_{a=1}^{6} x_{ia} + x_{i7}\cdot x_{i8} + x_{i9}\cdot x_{i10} + t_i) + p_i v_i + e_\mathrm{i}$ for Setup 2,
$y_i = c_{g3}\cdot(x_{i1}\cdot x_{i2} + x_{i3}\cdot x_{i4} + x_{i5}\cdot x_{i6} + x_{i7}\cdot x_{i8} + x_{i9}\cdot x_{i10}+ t_i) + p_i v_i + e_\mathrm{i}$ for Setup 3,
$y_i = c_{g4}\cdot(\sum_{a=1}^{8} x_{ia} + \sum_{a=9}^{10} x_{ia}\cdot t_i) + p_i v_i + e_\mathrm{i}$ for Setup 4,
$y_i = c_{g5}\cdot(\sum_{a=1}^{5} x_{ia} + \sum_{a=6}^{10} x_{ia}\cdot t_i) + p_i v_i + e_\mathrm{i}$ for Setup 5 and
$y_i = c_{g6}\cdot(\sum_{a=1}^{10} x_{ia}\cdot t_i) + p_i v_i + e_\mathrm{i}$ for Setup 6. In our simulations, c_{gj} (j = 1, ... , 6) were varied to ensure that trait-heritability to 40%. To display time-dependent SNP effects for Setups 4 and 5, we compared the time-dependent curves of averaged phenotype values for three different genotypes (0, 1, 2) at the first causal SNP (with no SNP-time interaction) and 10th one (with SNP-time interaction). Supplementary Fig. 1 clearly showed that the first causal SNP had only a main effect, but the 10th causal SNP interacted with time. We first conducted gridbayes [33] using all the data. For model comparisons, we then conducted qtlbim [29] in two ways: once on a subset of each simulated data, where only one measurement from each subject was randomly selected, and once with all the data by (incorrectly) assuming that all the measurements were independent. We named the two qtlbim analyses "qtlbim-sub" and "qtlbim-all, " respectively.

The one-dimensional genome-wide profiles of $2log(BF)$ for the combined main, epistatic effects, and SNP-time interactions of each SNP under the six setups were presented in Figs. 1 and 2. The dashed vertical lines indicate the locations of the 10 causal SNPs. The gridbayes analysis of all the data and qtlbim-sub detected the causal SNPs reasonably well, but gridbayes clearly outperformed qtlbim-sub in general. The qtlbim-all method occasionally identified the true causal SNPs, but it produced far more false-positive findings than gridbayes and qtlbim-sub.

To evaluate the performance of our Bayesian model, we further calculated the receiver operating characteristic (ROC) curves. For each setup, we conducted 100 simulations. The ROC curves with a false-positive rate less than 0.2 are presented in Fig. 3. The solid lines represent the results of gridbayes, the dot-dashed lines correspond to qtlbim-sub and the results from qtlbim-all are summarized by the long-dashed lines. The ROC curves demonstrated that gridbayes with all measurements appeared to outperform the qtlbim analyses in terms of improved true positive rates.

To diagnose the convergence of the MCMC samples, we conducted 10 parallel chains with different, over-dispersed initial values with respect to the true posterior distribution. Using 10^4 iterations, Geweke's Z-scores [35] for each chain based on the first 10% and last 50% of the samples indicated good convergence of all parameters. Based on 10 chains, Gelman and Rubin's potential scale reduction factors [36] were calculated, and the upper limits were less than 1.01 for all parameters. Supplementary Fig. 2 presents the trace plots of σ^2, δ_1, δ_2, δ_3, ψ_{21}, ψ_{31} and ψ_{32} for each setup, showing that all chains moved around the true values for all parameters, indicating good convergence. We plotted the marginal posterior and prior densities of all parameters based on 10, 000 random draws (Supplementary Fig. 3). It appeared that the random draws were approximately normal, with means close to the simulated values. Supplementary Fig. 4 displays the 95% highest posterior density (HPD) intervals for σ^2, δ_1, δ_2, δ_3, ψ_{21}, ψ_{31} and ψ_{32} for each setup. Most of the 95% HPD intervals contained the corresponding true values. Table 1 summarizes the posterior estimates of all parameters. The posterior means and medians were close to the true values and all the 95% HPD intervals contained the true values, demonstrating the good performance of our algorithm.

Simulation II

We conducted another simulation to estimate the number of true grid points using the DIC [30] and simplified BPIC [32,37]. The settings were almost the same as those in the previous simulations, except that the true number of grid points now varied from 2 to 4 (i.e., true k = 2, 3, 4). We simulated 100 datasets with 400 individuals and 1, 000 SNPs containing 10 causal SNPs (i.e., the proportion of causal SNPs = 1%) with only main effects. The causal SNPs were randomly assigned. The trait-heritability was set to 40%. The phenotype values were generated from the model:

$y_\mathrm{i} = c_{g1}\cdot(\sum_{a=1}^{10} x_{ia}\cdot t_i) + p_i v_i + e_\mathrm{i}$, where x_{ia} (a = 1, ..., 10) are genotypes of the causal SNPs and $t_i = (t_{i1}, ..., t_{in_i})^T$ are the time points of the ith individual. We set $(\delta_1, \delta_2, \delta_3, \delta_4) = (1, 1.2, 0.8, 0.7)$ and $(\psi_{21}, \psi_{31}, \psi_{32}, \psi_{41}, \psi_{42}, \psi_{43}) = (0.6, 0.4, 0.6, 0.2, 0.4, 0.6)$. Table 2 shows the average DIC, simplified BPIC scores over 100 simulations, and the proportion of times that the number of true grid points was correctly selected. All average DIC and average BPIC scores achieved the minimums at the true grid point number, and the percentages correctly selecting the true number of true grid points were 79%, 91%, and 100% for setups with 2, 3, and 4 true grid points using the DIC, and 94%, 98%, and 93% using the simplified BPIC. This illustrated the usefulness of the DIC and simplified BPIC in selecting the true number of grid points.

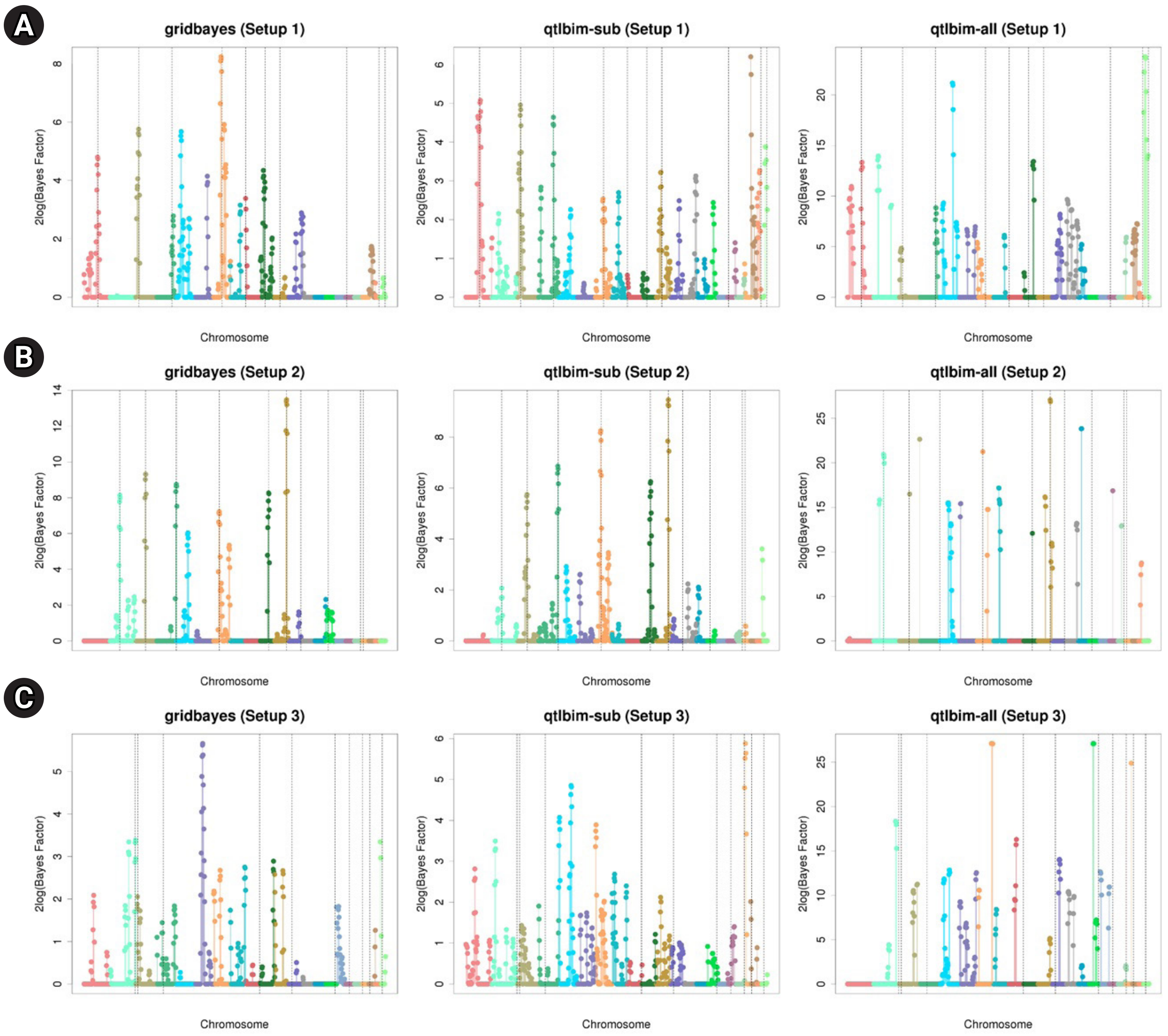

Fig. 1. Genome-wide profiles of $2log(BF)$ for all combined effects using gridbayes with all time points, qtlbim with one randomly-selected time point (qtlbim-sub) and qtlbim with all time points (qtlbim-all) for Setups 1 (A), 2 (B), and 3 (C).

Simulation III

For a more detailed evaluation of our Bayesian method, we conducted the following simulations with 100 replications for each scenario. We first considered 400 individuals with three to seven time points, resulting in 2,000 observations, and decreased the sample size from 400 to 100 to assess the effect of sample size in ROC curves (Fig. 4A). The simulation data contained 1,000 SNPs with 1% causal SNPs (i.e., 10 causal SNPs) with only main effects. The trait values were generated from the model: $y_i = c_{g1}\cdot(\sum_{a=1}^{10} x_{ia}\cdot t_i) + p_i v_i + e_i$, where x_{ia} are genotypes of the causal SNPs and $t_i = (t_{i1}, ..., t_{in_i})^T$ are the time points of the ith individual. As in the previous simulations, we set the number of grid points to $k = 3$ and $\sigma^2 = 1$, $\delta = (\delta_1, \delta_2, \delta_3) = (1, 1.2, 0.8)$, $\psi = (\psi_{21}, \psi_{31}, \psi_{32}) = (0.6, 0.4, 0.6)$. The trait-heritability was set to 40%. As the sample size decreased from 400 to 100, the true positive rates decreased in ROC curves, indicating that including more samples increased the true positive rates with fixed false positive rates. Next, we evaluated the effect of the number of SNPs (Fig. 4B). The simulation data were

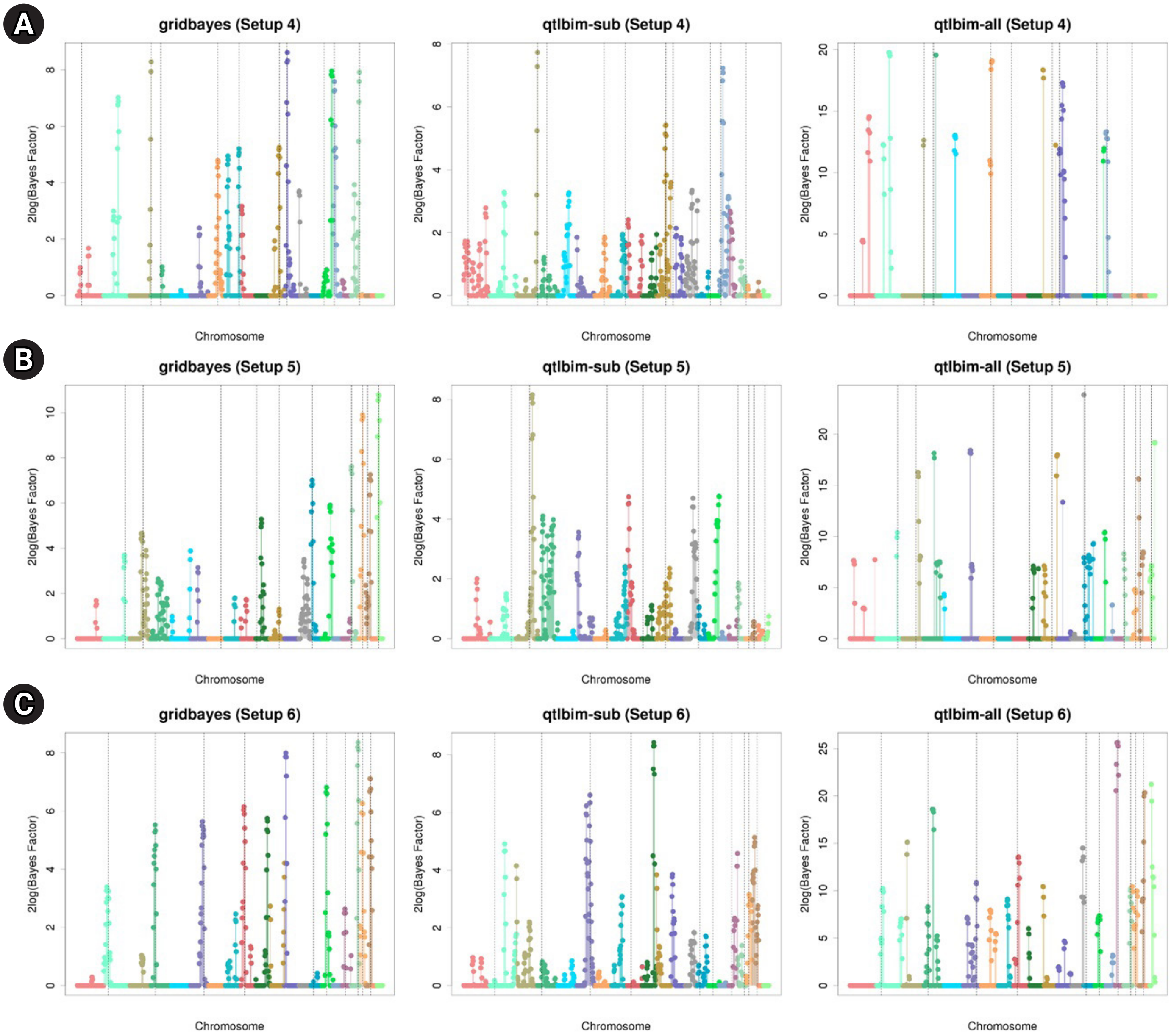

Fig. 2. Genome-wide profiles of 2*log*(*BF*) for all combined effects using gridbayes with all time points, qtlbim with one randomly-selected time point (qtlbim-sub) and qtlbim with all time points (qtlbim-all) for Setups 4 (A), 5 (B), and 6 (C).

generated with 400 individuals, 1% causal SNPs, and 40% trait-heritability. As the number of SNPs increased from 1,000 to 5,000 (i.e., the corresponding number of causal SNPs increased from 10 to 50), the true positive rates decreased, meaning that the inclusion of more SNPs deceased the true positive rates. We then examined the effect of the proportion of causal SNPs (Fig. 4C). The sample size and number of SNPs were fixed to 400 and 1,000, and the trait-heritability was set to 40%. The true positive rates decreased as the proportion of causal SNPs increased from 1% to 5% (i.e., the corresponding number of causal SNPs increased from 10 to 50) because per-SNP heritability—or the average proportion of phenotypic variation explained by a single SNP—decreased as the proportion of causal SNPs increased while keeping trait-heritability constant. Lastly, to demonstrate the effect of trait-heritability, we considered a setting where the sample size, number of SNPs, and proportion of causal SNPs were 400, 1,000, and 1%, respectively. We then changed trait-heritability from 40% to 10% in Fig. 4D. The true-positive rates decreased as trait-heritability decreased, showing

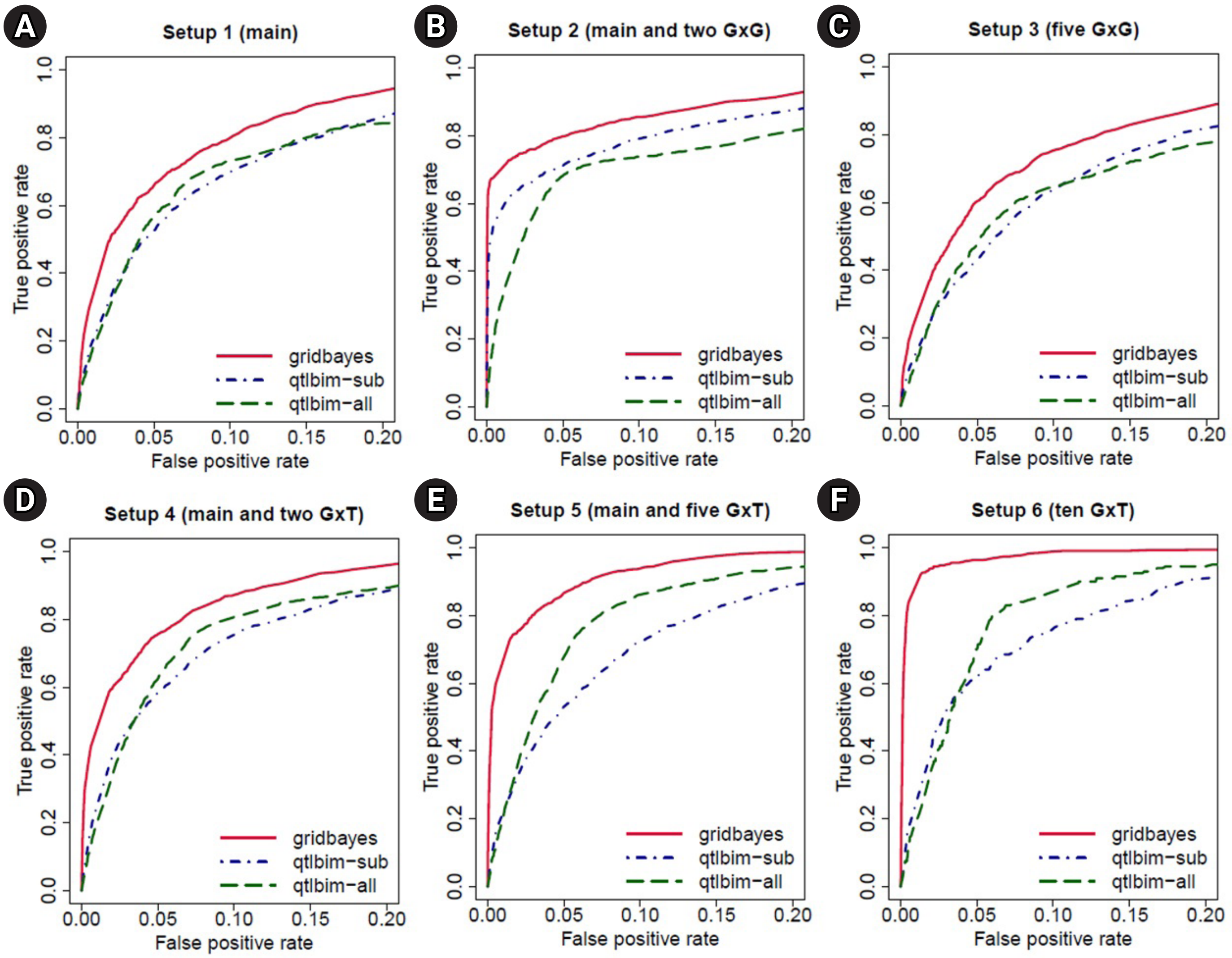

Fig. 3. Receiving operating characteristic curve analyses in the simulation study for Setup 1 (A), 2 (B), 3 (C), 4 (D), 5 (E) and 6 (F). The red solid lines represent the results of gridbayes, the blue dot-dashed lines correspond to qtlbim-sub and the green long-dashed lines display the results from qtlbim-all. gridbayes: grid-based Bayesian mixed models with all the data; qtlbim-sub qtlbim with a subset of each simulated data where only one measurement from each subject was randomly selected; qtlbim-all: qtlbim with all the data by (incorrectly) assuming that all the measurements were independent. G, gene; T, time.

that larger heritability increased the true positive rates. Supplementary Tables 1, 2, 3, and 4 summarize the posterior means, medians, standard deviations and 95% HPD intervals of all parameters in the simulations for sample size, number of SNPs, proportion of causal SNPs, and heritability, respectively. The posterior means and medians were close to the true values, and all the 95% HPD intervals contained the true values, indicating that our Bayesian method performed well.

Supplementary Table 5 showed the average DIC and simplified BPIC scores over 100 replications for all simulations. Table 3 summarizes the simulation settings for all simulation setups based on genetic effect terms, the number of grid points, sample size, number of observations, number of SNPs, number of causal SNPs, and trait-heritability.

Discussion

We developed a grid-based Bayesian mixed model for longitudinal genetic data with a built-in variable selection feature. The proposed Bayesian method modeled multiple candidate SNPs simultaneously and allowed SNP-SNP and SNP-time interactions, which enabled us to identify SNPs with time-varying effects. Such

Table 1. Posterior means, medians, standard deviations, and 95% HPD intervals of the parameters for random errors and random effects in the simulation study

Setup	Par	True	Mean	Med	SD	95% HPD	Setup	Par	True	Mean	Med	SD	95% HPD
1	σ^2	1	1.01	1.01	0.04	0.92 to 1.09	2	σ^2	1	1.01	1.01	0.04	0.93 to 1.10
	δ_1	1	0.98	0.98	0.11	0.77 to 1.19		δ_1	1	0.98	0.98	0.10	0.78 to 1.19
	δ_2	1.2	1.19	1.19	0.13	0.92 to 1.43		δ_2	1.2	1.19	1.19	0.13	0.92 to 1.43
	δ_3	0.8	0.76	0.76	0.13	0.5 to 1.00		δ_3	0.8	0.72	0.73	0.13	0.48 to 0.97
	ψ_{21}	0.6	0.65	0.63	0.21	0.31 to 1.13		ψ_{21}	0.6	0.67	0.64	0.21	0.33 to 1.14
	ψ_{31}	0.4	0.52	0.51	0.24	0.09 to 1.04		ψ_{31}	0.4	0.51	0.49	0.24	0.07 to 1.01
	ψ_{32}	0.6	0.56	0.53	0.27	0.11 to 1.18		ψ_{32}	0.6	0.67	0.64	0.29	0.20 to 1.34
3	σ^2	1	1.00	1.00	0.04	0.92 to 1.08	4	σ^2	1	1.00	1.00	0.04	0.92 to 1.09
	δ_1	1	1.06	1.06	0.10	0.85 to 1.26		δ_1	1	0.99	0.99	0.10	0.78 to 1.19
	δ_2	1.2	1.20	1.20	0.13	0.94 to 1.44		δ_2	1.2	1.18	1.19	0.13	0.92 to 1.42
	δ_3	0.8	0.74	0.74	0.12	0.49 to 0.98		δ_3	0.8	0.74	0.74	0.13	0.48 to 0.98
	ψ_{21}	0.6	0.69	0.67	0.20	0.36 to 1.14		ψ_{21}	0.6	0.62	0.60	0.20	0.29 to 1.07
	ψ_{31}	0.4	0.65	0.64	0.24	0.22 to 1.17		ψ_{31}	0.4	0.47	0.45	0.24	0.03 to 0.96
	ψ_{32}	0.6	0.65	0.62	0.27	0.19 to 1.27		ψ_{32}	0.6	0.65	0.61	0.29	0.18 to 1.31
5	σ^2	1	1.00	1.00	0.04	0.92 to 1.09	6	σ^2	1	1.00	1.00	0.04	0.92 to 1.09
	δ_1	1	0.98	0.98	0.10	0.77 to 1.18		δ_1	1	0.96	0.96	0.11	0.75 to 1.17
	δ_2	1.2	1.20	1.20	0.13	0.93 to 1.43		δ_2	1.2	1.18	1.19	0.13	0.92 to 1.41
	δ_3	0.8	0.72	0.72	0.13	0.47 to 0.97		δ_3	0.8	0.75	0.75	0.13	0.48 to 1.01
	ψ_{21}	0.6	0.63	0.60	0.20	0.29 to 1.09		ψ_{21}	0.6	0.61	0.58	0.20	0.27 to 1.07
	ψ_{31}	0.4	0.46	0.45	0.25	0.01 to 0.99		ψ_{31}	0.4	0.32	0.31	0.24	–0.14 to 0.81
	ψ_{32}	0.6	0.65	0.61	0.29	0.18 to 1.31		ψ_{32}	0.6	0.67	0.63	0.30	0.19 to 1.35

HPD, highest posterior density; Par, parameters; True, true values of parameters; Med, median; SD, standard deviation.

Table 2. Average DIC scores and simplified BPIC scores over 100 replications and the proportion selecting the model with the correct number of grid points using the proposed Bayesian model

True k	k	Avg DIC	#Sel (%)	Avg Sim BPIC	#Sel (%)	Avg PD
2	2	6,614.47	79	6,681.78	94	67.32
	3	6,617.97	15	6,690.51	6	72.54
	4	6,623.02	6	6,700.99	0	77.98
3	2	6,745.26	0	6,812.49	0	67.23
	3	6,696.53	91	6,769.83	98	73.30
	4	6,707.12	9	6,785.14	2	78.02
4	2	6,745.07	0	6,814.76	0	69.70
	3	6,718.66	0	6,792.75	7	74.09
	4	6,695.41	100	6,775.37	93	79.96

DIC, deviance information criterion; BPIC, Bayesian predictive information criterion; Avg DIC, average deviance information criterion scores over 100 replications; #Sel (%), proportion selecting the model with the correct number of grid points; Avg Sim BPIC, average simplified Bayesian predictive information criterion scores over 100 replications; Avg PD, average P_D.

SNPs are of great scientific and medical interest. In addition, we proposed a new grid-based method to model the covariance structure nonparametrically. Not only is the proposed method parsimonious in estimating the covariance matrix, but also by employing a reasonable number of grid-points, it can flexibly approximate any type of covariance structure. The number of grid points was preset, but DIC and simplified BPIC can be used to select the optimal number.

The simulation studies showed that the proposed Bayesian method using all time points outperformed the ordinary Bayesian method with one or all time points included. As expected, the proposed method that utilized the full data was more powerful than the cor-

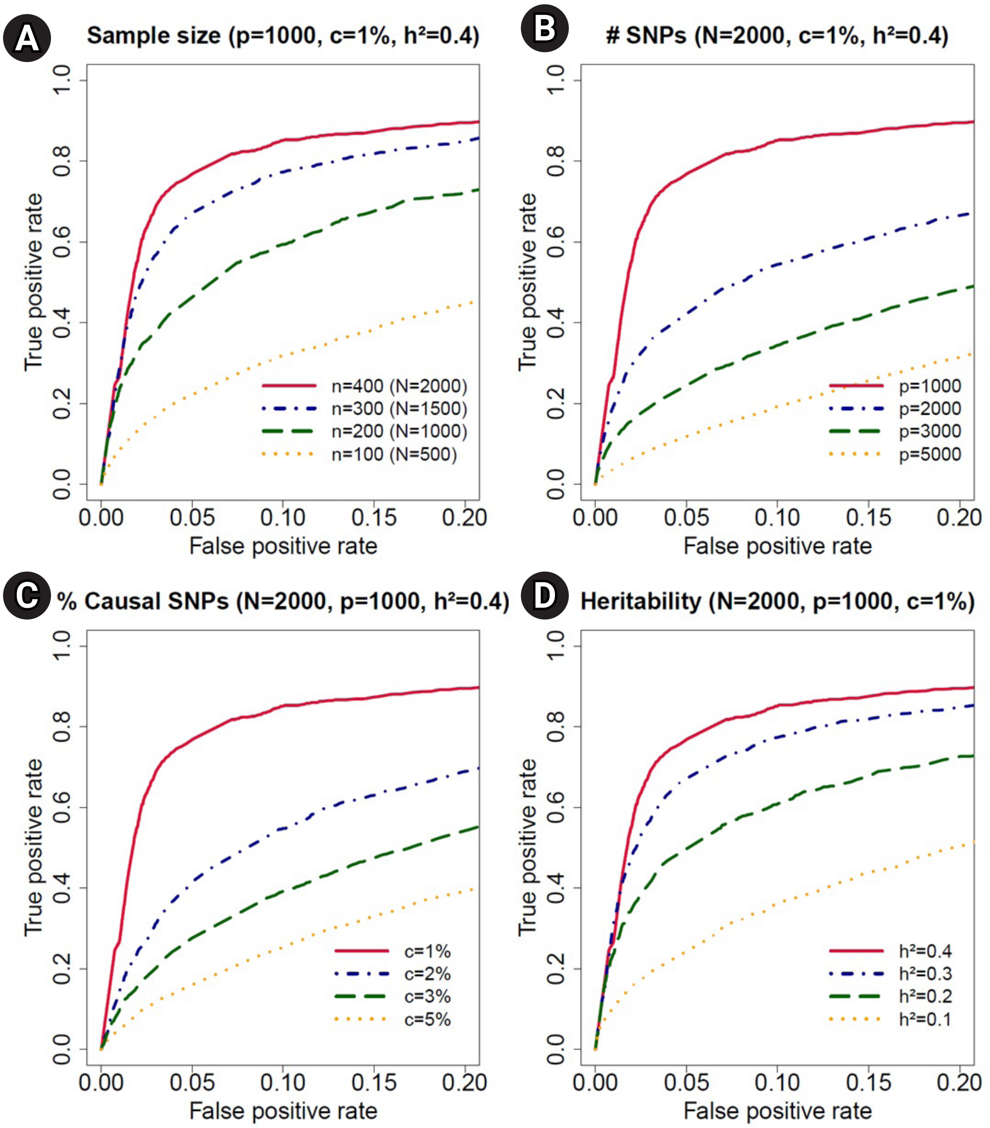

Fig. 4. Receiving operating characteristic curve analyses in the simulation study for sample size, number of single-nucleotide polymorphisms (SNPs), proportion of causal SNPs, and heritability. (A) We decreased the sample size from n = 400 (total number of observations, N = 2, 000) to n = 100 (N = 1, 000) for accessing the effect of sample size in receiver operating characteristic curves. The simulation data contained p = 1, 000 SNPs, c = 1% causal SNPs and h^2 = 40% trait-heritability. (B) We increased the number of SNPs from p = 1, 000 to p = 5, 000 to evaluate the effect of number of SNPs. The simulation data contained N = 2, 000 observations, c = 1% causal SNPs, and h^2 = 40% trait-heritability. (C) We increased the proportion of causal SNPs from c = 1% to 5%. The simulation data contained N = 2, 000 observations, p = 1, 000 SNPs, and h^2 = 40% trait-heritability. (D) We decreased the trait-heritability from h^2 = 40% to h^2 = 10%. The simulation data contained N = 2, 000 observations, p = 1, 000 SNPs, and c = 1% causal SNPs.

responding univariate analysis method that only used a subset of the data. Furthermore, the proposed Bayesian method performed better than the ordinary Bayesian method because our method modeled the within-subject correlation. Further simulation studies showed that statistical power increased as the data had more samples, a smaller number of SNPs, a lower proportion of causal SNPs, and larger trait-heritability. For our simulation studies, we utilized data from the 1000 Genome Project. With only 400 independent samples of EAS ancestry, we restricted out analysis with up to 5, 000 SNPs. With a sufficient sample size, our method can be applied to all available SNPs. We are currently developing a parallel computing algorithm based on the message passing interface to execute multiple groups of SNPs simultaneously. This will make it feasible to apply our method to large-sample GWAS data.

Table 3. Simulation settings for all simulation setups in the Results section based on genetic effect terms, the number of grid points (k), sample size (n), number of observations (N), number of SNPs (p), number of causal SNPs (c), and trait-heritability (h^2)

Simulation	Setup	Genetic effect terms	k	n	N	p	c (%)	h^2 (%)
I	1	Only main effects	3	400	2,000	,000	1	40
	2	Two SNP-SNP interactions	3	400	2,000	1,000	1	40
	3	Five SNP-SNP interactions	3	400	2,000	1,000	1	40
	4	Two SNP-time interactions	3	400	2,000	1,000	1	40
	5	Five SNP-time interactions	3	400	2,000	1,000	1	40
	6	Ten SNP-time interactions	3	400	2,000	1,000	1	40
II	1	Only main effects	2	400	2,000	1,000	1	40
	2	Only main effects	3	400	2,000	1,000	1	40
	3	Only main effects	4	400	2,000	1,000	1	40
III (a)	1	Only main effects	3	300	1,500	1,000	1	40
	2	Only main effects	3	200	1,000	1,000	1	40
	3	Only main effects	3	100	500	1,000	1	40
III (b)	1	Only main effects	3	400	2,000	2,000	1	40
	2	Only main effects	3	400	2,000	3,000	1	40
	3	Only main effects	3	400	2,000	5,000	1	40
III (c)	1	Only main effects	3	400	2,000	1,000	2	40
	2	Only main effects	3	400	2,000	1,000	3	40
	3	Only main effects	3	400	2,000	1,000	5	40
III (d)	1	Only main effects	3	400	2,000	1,000	1	40
	2	Only main effects	3	400	2,000	1,000	1	30
	3	Only main effects	3	400	2,000	1,000	1	20

SNP, single nucleotide polymorphism.

Another important issue to mention is Bayesian model identifiability. In the Bayesian community, there is a wide diversity of views on the identifiability issue. Lindley [38] remarked that non-identifiability causes no real difficulty in Bayesian approaches. Poirier [39] and Eberly and Carlin [40] argued that a Bayesian analysis of a non-identifiable model is always possible if priors on all of the parameters are proper, since proper priors yield proper posterior distributions, and hence every parameter can be well-estimated. However, if the priors imposed on any non-identifiable model are not proper, or too close to being improper, ill-behaved posterior distributions may be generated such that the trajectory of the parameters can drift to extreme values, as demonstrated by Gelfand and Sahu [41]. In this paper, we investigated the identifiability of our Bayesian model, which motivated us to utilize only proper priors (see the Methods section). Non-identifiability occurred when the number of the grid points equaled the number of observed time points (see Supplementary Data 1), but we found that the posterior distribution behaved well due to the proper priors employed.

ORCID

Wonil Chung: https://orcid.org/0000-0002-5766-6247
Youngkwang Cho: https://orcid.org/0000-0002-7980-4145

Authors' Contribution

Conceptualization: WC. Data curation: WC, YC. Formal analysis: WC, YC. Funding acquisition: WC. Methodology: WC. Writing - original draft: WC, YC. Writing - review & editing: WC.

Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (2020R1C1C1A01012657) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1A6A1A10044154). This work was supported by the Soongsil University Research Fund.

Supplementary Materials

Supplementary data can be found with this article online at http://www.genominfo.org.

References

1. Sabatti C, Service SK, Hartikainen AL, Pouta A, Ripatti S, Brodsky J, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 2009;41:35-46.
2. Mei H, Chen W, Jiang F, He J, Srinivasan S, Smith EN, et al. Longitudinal replication studies of GWAS risk SNPs influencing body mass index over the course of childhood and adulthood. PLoS One 2012;7:e31470.
3. Furlotte NA, Eskin E, Eyheramendy S. Genome-wide association mapping with longitudinal data. Genet Epidemiol 2012;36:463-471.
4. Das K, Li J, Fu G, Wang Z, Li R, Wu R. Dynamic semiparametric Bayesian models for genetic mapping of complex trait with irregular longitudinal data. Stat Med 2013;32:509-523.
5. Couto Alves A, De Silva NM, Karhunen V, Sovio U, Das S, Taal HR, et al. GWAS on longitudinal growth traits reveals different genetic factors influencing infant, child, and adult BMI. Sci Adv 2019;5:eaaw3095.
6. Gouveia MH, Bentley AR, Leonard H, Meeks KA, Ekoru K, Chen G, et al. Trans-ethnic meta-analysis identifies new loci associated with longitudinal blood pressure traits. Sci Rep 2021;11:4075.
7. Chung W. Statistical models and computational tools for predicting complex traits and diseases. Genomics Inform 2021;19:e36.
8. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-753.
9. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet 2012;90:7-24.
10. Chung W, Chen J, Turman C, Lindstrom S, Zhu Z, Loh PR, et al. Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. Nat Commun 2019;10:569.
11. Clarke AJ, Cooper DN. GWAS: heritability missing in action? Eur J Hum Genet 2010;18:859-861.
12. Smith EN, Chen W, Kahonen M, Kettunen J, Lehtimaki T, Peltonen L, et al. Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa heart study. PLoS Genet 2010;6:e1001094.
13. Satagopan JM, Yandell BS, Newton MA, Osborn TC. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 1996;144:805-816.
14. Yi N, Xu S. Mapping quantitative trait loci with epistatic effects. Genet Res 2002;79:185-198.
15. Yi N, George V, Allison DB. Stochastic search variable selection for identifying multiple quantitative trait loci. Genetics 2003; 164:1129-1138.
16. Yi N. A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. Genetics 2004;167:967-975.
17. Yi N, Yandell BS, Churchill GA, Allison DB, Eisen EJ, Pomp D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. Genetics 2005;170:1333-1344.
18. Yi N, Shriner D, Banerjee S, Mehta T, Pomp D, Yandell BS. An efficient Bayesian model selection approach for interacting quantitative trait loci models with many effects. Genetics 2007;176: 1865-1877.
19. Banerjee S, Yandell BS, Yi N. Bayesian quantitative trait loci mapping for multiple traits. Genetics 2008;179:2275-2289.
20. Wu WR, Li WM, Tang DZ, Lu HR, Worland AJ. Time-related mapping of quantitative trait loci underlying tiller number in rice. Genetics 1999;151:297-303.
21. Wu W, Zhou Y, Li W, Mao D, Chen Q. Mapping of quantitative trait loci based on growth models. Theor Appl Genet 2002;105:1043-1049.
22. Ma CX, Casella G, Wu R. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics 2002;161:1751-1762.
23. Yap JS, Fan J, Wu R. Nonparametric modeling of longitudinal covariance structure in functional mapping of quantitative trait loci. Biometrics 2009;65:1068-1077.
24. Yang R, Tian Q, Xu S. Mapping quantitative trait loci for longitudinal traits in line crosses. Genetics 2006;173:2339-2356.
25. Chung W, Zou F. Mixed-effects models for GAW18 longitudinal blood pressure data. BMC Proc 2014;8(Suppl 1):S87.
26. Chen Z, Dunson DB. Random effects selection in linear mixed models. Biometrics 2003;59:762-769.
27. Lehmann EL, Casella G. Theory of Point Estimation. New York: Springer, 2006.
28. Jeffreys H. Theory of Probability. 3rd ed. Oxford: Clarendon, 1961.
29. Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R,

Moon JY, et al. R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses. Bioinformatics 2007;23:641-643.

30. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. J R Stat Soc Series B Stat Methodol 2002;64:583-639.

31. Robert CP, Titterington DM. Discussion of a paper by D. J. Spiegelhalter et al. J. R. Stat. Soc. Ser. B 2002;64:621-622.

32. Ando T. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. Biometrika 2007;94:443-458.

33. Chung W. Bayesian Parametric and Nonparametric Methods for Multiple QTL Mapping and SNP-Set Analysis. Chapel Hill: University of North Carolina at Chapel Hill, 2013.

34. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics 2003;19:889-890.

35. Geweke JF. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Minneapolis: Federal Reserve Bank of Minneapolis, 1991.

36. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci 1992;7:457-472.

37. Ando T. Predictive Bayesian model selection. Am J Math Manag Sci 2011;31:13-38.

38. Lindley DV. Bayesian Statistics: A Review. Philadelphia: Society for Industrial and Applied Mathematics, 1972.

39. Poirier DJ. Revising beliefs in nonidentified models. Econ Theor 1998;14:483-509.

40. Eberly LE, Carlin BP. Identifiability and convergence issues for Markov chain Monte Carlo fitting of spatial models. Stat Med 2000;19:2279-2294.

41. Gelfand AE, Sahu SK. Identifiability, improper priors, and Gibbs sampling for generalized linear models. J Am Stat Assoc 1999; 94:247-253.