Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2012)
61, *Part* 1, *pp.* 1–24

Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing

Yuanjia Wang and Chiahui Huang,

Columbia University, New York, USA

Yixin Fang,

New York University, USA

Qiong Yang

Boston University, USA

and Runze Li

Pennsylvania State University, University Park, USA

[Received August 2010. Final revision April 2011]

Summary. In longitudinal genetic studies, investigators collect repeated measurements on a trait that changes with time along with genetic markers. For family-based longitudinal studies, since repeated measurements are nested within subjects and subjects are nested within families, both the subject level and the measurement level correlations must be taken into account in the statistical analysis to achieve more accurate estimation. In such studies, the primary interests include testing for a quantitative trait locus effect, and estimating the age-specific quantitative trait locus effect and residual polygenic heritability function. We propose flexible semiparametric models and their statistical estimation and hypothesis testing procedures for longitudinal genetic data. We employ penalized splines to estimate non-parametric functions in the model. We find that misspecifying the baseline function or the genetic effect function in a parametric analysis may lead to a substantially inflated or highly conservative type I error rate on testing and large mean-squared error on estimation. We apply the proposed approaches to examine age-specific effects of genetic variants reported in a recent genomewide association study of blood pressure collected in the Framingham Heart Study.

Keywords: Genomewide association study; Penalized splines; Quantitative trait locus

1. Introduction

For quantitative traits that change with age, such as blood pressure and level of cholesterol, longitudinal genetic studies offer a valuable opportunity to detect genes that have a time varying effect and examine how genes affect developmental features of these traits. One example of a longitudinal genetic study is the Framingham Heart Study (FHS) (Dawber *et al.*, 1951), which is a large on-going prospective study of risk factors for cardiovascular disease originated in 1948. In the FHS, repeated measurements are collected on subjects' clinical characteristics such as level of cholesterol, blood pressure and level of blood glucose. To understand genetic

Address for correspondence: Yuanjia Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY 10032, USA.
E-mail: yuanjia.wang@columbia.edu

© 2011 Royal Statistical Society 0035–9254/12/61001

underpinning of risk factors for cardiovascular disease, dense single-nucleotide polymorphism (SNP) genotyping was performed using approximately 550 000 SNPs in nearly 10 000 individuals from three-generation families in the FHS. The FHS provides an opportunity to discover not only the genes that affect the mean value of a risk factor, but also those that affect time varying features such as the rate of change over time in a trait.

Theories and evidence for genetic factors controlling time varying developmental features of a phenotype are noted in the plant, animal and human genetics literature. For example, complex biological organisms such as plants and animals have evolved through mutations in genes that control the developmental processes that lead to their mature forms (Rice, 2002; Raff, 2000; He *et al.*, 2010). Rice (2002) described general population genetic models to relate developmental features of a trait to a quantitative trait locus (QTL). From an evolutionary and developmental biology perspective, Raff *et al.* (2000) discussed mechanisms of regulatory genes that control developmental features of complex organisms. Zhao *et al.* (2004) attempted to map genes controlling rice plant growth. In human genetics, Province and Rao (1985) observed temporal trends in heritability of systolic blood pressure (SBP) in a Japanese-American family study, and Jarvik *et al.* (1997) demonstrated an age-dependent effect of the apo-E genotype on lipid levels.

Despite this evidence, however, interactions between gene and age or age-dependent genetic effects are routinely ignored in genetic analysis (Lasky-Su *et al.*, 2008). One disadvantage of this practice is that it may make the discovery of individual genes with moderate effects more difficult owing to a loss of power (Shi and Rao, 2008). Another limitation is that it may contribute to inconsistent replication of genetic association findings (Lasky-Su *et al.*, 2008). For example, when gene–age interaction exists, subjects in a replication sample may be in a different age range from the initial study sample, so the replication study may fail to discover a gene that has an effect in the original study age range.

A naive way of analysing genetic data with longitudinal phenotypes is to perform a set of genetic analyses at each age separately (Atwood *et al.*, 2002). However, this approach ignores rich information in the longitudinal structure and may not detect genes affecting the time varying features of a trait. Strauch *et al.* (2003) reviewed several two-step methods: the first step is either to take the average of trait measurements on a subject or to fit a longitudinal model without consideration of genetic markers or family structures; the second step is to perform genetic analysis on one or more summary statistics derived from the first step. This method may be improved by a joint approach that fits longitudinal and genetic parameters simultaneously. Zhang and Zhong (2006) and Shi and Rao (2008) used a parametric function such as exponential or Gaussian to accommodate time varying genetic effects in linkage studies. Shi and Rao (2008) showed that ignoring temporal trends in genetic effects can reduce power substantially. Although the major advantage of parametric models is parsimony, they may not be sufficiently flexible to describe the complicated underlying relationship between the gene and the trait over time. It is therefore desirable to consider more flexible models to analyse data from longitudinal genetic studies. For this, Zhao and Wu (2008) developed a wavelet-based non-parametric approach.

In this paper, we first present a semiparametric regression model for overall polygenic effect with longitudinal data collected from family-based genetic studies such as the FHS. Next, we extend the polygenic model to accommodate genetic markers and model age-dependent associations. For family-based designs, subjects are nested in families and repeated measurements are nested within subjects. Therefore it is critical to account for both the subject level and the measurement level correlations in the statistical analyses to achieve more accurate estimation. One of the key features of the family-based longitudinal genetic studies is that subjects in the

same family may not be independent, given any genetic marker. This is because the marker under consideration may not fully account for familial correlation between family members; therefore there may be residual familial correlation. The unexplained residual correlation aside from the marker is termed an unspecified residual polygenic effect and is modelled as a random effect.

The mixed effects model naturally lends itself to account for residual polygenic effects between subjects in a family, as well as serial correlation between repeated measurements of the outcome on the same subject. Meanwhile, it is desirable to model the baseline function and the genetic effect function non-parametrically, because there is usually limited information about the parametric forms of these functions. For this, we propose to use penalized splines (Eilers and Marx, 1996) to estimate non-parametric functions in the model. Penalized-spline-based methods have become popular in the recent literature (Ruppert *et al.*, 2003). In a penalized splines regression, an unknown smooth function is estimated by assuming a high dimensional spline basis and imposing a penalty on the spline coefficients to control overfitting and to achieve a smooth fit. Empirical and theoretical work has shown that the penalized spline as a reduced rank smoother can achieve a similar quality of fit to that of full rank estimators such as smoothing splines (Ruppert, 2002; Li and Ruppert, 2008). Additionally, its mixed model representation makes it particularly suitable for analysing longitudinal genetics data. Using this representation, it is easy to handle random polygenic effects and all approaches that are developed here can be implemented by standard statistical software packages such as procedure `MIXED` in SAS or `LME` in R, allowing researchers to use these methods routinely.

The primary interests in this work are to estimate baseline function, age-specific QTL effects and residual polygenic heritability, and to test for the QTL effect. The remainder of the paper is organized as follows. In Section 2, we propose two semiparametric regression models for family-based longitudinal genetic studies to estimate the baseline function, to test and estimate time varying QTL effects and to estimate residual polygenic heritability. In Section 3, we develop statistical methods for these two classes of models. In Section 4, we perform simulation studies to investigate properties of the methods proposed. In Section 5, we apply the developed methods to analyse the FHS blood pressure data. In Section 6, we discuss implications of our findings on the FHS and possible extensions of the methods proposed.

2. Models for longitudinal genetic studies

2.1. Partially linear mixed effects model for polygenic effect

The first step in a genetic epidemiological study is to assess polygenic heritability of a trait by examining the similarity of a trait in family members before using any genetic markers. Polygenic heritability quantifies the overall genetic effect on a trait. If there is no evidence of a polygenic effect or familial aggregation, it may not be necessary to pursue further study such as linkage or association analysis that aims at locating the underlying loci affecting the trait. In contrast if evidence for a genetic contribution to a trait is observed, then, to locate this factor along the genome, investigators decompose the polygenic effect into a major gene effect at a specific locus and a residual polygenic effect contributed by other unlinked loci.

A polygenic effect is treated as an unobserved random variable with covariance matrix specified by relationships between relatives (Lynch and Walsh, 1998; Khoury *et al.*, 1993). To be specific, let Y_{ijh} be the phenotype measurement for subject j in family i at visit h, and let t_{ijh} denote the subject's age at this visit. Let n_i denote the number of subjects in family i, let n_{ij} denote the number of measurements on subject (i, j) and let $N_i = \Sigma_j n_{ij}$. A partially linear mixed effects model for Y_{ijh} is defined to be

$$Y_{ijh} = \mu(t_{ijh}) + x_{ij}^{\mathrm{T}}\beta + \alpha_i + z_{ijh}^{\mathrm{T}}\gamma_{ij} + \varepsilon_{ijh}, \qquad \alpha_i \sim N(0, \sigma_\alpha^2), \quad \gamma_i \sim N(0, \Gamma_i), \quad \varepsilon_{ij} \sim \mathrm{GP}(0, \vartheta_{ij}) \tag{1}$$

where $\mu(t)$ is an unspecified baseline function and x_{ij} are time invariant environmental exposures such as sex with effects β, α_i are random shared environmental effects such as diet shared among family members, $\gamma_i = (\gamma_{i1}^{\mathrm{T}}, \ldots, \gamma_{in_i}^{\mathrm{T}})^{\mathrm{T}}$ are vectors of random polygenic effects, z_{ijh} are design vectors for γ_{ij} which can be time dependent to capture an age-related polygenic effect, $\varepsilon_{ij} = (\varepsilon_{ij1}, \ldots, \varepsilon_{ijn_{ij}})^{\mathrm{T}}$ are random measurement errors with possible serial correlation and $\mathrm{GP}(0, \vartheta_{ij})$ is a Gaussian process with covariance matrix ϑ_{ij}. The inclusion of exposures with time varying effects is deferred to the next section where we introduce the time varying QTL model. We assume that α_i, γ_{ij} and ε_{ij} are independent. The random polygenic effect reflects overall genetic information in a trait. Their covariance structure depends on the relationship between family members (Khoury *et al.* (1993), chapter 7). Specifically,

$$\Gamma_i = \mathrm{cov}(\gamma_i, \gamma_i^{\mathrm{T}}) = 2K^i \otimes \Omega_\gamma, \qquad \mathrm{cov}(\gamma_i, \gamma_{i'}^{\mathrm{T}}) = 0 \qquad \text{for } i \neq i', \tag{2}$$

where K^i is an $n_i \times n_i$ known kinship coefficient matrix whose (j, j')th element is determined by the relationship between subjects j and j' in family i, and Ω_γ is an unknown covariance of the polygenic effect. The kinship coefficient is defined as the probability of randomly drawing an allele in subject j that is identical by descent to an allele at the same locus randomly drawn from subject j'. For example, twice the kinship coefficient, $2K^i_{jj'}$, for a full sibling pair is $\frac{1}{2}$ and for a half-sibling pair is $\frac{1}{4}$ (Khoury *et al.* (1993), page 211). Parameters in Ω_γ represent the unknown polygenic variances. We use Σ_i to denote the covariance of Y_i, i.e. $\Sigma_i = \sigma_\alpha^2 \mathbf{1}_{N_i}\mathbf{1}_{N_i}^{\mathrm{T}} + 2Z_i\Gamma_i Z_i^{\mathrm{T}} + \mathrm{cov}(\varepsilon_i)$.

Heritability is defined as the ratio of the genetic variance to the total variance, i.e.

$$h_\gamma^2(t) = \sigma_\gamma^2(t)/\sigma_T^2(t), \tag{3}$$

where $\sigma_T^2(t) = \sigma_\alpha^2 + \sigma_\gamma^2(t) + \sigma_\varepsilon^2(t)$, and $\sigma_\gamma^2(t) = \omega_{11} + 2\omega_{12}t + \omega_{22}t^2$. For a linear design vector, $z_{ijh} = (1, t_{ijh})^{\mathrm{T}}$, ω_{ls} is the (l, s)th element of Ω_γ. Here $\sigma_\varepsilon^2(t)$ is the variance of the residual random measurement error. Whereas the linear random polygenic effect here serves as a parsimonious approximation to the underlying truth, we discuss more flexible ways to model the QTL genetic effect and heritability function in the next section. In addition, a test of $h_\gamma^2(t)$ based on functional principal components of heritability was proposed in Fang and Wang (2009).

Although the mixed effects model formulation of penalized splines allows the baseline and the QTL functions to be fitted by standard statistical software, one practical complication in genetic studies is to impose the correlation structure of polygenic effects predicted by kinship coefficients as shown in expression (2). In behavioural genetics, decomposing phenotypic variance into genetic and environmental components is typically done by structural equation models by using specialized software such as Mx (Neale *et al.*, 2004). Guo and Wang (2002) ignored the kinship correlation to use standard software to fit a multilevel model.

Rabe-Hesketh *et al.* (2008) showed that, for most family designs, we can reparameterize the polygenic effect into a few family-specific and subject-specific random effects allowing for easy handling of polygenic effects by standard software. For example, for nuclear families, we replace the polygenic effect γ_{ij} in model (1) by two family-specific and a subject-specific random effects as

$$\gamma_{ij} = a_{i1}\left(M_{ij} + \frac{C_{ij}}{2}\right) + a_{i2}\left(F_{ij} + \frac{C_{ij}}{2}\right) + a_{ij}\frac{C_{ij}}{\sqrt{2}},$$

where M_{ij} is a binary indicator for mother, F_{ij} for father and C_{ij} for children. The family-specific

random effects a_{i1} and a_{i2} induce required correlation between parents and each child and between the children. However, the induced variance for children from these two random effects is only half of the desirable variance and the other half is induced by the subject-specific random effects a_{ij}. By this reparameterization, we can fit a semiparametric model with polygenic effect by standard software.

2.2. Semivarying coefficient partially linear mixed effects model for quantitative trait locus effect

When genetic markers such as SNPs are available, we add marker genotypes to model (1) to assess association between a marker and a trait. Owing to dense SNP genotyping, we assume that the QTL is either at the SNP marker under consideration or tightly linked to it.

Let g_{ij} denote the SNP marker genotype for subject j in family i coded as the copies of minor alleles which take value 0, 1 or 2. Let $w_{ij}(t)$ denote time-dependent exposures with a potentially time varying effect such as body mass index. A semivarying coefficient partially linear mixed effects model for Y_{ijh} is

$$Y_{ijh} = \mu(t_{ijh}) + x_{ij}^{\mathrm{T}}\beta + \alpha_i + z_{ijh}^{\mathrm{T}}\tilde{\gamma}_{ij} + \beta_g(t_{ijh})g_{ij} + w_{ij}^{\mathrm{T}}(t_{ijh})\theta(t_{ijh}) + \varepsilon_{ijh}, \tag{4}$$

where $\tilde{\gamma}_{ij}$ is the residual polygenic effect aside from the QTL effect contributed by other unlinked loci, and $\theta(t)$ is the coefficient vector for covariates $w_{ij}(t)$. In this model, in addition to the baseline function $\mu(t)$ and other covariate effects, we are interested in estimating the time varying genetic function $\beta_g(t)$.

The age-specific QTL heritability is then defined as (Falconer, 1985)

$$h_g^2(t) = \sigma_g^2(t)/\sigma_T^2(t), \tag{5}$$

where $\sigma_g^2(t) = \mathrm{var}\{\beta_g(t)g_{ij}\} = \beta_g^2(t)\,\mathrm{var}(g_{ij})$ is the QTL genetic component, $\sigma_A^2(t) = \sigma_\alpha^2 + \sigma_\gamma^2(t) + \sigma_\varepsilon^2(t)$ is the sum of remaining components and $\sigma_T^2(t) = \sigma_g^2(t) + \sigma_A^2(t)$. The QTL heritability can be interpreted as the proportion of total variation explained by the QTL. The residual polygenic heritability contributed by other unlinked loci is $h_\gamma^2(t) = \sigma_\gamma^2(t)/\sigma_T^2(t)$. The total heritability in a trait is the sum of the QTL heritability and the residual polygenic heritability.

To test for association between a genetic marker and a trait, we consider the null hypothesis $H_0 : \beta_g(t) = 0$. To test for a constant genetic effect (i.e. the genetic effect does not change over time), we consider the null hypothesis $H_0 : \beta_g(t) = \beta_g$.

3. Statistical methods for longitudinal genetic studies

3.1. Estimation procedure for the partially linear mixed effects model

For simplicity, we use a truncated polynomial basis in our estimation procedure. Extension to other bases such as B-splines is discussed in Section 6. We approximate the mean function by a linear combination of spline basis functions

$$\mu(t) \approx \eta_0 + \eta_1 t + \ldots + \eta_q t^q + \sum_{m=1}^{M} \eta_{q+m}(t - \tau_m)_+^q,$$

where τ_m, $m = 1, \ldots, M$, is a given sequence of knots and q is the order of the splines. For given variance components, we estimate η and β as minimizers of a penalized weighted least square (Wu and Zhang (2006), chapter 7.3),

$$\tfrac{1}{2}(Y - X\beta - W\eta)^{\mathrm{T}}\Sigma^{-1}(Y - X\beta - W\eta) + \tfrac{1}{2}\lambda\eta^{\mathrm{T}}J\eta, \tag{6}$$

where Y is a vector of outcomes, Σ is the covariance of Y, $J = \mathrm{diag}(\mathbf{0}_{q+1}, \mathbf{1}_M)$ is a penalty matrix for truncated polynomial basis, X and W are design matrices specified in the supporting information that is associated with this paper and available on line and λ is a smoothing parameter. When λ goes to ∞, the spline coefficients are shrunk towards 0 and the fit converges to a polynomial function. When λ goes to 0, the fit converges to a weighted least square. The estimating equations for η and β are constructed in the supporting information. The solution for η takes the form of a ridge regression estimate. The variance components in Σ are estimated by maximizing a restricted likelihood as in a mixed effects model analysis.

It is well known that there is a mixed effects model representation of penalized splines (Ruppert *et al.*, 2003; Wand, 2003). We explore this connection to facilitate computation using standard software. For penalized splines, Wand (2003) showed that with a proper choice of smoothing parameter which we describe in the supporting information the solution to expression (6) is identical to the best linear unbiased predictor from a linear mixed effects model. The key is to specify the spline coefficients $\eta_{q+1}, \ldots, \eta_{q+M}$ as random effects with the same variance and to construct appropriate design matrices for the fixed and random effects.

The tuning parameters for penalized splines include number and placement of knots and smoothing parameter λ. Once the number of knots has been chosen, we place them at equal sample quantiles of the observed t_{ijh}s. The smoothness of the fit is controlled by both M and λ. Ruppert (2002) showed that, when M is adequately large, further increasing M does not improve the fit and can sometimes deteriorate the fit. For smooth functions that are either monotonic or unimodal, a moderate number of knots is usually sufficient (Ruppert, 2002). The smoothing parameter λ controls overfitting for a moderate to large number of knots and plays a more critical role than M.

For a given number of knots, λ can be chosen by generalized cross-validation, minimizing Akaike's information criterion AIC or estimating by restricted maximum likelihood. Krivobokova and Kauermann (2007) investigated the behaviour of several data-driven smoothing parameter selectors including restricted maximum likelihood and AIC with correlated data. They found through theoretical derivation and simulations that, when the correlation structure is misspecified, the AIC-based choice failed to estimate a function properly and the choice based on restricted maximum likelihood provides a much more satisfactory fit and exhibits less variability (Krivobokova and Kauermann, 2007). Here we use restricted maximum likelihood to estimate the smoothing parameters as shown in the on-line appendix.

3.2. Estimation procedure for the semivarying-coefficient linear mixed effects model

For model (4), we also approximate $\beta_g(t)$ by a linear combination of basis functions, i.e.

$$\beta_g(t) \approx \xi_0 + \xi_1 t + \ldots + \xi_q t^q + \sum_{m=1}^{M} \xi_{q+m}(t - \tau_m)_+^q. \tag{7}$$

Varying coefficients $\theta(t)$ can be handled in a similar fashion by the approximation

$$\theta(t) \approx \theta_0 + \theta_1 t + \ldots + \theta_q t^q + \sum_{m=1}^{M} \theta_{q+m}(t - \tau_m)_+^q.$$

For given variance components, the penalized weighted least square of β, η and ξ is

$$-\tfrac{1}{2} r' \Sigma^{-1} r - \tfrac{1}{2}\lambda_1 \eta' J \eta - \tfrac{1}{2}\lambda_2 \xi' J \xi - \tfrac{1}{2}\lambda_3 \theta' J \theta,$$

where $r = (Y - X\beta - W\eta - S_1\xi - S_2\theta)$, the design matrices S_1 and S_2 are defined in the supporting information, and λ_1, λ_2 and λ_3 are smoothing parameters for the baseline, the genetic

effect function and the varying coefficient for other covariates respectively. In the supporting information, we expand the mixed effects model that is used to fit expression (1) to obtain the coefficients for time varying genetic effects. As described there, we select λ_1, λ_2 and λ_3 by treating them as extra variance components and estimating by restricted maximum likelihood.

3.3. Estimating the total variance

Since the total phenotypic variance function $\sigma_T^2(t)$ is involved in heritability function (3), non-parametric estimation is desirable. Fan *et al.* (2007) proposed a semiparametric estimator of the covariance function $\vartheta(s,t)$. They assumed that the correlation function has a parametric form, i.e. $\vartheta(s,t) = \mathrm{cov}\{\varepsilon_{ij}(s), \varepsilon_{ij}(t)\} = \rho_\varepsilon(s,t;\nu)$, where ρ is a known function and ν is a vector of parameters. They estimated the variance function $\vartheta(t,t) = \mathrm{var}\{\varepsilon_{ij}(t)\}$ non-parametrically through local kernel smoothing. Here although the total variance is the summation of all variance components, we take a semiparametric approach by estimating it directly from observed data to protect further against potential model misspecification of some of the components. We propose a penalized-spline-based approach.

Recall that the total variance function equals the sum of the major QTL genetic component and remaining variance components ($\sigma_T^2(t) = \sigma_g^2(t) + \sigma_A^2(t) = \beta_g^2(t)\,\mathrm{var}(g_{ij}) + \sigma_A^2(t)$). We estimate the non-QTL components, i.e. $\sigma_A^2(t) = \sigma_\alpha^2 + \sigma_\gamma^2(t) + \sigma_\varepsilon^2(t)$, through penalized splines regression based on the residuals after subtracting the fitted mean curves and fixed effects from Y_{ijh}, but not any of the random-variance components. Therefore they retain the variability in the random-variance components in $\sigma_A^2(t)$. Let

$$\hat{\varepsilon}_{ijh} = Y_{ijh} - \hat{\mu}(t_{ijh}) - x_{ijh}\hat{\beta} - \hat{\beta}_g(t_{ijh})g_{ij} - \hat{\theta}(t_{ijh})\, w_{ij}(t_{ijh}),$$

where $\hat{\mu}(\cdot)$, $\hat{\beta}_g(\cdot)$ and $\hat{\theta}$ are the fitted values of the mean, the QTL genetic function and varying coefficient. Similarly to the estimation of η and β_g, we express $\log\{\sigma_A^2(t)\}$ as a linear combination of basis functions,

$$\log\{\sigma_A^2(t)\} \approx \rho_0 + \rho_1 t + \ldots + \rho_q t^q + \sum_{m=1}^{M} \rho_{q+m}(t-\tau_m)_+^q.$$

We then estimate ρ by fitting a penalized splines regression to $\log(\hat{\varepsilon}_{ijh}^2)$. Using the estimated fixed coefficients and the best linear unbiased predictor of the random effects, the fitted value of the sum of $\sigma_A^2(t)$ will be

$$\hat{\sigma}_A^2(t_{ijh}) = \exp\left\{\hat{\rho}_0 + \ldots + \hat{\rho}_q t_{ijh}^q + \sum_{m=1}^{M} \hat{\rho}_{q+m}(t_{ijh}-\tau_m)_+^q\right\}. \tag{8}$$

The estimated total variance is $\hat{\sigma}_T^2(t) = \hat{\sigma}_A^2(t) + \hat{\sigma}_g^2(t) = \hat{\sigma}_A^2(t) + \hat{\beta}^2(t)\,\mathrm{var}(g_{ij})$. Since the estimated total variance is used to calculate heritability in equations (3) and (5), we evaluate the performance of this procedure through examining mean average squared errors (MASEs), the mean bias and the variance of heritability estimates in Section 4.

3.4. Testing for association between a marker and a trait

When the QTL genetic effect is time invariant, the hypothesis of no association between a marker and a trait is specified by $H_0 : \beta_g = 0$ *versus* $H_a : \beta_g \neq 0$, which can be examined by a standard Wald test. When fitting a time varying QTL model, the hypothesis of no association is

$$H_0 : \xi_0 = \xi_1 = \ldots \xi_q = 0, \qquad \text{and } \sigma_\xi^2 = 0, \tag{9}$$

where $\xi_1, \ldots, \xi_q$ are coefficients for polynomial terms defined in approximation (7) and σ_ξ^2 is the variance of the random-spline coefficients $\xi_{q+1}, \ldots, \xi_{q+M}$ as described in the supporting information. This hypothesis can be examined by a likelihood ratio test. Crainiceanu and Ruppert (2004) showed that the distribution of the likelihood ratio test of hypothesis (9) for a penalized splines mixed model involving a variance component is non-standard owing to lack of independence, and using a conventional 50:50 mixture of χ^2-distributions may be conservative.

Here we propose to compute the p-value by a permutation- and simulation-based procedure. Since under the null hypothesis the marker genotypes are not associated with the trait, we can permute genotypes among subjects. However, it is not straightforward to randomize genotypes in a family sample. Even though there is no major QTL effect under the null hypothesis, there may be a residual polygenic effect causing family members to be correlated. Therefore family members are not exchangeable under the null hypothesis and simple permutation would not maintain phenotype correlation between related individuals.

Yang *et al.* (2010) proposed permuting genotypes among founders and then simulating offspring genotypes conditionally on permuted founders' genotypes based on a Mendelian law while keeping the phenotypes as observed. Specifically, we first permute the genotypes of founders (subjects who do not have parents) in all families. Given a set of permuted founders' genotypes, we generate an offspring's genotype by randomly selecting an allele from each parent of the offspring following the Mendelian law. Genotypes of siblings in the same family are assigned independently given their permuted parental genotypes. For each copy of permuted genotype data, the same model fitting procedure is carried out. In a genomewide association study, it is computationally challenging to conduct permutations for every SNP. Since the null distribution of the test statistic is the same for SNPs with the same founder genotype frequency for a given family data, we can group SNPs into strata that have the same or similar founder genotype frequency, and only one permutation null distribution is needed for each group (Yang *et al.*, 2010).

If the data consist of only one pedigree with all founders carrying the same homozygote genotypes, then all descendants will carry the same genotype and the design matrix of genetic effect will be singular with rank 1. Thus no association test can be performed and no permutation–simulation procedure can be applied. When data consist of multiple homozygous pedigrees as well as other pedigrees containing different founder genotypes, assuming random mating in founder generation we permuted founders from all pedigrees before simulating transmissions and therefore the size of the type I error will be maintained. The key assumptions for this procedure to be valid include random mating of founders and Mendelian transmission of descendants' alleles.

4. Simulations

In this section, we investigate the performance of our proposed estimation and testing procedures through Monte Carlo simulations. We simulated 100 nuclear families among which 50 had two children, 30 had three and 20 had four. The number of observations of each parent ranged from 4 to 8, the number of observations for children ranged from 2 to 4 and each subject was examined every 2 or 4 years. The total number of observations was 1749. Subjects' ages ranged from 10 to 75 years with a mean of 39.5 years. These settings were close to the assessment schedule in the FHS. For the analysis involving genetic markers, we simulated a fully linked genetic polymorphism with a dominant effect and a minor allele frequency of 0.25. We assumed that the transmission of allele from parental generation to offspring generation follows a Mendelian law.

4.1. Time invariant genetic effect

In the first few simulations, the baseline function $\mu(t)$ was a logarithm function, $-34.2 + 81.7 \log\{0.25(t+21.7)\}$, where the parameters were estimated from fitting a logarithm function to the FHS cholesterol data. This function was used to simulate the baseline and the genetic effect function on several traits at Genetic Analysis Workshop 13 (Daw *et al.*, 2003), where the simulations were designed to mock the actual FHS data that were provided at the workshop. The random shared familial environmental factor α_i had a variance of 16, and the polygenic effect γ_{ij} had a variance of 4. These parameters were chosen so that the polygenic heritability is in the range of that estimated by the FHS investigators (Levy *et al.*, 2000). The variance function of residuals was an exponential function, $\text{var}\{\varepsilon_{ij}(t)\} = \exp(0.02t)$. The correlation of the residuals was auto-regressive AR(1) with auto-correlation parameter 0.6. We also examine other functional forms of $\mu(t)$ such as the Gaussian function and the sine function. The baseline function was estimated by cubic truncated polynomials with 15 knots.

In simulation setting 1, we assumed $\beta_g(t) = \beta_g$ in model (4), where the true values of β_g are shown in Table 1. We computed the MASE of $\hat{\mu}(t)$ as the mean across the 500 simulations of the average squared error,

$$\text{ASE}(\mu) = \frac{1}{K} \sum_{t_k \in T_\kappa} \{\hat{\mu}(t_k) - \mu(t_k)\}^2,$$

where T_κ is a set of grid points over time and K is the cardinality of T_κ. Define the MASE of $\hat{h}^2_\gamma(t)$ and $\hat{h}^2_g(t)$ similarly. We summarize the maximal absolute relative bias, the mean bias and the mean variance averaged over grid points T_κ, and the MASE in the second to fifth columns of Table 1 (setting 1), which showed a small relative bias. The estimated time invariant genetic effect was 9.99 (true value 10), with a mean estimated standard error of 0.26 (empirical standard error 0.27).

We compare proposed semiparametric analyses where $\mu(t)$ was estimated through penalized splines with a correctly specified non-linear mixed effects model analysis and a misspecified parametric analysis where $\mu(t)$ was assumed to be a quadratic polynomial function. The results are recorded in the last six columns of Table 1 (setting 1). As expected, it is evident that when $\mu(t)$ is misspecified its estimation had large bias. It may be of interest to note that misspecification of the baseline function also affects estimation of the heritabilities. The mean bias in the marker-specific and the total heritabilities ($\hat{h}^2_g(t)$ and $\hat{h}^2_T(t)$) increased by 43% and 54% respectively, when $\mu(t)$ was misspecified. In terms of estimating the baseline function, the average variance of $\hat{\mu}(t)$ and $\hat{\beta}_g(t)$ is larger for the semiparametric analysis than for the parametric analysis under a correctly specified model. For the heritability estimators, the loss of efficiency of the semiparametric estimators is less notable.

For the variance components, the estimated polygenic variance was 4.06 (true value 4), and the family-specific variance component was 15.88 (true value 16). The asymptotic distribution of the heritability estimates is not straightforward to derive because the definition of the heritability is the ratio of two non-independent variance estimators. To compute the confidence interval (CI), we use non-parametric bootstrap resampling: we first resample family and then resample subjects within a family. As seen from Table 1, the maximal relative bias and MASE of the QTL heritability and the total heritability were small. We present the estimated marker-specific heritability, the total heritability and their CIs in Figs 1(a) and 1(b). The empirical and bootstrap standard errors were compared in Figs 1(c) and 1(d). The bootstrap standard error tracked the empirical standard error closely.

Our next simulation experiments examined effects of different baseline function estimators on testing a genetic effect. We simulated data under the same model (4) with $\beta_g(t) = \beta_g$, various

Table 1. Bias and MASE of the estimated functions in the time invariant (setting 1) and time varying (setting 2) analyses

Function	*Results for non-parametric model*†				*Results for misspecified model*‡			*Results for correctly specified model*§		
	Maximum relative bias§§	*Mean bias**	*Mean variance***	*MASE*$(\hat{f})$	*Maximum relative bias*	*Mean bias*	*Mean variance*	*Maximum relative bias*	*Mean bias*	*Mean variance*
Setting 1										
$\hat{\mu}(t)$	0.001	0.057	0.45	0.46	0.044	1.001	0.256	0.003	0.011	0.25
$\hat{h}_g^2(t)$	0.07	0.007	0.004	0.0032	0.18	0.01	0.0038	0.134	0.008	0.003
$\hat{h}_T^2(t)$	0.04	0.013	0.005	0.0041	0.19	0.02	0.0049	0.108	0.012	0.0043
Setting 2										
$\hat{\mu}(t)$	0.001	0.056	0.48	0.49	1.01	0.91	0.31	0.0002	0.012	0.25
$\hat{\beta}_g(t)$	0.007	0.036	0.638	0.92	0.45	0.23	0.36	0.007	0.0024	0.18
$\hat{h}_g^2(t)$	0.02	0.004	0.005	0.0043	0.42	0.34	0.30	0.012	0.0038	0.004
$\hat{h}_T^2(t)$	0.01	0.006	0.005	0.0049	0.51	0.37	0.36	0.001	0.0023	0.003

†$\mu(t)$ estimated non-parametrically by penalized splines.
‡$\mu(t)$ misspecified as $\mu(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$.
§$\mu(t)$ estimated in a non-linear mixed effects model with a correctly specified functional form.
§§Maximum relative bias is defined as $\max_{t_k \in T_\kappa} |\text{mean}\{\hat{f}(t_k)\} - f(t_k)|/f(t_k)$, where T_κ is a set of grid points and the average is taken over all repetitions of the simulation.
*Mean bias is defined as $(1/K)\Sigma_{t_k \in T_\kappa} \text{mean}\{\hat{f}(t_k) - f(t_k)\}$, where K is the cardinality of T_κ and the average is taken over all repetitions of the simulation.
**Mean empirical variance is defined as $(1/K)\Sigma_{t_k \in T_\kappa} \text{var}\{\hat{f}(t_k) - f(t_k)\}$, where K is the cardinality of T_κ and the variance is taken over all repetitions of the simulation.

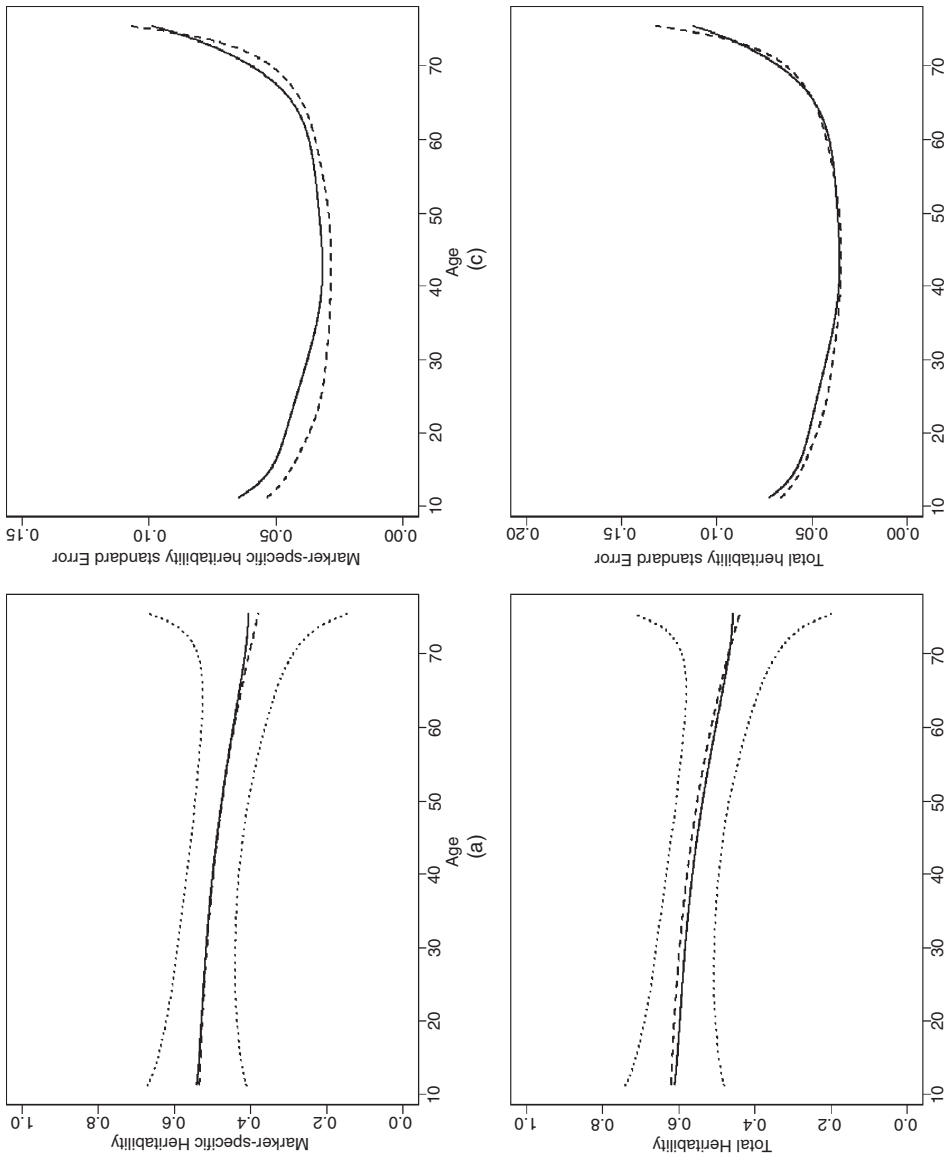

Fig. 1. Time invariant genetic effect model: (a) estimated marker-specific heritability (———, true; - - - - - - , estimated; · · · · · · , 95% CI); (b) total heritability (———, true; - - - - - - , estimated; · · · · · · , 95% CI); (c) marker-specific heritability (———, empirical standard errors; - - - - - - , bootstrap standard error); (d) total heritability (———, empirical standard error; - - - - - - , bootstrap standard error)

effect sizes of the genetic marker and various functional forms of $\mu(t)$ (see Table 2 for these specifications). The random measurement errors were simulated from a normal distribution with mean 0 and variance 10. We tested the significance of $\hat{\beta}_g$ by a standard Wald test. We compared the performance of the proposed semiparametric analysis where $\mu(t)$ is estimated through penalized splines with three other analyses:

(a) misspecifying $\mu(t)$ as a linear function;
(b) misspecifying $\mu(t)$ as a quadratic function;
(c) correctly specifying $\mu(t)$ as a non-linear function and estimating by fitting a non-linear mixed effects model.

First, we examine the type I error of all four analyses. From the second, sixth and 10th rows of Table 2, we see that the semiparametric analysis and the correctly specified non-linear analysis maintain the nominal level of type I errors. However, the two misspecified analyses reported either substantially inflated or highly conservative type I error rates depending on the true form of $\mu(t)$ and how it is specified. For example, when the true baseline function is a sine function but is misspecified as a linear or a quadratic polynomial, the type I error rate for a test for β_g at 5% level can be as high as 0.99. The erroneous type I error may happen for two reasons: first, incorrect estimation of the baseline function under a misspecified model may lead to an incorrect standard error estimate of $\hat{\beta}_g$; second, the mean of $\mu(T)$ across observed time points is not constant across different genotype groups, i.e. $E[\mu(T)|G=g]$ differs across levels of G in a partially linear model which may lead to an inconsistent estimate of $\hat{\beta}_g$.

Next we compared the power of the test when $\mu(t)$ is estimated non-parametrically with the correctly specified non-linear analysis. From Table 2, we see that the power for testing genetic

Table 2. Power for testing $\beta_g = 0$ assuming $\mu(t)$ to be a non-parametric function, misspecified parametric functions and a correctly specified non-linear function (α-level 0.05)

$\mu(t)$	*Analysis*	*β_g*	*Results for the following functions:*			
			Non-parametric†	*Misspecified: linear*‡	*Misspecified: quadratic*§	*Correctly specified*§
Logarithm*	Type I error	0	0.048	0.012	0.024	0.048
Logarithm	Power	0.5	0.51	0.06	0.46	0.54
Logarithm	Power	0.75	0.79	0.2	0.76	0.79
Logarithm	Power	1	0.94	0.53	0.93	0.96
Gaussian**	Type I error	0	0.046	0.85	0.004	0.058
Gaussian	Power	0.5	0.49	—	0	0.53
Gaussian	Power	0.75	0.93	—	0	0.93
Gaussian	Power	1	0.94	—	0	0.99
Sine††	Type I error	0	0.045	0.99	0.99	0.048
Sine	Power	0.5	0.46	—	—	0.49
Sine	Power	0.75	0.84	—	—	0.86
Sine	Power	1	0.95	—	—	0.96

†$\mu(t)$ estimated non-parametrically by penalized splines.
‡$\mu(t)$ misspecified as $\mu(t)=\alpha_0+\alpha_1 t$.
§$\mu(t)$ misspecified as $\mu(t)=\alpha_0+\alpha_1 t+\alpha_2 t^2$.
§§$\mu(t)$ estimated in a non-linear mixed effects model with a correctly specified functional form.
*True $\mu(t)=-34.2+81.7\log\{0.3(t+21.7)\}$.
**True $\mu(t)=200\exp\{-0.002(t-39)^2\}$.
††True $\mu(t)=150+50\sin(0.2t)$.

effects is slightly larger with a correctly specified non-linear baseline function as compared with the semiparametric analysis, with a difference up to 5%. For the scenarios in Table 2 where the two misspecified analyses had conservative type I errors, we also examined their power. As expected, the power was greatly reduced with a power loss up to 95% compared with the semiparametric analysis. For example, when the true $\mu(t)$ is a Gaussian function but misspecified as a linear or a quadratic function, the power for detecting a genetic effect was 0. In addition to a highly conservative type I error rate, this may also be due to a substantial increase in variability of the estimator $\hat{\beta}_g$ when the baseline function was misspecified in these cases.

To summarize, the first set of simulations suggest that misspecification of the baseline function has a non-ignorable effect on the type I error of testing the genetic effect even when the genetic effect does not change with time. Furthermore, the power of testing β_g when treating $\mu(t)$ as a non-parametric function is comparable with correctly specifying $\mu(t)$ as a non-linear function.

4.2. Time varying genetic effect

The second simulation setting examines properties of our methods when $\beta_g(t)$ changes with time. The performance of the baseline function estimator was comparable with the time invariant case (Table 1, setting 2). From Table 1 (setting 2), we see that the time varying genetic effect $\hat{\beta}_g(t)$ was estimated well with small MASE. We show the true and the estimated genetic effect and its confidence interval in Fig. 2(a). The bootstrap standard error and the empirical standard error shown in Fig. 2(b) were very close. The age-specific QTL heritability and total heritability were estimated well with maximal relative biases 0.02 and 0.01 respectively (Table 1, setting 2). The bootstrap and empirical standard error were close, which suggests satisfactory performance of the bootstrap procedure on assessing variabilities of heritability estimates (the results are similar to Fig. 1 and so are not shown).

Next, we compared the estimation bias and MASE of $\hat{\beta}_g(t)$ in a semiparametric analysis with a misspecified parametric analysis where we assumed $\beta_g(t)$ to be a quadratic polynomial and with a correctly specified non-linear mixed effects model analysis. In all analyses, we kept the estimation of $\mu(t)$ non-parametric because the analyses in the previous section showed a profound effect of misspecifying $\mu(t)$ on testing β_g. From Table 1 (setting 2), we see that the mean bias of $\hat{\beta}_g(t)$ over time increased from 0.036 in a non-parametric method to 0.23 for a misspecified quadratic model. The mean bias of the estimated marker-specific heritability, $\hat{h}^2_m(t)$, increased from 0.004 to 0.34, which is substantial. The mean bias of the estimated total heritability increased from 0.006 to 0.37.

The rest of the simulations concern testing of $\beta_g(t)$. The random measurement errors were simulated from a normal distribution with mean 0 and variance 6. The hypothesis $\beta_g(t) = 0$ was tested by the permutation procedure that was described in Section 3.5 in the semiparametric analysis. In all analyses, the baseline function was again estimated non-parametrically. We examine several functional forms for $\beta_g(t)$ including logarithmic, Gaussian and sine. The Gaussian function was used to model genetic effects on blood pressure in Shi and Rao (2008).

We first examine the type I error of the semiparametric analysis and two parametric analyses under a misspecified model. From Table 3, we see that all three analyses maintain the correct nominal level of type I error. We then examine the power of testing $\beta_g(t) = 0$. Again as expected, we see from Table 4 that the power is greatest for the non-linear mixed effects model analysis with a correctly specified model. However, in real applications such a true function is unknown and the computational algorithm in a non-linear analysis may not converge, especially when

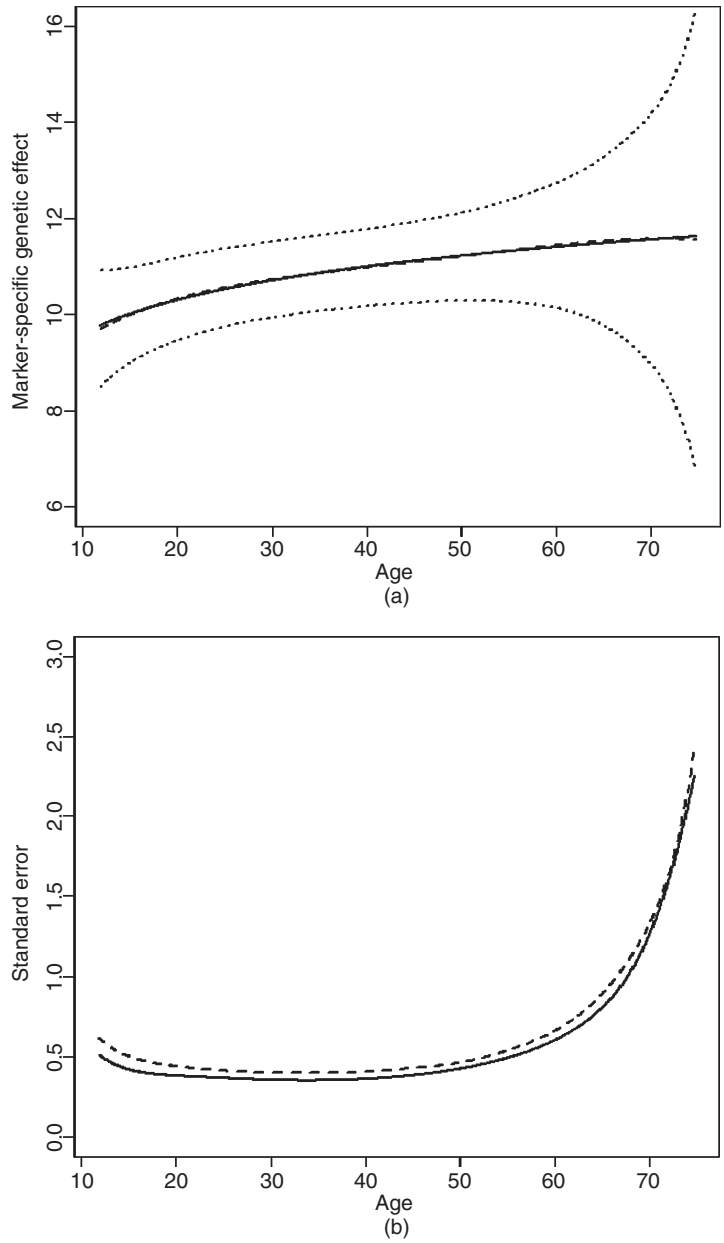

Fig. 2. Time-varying genetic effect model: (a) estimated age-specific genetic effect (———, true; - - - - -, estimated; ·····, 95% CI); (b) bootstrap (- - - - -) and empirical (———) standard error

starting values are poor or the sample size is small or moderate. Comparing the semiparametric approach with the misspecified parametric approaches, the power loss for the latter ranges from 0% to 55%. The power loss was more substantial for the Gaussian and sine functions, compared with the logarithm function. This suggests that power depends on the unknown functional form of the true genetic effect and the assumed parametric model. For the genetic effect that changes with time but has an average effect of zero across all time points, i.e.

Table 3. Type I error of the permutation test and the misspecified parametric analyses for testing $\beta_g(t) = 0$ in model (4)

α-level	*Results for the following functions:*		
	Non-parametric†	*Misspecified: linear*‡	*Misspecified: quadratic*§
0.005	0.0054	0.0045	0.0055
0.01	0.0128	0.01	0.008
0.05	0.0488	0.0515	0.0505
0.1	0.0914	0.1	0.1015

†$\mu(t)$ estimated non-parametrically by penalized splines.
‡$\mu(t)$ misspecified as $\mu(t) = \alpha_0 + \alpha_1 t$.
§$\mu(t)$ misspecified as $\mu(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$.

Table 4. Power for testing $\beta_g(t) = 0$ assuming $\beta_g(t)$ to be a non-parametric function, misspecified parametric functions and a correctly specified non-linear function† (α-level 0.05)

$\beta_g(t)$	*Mean* $\beta_g(t)$ *over t*	*Results for the following functions:*			
		Non-parametric‡	*Misspecified: linear*§	*Misspecified: quadratic*§§	*Correctly specified**
$\log(0.2t)/10 - 0.2$	−0.004	0.19	0.19	0.08	0.19
$\log(0.5t)/10 + 0.2$	0.49	0.39	0.39	0.34	0.39
$0.8 + 0.1\log(0.5t)$	1.09	0.98	0.98	0.98	0.98
$3\exp\{-0.075(t-39)^2\} - 0.5$	−0.001	0.73	0.04	0.42	0.99
$0.9\exp\{-0.025(t-39)^2\} + 0.2$	0.45	0.35	0.34	0.37	0.77
$1.5\exp\{-0.075(t-39)^2\} + 0.6$	0.85	0.86	0.85	0.85	0.99
$0.85\sin(0.2t) + 0.02$	0.01	0.60	0.06	0.05	0.94
$0.85\sin(0.2t) + 0.5$	0.49	0.51	0.36	0.3	0.86
$0.85\sin(0.2t) + 0.85$	0.84	0.87	0.81	0.78	0.96

†In all analyses $\mu(t)$ was estimated non-parametrically by penalized splines.
‡$\beta_g(t)$ estimated non-parametrically.
§$\beta_g(t)$ misspecified as $\beta_g(t) = \alpha_0 + \alpha_1 t$.
§§$\beta_g(t)$ misspecified as $\beta_g(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2$.
*$\beta_g(t)$ estimated in a non-linear mixed effects model with a correctly specified functional form.

$$\frac{1}{\sum_{ij} T_{ij}} \sum_{ijh} \beta_g(t_{ijh}) \approx 0,$$

the linear or quadratic analysis has very low power (close to zero) to detect the genetic effect.

To summarize, these simulations suggest that misspecifying $\beta_g(t)$ while estimating $\mu(t)$ non-parametrically does not affect the type I error rate of testing $\beta_g(t) = 0$ but may reduce power substantially.

5. Application to the Framingham Heart Study

In this section, we apply the proposed methods to analyse the FHS longitudinal blood pressure data and SNP genotype data. High blood pressure is considered a major risk factor for stroke

and heart disease and it affects about a third of the US adult population (Levy *et al.*, 2009). SBP and diastolic blood pressure (DBP) are complex traits that may be influenced by both environmental and genetic factors. The heritability of SBP is estimated to be high (30–60%; Levy *et al.* (2000)), which suggests a substantial genetic contribution. Recently, large-scale genomewide association studies have emerged as powerful tools to identify genes that are associated with complex traits such as blood pressure. Levy *et al.* (2009) performed a prospective meta-analysis on six genomewide association studies including the FHS and identified multiple SNPs significantly associated with SBP and DBP at the genomewide significance level. However, nonparametric estimation of the time varying polygenic effect or the age-specific QTL effect of blood pressure has not been examined in the literature. We analyse a subset of the FHS subjects (about 6000 subjects) and a subset of SNPs in four candidate regions.

In the FHS, the phenotype and the genotype data are collected from three cohorts. The original Framingham cohort (cohort 1) was first examined in 1948 and has been examined every 2 years thereafter. The offspring cohort (cohort 2), composed primarily of offspring of the original cohort and the spouses of these offspring, was examined first in 1971 and has been examined approximately every 4 years by using protocols that are similar to those used for study of the original cohort. Between 2002 and 2005 the study enrolled the third generation of the FHS. At each examination, the physician measured SBP and DBP twice and the average of the two measurements was used as the phenotype in the analysis.

Although the FHS started in an era when no antihypertensive treatment was available, as the study progressed, antihypertensive treatment became available and was prescribed to some of the subjects with hypertension. It is known that the treatment effect is a confounder for genetic effect which may lead to an underestimated genetic effect without any adjustment (Levy *et al.*, 2000; Tobin *et al.*, 2005). Tobin *et al.* (2005) examined the bias and variance of 10 methods of adjusting for treatment effect and found that one of the best methods is to add a reasonable number to observed SBP for subjects on antihypertensive treatment. Following Tobin *et al.* (2005) and Levy *et al.* (2009), we added 10 mmHg to observed SBP values and 5 mmHg to observed DBP values for participants who were in treatment.

The majority of the observations in the FHS were measured between age 30 and 75 years and this age range is of most scientific interest. To obtain stable estimates, we restricted the analysis sample to between age 30 and 75 years. The total sample size in our analysis was 6082 from 930 pedigrees (including 2934 nuclear families) and the mean number of subjects was 6.54 per extended family. There were 14505 observations and each subject had an average of 2.38 measurements of SBP and DBP. The age of the participants at the first visit ranged from 25 to 72 years. The mean age for all subjects at all visits was 45.7 years. The mean observed SBP was 121.2 mmHg and the mean observed DBP was 76.1 mmHg. 11% of subjects were on antihypertensive treatment in at least one examination and 12% of observations were taken when subjects were on treatment. The mean body mass index was 23.54.

In all our analyses, we included gender as a covariate with time invariant effect and body mass index as a time varying covariate with varying coefficient and used a linear design matrix for the random polygenic effect. We estimated the baseline function by a cubic truncated polynomial with 10 knots. We split pedigrees into nuclear families for easy handling of familial correlations. We first computed the baseline function and the polygenic heritability without using SNP markers as in model (1). The estimated age-specific baseline function and its 95% CI are superimposed on SBP measurements of 150 randomly selected subjects in Fig. 3. There is an increasing trend of mean SBP over time. The mean SBP was 123.5 mmHg (CI 122.9, 124.2) at age 30 years and increased to 138.6 mmHg (CI 134.8, 142.4) at age 75 years. The corresponding plot for DBP is shown in Fig. 4. The polygenic heritability of SBP has a decreasing trend. Heritability was

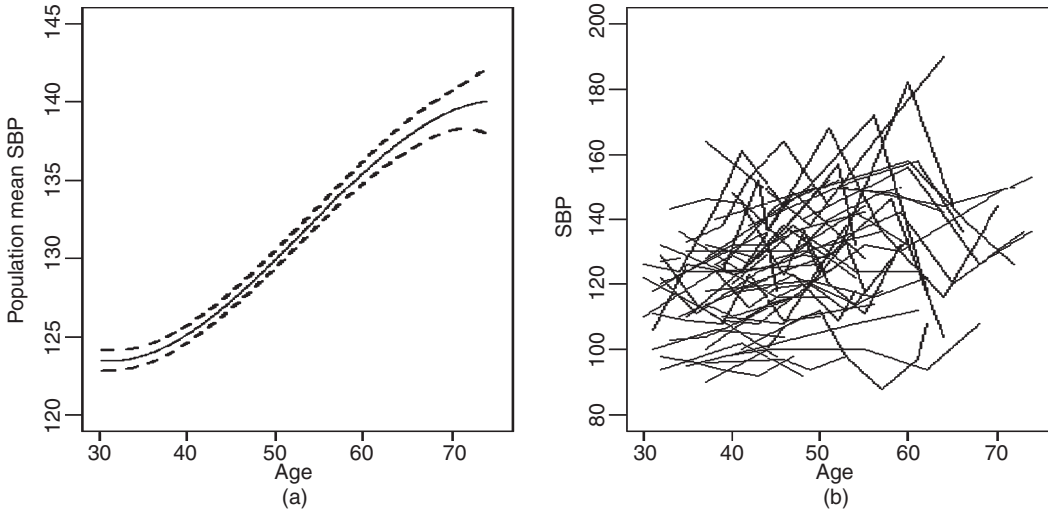

Fig. 3. (a) Estimated population mean function of SBP (———, estimates; – – –, 95% CI) and (b) scatter plot for 150 randomly selected subjects in the FHS

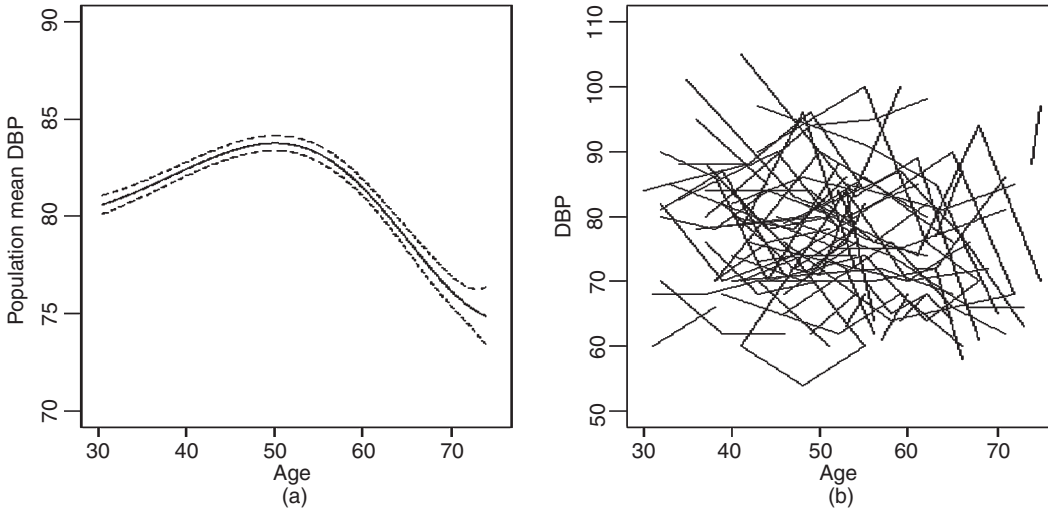

Fig. 4. (a) Estimated population mean function of DBP (———, estimates; – – –, 95% CI) and (b) scatter plot for 150 randomly selected subjects in the FHS

highest at age 35 years and it then decreased to 0.44 (CI 0.40, 0.50) at age 50 years and 0.23 at age 65 years (CI 0.18, 0.27). The long-term average heritability was reported to be between 0.3 and 0.6 (Levy *et al.*, 2000, 2009), which is in the range of our age-specific estimates. The total variance function increases over time. For DBP, the polygenic heritability also decreases with age. It was 0.44 (CI 0.37, 0.54) at age 35.4 years and then decreased to 0.29 (CI 0.17, 0.47) at age 75 years. Overall, DBP exhibits lower heritability than SBP. The gender effect was estimated as 1.57 (CI 0.80, 2.34) with men having higher SBP, on average.

Levy *et al.* (2009) conducted a meta-analysis of six genomewide association studies of blood pressure and reported several promising regions which may harbour genes predisposing blood pressure. We selected four promising candidate regions containing significant SNPs reported in

Levy *et al.* (2009) to analyse. There were 265 SNPs in the four regions. Among these, 109 SNPs were from two regions on chromosome 12 (86 from region 88300 kilobases to 88800 kilobases and 23 from region 110200 kilobases to 110600 kilobases), 104 were from a region on chromosome 11 (16600 kilobases to 17100 kilobases) and 52 were from a region on chromosome 3 (41700 kilobases to 42100 kilobases). Each of these regions spans about 500 kilobases on a chromosome. We first fit a time invariant model with a non-parametric baseline function but a constant genetic effect (i.e. $\beta_g(t) = \beta_g$) in model (4). Since adjusting for multiple comparisons by Bonferroni correction is conservative for dense SNPs in linkage disequilibrium, we used methods that were proposed in Gao *et al.* (2008). Specifically, we used principal components analysis to compute the effective number of SNPs needed to explain 99.5% of variability of all 234 SNPs and then divided the overall significance level (0.05) by this number. The resulting effective number of SNPs needed is 104, and the adjusted significance level is 4.81×10^{-4}. There was one SNP on chromosome 12 significant for SBP at this level and none for DBP (Table 5).

In addition to the time invariant analysis, we also fitted a time varying genetic effect model and tested hypothesis (9) on all SNPs. We found four significant SNPs for DBP and five for SBP after adjusting for multiple comparisons. None of these SNPs were identified through the time invariant analyses. For some SNPs, *p*-values in the time invariant model suggested association (for example, the *p*-value for rs1052501 in a time invariant analysis was 0.002). However, they did not reach the significance level. Other SNPs would not have been identified from a time invariant analysis (for example, the *p*-value for rs10858911 in a time invariant analysis was 0.32). As an example, we show the age-specific effects and confidence intervals of two SNPs in Fig. 5. The SNP rs1052501 is in linkage disequilibrium with three other SNPs in the same region identified for DBP through the time varying analysis. Two SNPs identified for SBP, rs4757448 and rs17700056, are in linkage disequilibrium. The time varying analysis suggests that there may be genes that not only affect the long-term average SBP but also affect a change of SBP with time. We discuss implications of these findings and compare with the time invariant analysis in the next section.

Table 5. Top ranking SNPs in the time invariant and time varying analyses with FHS data†

Trait	*SNP*	*Analysis*	*Chromosome*	*Location*	*Gene*	*Minor allele frequency*	*LRT*‡	*p-value*‡	*LRT-lin*§	*p-value*§
SBP	rs11065951	Invariant§§	12	110479861	ATXN2	0.052	12.36	4.4×10^{-4}	—	—
DBP	rs1052501	Varying*	3	41900402	ULK4	0.192	16.50	1.0×10^{-4}	8.87	0.01
DBP	rs7648578	Varying	3	41833735	ULK4	0.187	16.81	9.8×10^{-5}	9.38	0.01
DBP	rs2128834	Varying	3	41837649	ULK4	0.187	16.30	1.2×10^{-4}	8.57	0.01
DBP	rs3774372	Varying	3	41852418	ULK4	0.183	16.29	1.2×10^{-4}	7.73	0.02
SBP	rs10858911	Varying	12	88487272	—	0.396	24.56	1.0×10^{-5}	2.87	0.24
SBP	rs4757448	Varying	11	16954369	PLEKHA7	0.340	40.12	1.0×10^{-6}	1.23	0.54
SBP	rs17700056	Varying	11	16975383	PLEKHA7	0.341	38.39	1.0×10^{-6}	1.39	0.50
SBP	rs7943587	Varying	11	16812381	PLEKHA7	0.373	32.41	5.0×10^{-6}	0.49	0.78
SBP	rs7121911	Varying	11	16977903	PLEKHA7	0.208	24.75	1.8×10^{-5}	24.75	4.2×10^{-6}

†Significance level 4.81×10^{-4} adjusting for multiple comparisons of 265 SNPs by Gao *et al.* (2008).
‡Likelihood ratio statistic LRT and *p*-value in a time varying analysis treating genetic effects as a non-parametric function and estimated by linear splines.
§Likelihood ratio statistic LRT-lin and *p*-value in a time varying analysis treating genetic effects as a parametric linear function.
§§Treating genetic effects as time invariant.
*Treating genetic effects as time varying.

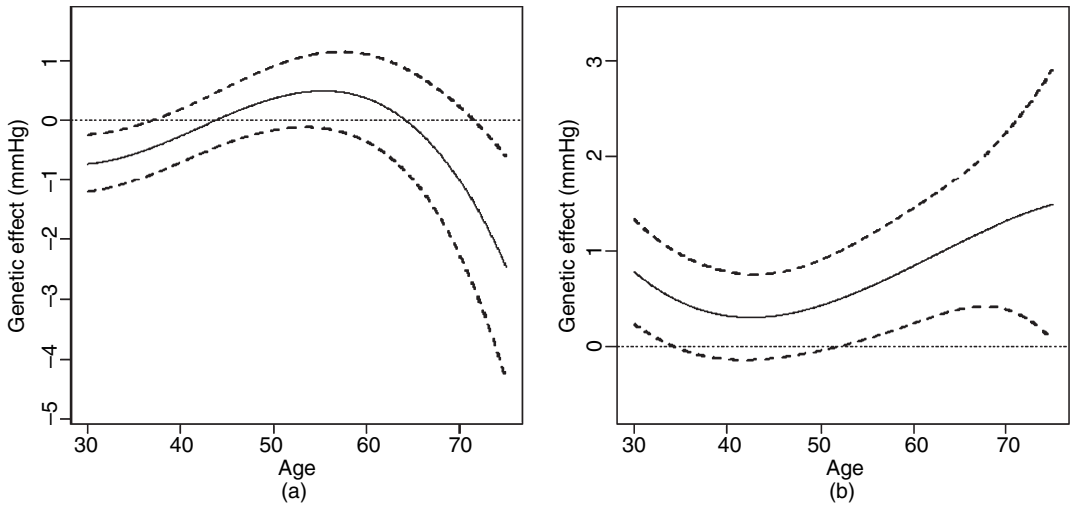

Fig. 5. Age-specific effects of two significant SNPs identified from the time varying analysis: (a) SBP, SNP rs10858911; (b) DBP, SNP rs1052501

6. Discussion

In this work, we propose semiparametric regression analysis of genetic studies with longitudinal phenotypes by penalized splines. A mixed effects model representation of penalized splines provides a convenient way to handle polygenic effects and shared environmental effects in genetic studies. Our simulations show that misspecifying the baseline function in a parametric analysis has a substantial effect on the type I error rate and power of testing genetic association regardless of whether the QTL effect changes with time. Furthermore, when the true genetic effect is a constant, the semiparametric analysis has power that is comparable with a non-linear analysis under a correctly specified model of the baseline function. It is therefore beneficial to model the baseline function non-parametrically. Misspecifying the genetic effect when the true effect varies with time in a parametric analysis can reduce the power significantly, especially when the average genetic effect over time is small. The semiparametric procedure proposed provides an alternative to existing time invariant analysis and parametric linear models for longitudinal genetic studies.

Although here the statistical procedures are developed for longitudinal data, they are also applicable to cross-sectional data when subjects' ages are recorded. In addition, for population-based case–control studies, the outcome is a binary variable. The penalized splines regression that is introduced here can be extended to generalized outcomes through a connection with generalized mixed effects models as discussed in Ruppert *et al.* (2003).

Population stratification is a potential confounder in genetic association studies. However, for the FHS all the study subjects are recruited from Framingham, Massachusetts, where the majority of the population is Caucasian and population stratification is found to be negligible (Wilk *et al.*, 2005). Nevertheless, one approach to adjust for population admixture is to estimate it by a principal components analysis and to include the first few principal components as covariates in the model (Price *et al.*, 2006), which can be readily incorporated in the framework of our proposed methods. The principal component weights are computed from founders in families and projected onto offspring to create principal component scores which are then included in a regression analysis. Another method is to incorporate the permutation procedure implemented

in the family-based association test (Rabinowitz and Laird, 2000) to our permutation test of the genetic effect. To be specific, one permutes offspring's genotypes given minimal sufficient statistics of the genetic model under the null hypothesis. A third strategy to adjust for population admixture in a regression-based analysis with family data is to include expected values of the genotype-related covariates given the minimal sufficient statistic for the genetic model under the null hypothesis as additional covariates (Yang *et al.*, 2000). This approach is the estimation analogy of the family-based association test. Wang *et al.* (2011) discussed an improvement to the procedure described in Yang *et al.* (2000) that computes the optimal covariate to minimize the estimation variance and to include these covariates in a regression analysis.

We implemented our methods with a truncated polynomial basis. Other bases such as *B*-splines can also be used. Models based on *B*-splines are equivalent to truncated polynomials through a reparameterization. The penalty matrix in expression (6) for *B*-splines, however, does not have the simple ridge penalty form and needs to be adapted. Eilers and Marx (1996) proposed a difference-based penalty. Wand and Ormerod (2008) considered a penalty matrix that is a direct generalization of smoothing splines (O'Sullivan penalized splines) and provided a mixed model representation. These works allow our methods to be extended to *B*-splines.

Although the methods proposed are illustrated through a candidate region analysis of the FHS data, the semiparametric analysis based on the mixed effect model can be implemented for analyses on a much larger scale, such as a genomewide association study. In our application of the FHS data with 6082 subjects and 14 505 observations, on average for each SNP the procedure proposed took 1.5 min to run on a Dell desktop computer with 2.00 GHz central processor unit and 3.25 Gbytes memory using R package `NLME` (Pinheiro and Bates, 2000). To complete a genomewide association study with 500 000 SNPs with parallel computing, this amounts to 2.6 days on a computing cluster with 200 nodes each with a 2.00 GHz central processor unit (about 10 days for a cluster of 50 nodes or about 500 days for a single node). In addition, in our experience, using SAS `PROC MIXED` (SAS Institute, 2004) improves computational efficiency.

Our analyses identified six SNPs for SBP and four SNPs for DBP residing in three genes. SNP rs11065951 locates within the gene ATXN2, which is a cytoplasmic protein. Lastres-Becker (2008) found that ATXN2 knock-out mice exhibited reduced fertility, locomotor hyperactivity and abdominal obesity and hepatosteatosis at the age of 6 months. ATXN2 was also reported to associate with neurological disorders (Huynh *et al.*, 1999), renal functions (Kottgen *et al.*, 2010) and obesity (Figueroa *et al.*, 2009) which may share some pathways with blood pressure. Four SNPs (rs4757448, rs17700056, rs7943587 and rs7121911) located in a protein coding gene, PLEKHA7, were reported to be linked to blood pressure at a genomewide level in another joint meta-analysis of genomewide association studies for blood pressure (Cohorts for Heart and Aging Research in Genetic Epidemiology consortium and Global BPgen; Newton-Cheh *et al.* (2009)). Four linked SNPs (rs1052501, rs7648578, rs2128834 and rs3774372) were located in the gene ULK4, which is an Unc-51-like kinase. This gene was also identified in the Cohorts for Heart and Aging Research in Genetic Epidemiology consortium study (Newton-Cheh *et al.*, 2009) as a candidate locus for blood pressure. However, little has been reported on the relationship between the function of this gene and blood pressure. Gene expression analysis has confirmed that SNPs in ULK4 alter gene expression levels in liver and lymphoblastoid cell lines (Levy *et al.*, 2009). Our analysis showed a potential time varying effect at this locus. This may deserve further functional research.

Aging is a complex process during which many biological and physiological changes take place which in turn may change a range of phenotypes, including blood pressure, and may change the interplay between environmental and genetic factors. Therefore, age may represent a surrogate of constellations of unmeasured factors. Taking into account the gene–age interaction

in a genetic association study may help to overcome some of the inconsistencies in replicating a genetic finding and may boost power (Lasky-Su *et al.*, 2008). Our time invariant analysis identifies two SNPs for SBP and the time varying analysis identifies a distinct SNP for SBP and four SNPs for DBP. None of the SNPs were identified by both analyses. Some of the SNPs may be missed if only the time invariant analysis was carried out. These results illustrate the complementary feature of the two analyses. When the true genetic effect does not vary with time, a time-invariant model may identify more SNPs owing to parsimony of the model. However, when the genetic effect does change with time or when age acts as a surrogate of unmeasured factors causing varying genetic effects, failure to acknowledge the time trend may reduce the power or lead to irreproducible results (Shi and Rao, 2008; Lasky-Su *et al.*, 2008).

Despite large efforts on gene mapping through genomewide association studies, until recently few genetic variants were known to be reproducibly associated with common disease. Part of the inconsistency may be explained by the dominant time invariant analyses strategy (Lasky-Su *et al.*, 2008). The general semiparametric approaches that we develop here may be applied to model age-dependent genetic effects, leading to more powerful genetic data analysis and potentially more consistent results. Our results also suggest a new hypothesis of possible time varying genetic effect on blood pressure at several loci. This needs to be confirmed by a future larger study. In addition, estimating age-specific heritability and genetic effects has implications for designing subsequent studies and developing treatment of a disease: sampling subjects at the age where heritability is at its peak would enhance the power of an association study, which is very important for detecting genes with moderate effects; effectively designing a future genomewide association study requires accurate estimation of the potential time varying effect size of a gene; and interventions may target different environmental or genetic factors at different ages, depending on which factor is dominant.

Acknowledgements

The authors thank the Associate Editor and the reviewers for their constructive and helpful comments on this paper and Ms Amanda Applegate for editorial assistance. The Framingham data were obtained from the FHS of the National Heart, Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine (contract N01-HC-25195). Yuanjia Wang's research is supported by National Institutes of Health grants AG031113-01A2 and NS073670-01. Runze Li's research was supported by National Institutes of Health grants R21 DA024260 and P50 DA-10075, National Science Foundation grant DMS 0348869 and National Natural Science Foundation of China grant 11028103. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute on Drug Abuse or the National Institutes of Health.

References

Atwood, L., Heard-Costa, N., Cupples, L., Jaquish, C., Wilson, P. and D'Agostine, R. (2002) Genome-wide linkage analysis of body mass index across 28 years of the Framingham Heart Study. *Am. J. Hum. Genet.*, **71**, 1044–1050.

Crainiceanu, C. and Ruppert, D. (2004) Restricted likelihood ratio tests in nonparametric longitudinal models. *Statist. Sin.*, **14**, 713–729.

Daw, E. W., 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, article S3.

Dawber, T. R., Meadors, G. F. and Moore, F. E. J. (1951) Epidemiological approaches to heart disease: the Framingham Study. *Am. J. Publ. Hlth*, **41**, 279–286.

Eilers, P. and Marx, B. (1996) Flexible smoothing with B-splines. *Statist. Sci.*, **11**, 89–121.

Falconer, D. S. (1985) *Introduction to Quantitative Genetics*, 2nd edn. New York: Longman.

Fan, J., Huang, T. and Li, R. (2007) Analysis of longitudinal data with semiparametric estimation of covariance function. *J. Am. Statist. Ass.*, **102**, 632–641.

Fang, Y. and Wang, Y. (2009) Testing for genetic effect on functional traits by functional principal components analysis based on heritability. *Statist. Med.*, **28**, 3611–3625.

Figueroa, K. P., Farooqi, S., Harrup, K., Frank, J., O'Rahilly, S. and Pulst, S. M. (2009) Genetic variance in the spinocerebellar ataxia type 2 (ATXN2) gene in children with severe early onset obesity. *PLOS One*, **4**, no. 12, article e8280.

Gao, X., Starmer, J. and Martin, E. (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genet. Epidem.*, **32**, 361–369.

Gudbjartsson, D. F., Bjornsdottir, U. S., Halapi, E., Helgadottir, A., Sulem, P., Jonsdottir, G. M., Thorleifsson, G., Helgadottir, H., Steinthorsdottir, V., Stefansson, H., Williams, C., Hui, J., Beilby, J., Warrington, N. M., James, A., Palmer, L. J., Koppelman, G. H., Heinzmann, A., Krueger, M., Boezen, H. M., Wheatley, A., Altmuller, J., Shin, H. D., Uh, S. T., Cheong, H. S., Jonsdottir, B., Gislason, D., Park, C. S., Rasmussen, L. M., Porsbjerg, C., Hansen, J. W., Backer, V., Werge, T., Janson, C., Jönsson, U. B., Ng, M. C., Chan, J., So, W. Y., Ma, R., Shah, S. H., Granger, C. B., Quyyumi, A. A., Levey, A. I., Vaccarino, V., Reilly, M. P., Rader, D. J., Williams, M. J., van Rij, A. M., Jones, G.T., Trabetti, E., Malerba, G., Pignatti, P. F., Boner, A., Pescollderungg, L., Girelli, D., Olivieri, O., Martinelli, N., Ludviksson, B. R., Ludviksdottir, D., Eyjolfsson, G. I., Arnar, D., Thorgeirsson, G., Deichmann, K., Thompson, P. J., Wjst, M., Hall, I. P., Postma, D. S., Gislason, T., Gulcher, J., Kong, A., Jonsdottir, I., Thorsteinsdottir, U. and Stenfansson, K. (2009) Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.*, **41**, 342–347.

Guo, G. and Wang, J. (2002) The mixed model or multilevel model for behavior genetics analysis. *Behav. Genet.*, **32**, 37–49.

He, Q., Berg, A., Li, Y., Vallejos, C. E. and Wu, R. (2010) Mapping genes for plant structure, development and evolution: functional mapping meets ontology. *Trends Genet.*, **26**, 39–46.

Huynh, D. P., Del Bigio, M. R., Ho, D. H. and Pulst, S. M. (1999) Expression of ataxin-2 in brains from normal individuals and patients with Alzheimer's disease and spinocerebellar ataxia 2. *Ann. Neur.*, **45**, 232–241.

Jarvik, G. P., Goode, E. L., Austin, M. A., Auwerx, J., Deeb, S., Schellenberg, G. D. and Reed, T. (1997) Evidence that the apolipoprotein E-genotype effects on lipid levels can change with age in males: a longitudinal analysis. *Am. J. Hum. Genet.*, **61**, 171–181.

Khoury, M., Beaty, H. and Cohen, B. (1993) *Fundamentals of Genetic Epidemiology*. New York: Oxford University Press.

Köttgen, A., Pattaro, C., Böger, C. A., Fuchsberger, C., Olden, M., Glazer, N. L., Parsa, A., Gao, X., Yang, Q., Smith, A. V., O'Connell, J. R., Li, M., Schmidt, H., Tanaka, T., Isaacs, A., Ketkar, S., Hwang, S. J., Johnson, A. D., Dehghan, A., Teumer, A., Paré, G., Atkinson, E. J., Zeller, T., Lohman, K., Cornelis, M. C., Probst-Hensch, N. M., Kronenberg, F., Tönjes, A., Hayward, C., Aspelund, T., Eiriksdottir, G., Launer, L. J., Harris, T. B., Rampersaud, E., Mitchell, B. D., Arking, D. E., Boerwinkle, E., Struchalin, M., Cavalieri, M., Singleton, A., Giallauria, F., Metter, J., de Boer, I. H., Haritunians, T., Lumley, T., Siscovick, D., Psaty, B. M., Zillikens, M. C., Oostra, B. A., Feitosa, M., Province, M., de Andrade, M., Turner, S. T., Schillert, A., Ziegler, A., Wild, P. S., Schnabel, R. B., Wilde, S., Munzel, T. F., Leak, T. S., Illig, T., Klopp, N., Meisinger, C., Wichmann, H. E., Koenig, W., Zgaga, L., Zemunik, T., Kolcic, I., Minelli, C., Hu, F. B., Johansonn, A., Igl, W., Zaboli, G., Wild, S. H., Wright, A. F., Campbell, H., Ellinghaus, D., Schreiber, S., Aulchenko, Y. S., Felix, J. F., Rivadeneira, F., Uitterlinden, A. G., Hofman, A., Imboden, M., Nitsch, D., Brandstätter, A., Kollerits, B., Kedenko, L., Mägi, R., Stumvoll, M., Kovacs, P., Boban, M., Campbell, S., Endlich, K., Völzke, H., Kroemer, H. K., Nauck, M., Völker, U., Polesek, O., Vitart, V., Badola, S., Parker, A. N., Ridker, P. M., Kardia, S. L., Blankenberg, S., Liu, Y., Curhan, G. C., Franke, A., Rochat, T., Paulweber, B., Prokopenko, I., Wang, W., Gudnason, V., Shuldiner, A. R., Coresh, J., Schmidt, R., Ferrucci, L., Shlipak, M. G., van Duijn, C. M., Borecki, I., Krämer, B. K., Rudan, I., Gyllensten, U., Wilson, J. F., Witteman, J. C., Pramstaller, P. P., Rettig, R., Hastie, N., Chasman, D. I., Kao, W. H., Heid, I. M. and Fox, C. S. (2010) New loci associated with kidney function and chronic kidney disease. *Nat. Genet.*, **42**, 376–384.

Krivobokova, T. and Kauermann, G. (2007) A note on penalized splines with correlated errors. *J. Am. Statist. Ass.*, **102**, 1328–1337.

Lasky-Su, J., Lyon, H. N., Emilsson, V., Heid, I. M., Molony, C., Raby, B. A., Lazarus, R., Klanderman, B., Soto-Quiros, M. E., Avila, L., Silverman, E. K., Thorleifsson, G., Thorsteinsdottir, U., Kronenberg, F., Vollmert, C., Illig, T., Fox, C. S., Levy, D., Laird, N., Ding, X., McQueen, M. B., Butler, J., Ardlie, K., Papoutsakis, C., Dedoussis, G., O'Donnell, C. J., Wichmann, H. E., Celedón, J. C., Schadt, E., Hirschhorn, J., Weiss, S. T., Stefansson, K. and Lange, C. (2008) On the replication of genetic associations: timing can be everything! *Am. J. Hum. Genet.*, **82**, 849–858.

Lastres-Becker, I., Brodesser, S., Lütjohann, D., Azizov, M., Buchmann, J., Hintermann, E., Sandhoff, K., Schürmann, A., Nowock, J. and Auburger, G. (2008) Insulin receptor and lipid metabolism pathology in ataxin-2 knock-out mice. *Hum. Molec. Genet.*, **17**, 1465–1481.

Levy, D., DeStefano, A. L., Larson, M. G., O'Donnell, C. J., Lifton, R. P., Gavras, H., Cupples, L. A. and Myers, R. H. (2000) Evidence for a gene influencing blood pressure on chromosome 17: genome scan linkage results

for longitudinal blood pressure phenotypes in subjects from the Framingham Heart Study. *Hypertension*, **36**, 477–483.

Levy, D., Ehret, G. B., Rice, K., Verwoert, G. C., Launer, L. J., Dehghan, A., Glazer, N. L., Morrison, A. C., Johnson, A. D., Aspelund, T., Aulchenko, Y., Lumley, T., Köttgen, A., Vasan, R. S., Rivadeneira, F., Eiriksdottir, G., Guo, X., Arking, D. E., Mitchell, G. F., Mattace-Raso, F. U., Smith, A. V., Taylor, K., Scharpf, R. B., Hwang, S. J., Sijbrands, E. J., Bis, J., Harris, T. B., Ganesh, S. K., O'Donnell, C. J., Hofman, A., Rotter, J. I., Coresh, J., Benjamin, E. J., Uitterlinden, A. G., Heiss, G., Fox, C. S., Witteman, J. C., Boerwinkle, E., Wang, T. J., Gudnason, V., Larson, M. G., Chakravarti, A., Psaty, B. M. and van Duijn, C. M. (2009) Genome-wide association study of blood pressure and hypertension. *Nat. Genet.*, **41**, 677–687.

Li, Y. and Ruppert, D. (2008) On the asymptotics of penalized splines. *Biometrika*, **95**, 415–436.

Lynch, M. and Walsh, B. (1998) *Genetics and Analysis of Quantitative Traits*. Sunderland: Sinauer.

Neale, M. C., Boker, S. M., Xie, G. and Maes, H. H. (2004) *Mx: Statistical Modeling*, 6th edn. Richmond: Virginia Commonwealth University. (Available from `http://www.vipbg.vcu.edu/mxgui/`.)

Newton-Cheh, C., Johnson, T., Gateva, V., Tobin, M. D., Bochud, M., Coin, L., Najjar, S. S., Zhao, J. H., Heath, S. C., Eyheramendy, S., Papadakis, K., Voight, B. F., Scott, L. J., Zhang, F., Farrall, M., Tanaka, T., Wallace, C., Chambers, J. C., Khaw, K. T., Nilsson, P., van der Harst, P., Polidoro, S., Grobbee, D. E., Onland-Moret, N. C., Bots, M. L., Wain, L.V., Elliott, K. S., Teumer, A., Luan, J., Lucas, G., Kuusisto, J., Burton, P. R., Hadley, D., McArdle, W. L., Wellcome Trust Case Control Consortium, Brown, M., Dominiczak, A., Newhouse, S. J., Samani, N. J., Webster, J., Zeggini, E., Beckmann, J. S., Bergmann, S., Lim, N., Song, K., Vollenweider, P., Waeber, G., Waterworth, D. M., Yuan, X., Groop, L., Orho-Melander, M., Allione, A., Di Gregorio, A., Guarrera, S., Panico, S., Ricceri, F., Romanazzi, V., Sacerdote, C., Vineis, P., Barroso, I., Sandhu, M. S., Luben, R. N., Crawford, G. J., Jousilahti, P., Perola, M., Boehnke, M., Bonnycastle, L. L., Collins, F. S., Jackson, A. U., Mohlke, K. L., Stringham, H. M., Valle, T. T., Willer, C. J., Bergman, R. N., Morken, M. A., Döring, A., Gieger, C., Illig, T., Meitinger, T., Org, E., Pfeufer, A., Wichmann, H. E., Kathiresan, S., Marrugat, J., O'Donnell, C. J., Schwartz, S. M., Siscovick, D. S., Subirana, I., Freimer, N. B., Hartikainen, A. L., McCarthy, M. I., O'Reilly, P. F., Peltonen, L., Pouta, A., de Jong, P. E., Snieder, H., van Glist, W. H., Clarke, R., Goel, A., Hamsten, A., Peden, J. F., Seedorf, U., Syvänen, A. C., Tognoni, G., Lakatta, E. G., Sanna, S., Scheet, P., Schlessinger, D., Scuteri, A., Dörr, M., Ernst, F., Felix, S. B., Homuth, G., Lorbeer, R., Reffelmann, T., Rettig, R., Völker, U., Galan, P., Gut, I. G., Hercberg, S., Lathrop, G. M., Zelenika, D., Deloukas, P., Soranzo, N., Williams, F. M., Zhai, G., Salomaa, V., Laakso, M., Elosua, R., Forouhi, N. G., Völzke, H., Uiterwaal, C. S., van der Schouw, Y. T., Numans, M. E., Matullo, G., Navis, G., Berglund, G., Bingham, S. A., Kooner, J. S., Connell, J. M., Bandinelli, S., Ferrucci, L., Watkins, H., Spector, T. D., Tuomilehto, J., Altshuler, D., Strachan, D. P., Laan, M., Meneton, P., Wareham, N. J., Uda, M., Jarvelin, M. R., Mooser, V., Melander, O., Loos, R. J., Elliott, P., Abecasis, G. R., Caulfield, M. and Munroe, P. B. (2009) Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.*, **41**, 666–676.

Pinheiro, J. C. and Bates, D. M. (2000) *Mixed-effects Models in S and S-PLUS*. New York: Springer.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Province, M. A. and Rao, D. C. (1985) Path analysis of family resemblance with temporal trends: applications to height, weight, and Quetelet index in Northeastern Brazil. *Am. J. Hum. Genet.*, **37**, 178–192.

Rabe-Hesketh, S., Skrondal, A. and Gjessing, H. K. (2008) Biometrical modelling of twin and family data using standard mixed model software. *Biometrics*, **64**, 280–288.

Rabinowitz, D. and Laird, N. (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum. Hered.*, **50**, 211–223.

Raff, R. A. (2000) Evo-devo: the evolution of a new discipline. *Nat. Rev. Genet.*, **1**, 74–79.

Rice, S. H. (2002) A general population genetic theory for the evolution of developmental interactions. *Proc. Natn. Acad. Sci. USA*, **99**, 15518–15523

Ruppert, D. (2002) Selecting the number of knots for penalized splines. *J. Computnl Graph. Statist.*, **11**, 735–757.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. New York: Cambridge University Press.

SAS Institute (2004) *SAS 9.1.3 Help and Documentation*. Cary: SAS Institute.

Shi, G. and Rao, D. C. (2008) Ignoring temporal trends in genetic effects substantially reduces power of quantitative trait linkage analysis. *Genet. Epidem.*, **32**, 61–72.

Strauch, J., Golla, A., Wilcox, M. A. and Baur, M. P. (2003) Genetic analysis of phenotypes derived from longitudinal data: Presentation Group 1 of Genetic Analysis Workshop 13. *Genet. Epidem.*, **25**, suppl. 1, S5–S17.

Tobin, M. D., Sheehan, N. A., Scurrah, K. J. and Burton, P. R. (2005) Adjusting for treatment effects in studies: antihypertensive therapy and systolic blood pressure. *Statist. Med.*, **24**, 2911–2935.

Wand, M. P. (2003) Smoothing and mixed models. *Computnl Statist.*, **18**, 223–249.

Wand, M. P. and Ormerod, J. T. (2008) On O'Sullivan penalised splines and semiparametric regression. *Aust. New Zeal. J. Statist.*, **50**, 179–198.

Wang, Y., Yang, Q. and Rabinowitz, D. (2011) Unbiased and efficient estimation of the effect of candidate genes on quantitative traits in the presence of population admixture. *Biometrics*, **67**, 331–343.

Wilk, J. B., Manning, A. K., Dupuis, J., Cupples, L. A., Larson, M. G., Newton-Cheh, C., Demissie, S., DeStefano, A. L., Hwang, S. J., Liu, C., Yang, Q. and Lunetta, K. L. (2005) No evidence of major population substructure in the Framingham Heart Study. *Genet. Epidem.*, **29**, article 286.

Wu, H. and Zhang, J. (2006) *Nonparametric Regression Methods for Longitudinal Data Analysis Mixed-effects Modeling Approaches*. New York: Wiley.

Yang, Q., Rabinowitz, D., Isasi, C. and Shea, S. (2000) Adjusting for confounding due to population admixture when estimating the effect of candidate genes on quantitative traits. *Hum. Hered.*, **50**, 227–233.

Yang, Q., Wu, H., Guo, C. and Fox, C. (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet. Epidem.*, **34**, 444–454.

Zhang, H. and Zhong, X. (2006) Linkage analysis of longitudinal data and design consideration. *BMC Genet.*, **7**, article 37.

Zhao, W. and Wu, R. (2008) Wavelet-based nonparametric functional mapping of longitudinal curves. *J. Am. Statist. Ass.*, **103**, 714–725.

Zhao, W., Zhu, J., Gallo-Meagher, M. and Wu, R. (2004) A unified statistical model for functional mapping of environment-dependent genetic expression and genotype × environment interactions for ontogenetic development. *Genetics*, **168**, 1751–1762.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Appendix for "Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing"'.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the author for correspondence for the article.