Statistics and Its Interface Volume 15 (2022) 181–196

Stochastic functional linear models for gene-based association analysis of quantitative traits in longitudinal studies

Bingsong Zhang, Shuqi Wang, Xiaohan Mei, Yue Han, Runqiu Wang, Hong-Bin Fang, Chi-Yang Chiu, Jun Ding, Zuoheng Wang, Alexander F. Wilson, Joan E. Bailey-Wilson, Momiao Xiong, and Ruzong Fan*

Longitudinally measured phenotypes are important for exploring genetic and environmental factors that affect complex traits over time. Genetic analysis of multiple measures in longitudinal studies provides a valuable opportunity to understand genetic architecture and biological variations of complex diseases. In this paper, stochastic functional linear models are developed for temporal association analysis at gene levels to analyze sequence data and longitudinally measured quantitative traits. Functional data analysis techniques are utilized to reduce high dimensionality of sequence data and draw useful information. A variance-covariance structure is constructed to model the measurement variation and correlations of the traits based on the theory of stochastic processes. Spline models are used to estimate the time-dependent trajectory mean function. By intensive simulation studies, it is shown that the proposed stochastic models control type I errors well, and have higher power levels than those of the perturbation tests. In addition, the proposed methods are robust when the correlation function is mis-specified. We test and refine the models and related software using real data sets of Framingham Heart Study.

Keywords and phrases: Rare variants, Sequence data, Association mapping, Quantitative trait loci, Longitudinal studies, Stochastic models, Functional data analysis.

1. INTRODUCTION

Longitudinally measured phenotypes are important for exploring key genetic and environmental factors that affect complex traits over time. Genetic analysis of multiple measures in longitudinal studies provides a valuable opportunity to understand genetic architecture and biological variations of complex diseases. Many genetic studies have been conducted in cohorts in which repeated measures on the trait of interest are collected on each participant over a period of time and sequence data are available [1, 2, 3]. Such studies not only provide a more accurate assessment of disease condition, but enable us to investigate gene's influencing on the trajectory of a trait and disease progression, which are likely to help reduce the remaining missing heritability of these traits [4, 5].

Although they are important, there is a paucity of statistical methods to analyze human genetic sequence data in longitudinal studies. The sequence data consist of rare variants, or common variants, or a combination of both, where the minor allele frequencies (MAFs) of rare variants are less than 0.01∼0.05. It is important to develop powerful methods to analyze sequence data in longitudinal studies. The genetic variants of an individual are assumed to be fixed due to low probability of mutation. However, the phenotypic traits and associated genetic effect vary with time. Statistical models which may better use longitudinal data and may reflect temporal trends of traits are needed. For many traits of complex diseases, genetic determinants can be important at some time period. At other time periods, environmental factors can be more important. It is important to develop models which can reflect the genetic effect as well as environmental effect on the traits over the time.

In this paper, stochastic functional linear models are developed for temporal association analysis at the gene level to analyze quantitative traits in longitudinal studies. In the presence of a large number of rare variants, gene-based analysis is a more powerful tool for gene mapping than testing of individual genetic variants. In the analysis of a single time measurement, functional regression models were found to perform markedly better than sequence kernel association tests (SKAT) procedure when the genetic effects of a gene are relatively large while SKAT procedure performs better in analysis of polygenes [6, 7, 8, 9, 10].

The functional models have not been developed to analyze longitudinal traits. To perform gene-based analysis of sequencing data for longitudinal traits, there are two major difficulties: (1) high dimension sequence data and (2) variation and correlation of longitudinally measured phenotypes. For unrelated individuals in the sample, it is reasonable to

*Corresponding author.

assume that their quantitative traits are independent. For the same individual, the quantitative traits at different times depend on each other, however. Hence, it is necessary to consider the variance-covariance structure carefully. In the literature of functional regression models, there are no methods which model variation and correlation of longitudinally measured phenotypes [11, 12, 13, 14, 15, 16].

To fill the gap, stochastic functional regression models are developed to analyze longitudinally measured quantitative traits. To analyze sequence data, functional data analysis techniques are utilized to reduce high dimensionality of sequencing data. A variance-covariance structure is constructed to model the measurement variation and correlations of an individual based on the theory of stochastic processes [17, 18]. The stochastic models can capture the temporal trend and detect the temporal genetic effects of complex traits. Spline models are used to estimate the time-dependent mean function [19, 20, 21, 22].

To evaluate the performance of the proposed stochastic models, extensive simulations are performed to calculate the empirical type I errors to check if false positives are well controlled. In addition we compare power performance with the perturbation tests proposed in He et al. (2017) [23], i.e., P-Dispersion, P-Burden, P-Fisher, and P-MinP. We test and refine the models and related software using real data sets of Framingham Heart Study (FHS).

2. MODELS AND METHODS

In the following, we are going to present a stochastic functional linear model to analyze quantitative traits. The variance-covariance structure is constructed to describe the trait variation and to properly account for correlation between multiple measurements on the same subject. Spline models are used to approximate temporal mean function and regression coefficients.

2.1 Stochastic functional linear models

Consider a population sample with n individuals. Assume that the n individuals are sequenced in a genomic region that has m variants. We assume that the m variants are located in a region with ordered physical positions $0 \leq u_1 < \cdots < u_m$, and that each variant's physical position u_ℓ is known, e.g., in terms of base pair positions. To make the notation simple, we normalize the region $[u_1, u_m]$ to be [0, 1]. For individual i, let $y_i(t)$ be his/her quantitative trait value at time t and the time t can be age of the individual. In addition, let $G_i = (g_i(u_1), \cdots, g_i(u_m))'$ denote the genotype of the m variants, and $Z_i(t) = (z_{i1}, \cdots, z_{ip}, z_{i,p+1}(t), \cdots, z_{i,p+q}(t))'$ denote a $(p+q) \times 1$ vector of covariates such as gender and age at the time t. Note that the covariates include time invariant variables $(z_{i1}, \cdots, z_{ip})'$ and time varying variables $(z_{i,p+1}(t), \cdots, z_{i,p+q}(t))'$. For the genotypes, we assume that $g_i(u_\ell)$ $(= 0, 1, 2)$ is the number of minor alleles of the individual at the ℓ-th variant located at the position u_ℓ.

We denote the i-th individual's genetic variant function (GVF) as $X_i(u), u \in [0,1]$. Note that the data set includes n discrete realizations or observations G_i of the genotypes, one for each individual. Using the genetic variant information G_i, we may estimate the related genetic variant function $X_i(u)$, which will be discussed below. A stochastic functional linear model at the time t can be defined as

(1)

$$y_i(t) = \mu(t) + Z_i(t)'\alpha(t) + \int_0^1 X_i(u)\beta(u)du + U_i(t) + \varepsilon_i(t),$$

where $\mu(t)$ is an overall mean at time t, $\alpha(t)$ is a vector of regression coefficient functions of the covariates $Z_i(t)$ at time t, $\beta(u)$ is a genetic effect function of the GVF $X_i(u)$. In model (1), $U_i(t)$ is an unknown random function due to both genetic and environmental factors of an individual, and $\varepsilon_i(t)$ is a random residual function. Assume that $U_i(t)$ and $\varepsilon_i(t)$ are independent. Moreover, assume that $\varepsilon_i(t)$ is normal $N(0, \sigma_e^2)$.

The model (1) has three major features. First, the overall mean $\mu(t)$ is a function of time t and it is unlikely a constant. In literature, it was found that mis-specification of $\mu(t)$ can lead to biased and unstable results [19, 23]. In practice, it is almost impossible to correctly specify the true mean function. In the previous study [19, 20], it was found that the non-parametric linear penalized spline model is a good choice to estimate $\mu(t)$. Second, the integration term $\int_0^1 X_i(u)\beta(u)du$ is used to model the genetic effect of the variants as previous work [6, 7, 8, 9, 10]. Third, one major difference of model (1) from functional linear models in literature is that we use random function $U_i(t)$ to model variation and correlation of stochastic process $y_i(t)$ [11, 12, 13, 15, 16].

2.2 Estimation of genetic variant functions

To estimate genetic variant functions $X_i(u)$ from the genotypes G_i, we use an ordinary linear square smoother [6, 14]. Let $\phi_k(u), k = 1, \cdots, K$, be a series of K basis functions, such as the B-spline basis and Fourier basis functions. Let $\mathbf{A}$ denote the $m \times K$ matrix containing the values $\phi_k(u_\ell)$, and we let $\phi(u) = (\phi_1(u), \cdots, \phi_K(u))'$. Using the discrete realizations $G_i = (g_i(u_1), \cdots, g_i(u_m))'$, we estimate the genetic variant function $X_i(u)$ using an ordinary linear square smoother as follows

(2)

$$\hat{X}_i(u) = (g_i(u_1), \cdots, g_i(u_m))\mathbf{A}[\mathbf{A}'\mathbf{A}]^{-1}\phi(u).$$

We consider two types of basis functions: (1) the B-spline basis: $\phi_k(u) = B_k(u), k = 1, \cdots, K$; and (1) the Fourier basis: $\phi_1(u) = 1, \phi_{2r+1}(u) = \sin(2\pi ru)$, and $\phi_{2r}(u) = \cos(2\pi ru)$, $r = 1, \cdots, (K-1)/2$. Here for Fourier basis, K is taken as a positive odd integer [11, 12, 13].

2.3 Estimation of genetic effect function $\beta(u)$

The genetic effect function $\beta(u)$ in the stochastic functional linear model (1) is assumed to be smooth, i.e., $\beta(u)$ is a continuous function. One may expand it by B-spline or Fourier basis functions. Formally, let us expand the genetic effect function $\beta(u)$ by a series basis functions $\Phi'(u) = (\phi_1(u), \cdots, \phi_{K_\beta}(u))$ as

$$\beta(u) = \sum_{k=1}^{K_\beta} \phi_k(u)\beta_k \quad = \quad \Phi'(u)\beta, \tag{3}$$

where the coefficients for the expansion are in a vector $\beta = (\beta_1, \cdots, \beta_{K_\beta})'$.

2.4 Estimation of mean function $\mu(t)$

To estimate mean function $\mu(t)$ and genetic effect function $\beta(u)$, one may treat them by either non-random expansion or random spline estimations. We discuss it by only talking about the estimation of $\mu(t)$.

Penalized spline estimations We may approximate $\mu(t)$ and $\beta(u)$ by linear combinations of penalized spline functions [20]. For instance, the q-order penalized spline model for $\mu(t)$ is

$$\mu(t) = \mu_0 + t\mu_1 + \cdots + t^q\mu_q + \sum_{k=1}^{K_\mu} \nu_k(t-\kappa_k)_+^q, \tag{4}$$

where $\mu_i, i = 0, 1, \cdots, q, q \geq 1$, are fixed effects, and $\nu_k, k = 1, 2, \cdots, K_\mu$, are identically and independently normal distributed random variables, $\kappa_k, k = 1, 2, \cdots, K_\mu$, is a pre-assigned sequence of knots, K_μ is the number of knots, and q is the order of the spline. In addition, $(t-\kappa_k)_+^q = \begin{cases} (t-\kappa_k)^q & \text{if } t-\kappa_k > 0 \\ 0 & \text{else} \end{cases}$. Let $\nu = (\nu_1, \cdots, \nu_{K_\mu})^\tau$. Assume that $\text{Cov}(\nu) = \sigma_\nu^2 I_{K_\mu}$, where I_{K_μ} is the identity matrix of rank K_μ.

Non-random expansion by basis functions The mean function $\mu(t)$ is assumed to be continuous. One may expand it by a series of K_μ basis functions $\psi_1(t), \cdots, \psi_{K_\mu}(t)$ as

$$\mu(t) = (\psi_1(t), \cdots, \psi_{K_\mu}(t))(\mu_1, \cdots, \mu_{K_\mu})' \quad = \quad \psi(t)'\mu, \tag{5}$$

where $\mu = (\mu_1, \cdots, \mu_{K_\mu})'$ is a $K_\mu \times 1$ vector of coefficients and $\psi(t) = (\psi_1(t), \cdots, \psi_{K_\mu}(t))'$.

2.5 Variance-covariance structure

The variance-covariance structure of stochastic processes $y_i(t)$ depends on the time [24]. Let $\sigma_U^2(t) = \text{Var}(U_i(t))$ be the variance of $U_i(t)$ at the time t. For a pair of time points t and s, let us denote correlation between $U_i(t)$ and $U_i(s)$ by $\rho_U(s,t)$. Then, the covariance between $U_i(t)$ and $U_i(s)$ is

$$\sigma_U(t,s) = \text{Cov}(U_i(t), U_i(s)) = \sigma_U(t)\sigma_U(s)\rho_U(s,t).$$

The variance-covariance structure of stochastic process $y_i(t)$ is characterized by

$$\text{Cov}(y_i(t), y_i(s)) = \begin{cases} \sigma_U^2(t) + \sigma_e^2 & \text{if } t = s \\ \sigma_U(t,s) & \text{if } t \neq s \end{cases}. \tag{6}$$

In the above formulation, the covariance $\text{Cov}(y_i(t), y_i(s))$ is assumed to be equal to the covariance of $U_i(t)$ and $U_i(s), t \neq s$. In practice, the correlation between $y_i(t)$ and $y_i(s)$ can be from the genetic and environmental factors or their combinations. For the population data, it is impossible to distinguish them. Hence, we simply put it as the correlation effect.

Suppose that the correlation functions $\rho_U(s,t)$ is a function of $|t-s|$, i.e., they are functions of the time range. This is true if $U_i(t)$ is stationary or second-order stationary [17]. For instance, assume that the correlation effect is an Ornstein-Uhlenbeck Gaussian process $U_i(t) = \lambda \exp(-t/\rho) W_i(\frac{2}{\rho} e^{2t/\rho}), \rho > 0$, where $W_i(t)$ is a standard Brownian motion, ρ is a range parameter, and λ is a scaling parameter. Then clearly, $U_i(t)$ has zero mean at all times t and constant variance. Moreover, the correlation function is $\rho_U(t,s) = \exp(-\frac{|s-t|}{\rho}) = \theta^{|t-s|}$, where $\theta = \exp(-1/\rho)$ indicates that the correlation decreases exponentially with the time range [17]. In this case, the correlation effect $U_i(t)$ is a stationary Gaussian process. We are particularly interested in the Ornstein-Uhlenbeck Gaussian process $U_i(t)$ for three reasons. First, it basically assume that the correlation of two measurements of an individual declines exponentially with the time range. This is a reasonable assumption in many situations. Second, we can fit the models conveniently in R using linear mixed model functions [25]. Third, we fitted models by assuming linear correlation in data analysis of single SNP analysis, but they lead to higher Akaike information criterion (AIC) and Bayesian information criterion (BIC) values and so the models are not as good as the Ornstein-Uhlenbeck process modeling [19].

In addition, one may use a linear correlation function $\rho_U(s,t) = \left(1 - \frac{|s-t|}{\rho}\right) 1_{(|s-t|<\rho)}$ and a Gaussian correlation function $\rho_U(s,t) = \exp\left[-\left(\frac{|s-t|}{\rho}\right)^2\right], \rho > 0$, to fit the model. In practice, true correlation function is never known. Robust statistical methods are needed for data analysis.

In certain cases, however, the covariance or correlation functions may not be functions of the time range. In this case, the correlation effect $U_i(t)$ is a non-stationary process. For instance, assume that the correlation effect is a Wiener process $U_i(t) = \theta_1 W_i(t)$, where $W_i(t)$ is a standard Brownian motion. Then $U_i(t)$ has zero mean at all times t. The covariance function is $\sigma_U(t,s) = \theta_1^2 \min(t,s)$.

2.6 Beta-smooth only stochastic functional linear model

To remove the assumption of the continuity of the GVF $X_i(u)$ in the stochastic functional linear model (1), a simplified functional linear model is obtained by replacing the integration term $\int_0^1 X_i(u)\beta(u)du$ in model (1) by the summation term $\sum_{\ell=1}^m g_i(u_\ell)\beta(u_\ell)$. Then, the model (1) can be revised as a beta-smooth only model

$$y_i(t) = \mu(t)+Z_i(t)'\alpha(t)+\sum_{\ell=1}^m g_i(u_\ell)\beta(u_\ell)+U_i(t)+\varepsilon_i(t). \tag{7}$$

In models (1) and (7), the overall mean $\mu(t)$ and coefficient functions $\alpha(t)$ and random function $U_i(t)$ depend on time t and they can capture temporal trends of the traits $y_i(t)$.

2.7 Dealing with missing genotype data

If some genotype data are missing, the stochastic models (1) and (7) can be modified to analyze the data. For example, assume there is no genotype information at the first variant for the i-th individual, i.e., $g_i(u_1)$ is missing in G_i. Let $\mathbf{A_1}$ denote the $m-1$ by K matrix containing values $\phi_k(u_\ell)$, where $\ell \in 2, \cdots, m$. Then, we may revise the estimation (2) as

$$\hat{X}_i(u) = \left(g_i(u_2), \cdots, g_i(u_m)\right)' \left[\mathbf{A_1'A_1}\right]^{-1}\mathbf{A_1'}\phi(u). \tag{8}$$

Note that the estimation (8) only depends on the available genotype data $\left(g_i(u_2), \cdots, g_i(u_m)\right)'$. Hence, each individual's GVF is estimated by his/her own data, a practical advantage of functional data analysis approach. Using the estimation (8), one may revise stochastic model (1) accordingly. If $g_i(u_1)$ is missing in G_i, we may revise the beta-smooth only model (7) as

$$y_i(t) = \mu(t)+Z_i(t)'\alpha(t)+\sum_{\ell=2}^m g_i(u_\ell)\beta(u_\ell)+U_i(t)+\varepsilon_i(t). \tag{9}$$

The revised model (9) only depends on the available genotype data $\left(g_i(u_2), \cdots, g_i(u_m)\right))'$.

2.8 Revised stochastic functional linear models

Replacing $X_i(u)$ in the stochastic model (1) with $\hat{X}_i(u)$ in (2), $\beta(u)$ with the expansion (3), and $\mu(t)$ by penalized spline (4) or non-random expansion (5), we have the following revised model

$$\begin{aligned} y_i(t) &= \mu(t) + Z_i(t)'\alpha(t) + (g_i(u_1), \cdots, g_i(u_m)) \cdot \\ &\quad \mathbf{A}[\mathbf{A'A}]^{-1}\int_0^1 \phi(u)\Phi'(u)du\beta + U_i(t) + \varepsilon_i(t) \\ &= \mu(t) + Z_i(t)'\alpha(t) + W_i'\beta + U_i(t) + \varepsilon_i(t), \qquad (10) \end{aligned}$$

where $W_i' = (g_i(u_1), \cdots, g_i(u_m))\mathbf{A}[\mathbf{A'A}]^{-1}\int_0^1 \phi(u)\Phi'(u)du$. In the statistical packages R *fda* or Matlab, codes to calculate $\mathbf{A}[\mathbf{A'A}]^{-1}$ and $\int_0^1 \phi(u)\Phi'(u)du$ are readily available [14].

For the beta-smooth only stochastic model (7), $\beta(u_\ell)$ is introduced as the genetic effect at the position u_ℓ. Expanding $\beta(u_\ell)$ by B-spline or Fourier basis functions as above, the stochastic model (7) can be revised as

$$\begin{aligned} y_i(t) &= \mu(t) + Z_i(t)'\alpha(t) + \\ &\quad \sum_{\ell=1}^m g_i(u_\ell)\left(\phi_1(u_\ell), \cdots, \phi_{K_\beta}(u_\ell)\right)(\beta_1, \cdots, \beta_{K_\beta})' \\ &\quad +U_i(t) + \varepsilon_i(t) \\ &= \mu(t) + Z_i(t)'\alpha(t) + W_i'\beta + U_i(t) + \varepsilon_i(t), \qquad (11) \end{aligned}$$

where $W_i' = \sum_{\ell=1}^m g_i(u_\ell)\left(\phi_1(u_\ell), \cdots, \phi_{K_\beta}(u_\ell)\right)$.

2.9 Likelihood ratio test (LRT) statistics and test procedure

To test for association between the quantitative trait and the m genetic variants, the null hypothesis is $H_0 : \beta = (\beta_1, \cdots, \beta_{K_\beta})' = 0$. Under the null, the stochastic models (10) and (11) are simplified as

$$y_i(t) = \mu(t) + Z_i(t)'\alpha(t) + U_i(t) + \varepsilon_i(t). \tag{12}$$

The stochastic models (10) or (11) and the null model (12) are nested. By fitting (10) or (11) and the null model (12), we may test the null $H_0 : \beta = 0$ by a χ^2-distributed likelihood ratio test (LRT) statistic with K_β degrees of freedom using the *lme* R package [25].

In *lme* R package, the exponential correlation function can be fitted by *correlation* = *corExp*, the linear correlation function can be fitted by *correlation* = *corLin*, and the Gaussian correlation function can be fitted by *correlation* = *corGaus* [25]. Since the *lme* R package can be readily used to fit the data, we omitted the details and the readers can refer to Pinheiro and Bates (2000) [25].

3. RESULTS

3.1 Simulation study

To evaluate the performance of the proposed models, simulation studies were carried out to calculate empirical type I error rates and power. We simulated 600 individuals with an age range from 20 to 65 years. For each individual, the number of observations ranged from 4 to 8 and each individual was examined every 2 or 4 years. In all simulations, we assumed that phenotype was affected by gender such that male's trait value was larger than that of females by 5, and gender was a covariate.

For mean function, we used one logarithm function $\mu(t) = -34.2 + 81.7\log(0.3(t+21.7))$ and an exponential function $\mu(t) = 110\exp(0.0002(t-25)^2)$ utilized in Fan et al. (2012) [19]. The curves of the two functions were plotted in Figure 1. The logarithm function $\mu(t) = -34.2 + 81.7\log(0.3(t+21.7))$ was taken from Daw et al. (2003) [26] and Wang et al. (2012) [22] whose estimates

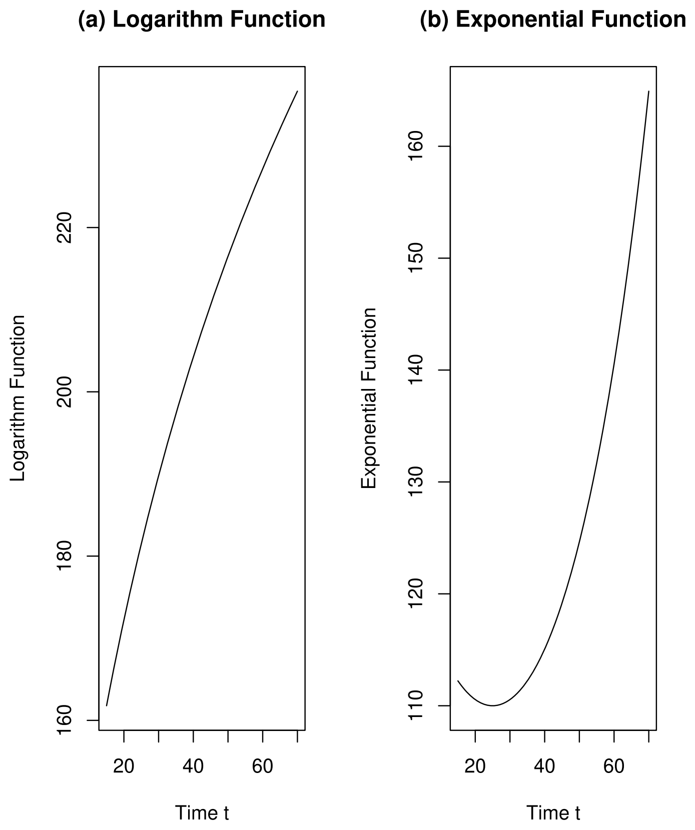

Figure 1. The mean curves of logarithm function $\mu(t) = -34.2 + 81.7 \log(0.3(t + 21.7))$ *and an exponential function* $\mu(t) = 110 \exp(0.0002 * (t - 25)^2))$ *utilized in Fan et al. (2012).*

were from the FHS cholesterol data, and the exponential function $\mu(t) = 110 \exp(0.0002(t - 25)^2)$ was used to mimic the FHS systolic blood pressure data. For the variance components, the subject variance σ_U^2 was 25 and error variance σ_e^2 was 10. To generate the correlation structure, we took $\theta = \exp(-1/\rho) = 0.2, 0.3, 0.4$, and defined correlation function as $\rho_U(t, s) = \exp(-\frac{|s-t|}{\rho}) = \theta^{|t-s|}$.

Type I error simulations The quantitative traits were generated by

$$y_i(t) = \mu(t) + Z_i\alpha + U_i(t) + \varepsilon_i(t),$$

where Z_i was the indicator function if a person is male and $\alpha = 5$. The mean function $\mu(t)$ was assumed to be unknown. We approximated it by the non-parametric linear penalized spline model (4) and the non-random expansion (5) and we set $K_\mu = 10$.

We simulated European-like (EUR) sequence data [27]. The EUR sequence data included 10,000 chromosomes covering 1 Mb regions and about 10% of the variants were common (MAF > 0.03) and the rest were rare. To calculate empirical type I error rates and power levels, genotypes were selected from variants in 6 kb subregions which were randomly selected from the 1 Mb region. On average, 117 variants were located in the 6kb regions if all variants are used in the analysis (i.e., some variants are common and the rest are rare). We also considered an analysis of rare variants (i.e., common variants were removed) and the average number of rare variants is 106 in the 6 kb regions. To fit model (1), we expanded the genetic variant functions and genetic effect function by relations (2) and (3), respectively. To fit model (7), the genetic effect function is expanded by relation (3). To create the basis functions, R package *fda* was used [14]. The order of B-spline basis was 4 and the number of B-spline basis functions was $K = K_\beta = 10$, and the number of Fourier basis functions was $K = K_\beta = 11$.

Note that the trait values were not related to the genotypes in the type I error simulations. For each simulation scenario, 10^7 phenotype-genotype datasets were generated. Models (1) and (7) were used to calculate test statistics and p-values using exponential correlation function $\rho_U(s, t) = \exp\left(-\frac{|s-t|}{\rho}\right)$. We reported type I error rates of perturbation tests proposed in He et al. (2017) [23], i.e., P-Dispersion, P-Burden, P-Fisher, and P-MinP. Then, an empirical type I error rate was calculated as the proportion of 10^7 p-values which were smaller than a given α level. In Tables 1 and 2, the empirical type I error rates were reported to test the null hypothesis of no association $H_0 : \beta_1 = \cdots = \beta_{K_\beta} = 0$, for the non-parametric linear penalized spline model (4) and the non-random expansion (5). Encouragingly, the empirical type I error rates of stochastic models (1) and (7) were close to the nominal levels 0.05, 0.01, 0.001, and 0.0001, suggesting that the type I errors are well-controlled in the stochastic models (1) and (7).

The perturbation tests control type I error rates correctly for logarithm function but inflate type one error rates for exponential function. Figure 1 shows that the logarithm function is relatively flat and so it is approximately close to a linear line, but the exponential function is unlikely close to a straight line. Hence, the perturbation tests may lead to high false positives for curved mean functions.

Empirical power simulations To evaluate the power of the LRT statistics of models (1) and (7), we simulated data sets under the alternative hypothesis. First, we generated genotypes of m variants in 6 kb subregions, similar to the type I error simulations, where the average number of m is 117 if some variants are common and the rest are rare, and the average number of m is 106 if all variants are rare. Then, an M causal variant subset of the m variants was randomly selected, yielding causal genotypes $(g_i(u_1), \cdots, g_i(u_M))$. For each dataset, the causal variants are the same for all the individuals in the dataset, but we allowed the causal variants to be different from dataset to dataset. Then, we generated the quantitative traits by

(13)
$$y_i(t) = \mu(t) + Z_i\alpha + \beta_1 g_i(u_1) + \cdots + \beta_M g_i(u_M) + U_i(t) + \varepsilon_i(t),$$

Table 1. **Empirical type I error rates of the LRT Statistics at nominal levels $\alpha = 0.05, 0.01, 0.001$, and 0.0001, when region size is 6 kb, some variants are common and the rest are rare, and the exponential correlation function $\rho_U(s,t) = \exp\left(-\frac{|s-t|}{\rho}\right)$ is exponential.** *The number of knots $K_\mu = 10$. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. Abbreviation: P-Disps: P-Dispersion; Exp: Exponential; Log: Logarithm.*

$\mu(t)$	θ	**Nominal**	**Stochastic Model (1)**				**Stochastic Model (7)**				**Perturbation Tests**			
		Level	$\mu(t)$: **Expansion (4)**		$\mu(t)$: **Expansion (5)**		$\mu(t)$: **Expansion(4)**		$\mu(t)$: **Expansion (5)**		**P-Disps**	**P-Burden**	**P-Fisher**	**P-MinP**
		α	**B-spline**	**Fourier**	**B-spline**	**Fourier**	**B-spline**	**Fourier**	**B-spline**	**Fourier**				
Exp	0.2	0.05	0.053762	0.054213	0.054847	0.055359	0.053762	0.054213	0.054847	0.055359	0.051121	0.052520	0.051261	0.050637
		0.01	0.011043	0.011199	0.011368	0.011551	0.011043	0.011199	0.011368	0.011551	0.010969	0.011494	0.011369	0.011103
		0.001	0.001156	0.001150	0.001204	0.001200	0.001156	0.001150	0.001204	0.001200	0.001289	0.001327	0.001495	0.001336
		0.0001	0.000115	0.000122	0.000119	0.000129	0.000115	0.000122	0.000119	0.000129	0.000165	0.000156	0.000209	0.000169
	0.3	0.05	0.053834	0.054243	0.054920	0.055406	0.053834	0.054243	0.054920	0.055406	0.051154	0.052495	0.051275	0.050587
		0.01	0.011054	0.011218	0.011385	0.011555	0.011054	0.011218	0.011385	0.011555	0.010951	0.011487	0.011401	0.011070
		0.001	0.001143	0.001176	0.001192	0.001221	0.001143	0.001176	0.001192	0.001221	0.001300	0.001344	0.001484	0.001334
		0.0001	0.000115	0.000121	0.000118	0.000127	0.000115	0.000121	0.000118	0.000127	0.000163	0.000161	0.000205	0.000168
	0.4	0.05	0.053839	0.054137	0.054923	0.055273	0.053839	0.054137	0.054923	0.055273	0.051158	0.052484	0.051268	0.050621
		0.01	0.011060	0.011223	0.011383	0.011556	0.011060	0.011223	0.011383	0.011556	0.010973	0.011491	0.011383	0.011089
		0.001	0.001138	0.001163	0.001181	0.001212	0.001138	0.001163	0.001181	0.001212	0.001292	0.001337	0.001489	0.001338
		0.0001	0.000112	0.000122	0.000120	0.000129	0.000112	0.000122	0.000120	0.000129	0.000165	0.000159	0.000203	0.000166
Log	0.2	0.05	0.053828	0.054218	0.054845	0.055347	0.053828	0.054218	0.054845	0.055347	0.050829	0.051430	0.050287	0.049762
		0.01	0.011048	0.011207	0.011361	0.011548	0.011048	0.011207	0.011361	0.011548	0.010611	0.010665	0.010497	0.010349
		0.001	0.001157	0.001148	0.001203	0.001196	0.001157	0.001148	0.001203	0.001196	0.001100	0.001092	0.001163	0.001068
		0.0001	0.000115	0.000123	0.000120	0.000131	0.000115	0.000123	0.000120	0.000131	0.000106	0.000114	0.000125	0.000107
	0.3	0.05	0.053841	0.054252	0.054896	0.055372	0.053841	0.054252	0.054896	0.055372	0.050815	0.051351	0.050301	0.049706
		0.01	0.011134	0.011246	0.011444	0.011569	0.011134	0.011246	0.011444	0.011569	0.010574	0.010731	0.010532	0.010408
		0.001	0.001172	0.001171	0.001213	0.001221	0.001172	0.001171	0.001213	0.001221	0.001089	0.001113	0.001142	0.001074
		0.0001	0.000120	0.000123	0.000125	0.000129	0.000120	0.000123	0.000125	0.000129	0.000106	0.000112	0.000128	0.000110
	0.4	0.05	0.053860	0.054290	0.054872	0.055366	0.053860	0.054290	0.054872	0.055366	0.050820	0.051361	0.050368	0.049757
		0.01	0.011084	0.011224	0.011401	0.011550	0.011084	0.011224	0.011401	0.011550	0.010591	0.010755	0.010528	0.010430
		0.001	0.001150	0.001174	0.001199	0.001219	0.001150	0.001174	0.001199	0.001219	0.001092	0.001108	0.001152	0.001073
		0.0001	0.000124	0.000129	0.000130	0.000136	0.000124	0.000129	0.000130	0.000136	0.000105	0.000108	0.000128	0.000113

Table 2. **Empirical type I error rates of the LRT Statistics at nominal levels $\alpha = 0.05, 0.01, 0.001$, and 0.0001, when region size is 6 kb, all variants are rare, and the exponential correlation function $\rho_U(s,t) = \exp\left(-\frac{|s-t|}{\rho}\right)$ is exponential.** *The number of knots $K_\mu = 10$. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. Abbreviation: P-Disps: P-Dispersion; Exp: Exponential; Log: Logarithm.*

$\mu(t)$	θ	**Nominal**	**Stochastic Model (1)**				**Stochastic Model (7)**				**Perturbation Tests**			
		Level	$\mu(t)$: **Expansion (4)**		$\mu(t)$: **Expansion (5)**		$\mu(t)$: **Expansion (4)**		$\mu(t)$: **Expansion (5)**		**P-Disps**	**P-Burden**	**P-Fisher**	**P-MinP**
		α	**B-spline**	**Fourier**	**B-spline**	**Fourier**	**B-spline**	**Fourier**	**B-spline**	**Fourier**				
Exp	0.2	0.05	0.053687	0.054091	0.054787	0.055265	0.053687	0.054091	0.054787	0.055265	0.052079	0.053019	0.052049	0.051591
		0.01	0.011112	0.011219	0.011440	0.011565	0.011112	0.011219	0.011439	0.011565	0.011375	0.011843	0.011814	0.011504
		0.001	0.001156	0.001168	0.001205	0.001217	0.001156	0.001168	0.001205	0.001217	0.001356	0.001439	0.001557	0.001406
		0.0001	0.000117	0.000126	0.000123	0.000133	0.000117	0.000126	0.000123	0.000133	0.000166	0.000177	0.000223	0.000177
	0.3	0.05	0.053719	0.054110	0.054831	0.055275	0.053719	0.054110	0.054832	0.055275	0.052040	0.053044	0.051990	0.051593
		0.01	0.011047	0.011236	0.011380	0.011574	0.011047	0.011236	0.011380	0.011574	0.011364	0.011791	0.011785	0.011495
		0.001	0.001152	0.001174	0.001193	0.001222	0.001152	0.001174	0.001193	0.001222	0.001333	0.001446	0.001557	0.001400
		0.0001	0.000119	0.000121	0.000126	0.000127	0.000119	0.000121	0.000126	0.000127	0.000167	0.000176	0.000221	0.000177
	0.4	0.05	0.053736	0.054098	0.054839	0.055226	0.053736	0.054098	0.054839	0.055226	0.052075	0.053043	0.052067	0.051632
		0.01	0.011047	0.011197	0.011371	0.011528	0.011047	0.011197	0.011371	0.011528	0.011353	0.011816	0.011762	0.011497
		0.001	0.001148	0.001151	0.001193	0.001200	0.001147	0.001151	0.001193	0.001200	0.001342	0.001436	0.001556	0.001404
		0.0001	0.000123	0.000124	0.000130	0.000131	0.000123	0.000124	0.000130	0.000131	0.000162	0.000175	0.000221	0.000180
Log	0.2	0.05	0.053763	0.054218	0.054791	0.055322	0.053763	0.054218	0.054791	0.055322	0.051480	0.051622	0.050787	0.050266
		0.01	0.011118	0.011282	0.011420	0.011607	0.011117	0.011282	0.011420	0.011607	0.010784	0.010796	0.010625	0.010505
		0.001	0.001164	0.001181	0.001208	0.001229	0.001164	0.001181	0.001208	0.001229	0.001073	0.001101	0.001137	0.001059
		0.0001	0.000118	0.000124	0.000126	0.000129	0.000118	0.000124	0.000126	0.000129	0.000098	0.000112	0.000118	0.000106
	0.3	0.05	0.053869	0.054273	0.054894	0.055352	0.053870	0.054273	0.054894	0.055352	0.051467	0.051541	0.050674	0.050250
		0.01	0.011088	0.011239	0.011395	0.011559	0.011087	0.011239	0.011395	0.011559	0.010778	0.010817	0.010648	0.010504
		0.001	0.001150	0.001155	0.001190	0.001207	0.001150	0.001155	0.001190	0.001207	0.001078	0.001109	0.001130	0.001061
		0.0001	0.000121	0.000124	0.000127	0.000129	0.000121	0.000124	0.000127	0.000129	0.000097	0.000116	0.000120	0.000104
	0.4	0.05	0.053785	0.054111	0.054835	0.055212	0.053786	0.054111	0.054836	0.055212	0.051483	0.051553	0.050799	0.050320
		0.01	0.011075	0.011244	0.011381	0.011560	0.011075	0.011244	0.011381	0.011560	0.010818	0.010854	0.010664	0.010518
		0.001	0.001136	0.001151	0.001182	0.001199	0.001136	0.001151	0.001182	0.001199	0.001081	0.001113	0.001132	0.001057
		0.0001	0.000119	0.000124	0.000126	0.000130	0.000119	0.000124	0.000126	0.000130	0.000097	0.000111	0.000121	0.000102

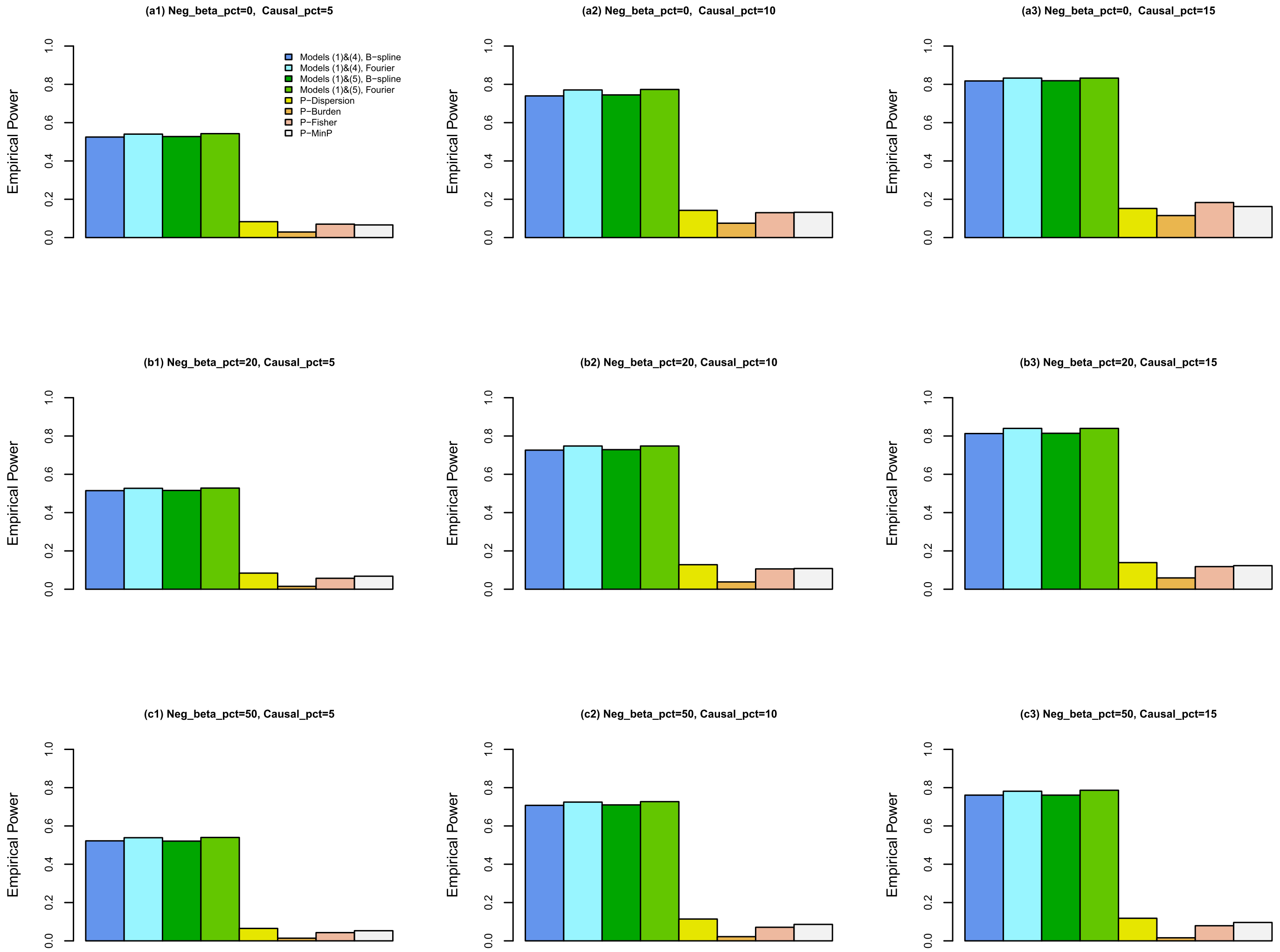

Figure 2. Empirical power levels of LRT statistics of stochastic model (1) using exponential correlation function $\rho_U(s,t) = \exp\left(-\frac{|s-t|}{\rho}\right)$ and perturbation tests proposed in He et al. (2017) (i.e., P-Dispersion, P-Burden, P-Fisher, and P-MinP) at nominal level $\alpha = 10^{-3}$ when some variants are common and the rest are rare, the mean curve is an exponential function $\mu(t) = 110\exp(0.0002 * (t-25)^2))$, $\theta = \exp(-1/\rho) = 0.2$, and $n = 600$ individuals. To fit model (1), we expanded the genetic variant functions and genetic effect function by relations (2) and (3), respectively. The mean function $\mu(t)$ was approximated by the non-parametric linear penalized spline model (4) and the non-random expansion (5) and we set $K_\mu = 10$. The legends from the top to the bottom in plot (a1) correspond to the bars from the left to the right columns in each of plots (a1)–(c3). Abbreviation: Neg_beta_pct means percentage of causal variants which have negative effects.

where $\beta_1, \cdots, \beta_M$ are additive effects for the causal variants defined as follows: $|\beta_j| = c|\log_{10}(MAF_j)|$, where MAF_j was the MAF of the j-th variant. Three different settings were considered: 5%, 10%, and 15% of variants are chosen as causal variants. When 5%, 10%, and 15% of the variants were causal, the constant c is taken as $c = 6$, 5, and 4, if some variants are common and the rest are rare. If all variants are rare, the constant c is taken as $c = 12$, 10, and 8, when 5%, 10%, and 15% of the variants were causal.

We calculated the LRT and p-values using the original genotypes of m variants by models (1) and (7). We compared the power of models (1) and (7) with perturbation tests proposed in He et al. (2017) [23]. The empirical power results shown in Figures 2, 3, 4, and 5 were based on 10^3 simulated replicates with $\alpha = 10^{-3}$. In addition to varying the percentage of causal variants in the subregion, we also varied the direction of effect. We considered situations where (i) all causal variants have positive effects; (ii) 20%/80% causal variants have negative/positive effects; and (iii) 50%/50% causal variants have negative/positive effects.

Figures 2, 3, 4, and 5 show that the LRT statistics of non-parametric penalized spline model (4) and non-random expansion (5) have similar power. The power levels are almost identical regardless of the choice of basis functions

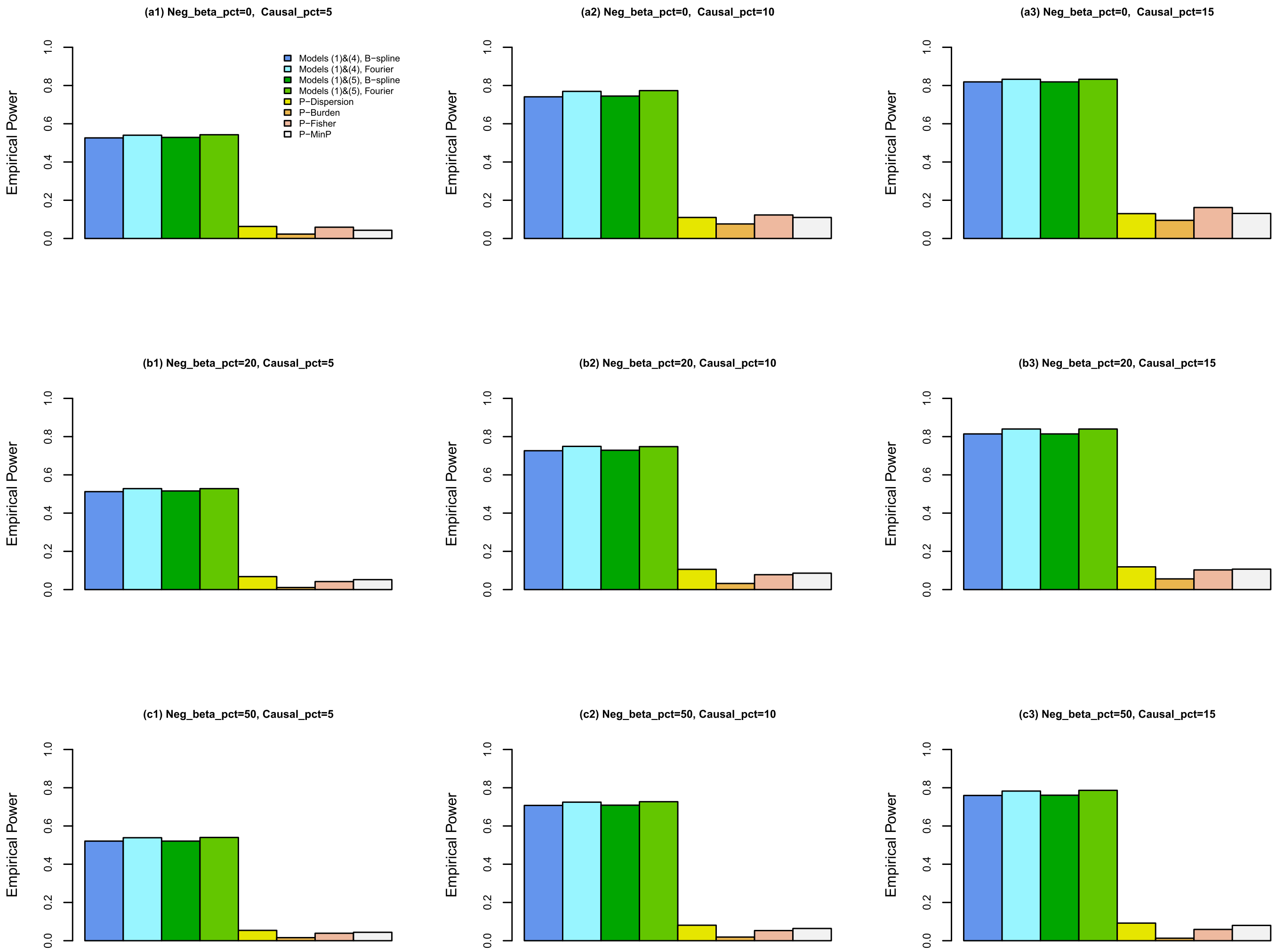

Figure 3. Empirical power levels of LRT statistics of stochastic model (1) using exponential correlation function $\rho_U(s,t) = \exp\left(-\frac{|s-t|}{\rho}\right)$ and perturbation tests proposed in He et al. (2017) (i.e., P-Dispersion, P-Burden, P-Fisher, and P-MinP) at nominal level $\alpha = 10^{-3}$ when some variants are common and the rest are rare, the mean curve is a logarithm function $\mu(t) = -34.2 + 81.7\log(0.3(t + 21.7))$, $\theta = \exp(-1/\rho) = 0.2$, and $n = 600$ individuals. Other parameters and notation are the same as those of Figure 2. The legends from the top to the bottom in plot (a1) correspond to the bars from the left to the right columns in each of plots (a1)–(c3).

to expand genetic effect function $\beta(u)$. Hence, the LRT statistics are very stable in terms of power performance and they are robust whether the mean function $\mu(t)$ is approximated by the spline model (4) or the non-random expansion (5), or which basis functions are used to expand the genetic effect function.

In Figures 2, 3, 4, and 5, the power levels of the LRT statistics of stochastic model (1) are higher than those the perturbation tests. Hence, the proposed stochastic models work well in the circumstances that we consider herein. The results of model (7) are similar to the model (1), but we did not present them.

In Supplementary Materials, http://intlpress.com/site/pub/files/_supp/sii/2022/0015/0002/SII-2022-0015-0002-s001.pdf, we provide simulation results using linear and Gaussian correlation functions. Note that the data were generated by exponential correlation structure. Thus, results of linear and Gaussian correlation functions can evaluate robustness of the proposed models when correlation function is mis-specified. To fit model (1), we expanded genetic variant functions and genetic effect function by relations (2) and (3) as in the main text, respectively. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. The mean function $\mu(t)$ was approximated by the non-parametric linear penalized spline model (4) and non-random expansion (5) and we set $K_\mu = 10$. For each simulation scenario, 10^6 datasets were generated to calculate type I error rates and 10^3 datasets were generated to calculate power levels. The results of Supplementary

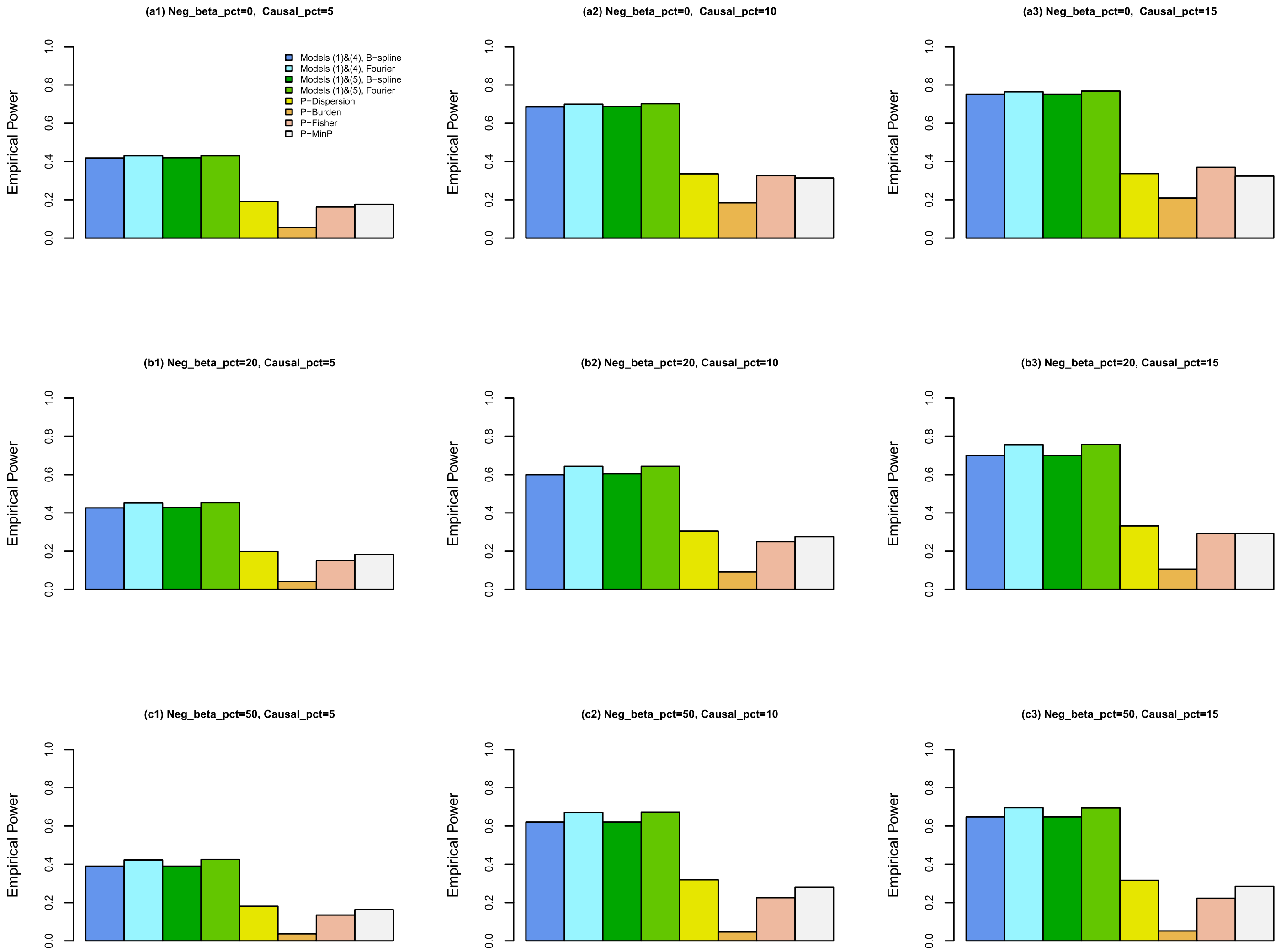

Figure 4. Empirical power levels of LRT statistics of stochastic model (1) using exponential correlation function $\rho_U(s,t) = \exp\left(-\frac{|s-t|}{\rho}\right)$ and perturbation tests proposed in He et al. (2017) (i.e., P-Dispersion, P-Burden, P-Fisher, and P-MinP) at nominal level $\alpha = 10^{-3}$ when all variants are rare, the mean curve is an exponential function $\mu(t) = 110\exp(0.0002 * (t-25)^2))$, $\theta = \exp(-1/\rho) = 0.2$, and $n = 600$ individuals. Other parameters and notation are the same as those of Figure 2. The legends from the top to the bottom in plot (a1) correspond to the bars from the left to the right columns in each of plots (a1)–(c3).

Materials show that the type I error rates and power levels are similar to these in the main tex using exponential correlation function to fit the model. Therefore, the proposed models are robust.

To make sure that the results are stable, we examined a wide range of parameters: $6 \leq K = K_\beta \leq 15$ for B-spline and Fourier basis functions (data not shown). We found the results do not strongly depend on the choices of the parameters.

3.2 Analysis of FHS BMI data

We applied the proposed methods to analyze FHS data. We investigate association with body mass index (BMI). The objective of the FHS was to identify the common factors that contribute to cardiovascular disease by following its development over a long period of time. The first cohort started in 1948 to recruit 5,209 subjects between the ages of 29 and 62 from the town of Framingham, MA. Since 1948, the subjects have continued to return to the study every two years. In 1971, the study enrolled a second-generation group to participate in similar examinations, i.e., Cohort 2. The original cohort has data from 28 examinations and the offspring cohort has data from 8 examinations. Between 2002 and 2005, the study enrolled the third generation of the FHS - 4095 offspring of the second generation [3, 28, 29]. The FHS sample consists of unrelated individuals as well as individ-

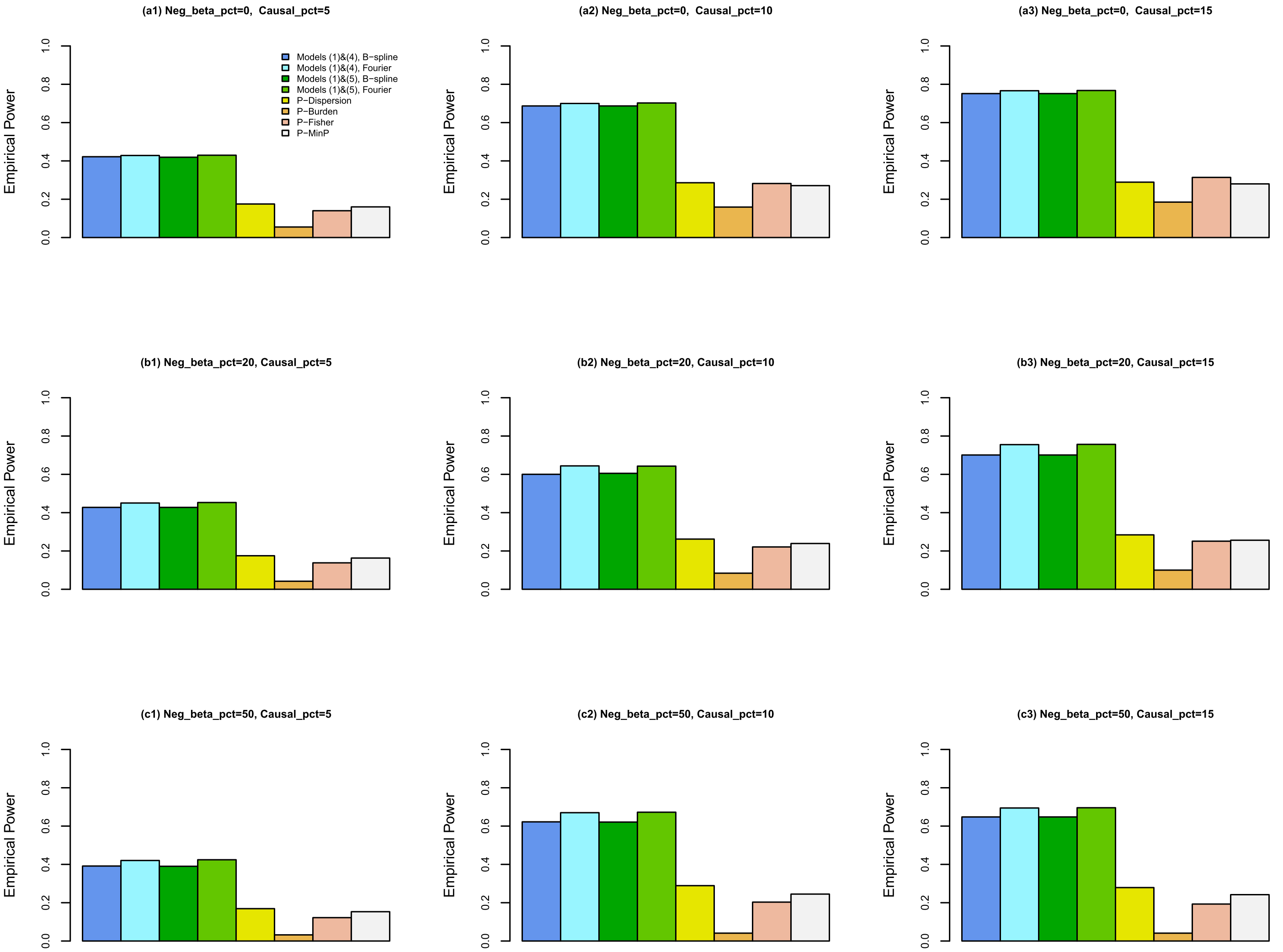

Figure 5. Empirical power levels of LRT statistics of stochastic model (1) using exponential correlation function $\rho_U(s,t) = \exp\left(-\frac{|s-t|}{\rho}\right)$ and perturbation tests proposed in He et al. (2017) (i.e., P-Dispersion, P-Burden, P-Fisher, and P-MinP) at nominal level $\alpha = 10^{-3}$ when all variants are rare, the mean curve is a logarithm function $\mu(t) = -34.2 + 81.7\log(0.3(t + 21.7))$, $\theta = \exp(-1/\rho) = 0.2$, and $n = 600$ individuals. Other parameters and notation are the same as those of Figure 2. The legends from the top to the bottom in plot (a1) correspond to the bars from the left to the right columns in each of plots (a1)–(c3).

uals from multi-generational pedigrees. In our analysis, we only use the information of unrelated individuals since our models are based on population data. For instance, the data of the two unrelated parents are used for a nuclear family but the data of the offspring are not. We analyze a subset of 1,898 unrelated individuals from the first two generation cohorts (original and offspring cohorts) with a total number of 13,171 measurements. In the dataset, 825 are males and 1,073 are females. The number of measures on each subject ranges from 1 to 8, and the intervals between measurements are highly variable among subjects. The 1,898 unrelated individuals were genotyped on an Affymetrix 500K array after quality checks of completeness (i.e., proportion of variants for which genotype is called) $> 95\%$, empirical inbreeding coefficient < 0.05, and Hardy-Weinberg equilibrium.

We perform a genome-wide gene-based longitudinal association analysis with BMI using exponential correlation function to model correlation. For each gene region, we extract all variants that are on the Affymetrix 500K chip of sample genotypes within 100 kb of the gene. All together we test association of 42,223 genes with BMI, including sex, age at each examination, and the top 5 principal components as covariates in the analysis [30].

The results of LRT statistics of models (1) and (7) using exponential correlation function and perturbation tests

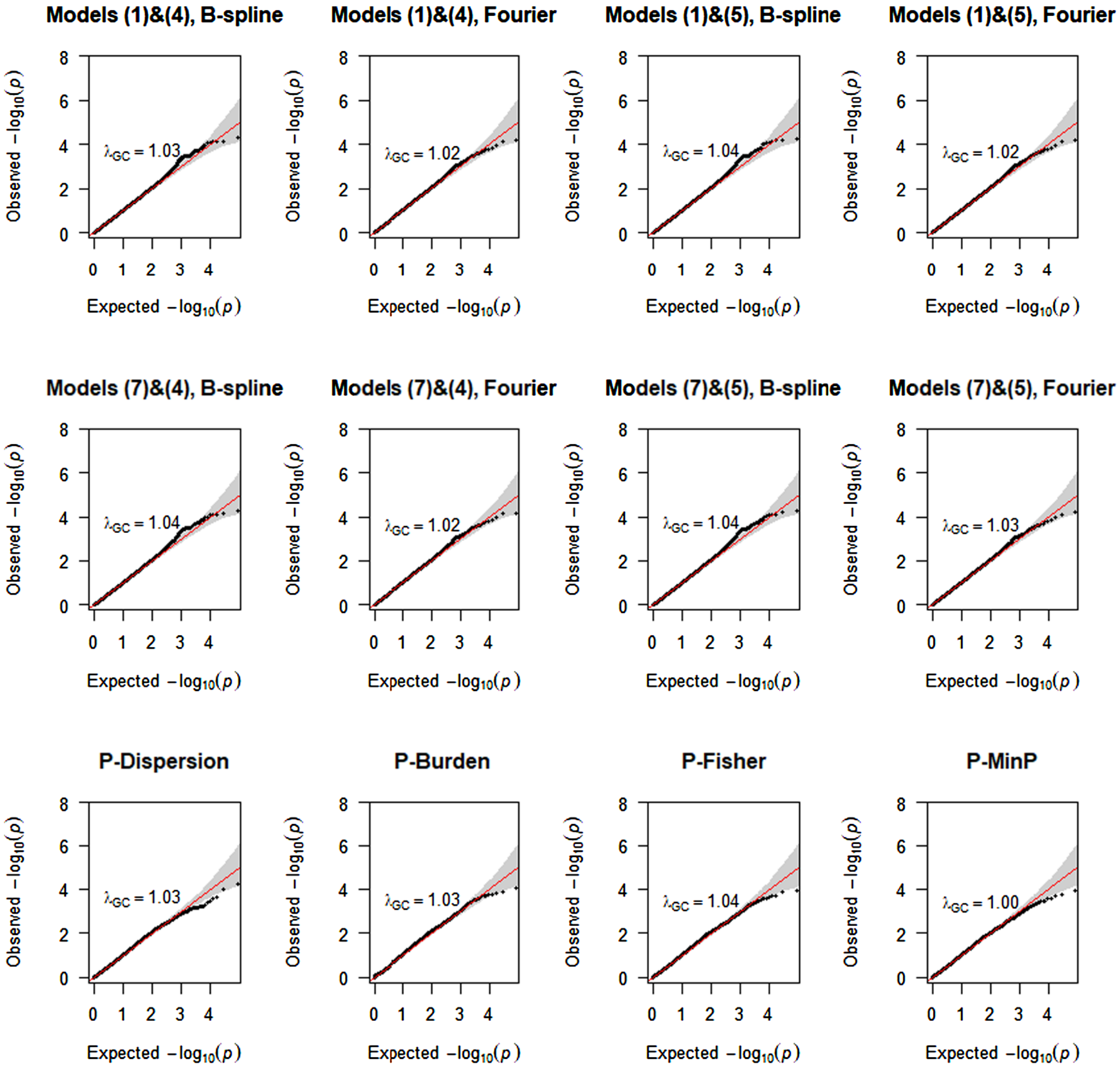

Figure 6. Q-Q plots for the LRT statistics of stochastic model (1) and perturbation tests proposed in He et al. (2017) (i.e., P-Dispersion, P-Burden, P-Fisher, and P-MinP) for FHS BMI data.

proposed in He et al. (2017) for FHS BMI data are reported in Table 3 and Figure 6. Quantile-quantile (QQ) plots in Figure 6 shows that genomic control inflation factors for the LRT statistics of stochastic model (1) and perturbation tests are all around 1.0. Table 3 reports 10 top gene regions when at least one of the tests provides a p-value $< 10^{-4}$. Interestingly, none of perturbation tests reaches a p-value $< 10^{-4}$ while the proposed LRT statistics provide association signals. Thus, perturbation tests are not appropriate to analyze the FHS data.

3.3 Computational time

In our type I error rate calculations, we divided 10^7 datasets into 1,000 independent jobs by different random seeds, and each job simulated and analyzed 10,000 datasets. Roughly, it takes 8–9 days to finish the calculations. Hence, it took about 1 day to simulate and analyze 1,200 datasets. In real data analysis, our software can be used to perform genome-wide association analysis by dividing the analysis into independent jobs.

4. DISCUSSION

In this paper, we develop stochastic functional linear models to analyze longitudinally measured quantitative traits and sequence data in longitudinal studies. To analyze sequence data, high dimensional genetic data are treated as realizations of a stochastic process and functional data analysis techniques are used to reduce the dimensionality. The quantitative traits are modeled by a continuous stochastic process. Based on the theory of stochastic processes, the variance-covariance structure of the trait values is constructed to analyze multiple measurement variation and cor-

Table 3. **Application to FHS BMI data using exponential correlation function to model correlations.** *In all subjects, 1,898 unrelated individuals were used for analysis. Genes with at least one p-value $< 10^{-4}$ was reported. Abbreviation: P-Disps: P-Dispersion.*

Gene	Chr	# of SNPs	Stochastic Model (1)				Stochastic Model (7)				Perturbation Tests			
			$\mu(t)$: **Expansion (4)**		$\mu(t)$: **Expansion (5)**		$\mu(t)$: **Expansion (4)**		$\mu(t)$: **Expansion (5)**		**P-Disps**	**P-Burden**	**P-Fisher**	**P-MinP**
			B-spline	**Fourier**	**B-spline**	**Fourier**	**B-spline**	**Fourier**	**B-spline**	**Fourier**				
ERG	21	128	5.02E-05	0.000340	5.36E-05	0.000355	5.02E-05	0.000340	5.36E-05	0.000355	0.282306	0.153643	0.188188	0.265644
CTD-3064M3.7	8	23	6.94E-05	0.000599	6.63E-05	0.000571	0.000140	0.000599	0.000134	0.000571	0.661380	0.327643	0.486859	0.473628
AC092159.3	2	65	6.98E-05	0.034134	6.37E-05	0.032278	6.98E-05	0.034134	6.37E-05	0.032278	0.080844	0.068355	0.056938	0.119446
RP11-332H18.3	17	31	7.64E-05	0.000231	7.79E-05	0.000234	7.64E-05	0.000231	7.79E-05	0.000234	0.938222	0.562148	0.801520	0.741850
AC007115.3	12	30	8.07E-05	0.000193	8.33E-05	0.000203	8.07E-05	0.000193	8.33E-05	0.000203	0.526067	0.571080	0.554481	0.615958
RP11-685G9.4	15	27	8.09E-05	0.000948	7.96E-05	0.000929	8.09E-05	0.000948	7.96E-05	0.000929	0.427664	0.285943	0.341504	0.423652
RPL21P108	13	67	0.000204	7.28E-05	0.000215	7.65E-05	0.000204	7.28E-05	0.000215	7.65E-05	0.058683	0.011582	0.018673	0.021595
CARTPT	5	43	0.000226	6.72E-05	0.000212	6.50E-05	0.000226	6.72E-05	0.000212	6.50E-05	1.000000	0.856336	1.000000	0.963872
AC092159.2	2	71	0.000105	0.000406	9.91E-05	0.000375	0.000105	0.000406	9.91E-05	0.000375	0.089734	0.062822	0.054538	0.112101
MCCC2	5	58	0.000107	0.000517	9.64E-05	0.000494	0.000107	0.000517	9.64E-05	0.000493	1.000000	1.000000	1.000000	1.000000

relations. Genetic effect functions are estimated by an ordinary linear square smoother. Non-parametric penalized and non-random spline models are used to approximate the time-dependent mean functions. To evaluate the performance of the stochastic models, simulation studies were carried out to calculate and to compare empirical type I error rates and power. The models are found to perform well in terms of reasonable type I error rates and power. In addition, the proposed methods are robust when the correlation function is mis-specified.

In the power comparison, it was found that the LRT statistics of proposed stochastic models are higher than those the perturbation tests proposed in He et al. (2017) [23]. In the power comparison, note the data was generated by model (13), in which the mean function $\mu(t)$ is either exponential or logarithm. The models (1) and (7) treated the mean as a function $\mu(t)$, but it was treated as a constant intercept in He et al. (2017) [23]. Therefore, the models (1) and (7) are the correct model to analyze the simulated data while the models of He et al. (2017) [23] are not. In short, the power comparison may not be interpreted as that the performance of the LRT statistics of the proposed stochastic models are always better. Nevertheless, the LRT statistics of stochastic models (1) and (7) perform better in the circumstance considered hereby.

As the previous paper of Fan et al. (2012) [19] to analyze common variants, one merit of the proposed stochastic models is that the number of parameters does not depend on the number of multiple measurements and the number of genetic variants. The number of parameters is fixed after carefully specifying the number of basis functions and variance-covariance structure. The parameters are specified through two components based on the theory of stochastic processes: (i) stochastic regression models (1) and (7); (ii) temporal variance-covariance functions given by equations (6). To estimate the mean function $\mu(t)$ and genetic effect functions, the parameters are specified by spline models.

The proposed approaches can only analyze population data. It will be very interesting and important to extend the methods to analyze family data or combinations of family data and population data. The stochastic regression models (1) and (7) can be used to model the trait means, which take care of the association information. The temporal variance-covariance functions given by equation (6) can be used for one individual's measurements. For family members, the temporal variance-covariance functions can be constructed in the same way as variance component models presented [31, 32]. Then, one may compare the method with kernel machine test in literature [33]. In addition, it is interesting in developing stochastic models to analyze qualitative genetic traits. This paper builds a foundation for future research to develop stochastic models for gene-based analysis in longitudinal studies.

This paper focuses on quantitative traits. The proposed models can be readily implemented by *lme* R package [25]. Similar idea can be applied to build stochastic models to analyze discrete traits. However, the estimation procedure for parameters is different from that of quantitative traits. More in-depth research is needed to extend the models to analyze the discrete traits.

ACKNOWLEDGEMENTS

This study was supported by U.S. National Science Foundation grant DMS-1915904 (Ruzong Fan and Bingsong Zhang) and DMS-1916246 (Zuoheng Wang), by National Institutes of Health grant Z01 HG000153 (Ruzong Fan, Chi-yang Chiu, Alexander F. Wilson, and Joan E. Bailey-Wilson), and by the Intramural Research Program of the National Human Genome Research Institute (Ruzong Fan, Chi-yang Chiu, Alexander F. Wilson and Joan E. Bailey-Wilson), National Institutes of Health, Maryland. This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov).

Computer program The methods proposed in this paper are implemented using functional data analysis (fda) procedures implemented in the statistical package R. The R codes and an example are available from https://sites.google.com/a/georgetown.edu/ruzong-fan/about.

Data sharing In the paper, real exome chip data were analyzed from Framingham Heart Study. The data were obtained from dbGAP of NIH. We have permission to analyze the data, but we are not allowed to release them.

Received 31 March 2020

REFERENCES

[1] Bild D, Bluemke D, Burke G, Detrano R, Roux A, Folsom A, et al. (2002). Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol.* **156** 871–881.

[2] Gottesman O, Kuivaniemi H, Tromp G, Faucett W, Li R, Manolio T, Sanderson S, Kannry J, Zinberg R, Basford M, et al. (2003). The Electronic Medical Records and Genomics (eMERGE) network: past, present, and future. *Genet Med.* **15** 761–771.

[3] Splansky G, Corey D, Yang Q, Atwood L, Cupples L, Benjamin E, D'Agostino R, Fox C, Larson M, Murabito J, O'Donnell C, Vasan R, Wolf P, and Levy D. (2007). The third generation cohort of the National Heart, Lung, Blood Institute's Framingham Heart Study: design, recruitment, initial examination. *Am J Epidemiol.* **165** 1328–1335.

[4] Eichler E, Flint J, Gibson G, Kong A, Leal S, Moore J, and Nadeau J. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* **11** 446–450.

[5] Manolio T, Collins F, Cox N, Goldstein D, Hindorff L, Hunter D, McCarthy M, Ramos E, Cardon L, Chakravarti A, et al. (2009). Finding the missing heritability of complex diseases. *Nature.* **461** 747–753.

[6] Fan R, Wang Y, Mills J, Wilson A, Bailey-Wilson J, and Xiong M. (2013). Functional linear models for association analysis of quantitative traits. *Genet Epidemiol.* **37** 726–742.

[7] Fan R, Chiu C, Jung J, Weeks D, Wilson A, Bailey-Wilson J, Amos C, Chen Z, Mills J, and Xiong M. (2016). A comparison study of fixed and mixed effect models for gene level association studies of complex traits. *Genet Epidemiol.* **40** 702–721.

[8] Luo L, Zhu Y, and Xiong M. (2012). Quantitative trait locus analysis for next-generation sequencing with the functional linear models. *J Med Genet.* **49** 513–524.

[9] Luo L, Zhu Y, and Xiong M. (2013). Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. *Eur J Hum Genet.* **21** 217–224.

[10] Vsevolozhskaya O, Zaykin D, Barondess D, Tong X, Jadhav S, and Lu Q. (2016). Uncovering local trends in genetic effects of multiple phenotypes via functional linear models. *Genet Epidemiol.* **40** 210–221.

[11] de Boor C. (2001). *A Practical Guide to Splines, Applied Mathematical Sciences 27*, revised version. New York: Springer. MR1900298

[12] Ferraty F, and Romain Y. (2010). *The Oxford Handbook of Functional Data Analysis.* New York: Oxford University Press. MR2917982

[13] Horváth L, and Kokoszka P. (2012). *Inference for Functional Data With Applications.* New York: Springer. MR2920735

[14] Ramsay J, Hooker G, and Graves S. (2009). *Functional Data Analysis With R and Matlab.* New York: Springer. MR3645102

[15] Ramsay J, and Silverman B. (2005). *Functional Data Analysis*, 2nd ed. New York: Springer. MR2168993

[16] Yao F, Müller H, and Wang J. (2005). Functional linear regression analysis for longitudinal data. *Ann Stat.* **33(6)** 2873–2903. MR2253106

[17] Ross S. (1996). *Stochastic Processes*, 2nd ed. New York: John Wiley & Sons. MR1373653

[18] Karlin S, and Taylor H. (1981). *A Second Course in Stochastic Process.* Academic Press, New York. MR0611513

[19] Fan R, Zhang Y, Albert P, Liu A, Wang Y, and Xiong M. (2012). Longitudinal association analysis of quantitative traits. *Genet Epidemiol.* **36** 856–869.

[20] Wang Y. (2011). *Smoothing Splines, Methods Applications.* CRC Press, A Chapman & Hall Book. MR2814838

[21] Wang Y, and Huang C. (2012). Semiparametric variance components models for genetic studies with longitudinal phenotypes. *Biostatistics* **13** 482–496.

[22] Wang Y, Huang C, Fang Y, Yang Q, and Li R. (2012). Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing. *Applied Statistics: J R Stat Soc Ser C Appl Stat.* **61** 1–24. MR2877582

[23] He Z, Lee S, Zhang M, Smith J, Guo X, Palmas W, Kardia S, Ionita-Laza I, and Mukherjee B. (2017). Rare-variant association test in longitudinal studies, with an application to the multi-ethnic study of atherosclerosis (MESA). *Genet Epidemiol.* **41** 801–810.

[24] Soler J, and Blangero J. (2003). Longitudinal familial analysis of blood pressure involving parametric (co)variance functions. *BMC Genet.* **4(Suppl 1)** S87.

[25] Pinheiro J, and Bates D. (2000). *Mixed-Effects Models in S and S-PLUS.* Springer.

[26] Daw E, Morrison J, Zhou X, and Thomas D. (2003). Genetic Analysis Workshop 13: Simulated longitudinal data on families for a system of oligogenic traits. *BMC Genet.* **4(Suppl 1)** S3.

[27] Schaffner S, Foo C, Gabriel S, Reich D, Daly M, and Altshuler D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15** 1576–1583.

[28] Levy D, Larson M, Benjamin E, Newton-Cheh C, Wang T, Hwang S, Vasan R, and Mitchell G. (2007). Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. *BMC Med Genet.* **8(Suppl 1)** S3.

[29] Levy D, Ehret G, Rice K, Verwoert G, Launer L, Dehghan A, Glazer N, et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nat Genet.* **41** 677–687.

[30] Wang Z, Xu K, Zhang X, Wu X, and Wang Z. (2017). Longitudinal snp-set association analysis of quantitative phenotypes. *Genet Epidemiol.* **41** 81–93.

[31] Lange K. (2002). *Mathematical and Statistical Methods for Genetic Analysis*, 2nd ed. Springer. MR1892279

[32] Wu X, and McPeek M. (2018). L-GATOR: Genetic association testing for a longitudinally measured quantitative trait in samples with related individuals. *Am J Hum Genet.* **102(4)** 574–591.

[33] Yan Q, Weeks D, Tiwari H, Yi N, Zhang K, Gao G, Lin W, Lou X, Chen W, and Liu N. (2016). Rare-variant kernel machine test for longitudinal data from population and family samples. *Hum Hered.* **80** 126–138.

Bingsong Zhang
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
US
E-mail address: bz117@georgetown.edu

Shuqi Wang
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
US
E-mail address: sw1080@georgetown.edu

Xiaohan Mei
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
US
E-mail address: xm74@georgetown.edu

Yue Han
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
US
E-mail address: yh507@georgetown.edu

Runqiu Wang
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
US
E-mail address: rw872@georgetown.edu

Hong-Bin Fang
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
US
E-mail address: hf183@georgetown.edu

Chi-Yang Chiu
Division of Biostatistics, Department of Preventive Medicine
University of Tennessee Health Science Center
Memphis, TN 38163
US
Computational and Statistical Genomics Branch, National Human Genome Research Institute
National Institutes of Health
Bethesda, MD 20892
US
E-mail address: chiu@uthsc.edu

Jun Ding
Laboratory of Genetics and Genomics, National Institute on Aging
National Institutes of Health
Bethesda, MD 20892
US
E-mail address: jun.ding@nih.gov

Zuoheng Wang
Department of Biostatistics
Yale University
New Haven, CT 06520
US
E-mail address: zuoheng.wang@yale.edu

Alexander F. Wilson
Computational and Statistical Genomics Branch, National Human Genome Research Institute
National Institutes of Health
Bethesda, MD 20892
US
E-mail address: afw@mail.nih.gov

Joan E. Bailey-Wilson
Computational and Statistical Genomics Branch, National Human Genome Research Institute
National Institutes of Health
Bethesda, MD 20892
US
E-mail address: jebw@mail.nih.gov

Momiao Xiong
Human Genetics Center
University of Texas - Houston
P. O. Box 20334
Houston, Texas 77225
US
E-mail address: momiao.xiong@uth.tmc.edu

Ruzong Fan
Department of Biostatistics, Bioinformatics, and Biomathematics
Georgetown University Medical Center
Washington, DC 20057
US
Computational and Statistical Genomics Branch, National Human Genome Research Institute
National Institutes of Health
Bethesda, MD 20892
US
E-mail address: rf740@georgetown.edu
url: http://biostatistics.georgetown.edu/news/ruzong-fan-joins-biostatistics-faculty