The Annals of Applied Statistics
2015, Vol. 9, No. 2, 640–664
DOI: 10.1214/15-AOAS808

BAYESIAN GROUP LASSO FOR NONPARAMETRIC VARYING-COEFFICIENT MODELS WITH APPLICATION TO FUNCTIONAL GENOME-WIDE ASSOCIATION STUDIES

BY JIAHAN LI[*], ZHONG WANG[†,1], RUNZE LI[‡,2] AND RONGLING WU[‡,†,3]

University of Notre Dame[*], Beijing Forestry University[†] and Pennsylvania State University[‡]

Although genome-wide association studies (GWAS) have proven powerful for comprehending the genetic architecture of complex traits, they are challenged by a high dimension of single-nucleotide polymorphisms (SNPs) as predictors, the presence of complex environmental factors, and longitudinal or functional natures of many complex traits or diseases. To address these challenges, we propose a high-dimensional varying-coefficient model for incorporating functional aspects of phenotypic traits into GWAS to formulate a so-called functional GWAS or *f*GWAS. The Bayesian group lasso and the associated MCMC algorithms are developed to identify significant SNPs and estimate how they affect longitudinal traits through time-varying genetic actions. The model is generalized to analyze the genetic control of complex traits using subject-specific sparse longitudinal data. The statistical properties of the new model are investigated through simulation studies. We use the new model to analyze a real GWAS data set from the Framingham Heart Study, leading to the identification of several significant SNPs associated with age-specific changes of body mass index. The *f*GWAS model, equipped with the Bayesian group lasso, will provide a useful tool for genetic and developmental analysis of complex traits or diseases.

1. Introduction. Phenotypic traits of paramount importance to agriculture and human health are quantitatively inherited, involving an unknown (usually very high) number of genes and undergoing a series of developmental pathways and events [Lynch and Walsh (1998); Wu and Lin (2006)]. These complexities have made the genetic analysis of quantitative traits one of the most difficult tasks in biological sciences. Recently emerging genome-wide association studies (GWAS) have provided a great promise to systematically characterize the genetic control of complex traits and have been increasingly instrumental for the identification of significant genetic variants that control phenotypic variation [Shuldiner et al. (2009); Takeuchi et al. (2009); Teichert et al. (2009); Yang et al. (2010)]. In human genetics, these results have started to gain a growing body of novel findings with

Received December 2012; revised January 2015.
[1]Supported in part by NSFC Grant 31470675.
[2]Supported in part by NIDA Grant P50-DA10075.
[3]Supported in part by Grants U01 HL119178, NIH/NHLBI-1U10HL098115 and UL1 TR000127.
Key words and phrases. Bayesian approach, group variable selection, longitudinal data, GWAS.

potential clinical relevance [Daly (2010)]. In plant and animal genetics, GWAS, with the advent of a continuously falling genotyping cost, have been considered more seriously than any time before [Filiault and Maloof (2012)]. Despite their powerful impact on genetic studies, however, GWAS also encounter tremendous challenges from statistical analysis and interpretation.

First, GWAS usually genotype hundreds of thousands of single-nucleotide polymorphisms (SNPs) on thousands of subjects, leading to a number of SNPs strikingly larger than the sample size used. Thus, to analyze these SNPs, simple univariate linear regression has to be used for individual tests. However, this method ignores the effects of other SNPs while assessing one particular SNP, and is subjected to a severe adjustment issue for multiple comparisons. Moreover, in biology and biomedicine, a phenotypic trait can always be better described by a dynamic trajectory because the trait undergoes a developmental process [Wu and Lin (2006)]. For example, human body height growth is a process from infancy to adulthood; the genetic study of adult height only, as conducted in many current GWAS [Lettre (2011)], provides limited information about the developmental genetics of height and its relationship with physical and mental characteristics at various stages of growth. In clinical trials, longitudinal measures are one of the most common data types, including HIV dynamics, cancer growth and drug response to varying doses [Wang et al. (2009)]. In this article, we address these issues by developing novel statistical models and algorithms that can analyze multiple SNPs simultaneously and integrate the developmental mechanisms of trait formation into a general GWAS framework through mathematical functions. The extension of the models to tackle genotype-environment interactions using GWAS is straightforward.

In a linear regression model for GWAS where SNPs are predictors, multiple regression breaks down when the number of predictors far exceeds the number of subjects. Alternatively, variable selection approaches could identify important genetic factors and enhance the predictive power of the final model. For example, in analyzing case-control cohorts, lasso regression [Tibshirani (1996)] and elastic-net regression [Zou and Hastie (2005)] were studied by Wu et al. (2009) and Cho et al. (2009), respectively. Li et al. (2012) and He and Lin (2011) further proposed two-stage variable selection approaches to identify disease susceptibility genes. These methods, however, are restricted to models with a single phenotypic measurement from each subject.

For genetic studies of dynamic traits that are measured repeatedly at multiple time points, Wu and Lin (2006) proposed a conceptual model called functional mapping by incorporating longitudinal and functional data analysis into a genetic design. Depending on the availability of explicit mathematical equations to describe a biological process, functional mapping uses parametric, nonparametric or semiparametric approaches for modeling nonlinear effects of genetic variants over time and further revealing a dynamic landscape of interplay between genes and developmental pattern. Das et al. (2011) implemented functional mapping into a

GWAS setting, leading to the birth of a so-called functional GWAS or *f*GWAS model. The basic principle of functional mapping and *f*GWAS is to model and predict the temporal pattern of genetic effects on a particular trait or disease in a quantitative manner. Time-varying change of gene expression has been found to be a ubiquitous phenomenon because different metabolic pathways, regulated by genes directly or indirectly, are required for an organism to best adapt to developmental alteration. In a genetic study of body mass index (BMI) by linkage mapping, Gorlova et al. (2003) identified different BMI susceptibility genes as well as different modes of inheritance triggered by these genes in children and adults. A common variant in the obesity-associated FTO gene, identified by a genome-wide search, was observed to be reproducibly associated with BMI and obesity from childhood into old age, but displayed varying magnitudes of genetic effects between child and adult stages [Fraying et al. (2007)].

To increase its applicability in clinical genomics, *f*GWAS could further accommodate irregular longitudinal data measured at subject-specific time points. But both functional mapping and *f*GWAS analyze SNPs individually or pairwise, and are incapable of depicting a comprehensive picture of the genetic architecture of dynamic traits. The motivation of this article is to develop a variable selection model for *f*GWAS, with a focus on nonparametric modeling of temporal genetic effects of SNPs. Variable selection in a nonparametrical setting is equivalent to selecting a subset of predictors with nonzero functional coefficients. Lin and Zhang (2006) developed COSSO for model selection in a smoothing spline ANOVA model, with the penalty term being the sum of component norms. Zhang and Lin (2006) further extended it to nonparametric regression in an exponential family. Wang, Li and Huang (2008) estimated time-varying effects using basis expansion and selected significant predictors by imposing SCAD penalty functions on the L_2-norm of these basis expansions.

We propose a Bayesian group lasso approach for variable selections in nonparametric varying-coefficient models. Group lasso was first proposed by Yuan and Lin (2006). They considered the problem of selecting important groups of independent variables in linear regression models and generalized lasso by encouraging sparsity at the group level. However, since the Hessian is not defined at the optimal solution, they did not provide variance estimates for the regression coefficients. Here, we express time-varying effects as a linear combination of Legendre polynomials, and in such a case, the selection of important predictors corresponds to the selection of groups of polynomials. We develop a Bayesian hierarchical model for group variable selection and estimate all parameters by MCMC algorithms. Our method provides not only point estimates but also interval estimates of all parameters, and the traditional Bayesian lasso [Park and Casella (2008)] is its special case in which the response variable is univariate.

In Section 2, we introduce the *f*GWAS model that connects genotypes and irregular longitudinal phenotypical data. Section 3 shows the Bayesian hierarchical representation for this nonparametric varying-coefficient model, where group

lasso penalties are applied to individual functional coefficients. The posterior computations as well as the interpretation of the results are described in Section 4. In Section 5, the statistical properties of the model are investigated through simulation studies. Section 6 provides the application to a real GWAS example from the Framingham Heart Study that analyzes age-specific changes of genetic effects on body mass index (BMI). BMI is a heuristic measure of body weight based on a person's weight and height, providing the most widely used diagnostic tool to identify whether individuals are underweighted, overweighted or obese, and, further, to examine their risk of developing obesity-related diseases, such as hypertension, type 2 diabetes and cardiovascular diseases [Frayling (2007)]. We use a nonparametric approach based on orthogonal polynomials to approximate age-specific change in BMI. The discussion about the new model is given in Section 7.

2. The *f*GWAS model. The model for functional genome-wide association studies (*f*GWAS) is the integration of functional data analysis and genome-wide association studies. The primary goal of the *f*GWAS is to study the dynamic pattern of genetic actions and interactions triggered by significant SNPs throughout the entire genome. Beyond traditional GWAS, *f*GWAS targets phenotypic traits that are measured longitudinally at repeated time points. Suppose in a genome-wide association study involving n subjects, a continuous longitudinal trait of interest is measured at irregularly spaced time points, which are not common to all subjects. Let $\mathbf{y}_i = (y_i(t_{i1}), \ldots, y_i(t_{iT_i}))^T$ be the T_i-dimensional vector of measurements on subject i where $\mathbf{t}_i = (t_{i1}, \ldots, t_{iT_i})^T$ is the corresponding vector of measurement time points after standardization. $\mathbf{y}_i$ can be described as

$$
\begin{aligned}
\begin{pmatrix} y_i(t_{i1}) \\ \vdots \\ y_i(t_{iT_i}) \end{pmatrix} &= \begin{pmatrix} \mu(t_{i1}) \\ \vdots \\ \mu(t_{iT_i}) \end{pmatrix} + \begin{pmatrix} \alpha_1(t_{i1}) & \cdots & \alpha_q(t_{i1}) \\ \vdots & & \vdots \\ \alpha_1(t_{iT_i}) & \cdots & \alpha_q(t_{iT_i}) \end{pmatrix} \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iq} \end{pmatrix} \\
&\quad + \begin{pmatrix} a_1(t_{i1}) & \cdots & a_p(t_{i1}) \\ \vdots & & \vdots \\ a_1(t_{iT_i}) & \cdots & a_p(t_{iT_i}) \end{pmatrix} \begin{pmatrix} \xi_{i1} \\ \vdots \\ \xi_{ip} \end{pmatrix} \\
&\quad + \begin{pmatrix} d_1(t_{i1}) & \cdots & d_p(t_{i1}) \\ \vdots & & \vdots \\ d_1(t_{iT_i}) & \cdots & d_p(t_{iT_i}) \end{pmatrix} \begin{pmatrix} \zeta_{i1} \\ \vdots \\ \zeta_{ip} \end{pmatrix} + \begin{pmatrix} e_i(t_{i1}) \\ \vdots \\ e_i(t_{iT_i}) \end{pmatrix}.
\end{aligned} \tag{2.1}
$$

We introduce matrix notation for a succinct presentation. Let $\boldsymbol{\alpha}(t_{i\ell}) = (\alpha_1(t_{i\ell}), \ldots, \alpha_q(t_{i\ell}))^T$ be the q-dimensional vector of covariate effects, $\mathbf{X}_i = (X_{i1}, \ldots, X_{iq})^T$ be the observed covariate vector for subject i, $\mathbf{a}(t_{i\ell}) = (a_1(t_{i\ell}), \ldots, a_p(t_{i\ell}))^T$ and

$\mathbf{d}(t_{i\ell}) = (d_1(t_{i\ell}), \ldots, d_p(t_{i\ell}))^T$ be the p-dimensional vectors of the additive and dominant effects of SNPs, respectively. Furthermore, let $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{ip})^T$ and $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{ip})^T$ be the indicator vectors of the additive and dominant effects of SNPs for subject i. Thus, at time point $t_{i\ell}$,

$$y_i(t_{i\ell}) = \mu(t_{i\ell}) + \boldsymbol{\alpha}(t_{i\ell})^T \mathbf{X}_i + \mathbf{a}(t_{i\ell})^T \boldsymbol{\xi}_i + \mathbf{d}(t_{i\ell})^T \boldsymbol{\zeta}_i + e_i(t_{i\ell}), \quad i = 1, \ldots, n, \ell = 1, \ldots, T_i, \tag{2.2}$$

where $\mu(t_{i\ell})$ is the overall mean and $e_i(t_{i\ell})$ is the residual error assumed to follow a $\mathrm{N}(0, \sigma^2(t_{i\ell}))$ distribution. The jth elements of $\boldsymbol{\xi}_i$ and $\boldsymbol{\zeta}_i$ are defined as

$$\xi_{i,j} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } AA, \\ 0, & \text{if the genotype of SNP } j \text{ is } Aa, \\ -1, & \text{if the genotype of SNP } j \text{ is } aa, \end{cases}$$

$$\zeta_{i,j} = \begin{cases} 1, & \text{if the genotype of SNP } j \text{ is } Aa, \\ 0, & \text{if the genotype of SNP } j \text{ is } AA \text{ or } aa. \end{cases}$$

In other words, $a_j(t_{i\ell})$ represents the average effect of substituting one allele for the other, and $d_j(t_{i\ell})$ represents how the average genotypic value of the heterozygote deviates from the mean of the homozygotes.

In the fGWAS model, the effects of covariates and SNPs are assumed to be functions of time. Many methods of estimating time-varying coefficients of a linear model in a longitudinal data setting have been proposed and studied, including basis expansion methods, local polynomial kernel methods and smoothing spline methods. Among these techniques, Legendre polynomials have been widely used by quantitative geneticists for modeling the growth curves [Lin and Wu (2006)], the programmed cell death (PCD) process [Cui et al. (2008)] or the genetic effects responsible for other traits [e.g., Suchocki and Szyda (2011); Yang and Xu (2007); Das et al. (2011)]. By approximating time-varying effects using Legendre polynomials, the expansion coefficients can be solved through regression. Moreover, the biological evidence or the prior belief about the time-dependency of genetic control can be integrated by just truncating the series. Motivated by these studies, we approximate the effect of the kth covariate by a Legendre polynomial of order $v - 1$:

$$\big(\alpha_k(t_{i1}), \ldots, \alpha_k(t_{iT_i})\big)^T = U_i \mathbf{r}_k, \qquad k = 1, \ldots, q, \tag{2.3}$$

where $\mathbf{r}_k = (r_{k0}, \ldots, r_{k(v-1)})^T$ are the Legendre polynomial coefficients, and

$$U_i = \begin{pmatrix} \mathbf{u}_{i1}^T \\ \vdots \\ \mathbf{u}_{iT_i}^T \end{pmatrix} = \begin{pmatrix} 1 & t_{i1} & \frac{1}{2}(3t_{i1}^2 - 1) & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_{iT_i} & \frac{1}{2}(3t_{iT_i}^2 - 1) & \cdots \end{pmatrix} \tag{2.4}$$

are Legendre polynomial functions. Similarly, other time-varying effects can be represented as

$$\big(a_j(t_{i1}),\ldots,a_j(t_{iT_i})\big)^T = U_i\mathbf{b}_j, \qquad j=1,\ldots,p, \tag{2.5}$$

$$\big(d_j(t_{i1}),\ldots,d_j(t_{iT_i})\big)^T = U_i\mathbf{c}_j, \qquad j=1,\ldots,p, \tag{2.6}$$

$$\big(\mu(t_{i1}),\ldots,\mu(t_{iT_i})\big)^T = U_i\mathbf{m}, \tag{2.7}$$

where $\mathbf{b}_j = (b_{j0},\ldots,b_{j(v-1)})^T$ are the Legendre polynomial coefficients for the additive effect of the jth SNP, $\mathbf{c}_j = (c_{j0},\ldots,c_{j(v-1)})^T$ are the Legendre polynomial coefficients for the dominant effect of the jth SNP, and $\mathbf{m} = (m_0,\ldots,m_{v-1})^T$ are the Legendre polynomial coefficients for the overall mean function.

After introducing Legendre polynomials to approximate time-varying effects of covariates and SNPs, the full model of fGWAS becomes

$$\begin{aligned} y_i(t_{i\ell}) &= \mathbf{u}_{il}^T\mathbf{m} + (\mathbf{u}_{il}^T\mathbf{r}_1,\ldots,\mathbf{u}_{il}^T\mathbf{r}_q)\mathbf{X}_i \\ &\quad + (\mathbf{u}_{il}^T\mathbf{b}_1,\ldots,\mathbf{u}_{il}^T\mathbf{b}_p)\boldsymbol{\xi}_i + (\mathbf{u}_{il}^T\mathbf{c}_1,\ldots,\mathbf{u}_{il}^T\mathbf{c}_p)\boldsymbol{\zeta}_i + e_i(t_{i\ell}), \\ &\qquad\qquad i=1,\ldots,n,\ \ell=1,\ldots,T_i. \end{aligned} \tag{2.8}$$

Last, since measurements within each subject are possibly correlated with one another, we assume that $\mathbf{e}_i = (e_i(t_{i1}),\ldots,e_i(t_{iT_i}))^T$ follows a multivariate normal distribution with zero mean and covariance matrix Σ_i. Both parametric and nonparametric methods have been developed to model the structure of covariance between longitudinal measurements [Ma, Casella and Wu (2002); Zhao et al. (2005); Yap, Fan and Wu (2009)]. In particular, we employ the first-order autoregressive [AR(1)] model to approximate the residual covariance matrix. This covariance structure allows different measurement time points for different subjects, and assumes a constant variance over time and an exponentially decaying correlation, $\rho^{|t_{i2}-t_{i1}|}$, $0<\rho<1$, between two measurements. Moreover, the matrix determinant in the likelihood function can be easily computed. In our real data example, the variance of repeated measurements is stable over time. In longitudinal data sets with variance heteroscedasticity, however, a Transform-Both-Sides (TBS) technique [Wu et al. (2004)] can be employed to satisfy the variance stationarity assumption in the AR(1) model.

3. Bayesian hierarchical representation for group Lasso penalties. In high-dimensional regression problems, such as GWAS, a regularized approach is preferred to identify predictors with nonzero effects and to achieve better out-of-sample predictive performance. When parameters that we would like to penalize are finite-dimensional, we may apply different penalty functions to them to perform variable selection. But when these parameters are nonparametric smooth functions, a traditional regularization procedure cannot be directly applied. In this

situation, regularized estimation for selecting important predictors is equivalent to selecting functional coefficients that are not identically zero.

Let $\|\mathbf{b}_j\|$ be the L_2 norm of the vector $\mathbf{b}_j$. The time-varying additive effect of the jth SNP is identically zero if and only if $\|\mathbf{b}_j\| = 0$. Therefore, if we estimate additive effects by a Legendre polynomial of order v, and would like to identify significant additive effects via penalized methods, we could partition all parameters of additive effects $(\mathbf{b}_1^T, \ldots, \mathbf{b}_p^T)$ into p groups of size v according to p SNPs, and encourage sparse solution at the group level or select a subset of groups with nonzero L_2 norms. That is, the group lasso minimizes the following penalized least square:

$$\frac{1}{2}\|\mathbf{y} - \boldsymbol{\mu}\|^2 + \lambda \sum_{j=1}^{p} \|\mathbf{b}_j\| + \lambda^* \sum_{j=1}^{p} \|\mathbf{c}_j\|, \tag{3.1}$$

where $\mathbf{y}^T = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)$, $\boldsymbol{\mu}^T = E\mathbf{y}^T = (\boldsymbol{\mu}_1^T, \ldots, \boldsymbol{\mu}_n^T)$ and λ and λ^* are two regularization parameters. λ and λ^* control the amount of shrinkage toward zero: the larger their values, the greater the amount of shrinkage. They should be adaptively determined from the data to minimize an estimate of expected prediction error.

From a Bayesian perspective, the group lasso estimates can be interpreted as posterior mode estimates when the regression parameters have multivariate independent and identical Laplace priors. Therefore, when group lasso penalties are imposed on the Legendre coefficients of additive and dominant effects, the conditional prior for $\mathbf{b}_j$ is a multivariate Laplace distribution with the scale parameter $(v\lambda^2/\sigma^2)^{-1/2}$:

$$\pi(\mathbf{b}_j|\sigma^2) = (v\lambda^2/\sigma^2)^{v/2} \exp\bigl(-(v\lambda^2/\sigma^2)^{-1/2}\|\mathbf{b}_j\|\bigr), \tag{3.2}$$

and the conditional multivariate Laplace prior for dominant effect $\mathbf{c}_j$ is

$$\pi(\mathbf{c}_j|\sigma^2) = (v\lambda^{*2}/\sigma^2)^{v/2} \exp\bigl(-(v\lambda^{*2}/\sigma^2)^{-1/2}\|\mathbf{c}_j\|\bigr). \tag{3.3}$$

To ensure the derived conditional distribution of $\mathbf{b}_j$ has a standard form, we rewrite the multivariate Laplace prior distribution as a scale mixture of a multivariate Normal distribution with a Gamma distribution, that is,

$$\begin{aligned}
&\text{M-Laplace}\bigl(\mathbf{b}_j|0, (v\lambda^2/\sigma^2)^{-1/2}\bigr)\\
&\quad\propto (v\lambda^2/\sigma^2)^{v/2} \exp\bigl(-(v\lambda^2/\sigma^2)^{1/2}\|\mathbf{b}_j\|\bigr)\\
&\quad\propto \int_0^\infty \text{MVN}\bigl(\mathbf{b}_j|\mathbf{0}, \text{diag}(\sigma^2\tau_j^2, \ldots, \sigma^2\tau_j^2)\bigr)\, \text{Gamma}\left(\tau_j^2 \middle| \frac{v+1}{2}, \frac{2}{v\lambda^2}\right) d\tau_j^2,
\end{aligned}$$

where $(v\lambda^2/\sigma^2)^{-1/2})$ is the scale parameter of the multivariate Laplace distribution, a v-by-v diagonal matrix $\text{diag}(\sigma^2\tau_j^2, \ldots, \sigma^2\tau_j^2)$ is the covariance matrix of the multivariate normal distribution with mean zero, $\frac{v+1}{2}$ is the shape parameter of the Gamma distribution, and $\frac{2}{v\lambda^2}$ is the scale parameter of the Gamma distribution.

After integrating out τ_j^2, the conditional prior on $\mathbf{b}_j$ has the desired form (3.2). Then, in a Bayesian hierarchical model, we can rewrite the multivariate Laplace priors on $\mathbf{b}_j$ as

$$\mathbf{b}_j|\tau_j^2, \sigma^2 \sim \text{MVN}(\mathbf{0}, \text{diag}(\sigma^2\tau_j^2, \ldots, \sigma^2\tau_j^2)),$$
$$\tau_j^2|\lambda \sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^2}\right).$$

Likewise, the multivariate-Laplacian prior on $\mathbf{c}_j$ can be replaced by

$$\mathbf{c}_j|\tau_j^{*2}, \sigma^2 \sim \text{MVN}(\mathbf{0}, \text{diag}(\sigma^2\tau_j^{*2}, \ldots, \sigma^2\tau_j^{*2})),$$
$$\tau_j^{*2}|\lambda \sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^{*2}}\right).$$

Then, given λ and λ^*, we have the following hierarchical representation of the penalized regression model:

$$\mathbf{y}|\mathbf{m}, \mathbf{r}_k, \mathbf{b}_j, \mathbf{c}_j, \rho, \sigma^2 \propto (2\pi)^{-(\sum_i^n T_i)/2}\left(\prod_i^n |\Sigma_i|^{-1/2}\right)e^{-1/2\sum_i^n(\mathbf{y}_i-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{y}_i-\boldsymbol{\mu}_i)},$$
$$\mathbf{m} \sim N_v(0, \Sigma_{m0}),$$
$$\mathbf{r}_k \sim N_v(0, \Sigma_{r0}), \qquad k = 1, \ldots, q,$$
$$\mathbf{b}_j|\tau_j^2, \sigma^2 \sim \text{MVN}(\mathbf{0}, \text{diag}(\sigma^2\tau_j^2, \ldots, \sigma^2\tau_j^2)), \qquad j = 1, \ldots, p,$$
$$\tau_j^2|\lambda \sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^2}\right), \qquad j = 1, \ldots, p,$$
$$\mathbf{c}_j|\tau_j^{*2}, \sigma^2 \sim \text{MVN}(\mathbf{0}, \text{diag}(\sigma^2\tau_j^{*2}, \ldots, \sigma^2\tau_j^{*2})), \qquad j = 1, \ldots, p,$$
$$\tau_j^{*2}|\lambda^* \sim \text{Gamma}\left(\frac{v+1}{2}, \frac{2}{v\lambda^{*2}}\right), \qquad j = 1, \ldots, p,$$
$$\rho \sim \text{U}(-1, 1),$$
$$\sigma^2 \sim \pi(\sigma^2),$$
$$\sigma^2, \lambda, \lambda^* > 0,$$

where λ and λ^* are regularization parameters or group lasso parameters that control the shrinkage intensities in estimating genetic effects. We assign a conjugate multivariate normal prior to $\mathbf{m}$ when estimating the overall mean function. We also assign conjugate multivariate normal priors to the Legendre coefficients of covariates $\mathbf{r}_k, k = 1, \ldots, q$, because covariates in GWAS are usually low dimensional and are not the target of variable selection. We assume a Uniform prior on $[-1, 1]$ for ρ, the autoregressive parameter in the assumed AR(1) covariance

matrix. Finally, since the data are usually sufficient to estimate σ, we can use a noninformative prior such as $\pi(\sigma^2) = 1/\sigma^2$ for σ^2.

Traditionally, two group lasso parameters λ and λ^* can be prespecified by cross-validation or generalized cross-validation. However, in the Bayesian group lasso setting, λ and λ^* can be estimated along with other parameters by assigning appropriate hyperpriors to them. This procedure determines the amount of regularization from the data and avoids refitting the model. In particular, the following conjugate gamma priors are considered,

$$\pi\left(\frac{\lambda^2}{2}\right) \sim \text{Gamma}(a, b) \quad \text{and} \quad \pi\left(\frac{\lambda^{*2}}{2}\right) \sim \text{Gamma}(a^*, b^*),$$

where a, b, a^* and b^* are small values so that the priors are essentially noninformative. With this specification, group lasso parameters can simply join the other parameters in the Gibbs sampler.

4. Posterior computation and interpretation. We estimate the unknown parameters and hyperparameters by sampling from their conditional posterior distributions through MCMC algorithms. Given the data likelihood and prior distributions, the posterior distributions of all unknowns can be obtained by Bayes' theorem. For most of the parameters, the conditional posterior distributions have closed forms by conjugacy, which facilitates drawing posterior samples.

Assuming that priors for different predictors are independent, we can express the joint posterior distribution of all parameters as

$$\begin{aligned}
&\pi\left(\mathbf{m}, \mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho|\mathbf{y}\right) \\
&\quad \propto \pi(\mathbf{y}|\cdot)\pi(\mathbf{m})\pi\left(\sigma^2\right)\pi(\rho) \prod_{k=1}^{q} \pi(\mathbf{r}_k) \\
&\qquad \times \prod_{j=1}^{p} \pi\left(\mathbf{b}_j|\tau_j^2\right)\pi\left(\tau_j^2|\lambda\right)\pi(\lambda)\pi\left(\mathbf{c}_j|\tau_j^{*2}\right)\pi\left(\tau_j^{*2}|\lambda^*\right)\pi\left(\lambda^*\right).
\end{aligned}$$

Conditional on the parameters $(\mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho)$, we derive the conditional posterior distribution of $\mathbf{m}$ as

$$\begin{aligned}
&\pi\left(\mathbf{m}|\mathbf{y}, \mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho\right) \\
&\quad \propto \pi(\mathbf{m})\pi(\mathbf{y}|\cdot) \\
&\quad \propto \exp\Bigg(-\frac{1}{2}\mathbf{m}^T \Sigma_{m0}^{-1}\mathbf{m} \\
&\qquad - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)} - U_i\mathbf{m})^T \Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)} - U_i\mathbf{m})\Bigg)
\end{aligned}$$

$$\propto \exp\Bigg(\mathbf{m}^T \Sigma_{m0}^{-1}\mathbf{m} + \sum_{i=1}^{n}(U_i\mathbf{m})^T \Sigma_i^{-1}(U_i\mathbf{m}) - 2\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)})^T \Sigma_i^{-1}(U_i\mathbf{m})\Bigg)$$

$$\propto \exp\Bigg(\mathbf{m}^T\Bigg(\Sigma_{m0}^{-1} + \sum_{i=1}^{n} U_i^T \Sigma_i^{-1} U_i\Bigg)\mathbf{m} - 2\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)})^T \Sigma_i^{-1}(U_i\mathbf{m})\Bigg).$$

Hence, the conditional posterior distribution of $\mathbf{m}$ is $\mathrm{MVN}_v(\boldsymbol{\mu}_m, \Sigma_m)$, where

$$\boldsymbol{\mu}_m = \Bigg(\Sigma_{m0}^{-1} + \sum_{i=1}^{n} U_i^T \Sigma_i^{-1} U_i\Bigg)^{-1}\Bigg(\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-m)})^T \Sigma_i^{-1} U_i\Bigg)^T,$$

and

$$\Sigma_m = \Bigg(\Sigma_{m0}^{-1} + \sum_{i=1}^{n} U_i^T \Sigma_i^{-1} U_i\Bigg)^{-1}.$$

Similarly, since $\mathbf{r}_k$, $\mathbf{b}_j$ and $\mathbf{c}_j$ have conjugate multivariate normal priors, the posterior distribution for $\mathbf{r}_k$ is $\mathrm{MVN}_v(\boldsymbol{\mu}_{r_k}, \Sigma_{r_k})$, with

$$\boldsymbol{\mu}_{r_k} = \Bigg(\Sigma_{r0}^{-1} + \sum_{i=1}^{n}(X_{ik}U_i)^T \Sigma_i^{-1}(X_{ik}U_i)\Bigg)^{-1} \times \Bigg(\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-r_k)})^T \Sigma_i^{-1}(X_{ik}U_i)\Bigg)^T,$$

and

$$\Sigma_{r_k} = \Bigg(\Sigma_{r0}^{-1} + \sum_{i=1}^{n}(X_{ik}U_i)^T \Sigma_i^{-1}(X_{ik}U_i)\Bigg)^{-1},$$

the posterior distribution for $\mathbf{b}_j$ is $\mathrm{MVN}_v(\boldsymbol{\mu}_{b_j}, \Sigma_{b_j})$, with

$$\boldsymbol{\mu}_{b_j} = \Bigg(\big(\sigma^2\tau_j^2\big)^{-1} + \sum_{i=1}^{n}(\xi_{ij}U_i)^T \Sigma_i^{-1}(\xi_{ij}U_i)\Bigg)^{-1} \times \Bigg(\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_{i(-b_j)})^T \Sigma_i^{-1}(\xi_{ij}U_i)\Bigg)^T,$$

and

$$\Sigma_{b_j} = \Bigg(\big(\sigma^2\tau_j^2\big)^{-1} + \sum_{i=1}^{n}(\xi_{ij}U_i)^T \Sigma_i^{-1}(\xi_{ij}U_i)\Bigg)^{-1},$$

and the posterior distribution for $\mathbf{c}_j$ is $\mathrm{MVN}_v(\boldsymbol{\mu}_{c_j}, \Sigma_{c_j})$, with

$$\boldsymbol{\mu}_{c_j} = \left((\sigma^2 \tau_j^{*2})^{-1} + \sum_{i=1}^{n} (\zeta_{ij} U_i)^T \Sigma_i^{-1} (\zeta_{ij} U_i) \right)^{-1} \times \left(\sum_{i=1}^{n} (\mathbf{y}_i - \boldsymbol{\mu}_{i(-c_j)})^T \Sigma_i^{-1} (\zeta_{ij} U_i) \right)^T,$$

and

$$\Sigma_{c_j} = \left((\sigma^2 \tau_j^{*2})^{-1} + \sum_{i=1}^{n} (\zeta_{ij} U_i)^T \Sigma_i^{-1} (\zeta_{ij} U_i) \right)^{-1}.$$

Now, we derive the conditional posterior distribution for τ_j^2 and λ^2 from the joint posterior distribution. Since

$$\begin{aligned} &\pi(\tau_j^2 | \mathbf{y}, \mathbf{m}, \mathbf{r}_k, \mathbf{b}_j, \lambda, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho) \\ &\quad \propto \pi(\tau_j^2 | \lambda) \pi(\mathbf{b}_j | \tau_j^2, \sigma^2) \\ &\quad \propto (\tau_j^2)^{((v+1)/2)-1} \exp\left(-\tau_j^2 \frac{v\lambda^2}{2} \right) (\tau_j^2)^{-v/2} \\ &\qquad \times \exp\left(-\frac{1}{2} \mathbf{b}_j^T (\sigma^2 \operatorname{diag}(\tau_j^2, \ldots, \tau_j^2))^{-1} \mathbf{b}_j \right) \\ &\quad \propto \exp\left(-\tau_j^2 \frac{v\lambda^2}{2} - \frac{1}{2\sigma^2 \tau_j^2} \|\mathbf{b}_j\|^2 \right) (\tau_j^2)^{-1/2}, \end{aligned}$$

and

$$\begin{aligned} &\pi(\lambda^2 | \mathbf{y}, \mathbf{m}, \mathbf{r}_k, \mathbf{b}_j, \tau_j^2, \mathbf{c}_j, \tau_j^{*2}, \lambda^*, \sigma^2, \rho) \\ &\quad \propto \pi(\lambda^2) \prod_{j=1}^{p} \pi(\tau_j^2 | \lambda) \\ &\quad \propto (\lambda^2)^{a-1} \exp(-b\lambda^2) \prod_{j=1}^{p} \left(\frac{v\lambda^2}{2} \right)^{(v+1)/2} \exp\left(-\frac{v\lambda^2}{2} \tau_j^2 \right), \end{aligned}$$

the posterior distribution for $\frac{1}{\tau_j^2}$ is inverse-Gaussian$(v\lambda^2, \sqrt{\frac{v\lambda^2\sigma^2}{\|\mathbf{b}_j\|^2}})$ and the posterior distribution for λ^2 is Gamma$(a + \frac{pv+p}{2}, b + \frac{v\sum_{j=1}^{p} \tau_j^2}{2})$.

Similarly, the posterior distribution for $\frac{1}{\tau_j^{*2}}$ is inverse-Gaussian$(v\lambda^{*2}, \sqrt{\frac{v\lambda^{*2}\sigma^2}{\|\mathbf{b}_j\|^2}})$, and the posterior distribution for λ^{*2} is Gamma$(a^* + \frac{pv+p}{2}, b^* + \frac{v\sum_{j=1}^{p} \tau_j^{*2}}{2})$. From

these posteriors, we can see that the hierarchical expansion of the Multivariate Laplace prior indeed gives closed forms of posterior distributions for efficient Gibbs sampling.

Last, if we assume a stationary AR(1) covariance structure, that is,

$$\Sigma_i = \sigma^2 \Gamma_i = \sigma^2 \begin{pmatrix} 1 & \rho^{|t_{i1}-t_{i2}|} & \dots & \rho^{|t_{i1}-t_{iT_i}|} \\ \rho^{|t_{i2}-t_{i1}|} & 1 & \dots & \rho^{|t_{i2}-t_{iT_i}|} \\ \vdots & \vdots & \vdots & \vdots \\ \rho^{|t_{i1}-t_{iT_i}|} & \rho^{|t_{i2}-t_{iT_i}|} & \dots & 1 \end{pmatrix},$$

the posterior distribution for σ^2 is an inverse chi-square distribution, or

$$\pi(\sigma^2|\cdot) \sim \text{Inv-}\chi^2\left(\sum_{i=1}^n T_i, \frac{\sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \Gamma_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)}{\sum_{i=1}^n T_i}\right),$$

where the first parameter is the degree of the freedom parameter and the second one is the scale parameter, and

$$\begin{aligned} \pi(\rho|\cdot) &\propto \pi(\mathbf{y}|\cdot)\pi(\rho) \\ &\propto \prod_{i=1}^n \left(|\Gamma_i|^{-1/2}\right) \exp\left(-\frac{1}{2}\sum_{i=1}^n (\mathbf{y}_i - \boldsymbol{\mu}_i)^T \Gamma_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right). \end{aligned}$$

Based on this expression, the corresponding Metropolis–Hastings algorithm can be developed to update ρ.

We use MCMC algorithms to estimate the posterior distribution of each parameter by drawing posterior samples from the corresponding conditional posterior distribution, given the current values of all other parameters and the observed data. We use the potential scale reduction factor [PSRF; Gelman and Rubin (1992); Gelman et al. (2004)] to access the convergence. Squared PSRF is defined as the ratio of the marginal posterior variance to the within-chain variance, and a PSRF less than 1.1 indicates good convergence. We run 4000 additional iterations after all chains converge.

5. Computer simulation. We first investigate the new Bayesian group lasso approach for selecting important time-varying effects through simulation studies. We generate data in the fGWAS setting according to the model (2.8) with the number of covariates $q = 1$, the number of SNPs $p = 3000$, and the number of individuals $n = 600$ or 800. Following the simulation techniques in the literature, genotypical data ξ_{ij} is derived from u_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, p$, where each u_{ij} has a standard normal distribution marginally, and $\text{cov}(u_{ij}, u_{ik}) = \rho_G = 0.1$ or 0.5, representing two levels of linkage disequilibrium. We set

$$\xi_{ij} = \begin{cases} 1, & u_{ij} > c, \\ 0, & -c \le u_{ij} \le c, \\ -1, & u_{ij} < -c, \end{cases}$$

TABLE 1
Parameters used in the simulated example

Time-varying effect	Parameter	Legendre coefficients			
		0	**1**	**2**	**3**
Mean effect	$\mathbf{m}$	13.40	−3.08	1.88	−3.20
Covariate effect	$\mathbf{r}_1$	3.00	0.15	−2.67	3.25
Additive effect	$\mathbf{b}_1$	1.04	0.88	−2.05	0.00
	$\mathbf{b}_2$	1.17	−0.22	0.74	−4.72
	$\mathbf{b}_3$	1.40	0.00	0.00	0.00
Dominant effect	$\mathbf{c}_3$	1.49	−2.13	4.82	1.42
	$\mathbf{c}_4$	1.00	1.32	1.90	1.50
	$\mathbf{c}_5$	1.26	−1.22	0.00	0.00

where c is used to determine the minor allele frequencies. Then, we derive the indicator matrix ζ_{ij} of dominant effects from ξ_{ij}.

We assume that the dynamic pattern of the trait is controlled by 5 SNPs and 1 covariate. In particular, we set $\mathbf{b}_j = \mathbf{0}$ for $j = 4, \ldots, p$, and $\mathbf{c}_j = \mathbf{0}$ for $j = 1, 2, 6, \ldots, p$. Sex is included as a covariate and is generated by randomly assigning a sex to each subject. The time-varying effects of overall mean, covariate and causal SNPs are generated by Legendre polynomials, with Legendre coefficients listed in Table 1. The true polynomial degrees for these causal SNPs could be 0, 1, 2 or 3, allowing constant genetic effects, linear genetic effects or more complicated patterns of genetic control.

To simulate irregular longitudinal phenotypical data, we assume that the number of measurements for each subject is between 5 and 12, and all subjects are in the age range of 30 to 80 years. For each subject with a specific number of measurements, traits of interest are observed at ages randomly drawn from 30 to 80. The residual covariance matrix among different time points was assumed to be AR(1) with $\rho = 0.4$ and $\sigma^2 = 4, 9$ or 16. The phenotypes observed at subject-specific time points and genotypes of all subjects are collected for Bayesian analysis.

For each simulated data set, we implement MCMC algorithms as described in Section 4. In practice, the degree of Legendre polynomials should be determined a priori. We recommend a procedure that analyzes all SNPs with different polynomial degrees, where group lasso penalties are used to regularize the estimation. When the polynomial degree is 0 (constant effect), the group lasso penalty reduces to a lasso penalty. Then the polynomial degree $\hat{v}$ that gives the lowest Bayesian information criterion (BIC) of the final model is chosen. In simulations, however, this is computationally expensive. Therefore, the polynomial degree is fixed at $\hat{v} = 3$ in simulation studies. Simulation results (see Table 2) suggest that, as long as the specified polynomial degree is greater than or equal to the largest degree of

TABLE 2
Variable selection performance in the simulated example

n	σ^2	No. of nonzeros		Proportion of			Time (h)
		C	IC	Under-fit	Correct-fit	Over-fit	
$\rho_G = 0.1$							
600	16	3.77	0.00	0.86	0.14	0.00	17.99
600	9	4.93	0.00	0.07	0.93	0.00	17.67
600	4	5.00	0.00	0.00	1.00	0.00	17.40
800	16	4.99	0.00	0.01	0.99	0.00	23.69
800	9	5.00	0.00	0.00	1.00	0.00	24.78
800	4	5.00	0.00	0.00	1.00	0.00	24.35
$\rho_G = 0.5$							
600	16	4.61	0.00	0.35	0.65	0.00	17.90
600	9	5.00	0.00	0.00	1.00	0.00	17.29
600	4	5.00	0.00	0.00	1.00	0.00	17.63
800	16	5.00	0.00	0.00	1.00	0.00	23.97
800	9	5.00	0.00	0.00	1.00	0.00	23.49
800	4	5.00	0.00	0.00	1.00	0.00	24.38

all nonzero effects, the proposed framework works well in selecting casual SNPs and estimating their time-varying effects.

Once all posterior samples are collected from MCMC algorithms, SNPs are selected in the following way: a time-varying additive effect $a_j(t)$ or dominant effect $d_j(t)$ is included in the final model if at least one of its four Legendre coefficients has a two-sided 95% interval estimate that does not cover zero. In the supplemental article [Li et al. (2015)], we plot the potential scale reduction factor against iterations for each parameter in $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$, $\mathbf{c}_3$, $\mathbf{c}_4$ and $\mathbf{c}_5$. This is a simulation randomly drawn from the specification $n = 600$ and $\sigma^2 = 16$. All chains converge very quickly and stay below the threshold of 1.05 (the red line).

To evaluate the variable selection performance of the proposed procedure, we calculate several measures of model sparsity for the final model, which are summarized in Table 2. Column "C" shows the average number of SNPs with nonzero varying-coefficients correctly included in the final model, and column "IC" is the average number of SNPs with no genetic effect incorrectly included in the final model. Column "Under-fit" represents the proportion of excluding any relevant SNP in the final model. Similarly, column "Correct-fit" represents the proportion that the extract true model was selected and column "Over-fit" gives the proportion of including all relevant SNPs as well as one or more irrelevant SNPs. Clearly, both sample size and the noise level play important roles in how well the Bayesian group lasso could select the exactly correct model. However, as sample size decreases and noise increases, our procedure tends to select fewer important SNPs

rather than produce more false positives. Moreover, the impact of linkage disequilibrium is limited, and our method works slightly better in the presence of high linkage disequilibrium.

Other than the performance of selecting truly important SNPs, we further investigate how well the procedure estimates the time-varying effects of selected SNPs. To ameliorate the bias of the parameter estimates introduced by group lasso penalties, we always refit the *f*GWAS model after variable selections, where only selected SNPs are included in the final model and all regularization parameters are set to zero. For each time-varying genetic effect of important SNPs, Tables 1 and 2 in the supplemental article [Li et al. (2015)] summarize the average estimates, standard errors and the mean squared errors (MSEs) of Legendre coefficients over replications where the effect is selected for $\rho_G = 0.1$. As can be seen from these tables, both bias and standard error decrease as noise level decreases. MSEs are slightly lower for additive effects and lower order Legendre coefficients.

To compare the parameter estimates with those produced by another strategy aimed at the same genetic model, we implement the univariate *f*GWAS approach by Das et al. (2011) using the same data set. Specifically, this single-SNP analysis extends the traditional GWAS analysis framework by allowing the phenotype to be collected repeatedly over time and approximating the time-varying genetic effects by Legendre polynomials. A Benjamini-Hochberg false discovery rate (FDR) controlling procedure is used to adjust for multiple comparisons in selecting significant SNPs. Table 3 in the supplemental article [Li et al. (2015)] shows that this single-SNP analysis produces biased estimates for all parameters.[4]

Finally, we compare the variable selection performance of four approaches: (1) a Bayesian group lasso; (2) a univariate *f*GWAS approach by Das et al. (2011); (3) a functional principal component analysis (fPCA) approach [Ramsay and Silverman (2005)] that analyzes the fPCA of the longitudinal phenotype; and (4) a slope model that simplifies the longitudinal phenotype to its slope.[5] In the third and the fourth model, the leading three fPCA scores and the slope calculated from each growth curve are tested against genetic predictors, respectively, where group lasso or lasso regressions with 5-fold cross-validation are used to select relevant SNPs.

For fairness of comparison, longitudinal phenotype data are not generated from our nonparametric genetic model (2.8). Instead, we use the same genotype data with $\rho_G = 0.1$ but assume the following time-varying genetic effects: $a_1(t) = 0.5 + \sin(0.2t)$, $a_2(t) = 1/(0.5 + \exp(-0.06t)) - 0.5$, $a_3(t) = \log(0.05t)$, $d_3(t) = -1.5$, $d_4(t) = 60/t$, and $d_5(t) = 0.2 - 0.035t$ for the first five SNPs. These functional forms are unknown to researchers. Table 3 presents

[4]Since this approach cannot identify if the significance is due to the additive effect or the dominant effect, both effects are reported for five important SNPs.

[5]We thank the Associate Editor and an anonymous referee for pointing out the fPCA method and the slope method, respectively.

TABLE 3
Variable selection performance of alternative methods in the simulated example

		Nonzeros		Proportion of			Nonzeros		Proportion of		
n	σ^2	**C**	**IC**	**U.-fit**	**C.-fit**	**O.-fit**	**C**	**IC**	**U.-fit**	**C.-fit**	**O.-fit**
Bayesian group lasso							*Das et al.* (2011)				
600	16	3.93	0.00	0.83	0.17	0.00	4.94	0.51	0.06	0.55	0.39
600	9	4.80	0.00	0.20	0.80	0.00	5.00	0.20	0.00	0.83	0.17
600	4	5.00	0.00	0.00	1.00	0.00	4.98	0.02	0.02	0.96	0.02
800	16	4.86	0.00	0.14	0.86	0.00	4.99	0.40	0.01	0.70	0.29
800	9	5.00	0.00	0.00	1.00	0.00	5.00	0.23	0.00	0.81	0.19
800	4	5.00	0.00	0.00	1.00	0.00	5.00	0.01	0.00	0.99	0.01
Functional PCA							*Slope model*				
600	16	0.51	4.46	1.00	0.00	0.00	1.06	10.28	1.00	0.00	0.00
600	9	1.19	7.01	0.98	0.00	0.02	2.53	16.42	0.99	0.00	0.01
600	4	2.75	13.53	0.78	0.00	0.22	3.65	21.22	0.97	0.00	0.03
800	16	0.77	5.02	1.00	0.00	0.00	1.82	10.91	1.00	0.00	0.00
800	9	1.79	11.14	0.88	0.00	0.12	3.06	14.90	0.98	0.00	0.02
800	4	2.90	17.16	0.61	0.00	0.39	3.82	19.00	0.99	0.00	0.01

variable selection results, where all measures strongly prefer the Bayesian group lasso. Among the alternative approaches, the univariate *f*GWAS approach has the best variable selection performance. For the fPCA approach and the slope approach, the probability of selecting casual SNPs increases with the signal-to-noise ratio (column "C"), but the proportion of under-fit is always substantial. Interestingly, as signal-to-noise ratio increases, the probability of identifying false positives also increases (columns "IC" and "Over-fit"), especially when σ^2 decreases from 16 to 9. The inconsistency of these procedures suggests the risk of inflated false positive rates when only the major movements of growth curves are captured and tested in association studies.

In the above simulation studies, the minor allele frequency is set to 0.3. Unreported simulations also demonstrate that as the minor allele frequency decreases, both statistical powers and false positive rates decrease. But our method is still much better than the alternative approaches. Despite the Bayesian framework's theoretical advantages in handling parameter uncertainty, practically it could be slower than frequentist methods. When $n = 600$, $\sigma^2 = 9$, $\rho_G = 0.1$ and the number of SNPs $p = 1000$, the Gibbs sampler's computational time is about 5.70 hours. Experiments show that a linear regression line[6] can describe almost perfectly the relationship between the computational time in hours and p: $\log_{10}(time) = 0.754 + \log_{10}(p/1000)$.

[6]We thank the Editor for sharing the idea of using this regression.

6. Worked example. We use the newly developed model to analyze a real GWAS data set from the Framingham Heart Study (FHS), a cardiovascular study based in Framingham, Massachusetts, supported by the National Heart, Lung, and Blood Institute, in collaboration with Boston University [Dawber, Meadors and Moore (1951)]. Recently, 550,000 SNPs have been genotyped for the entire Framingham cohort [Jaquish (2007)], from which 493 males and 372 females were randomly chosen for our data analysis. These subjects were measured for body mass index (BMI) at multiple time points from age 29 to age 61. The number of measurements for a subject ranges from 2 to 18, and the intervals of measurement are also highly variable among subjects. As is standard practice, SNPs with rare allele frequency $<10\%$ were excluded from data analysis. The numbers and percentages of nonrare allele SNPs vary among different chromosomes and range from 4417 to 28,771 and from 0.64 to 0.72, respectively.

A single-SNP analysis was used to analyze the phenotypic data of BMI for males and females separately. Figure 1 gives $-\log_{10}$ p-values for each SNP in the two sexes, from which 33,239 SNPs with $-\log_{10}$ p-values greater than 2.0 in at least one sex were selected. Before applying Bayesian group lasso analysis to this irregular longitudinal data set, we imputed missing genotypes for a small proportion of SNPs according to the distribution of genotypes in the population. Then, by treating the sex as a covariate, we imposed group lasso penalties on both additive effects and dominant effects in hopes of identifying SNPs with notable effects on BMI, where all effects are possibly functions of time. According to our discussions in Section 5, the whole procedure was repeated with polynomial degrees: 0, 1, 2, 3 and 4, and the corresponding BICs of the final model are as follows: 27,470, 27,444, 27,416, 27,408 and 27,426. Therefore, a polynomial degree of 3 is appropriate in this real data example.

The Bayesian group lasso selected 24 significant SNPs, located on chromosomes 1, 2, 3, 4, 6, 7, 12, 14, 16 and 23. Table 4 tabulates the names, positions, alleles and estimated Legendre coefficients of these SNPs. The first allele in the column "Alleles" represents the minor allele. Using the Legendre coefficient estimates, we plot their time-varying additive effects and dominant effects in Figures 2 and 3, respectively, where the associated interval estimates[7] are also provided. Some of these detected SNPs are located in a similar region of candidate genes for obesity. For example, the detected SNPs on chromosomes 4, 6 and 12 are close to candidate genes for BMI-related type 2 diabetes [Frayling (2007)].

Figures 2 and 3 show that the time courses of the genetic effects of some SNPs are relatively constant (magenta), monotonically increasing (black) or decreasing

[7]Suppose for one varying-coefficient, the interval estimate of the qth Legendre coefficient is $(b_{q,U}, b_{q,L})^T, q = 1, \ldots, 4$, and the Legendre polynomials are $(u_0, u_1, u_2, u_3)^T = (1, t, \frac{1}{2}(3t^2 - 1), \frac{1}{2}(5t^3 - 3t))^T$ for each standardized time point $t \in [-1, 1]$. Then the interval estimate of the varying-coefficient at time t is $(\sum_{q=1}^4 \tilde{b}_{q,U} u_q, \sum_{q=1}^4 \tilde{b}_{q,L} u_q)^T$, where $\tilde{b}_{q,U} = b_{q,U}$ if u_q is positive and $b_{q,L}$ otherwise, and $\tilde{b}_{q,L} = b_{q,L}$ if u_q is positive and $b_{q,U}$ otherwise.

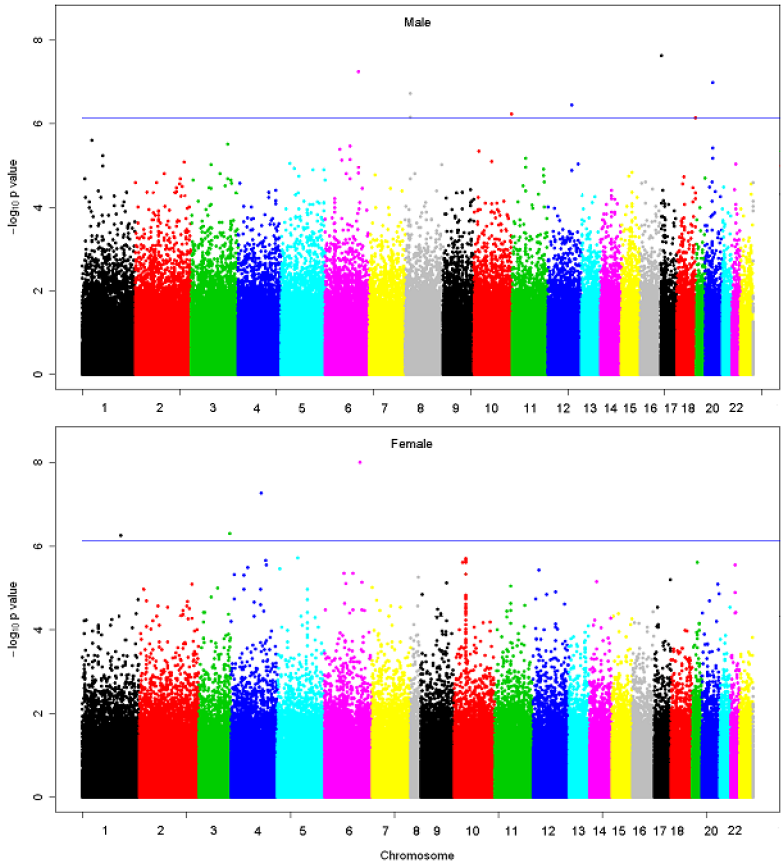

FIG. 1. *Manhattan plot of p-values for association by genomic position for male and female, where different colors across the x-axis represent different chromosomes, and the horizontal line indicates the significance level obtained by the Benjamini-Hochberg FDR adjustment at* $\alpha = 5\%$.

(blue). That is, given a population carrying one of these SNPs in the same environment, the expected BMI is different at different ages. Individuals carrying certain SNPs may have lower BMI in mid-life but tend to have higher BMI when they are younger or older (red). Conversely, individuals carrying certain SNPs tend to have higher BMI in mid-life (green), which may increase the risk for stroke later in life, according to a prospective study [Jood et al. (2004)].

7. Discussion. When the number of predictors p is much larger than the number of observations n, highly regularized approaches, such as penalized regression models, are favorable to identify nonzero coefficients, to enhance model predictability and to avoid overfitting [Hastie, Tibshirani and Friedman (2009)]. In this article, we proposed a Bayesian regularized estimation procedure for nonparametric varying-coefficient models that could simultaneously estimate time-

TABLE 4

Information about selected SNPs in the real data example

				Estimated legendre coefficients							
				Additive effect				Dominant effect			
Chr	**Name**	**Position**	**Alleles**								
1	ss66334458	79,393,823	C/T	−1.504	−1.656	−1.334	−0.143	1.136	2.645	2.480	0.780
1	ss66050888	93,240,623	A/G	−0.431	−0.532	−0.344	−0.111	0.737	0.745	−0.185	0.190
1	ss66275851	93,245,738	C/T	−0.128	−0.949	−1.096	−0.448	0.079	−0.254	0.278	−0.004
1	ss66048018	115,427,398	A/G	0.396	0.217	0.028	0.418	0.213	−0.007	0.571	0.294
1	ss66287256	221,051,934	G/A	0.497	0.788	0.934	−0.047	0.386	−1.065	−0.672	0.221
1	ss66104828	234,701,498	A/C	0.111	−0.620	−0.951	−0.461	1.445	1.833	1.307	0.205
2	ss66484730	103,489,666	G/A	−0.341	0.254	0.335	0.098	0.057	0.612	0.552	−0.565
2	ss66232775	103,493,541	T/C	0.476	−1.098	−0.806	−0.220	0.043	−0.816	−1.011	0.810
2	ss66185516	239,065,169	G/T	0.397	0.074	−0.053	0.228	1.039	0.852	0.687	0.269
3	ss66397464	73,251,862	C/T	0.415	0.183	−0.198	−0.192	0.677	0.895	0.437	−0.212
4	ss66402098	186,281,132	T/C	−0.225	0.630	0.565	−0.009	0.418	0.651	0.744	0.244
6	ss66218814	3,311,818	C/T	−0.724	0.043	0.159	0.182	0.237	−0.795	−1.056	−0.336
7	ss66083459	89,430,534	T/G	−0.744	0.518	0.336	0.141	0.377	−1.070	−1.555	−1.214
12	ss66288005	29,860,263	A/G	−0.342	0.724	0.541	0.322	0.096	0.563	0.875	0.806
14	ss66282595	24,339,998	G/A	1.461	1.471	1.217	−0.246	−0.588	−1.194	−0.973	0.022
14	ss66411959	24,340,175	G/A	−0.782	−1.357	−1.311	−0.080	0.307	0.323	0.068	−0.276
14	ss66416767	24,348,496	G/T	−0.232	−0.589	−0.117	−0.033	0.507	0.610	−0.098	−0.559
14	ss66281419	77,702,561	G/A	−0.802	0.151	0.488	0.252	0.438	−0.109	0.254	−0.189
16	ss66091573	57,829,089	C/T	1.402	2.674	1.450	0.400	0.999	2.747	1.859	0.426
16	ss66242525	57,935,351	C/T	−0.548	−0.516	0.465	0.579	−0.537	−0.479	0.093	1.234
16	ss66489647	57,938,934	A/G	−0.217	−2.058	−1.834	−1.059	0.595	−0.350	−1.070	−0.967
16	ss66444701	82,976,515	C/G	−0.672	−0.259	0.831	0.481	0.639	1.552	0.876	−0.143
16	ss66529263	84,383,030	G/T	0.478	0.539	−0.033	0.301	0.066	−0.313	−0.037	−0.111
23	ss66369851	121,966,143	G/T	−0.419	−0.108	−0.052	−0.025	0.050	−0.663	−0.454	−0.185

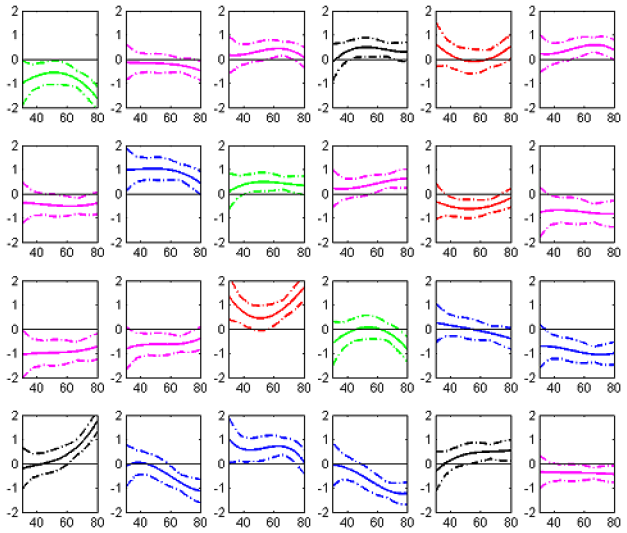

FIG. 2. *Additive effects of selected SNPs in the real data example.*

varying effects and implement variable selection. The procedure extends the standard Bayesian lasso [Park and Casella (2008)] and standard group lasso [Yuan and Lin (2006)] to a nonparametric setting, and is applicable to irregular longitudinal data.

We approximated time-varying effects by Legendre polynomials and presented a Bayesian hierarchical model with group lasso penalties that encourages sparse

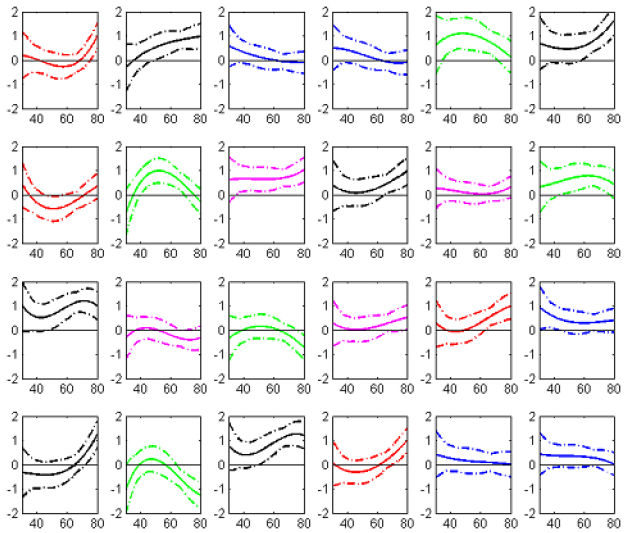

FIG. 3. *Dominant effects of selected SNPs in the real data example.*

solutions at the group level. The group lasso penalties are introduced by assigning multivariate Laplace priors to regression coefficients, and are implemented on the basis of its hierarchical expansion which yields an efficient Gibbs sampler in the MCMC estimation. Although computationally intensive, it outperforms the standard group lasso in the sense that it provides not only point estimates but also interval estimates of all parameters. In addition, the Bayesian group lasso treats the regularization parameters as unknown hyperparameters and estimates them along with other parameters. This technique avoids choosing the tuning parameters by cross-validation and automatically accounts for the uncertainty in its selection that affects the estimates of regression coefficients.

In one of the most powerful but challenging areas in genetics, we incorporated our new procedure to genome-wide association studies (GWAS) by testing a large number of SNPs simultaneously, particularly with $p \gg n$, based on the dynamic pattern of genetic effects on complex phenotypes or diseases. We first applied the new approach to *f*GWAS for age-specific changes of BMI and successfully identified several significant SNPs, some of which are confirmed by empirical genetic studies [Frayling (2007)]. For example, previous molecular studies have observed a candidate gene (*FTO*) coding alpha-ketoglutarate-dependent dioxygenase, a fat mass and obesity-associated protein. Our model detected SNPs ss66091573, ss66242525 and ss66489647 on chromosome 16 in a region of the FTO gene, suggesting the biological relevance of these SNPs in fat-related trait control. Our model also detected other SNPs in close proximity of different candidate genes; that is, SNP ss66397464 in peroxisome proliferator-activated receptor-γ gene (PPARG) on chromosome 3, SNP ss66402098 in the Wolfram syndrome 1 gene (WFSI) on chromosome 4, SNP ss66218814 in CDK5 regulatory-subunit-associated protein 1-like 1 gene (CDKAL1) on chromosome 6, and SNP ss66288005 in potassium inwardly-rectifying channel, subfamily J, member 11 gene (KCNJ11) on chromosome 12 [Frayling (2007)]. Among these four genes, PPARG and KCNJ were found to be associated with obesity [Vidal-Puig et al. (1997); Morgan et al. (2010)], while WFSI and CDKAL1 are believed to be associated with diabetes [Sandhu et al. (2007); Scott et al. (2007); Steinthorsdottir et al. (2007)]. Therefore, all these discoveries have well validated the biological relevance of the new model.

To address challenges for the post-GWAS era, genetic association studies began to focus on SNPs within a set of functional candidate genes. For instance, Michel et al. (2010) analyzed 566 SNPs from 14 candidate genes that are believed to be associated with asthma. Xu and Taylor (2009) developed tools to recommend SNPs based on information on gene expression studies, regulatory pathways and functional regions that appear to be linked to the disease. In their example, 1361 SNPs were recommended for a genetic association study on prostate cancer. These tools could be used as a preprocessing step for the proposed procedure in this article. Statistically, on the other hand, variable screening approaches [Fan and Lv (2008)] for longitudinal data can be developed to recommend a subset of SNPs.

From a theoretical point of view, the proposed method can also approximate varying-coefficients by nonparametric techniques other than Legendre polynomials, and model the within-subject correlation by other parametric or nonparametric covariance structures. Given its potential influence, an optimal model for longitudinal covariance structure should be chosen based on the nature of practical data [Zhao et al. (2005); Yap, Fan and Wu (2009)]. More generally, it can be easily extended to the problem where the number of variables in each group varies, such as the multi-factor ANOVA with each factor having several levels. Also, gene-gene interactions and gene-environment interactions can be incorporated to better decipher a detailed picture of the genetic architecture of a complex trait.

Acknowledgments. Jiahan Li and Zhong Wang contributed equally to this work. The authors are grateful to the Editor, the Associate Editor and three anonymous referees for providing valuable comments that significantly improved the paper.

SUPPLEMENTARY MATERIAL

Convergence diagnostics and summary of parameter estimates (DOI: 10.1214/15-AOAS808SUPP; .pdf). We plot the potential scale reduction factor (PSRF) against iterations and summarize the average estimates, standard errors and mean squared errors (MSEs) of corresponding Legendre coefficients for the first five genetic predictors.

REFERENCES

CHO, S., KIM, H., OH, S., KIM, K. and PARK, T. (2009). Elastic-net regularization approaches for genome-wide association studies of rheumatoid arthritis. *BMC Proc*. **3 Suppl 7** S25.

CUI, Y., WU, R., CASELLA, G. and ZHU, J. (2008). Nonparametric functional mapping of quantitative trait loci underlying programmed cell death. *Stat. Appl. Genet. Mol. Biol*. **7** Art. 4, 32. MR2386321

DALY, A. K. (2010). Genome-wide association studies in pharmacogenomics. *Nat. Rev. Genet*. **11** 241–246.

DAS, K., LI, J., WANG, Z., FU, G., LI, Y., MAUGER, D., LI, R. and WU, R. (2011). A dynamic model for genome-wide association studies. *Hum. Genet*. **129** 629–639.

DAWBER, T. R., MEADORS, G. F. and MOORE, F. E. JR. (1951). Epidemiological approaches to heart disease: The framingham study. *Am. J. Publ. Health* **41** 279–286.

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **70** 849–911. MR2530322

FILIAULT, D. L. and MALOOF, J. N. (2012). A genome-wide association study identifies variants underlying the arabidopsis thaliana shade avoidance response. *PLoS Genet*. **8** e1002589.

FRAYLING, T. M. (2007). Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat. Rev. Genet*. **8** 657–662.

FRAYLING, T. M., TIMPSON, N. J., WEEDON, M. N., ZEGGINI, E., FREATHY, R. M., LINDGREN, C. M., PERRY, J. R. B., ELLIOTT, K. S., LANGO, H., RAYNER, N. W., SHIELDS, B., HARRIES, L. W., BARRETT, J. C., ELLARD, S., GROVES, C. J., KNIGHT, B., PATCH, A.-M., NESS, A. R., EBRAHIM, S., LAWLOR, D. A., RING, S. M., BEN-SHLOMO, Y.,

JARVELIN, M.-R., SOVIO, U., BENNETT, A. J., MELZER, D., FERRUCCI, L., LOOS, R. J. F., BARROSO, I., WAREHAM, N. J., KARPE, F., OWEN, K. R., CARDON, L. R., WALKER, M., HITMAN, G. A., PALMER, C. N. A., DONEY, A. S. F., MORRIS, A. D., SMITH, G. D., HATTERSLEY, A. T. and MCCARTHY, M. I. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316** 889–894.

GELMAN, A. and RUBIN, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statist. Sci.* **7** 457–472.

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall, Boca Raton, FL. MR2027492

GORLOVA, O. Y., AMOS, C. I., WANG, N. W., SHETE, S., TURNER, S. T. and BOERWINKLE, E. (2003). Genetic linkage and imprinting effects on body mass index in children and young adults. *European Journal of Human Genetics* **11** 425–432.

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

HE, Q. and LIN, D.-Y. (2011). A variable selection method for genome-wide association studies. *Bioinformatics* **27** 1–8.

JAQUISH, C. E. (2007). The framingham heart study, on its way to becoming the gold standard for cardiovascular genetic epidemiology? *BMC Med. Genet.* **8** 63.

JOOD, K., JERN, C., WILHELMSEN, L. and ROSENGREN, A. (2004). Body mass index in mid-life is associated with a first stroke in men: A prospective population study over 28 years. *Stroke* **35** 2764–2769.

LETTRE, G. (2011). Recent progress in the study of the genetics of height. *Human Genetics* **129** 465–472.

LI, J., DAS, K., FU, G., LI, R. and WU, R. (2012). Bayesian lasso for genome-wide association studies. *Bioinformatics* **27** 516–523.

LI, J., WANG, Z., LI, R. and WU, R. (2015). Supplement to "Bayesian group Lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies." DOI:10.1214/15-AOAS808SUPP.

LIN, M. and WU, R. (2006). A joint model for nonparametric functional mapping of longitudinal trajectory and time-to-event. *BMC Bioinformatics* **7** 138.

LIN, Y. and ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34** 2272–2297. MR2291500

LYNCH, M. and WALSH, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer, Sunderland, MA.

MA, C. X., CASELLA, G. and WU, R. L. (2002). Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. *Genetics* **161** 1751–1762.

MICHEL, S., LIANG, L., DEPNER, M., KLOPP, N., RUETHER, A., KUMAR, A., SCHEDEL, M., VOGELBERG, C., VON MUTIUS, E., VON BERG, A., BUFE, A., RIETSCHEL, E., HEINZMANN, A., LAUB, O., SIMMA, B., FRISCHER, T., GENUNEIT, J., GUT, I. G., SCHREIBER, S., LATHROP, M., ILLIG, T. and KABESCH, M. (2010). Unifying candidate gene and GWAS approaches in asthma. *PLoS ONE* **5** e13894.

MORGAN, A. R., THOMPSON, J. M., MURPHY, R., BLACK, P. N., LAM, W. J., FERGUSON, L. R. and MITCHELL, E. A. (2010). Obesity and diabetes genes are associated with being born small for gestational age: Results from the auckland birthweight collaborative study. *BMC Medical Genetics* **11** 125.

PARK, T. and CASELLA, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103** 681–686. MR2524001

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

SANDHU, M. S., WEEDON, M. N., FAWCETT, K. A., WASSON, J., DEBENHAM, S. L., DALY, A., LANGO, H., FRAYLING, T. M., NEUMANN, R. J., SHERVA, R., BLECH, I., PHAROAH, P. D., PALMER, C. N. A., KIMBER, C., TAVENDALE, R., MORRIS, A. D., MCCARTHY, M. I., WALKER, M., HITMAN, G., GLASER, B., PERMUTT, M. A., HATTERSLEY, A. T., WAREHAM, N. J. and BARROSO, I. (2007). Common variants in WFS1 confer risk of type 2 diabetes. *Nat. Genet.* **39** 951–953.

SCOTT, L. J., MOHLKE, K. L., BONNYCASTLE, L. L., WILLER, C. J., LI, Y., DUREN, W. L., ERDOS, M. R., STRINGHAM, H. M., CHINES, P. S., JACKSON, A. U., PROKUNINA-OLSSON, L., DING, C.-J., SWIFT, A. J., NARISU, N., HU, T., PRUIM, R., XIAO, R., LI, X.-Y., CONNEELY, K. N., RIEBOW, N. L., SPRAU, A. G., TONG, M., WHITE, P. P., HETRICK, K. N., BARNHART, M. W., BARK, C. W., GOLDSTEIN, J. L., WATKINS, L., XIANG, F., SARAMIES, J., BUCHANAN, T. A., WATANABE, R. M., VALLE, T. T., KINNUNEN, L., ABECASIS, G. R., PUGH, E. W., DOHENY, K. F., BERGMAN, R. N., TUOMILEHTO, J., COLLINS, F. S. and BOEHNKE, M. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316** 1341–1345.

SHULDINER, A. R. et al. (2009). Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. *J. Am. Med. Assoc.* **302** 849–857.

STEINTHORSDOTTIR, V., THORLEIFSSON, G., REYNISDOTTIR, I., BENEDIKTSSON, R., JONSDOTTIR, T., WALTERS, G. B., STYRKARSDOTTIR, U., GRETARSDOTTIR, S., EMILSSON, V., GHOSH, S., BAKER, A., SNORRADOTTIR, S., BJARNASON, H., NG, M. C. Y., HANSEN, T., BAGGER, Y., WILENSKY, R. L., REILLY, M. P., ADEYEMO, A., CHEN, Y., ZHOU, J., GUDNASON, V., CHEN, G., HUANG, H., LASHLEY, K., DOUMATEY, A., SO, W.-Y., MA, R. C. Y., ANDERSEN, G., BORCH-JOHNSEN, K., JORGENSEN, T., VAN VLIET-OSTAPTCHOUK, J. V., HOFKER, M. H., WIJMENGA, C., CHRISTIANSEN, C., RADER, D. J., ROTIMI, C., GURNEY, M., CHAN, J. C. N., PEDERSEN, O., SIGURDSSON, G., GULCHER, J. R., THORSTEINSDOTTIR, U., KONG, A. and STEFANSSON, K. (2007). A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.* **39** 770–775.

SUCHOCKI, T. and SZYDA, J. (2011). Statistical modelling of growth using a mixed model with orthogonal polynomials. *J. Appl. Genet.* **52** 95–100.

TAKEUCHI, F., MCGINNIS, R., BOURGEOIS, S., BARNES, C., ERIKSSON, N., SORANZO, N., WHITTAKER, P., RANGANATH, V., KUMANDURI, V., MCLAREN, W., HOLM, L., LINDH, J., RANE, A., WADELIUS, M. and DELOUKAS, P. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. *PLoS Genet.* **5** e1000433.

TEICHERT, M., EIJGELSHEIM, M., RIVADENEIRA, F., UITTERLINDEN, A. G., VAN SCHAIK, R. H. N., HOFMAN, A., SMET, P. A. G. M. D., VAN GELDER, T., VISSER, L. E. and STRICKER, B. H. C. (2009). A genome-wide association study of acenocoumarol maintenance dosage. *Hum. Mol. Genet.* **18** 3758–3768.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

VIDAL-PUIG, A. J., CONSIDINE, R. V., JIMENEZ-LIÑAN, M., WERMAN, A., PORIES, W. J., CARO, J. F. and FLIER, J. S. (1997). Peroxisome proliferator-activated receptor gene expression in human tissues. Effects of obesity, weight loss, and regulation by insulin and glucocorticoids. *J. Clin. Invest.* **99** 2416–2422.

WANG, L., LI, H. and HUANG, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *J. Amer. Statist. Assoc.* **103** 1556–1569. MR2504204

WANG, Z., LI, Y., LI, Q. and WU, R. (2009). Joint functional mapping of quantitative trait loci for HIV-1 and $CD4^+$ dynamics. *Int. J. Biostat.* **5** Art. 9, 26. MR2491436

WU, R. and LIN, M. (2006). Functional mapping—How to map and study the genetic architecture of dynamic complex traits. *Nature Review Genetics* **7** 229–237.

WU, R., MA, C.-X., LIN, M., WANG, Z. and CASELLA, G. (2004). Functional mapping of quantitative trait loci underlying growth trajectories using a transform-both-sides logistic model. *Biometrics* **60** 729–738. MR2089449

WU, T. T., CHEN, Y. F., HASTIE, T., SOBEL, E. and LANGE, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25** 714–721.

XU, Z. and TAYLOR, J. A. (2009). SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res*. **37(suppl 2)** W600–W605.

YANG, R. and XU, S. (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics* **176** 1169–1185.

YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. and VISSCHER, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet*. **42** 565–569.

YAP, J. S., FAN, J. and WU, R. (2009). Nonparametric modeling of longitudinal covariance structure in functional mappings of quantitative trait loci. *Biometrics* **65** 1068–1077. MR2756494

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **68** 49–67. MR2212574

ZHANG, H. H. and LIN, Y. (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statist. Sinica* **16** 1021–1041. MR2281313

ZHAO, W., CHEN, Y. Q., CASELLA, G., CHEVERUD, J. M. and WU, R. L. (2005). A nonstationary model for functional mapping of complex traits. *Bioinformatics* **21** 2469–2477.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol*. **67** 301–320. MR2137327

J. LI
DEPARTMENT OF APPLIED AND COMPUTATIONAL MATHEMATICS AND STATISTICS
UNIVERSITY OF NOTRE DAME
NOTRE DAME, INDIANA 46556
USA
E-MAIL: jli7@nd.edu

R. LI
DEPARTMENT OF STATISTICS
THE METHODOLOGY CENTER
PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PENNSYLVANIA 16802
USA
E-MAIL: rzli@psu.edu

Z. WANG
CENTER FOR COMPUTATIONAL BIOLOGY
BEIJING FORESTRY UNIVERSITY
BEIJING
CHINA 100083
E-MAIL: zhongwang@bjfu.edu.cn

R. WU
CENTER FOR STATISTICAL GENETICS
PENNSYLVANIA STATE UNIVERSITY
HERSHEY, PENNSYLVANIA 17033
USA
AND
CENTER FOR COMPUTATIONAL BIOLOGY
BEIJING FORESTRY UNIVERSITY
BEIJING
CHINA 10008
E-MAIL: rwu@hes.hmc.psu.edu