HHS Public Access
Author manuscript
Genet Epidemiol. Author manuscript; available in PMC 2015 May 26.

Published in final edited form as:
Genet Epidemiol. 2012 December ; 36(8): 856–869. doi:10.1002/gepi.21673.

Longitudinal Association Analysis of Quantitative Traits

Ruzong Fan[1,*], **Yiwei Zhang**[2], **Paul S. Albert**[1], **Aiyi Liu**[1], **Yuanjia Wang**[3], and **Momiao Xiong**[4]

[1]Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health, Rockville, Maryland [2]Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota [3]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, New York [4]Human Genetics Center, University of Texas – Houston, Houston, Texas

Abstract

Longitudinal genetic studies provide a valuable resource for exploring key genetic and environmental factors that affect complex traits over time. Genetic analysis of longitudinal data that incorporate temporal variations is important for understanding genetic architecture and biological variations of common complex diseases. Although they are important, there is a paucity of statistical methods to analyze longitudinal human genetic data. In this article, longitudinal methods are developed for temporal association mapping to analyze population longitudinal data. Both parametric and nonparametric models are proposed. The models can be applied to multiple diallelic genetic markers such as single-nucleotide polymorphisms and multiallelic markers such as microsatellites. By analytical formulae, we show that the models take both the linkage disequilibrium and temporal trends into account simultaneously. Variance-covariance structure is constructed to model the single measurement variation and multiple measurement correlations of an individual based on the theory of stochastic processes. Novel penalized spline models are used to estimate the time-dependent mean functions and regression coefficients. The methods were applied to analyze Framingham Heart Study data of Genetic Analysis Workshop (GAW) 13 and GAW 16. The temporal trends and genetic effects of the systolic blood pressure are successfully detected by the proposed approaches. Simulation studies were performed to find out that the nonparametric penalized linear model is the best choice in fitting real data. The research sheds light on the important area of longitudinal genetic analysis, and it provides a basis for future methodological investigations and practical applications.

© 2012 Wiley Periodicals, Inc.

[*]Correspondence to: Ruzong Fan, Biostatistics and Bioinformatics Branch, Division of Epidemiology, Statistics and Prevention Research, 6100 Executive Boulevard, Room 7B05, MSC 7510, Bethesda, MD 20852. fanr@mail.nih.gov.

Supporting Information is available in the online issue at wileyonlinelibrary.com.

COMPUTER PROGRAM

The methods proposed in this article can be implemented by using procedure of linear mixed effect models in the statistical package R, i.e., lme function. The R codes for data analysis and simulations are available from the corresponding author, Dr. Fan, upon request.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

association mapping; quantitative trait loci; longitudinal analysis

INTRODUCTION

In the classical theory of statistical genetics, the research is limited to analyze data in which phenotypes and related covariate measurements are observed only one time. For longitudinal human studies, multiple measurements of some quantitative or qualitative traits are taken for each individual over time, in addition to the genotype information and covariates such as gender, age, and familial income. For example, multiple measurements such as blood pressure, age, and body mass index (BMI) were recorded for individuals in Framingham Heart Study (FHS) data, which were available via Genetic Analysis Workshop (GAW) 13 and GAW 16 [Cupples et al., 2003; MacCluer et al., 2009]. Hence, multiple measurements are available for a subject, which depend on the subject's age or time. However, there is very little research on statistical methods to analyze longitudinal genetic data. To our knowledge, there is no method for temporal linkage disequilibrium (LD) mapping or association analysis of longitudinal traits of population data and there is no handy software for the data analysis. Some studies propose to collapse the measurements to be a single value and then to run an analysis based on the classical theory of statistical genetics [Mukherjee et al., 2012]. Obviously, this method ignores the temporal nature of longitudinal data and it is not able to catch the temporal trend of the traits.

Due to the lack of statistical models and methods in analyzing longitudinal genetic data, investigators usually take a simple approach of averaging multiple response measurements of the same individual to analyze the longitudinal genetic traits. For instance, the sample averages of the blood pressure, age, and BMI were used in genome scan linkage studies for FHS data by Levy et al. [2000, 2009]. Although this approach makes it possible to apply the existing statistical methods and software to handle longitudinal traits, it is, essentially, not a longitudinal analysis of repeated genetic traits. The phenotype traits are usually varying with time, and so there are temporal variations [Fan et al., 2012; Mountz et al., 2001; Soler and Blangero, 2003]. After collapsing the multiple measurements to be a single value, no temporal variations can be detected in an analysis. This type of analysis may not always be able to get ideal results and to draw the best information from the data [Shi and Rao, 2008].

For certain traits of complex diseases, such as BMI and blood pressure, genetic determinants are important at some period of time of human development. At other time periods, environmental factors are more important, such as diet and family income. It is important to develop statistical models and methods which may better use the longitudinal data, and may reflect temporal trends [Lasky-Su et al., 2008]. For the phenotype traits which have an important genetic component, the power to localize the genetic location and to detect important genetic determinants of the traits can be high at the specific stage that genetic contribution is high. On the other hand, the power to localize the genetic location and to detect important genetic determinants can be low at the other stage that environmental factors are more important than genetic contribution. A better understanding of the temporal

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

variations of the traits and the temporal genetic contribution to phenotype traits may provide more insights in mapping the temporal genetic traits and in determining the important genetic determinants. In addition, longitudinal genetics analysis may help with the detection of gene-environment interactions.

In GAW 13 and GAW 16, a wide range of methods and models was developed to analyze the FHS data and the similarly structured simulated data [Almasy et al., 2003a,b]. The research mainly focused on linkage analysis using family data in GAW 13. Variance component approach was extended to incorporate temporal trends for linkage analysis of longitudinal traits to detect quantitative trait loci (QTL) in de Andrade et al. [2002] and de Andrade and Olswold [2003]. The methods suffer from a large number of parameters when the number of measurements increases since each measurement corresponds to a set of variance-covariance parameters. In addition, parametric variance component models were developed for linkage analysis of longitudinal genetic data, which may not reflect the temporal trends well in Zhang and Zhong [2006]. Overall, the existing longitudinal statistical models are problematic in one or more aspects.

In this article, temporal association analysis methods are developed for the population longitudinal data. Both parametric and nonparametric models are proposed. By utilizing multiple genetic markers which can be either diallelic or multiallelic, we develop temporal association mapping models based on the motivation of population genetics model. By analytical formulae, we show that the models take both the LD and temporal trends into account. To reduce the number of parameters, variance-covariance structure is constructed to model the single measurement variation and multiple measurement correlations of an individual based on the theory of stochastic processes. This is a key different part between our methods and those proposed by de Andrade et al. [2002] and de Andrade and Olswold [2003] since our models contain very small number of parameters to facilitate data analysis for robust results. Novel penalized spline models are used to estimate the time-dependent mean functions and regression coefficients nonparametrically which makes our methods different from the parametric models of Zhang and Zhong [2006]. To describe the usefulness of the proposed approaches, the methods were applied to analyze FHS data of GAW 13 and GAW 16 to detect the temporal trends of systolic blood pressure (SBP). Simulation was performed to evaluate robustness, power, and parameter estimation accuracy of the proposed nonparametric and parametric models.

MATERIALS AND METHODS

Before discussing methods and models, we present in Figure 1 a time plot of SBP and total plasma cholesterol level against age in years. The plot is based on a sample from the GAW 13 data, Cohort 2 of Problem 1 of the FHS. The cohort includes 330 pedigrees. The sample consists of 330 individuals, one from a pedigree. From the time plot, it can be seen that both SBP and total plasma cholesterol increase as age increases. Thus, one needs to take this temporal trend into account in modeling the longitudinal genetic traits. In addition, there are large trait variations. For SBP, the variation seems homogeneous, while the variation of total plasma cholesterol level increases as age gets old.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In the following, we are going to present a general temporal population quantitative genetics model. Then, we propose population temporal association models using typed genetic markers. The variance-covariance structure is constructed to describe the trait variation and to properly account for correlation between multiple measurements on the same subject. Penalized spline methods are used to approximate temporal mean function and regression coefficients.

A TEMPORAL POPULATION GENETICS MODEL

Consider a quantitative trait locus Q which has two alleles Q_1 and Q_2 with allele frequencies q_1 and q_2, respectively. To simplify the presentation, we first assume that QTL Q is the only major locus responsible for the trait value. In addition, neither covariates nor environmental nor polygenic effects are considered. For the trait value, let $\mu_{ij}(t)$ be the effect of genotype $Q_i Q_j$, $i, j = 1, 2$, $\mu_{12}(t) = \mu_{21}(t)$, at the time t. Here the time t is usually age of an individual. Let the genic effect of allele Q_i be $\alpha_i(t)$, $i = 1, 2$. The genotypic effects at the time t can be expressed as

$$\begin{aligned}\mu_{11}(t)&=\mu(t)+2\alpha_1(t)+d_1(t),\\ \mu_{12}(t)&=\mu(t)+\alpha_1(t)+\alpha_2(t)+d_2(t),\\ \mu_{22}(t)&=\mu(t)+2\alpha_2(t)+d_3(t),\end{aligned}$$

where $\mu(t)$ is the overall population mean, and $d_i(t)$ is the deviation of the related genotypic value from that of an additive effect model. Let us consider a fixed time t_0. Minimizing $F(\mu,\alpha_1,\alpha_2)=\sum_{i=1}^{2}\sum_{j=1}^{2}q_iq_j(\mu_{ij}(t_0)-\mu(t_0)-\alpha_i(t_0)-\alpha_j(t_0))^2$, classical theory of quantitative genetics provides the estimates of $\mu(t_0)$, $\alpha_1(t_0)$, $\alpha_2(t_0)$ as in other studies [Jacquard, 1974; Lange, 2002]

$$\begin{aligned}\hat{\mu}(t_0)&=\mu_{11}(t_0)q_1^2+2\mu_{12}(t_0)q_1q_2+\mu_{22}(t_0)q_2^2=\mu_0(t_0),\\ \hat{\alpha}_1(t_0)&=q_1\mu_{11}(t_0)+q_2\mu_{12}(t_0)-\mu(t_0),\\ \hat{\alpha}_2(t_0)&=q_1\mu_{21}(t_0)+q_2\mu_{22}(t_0)-\mu(t_0).\end{aligned}$$

Plugging these estimates into $\mu_{ij}(t_0)$, we can obtain the following expressions

$$\begin{aligned}\mu_{11}(t_0)&=\mu_0(t_0)+2q_2\alpha_Q(t_0)-q_2^2\delta_Q(t_0),\\ \mu_{12}(t_0)&=\mu_0(t_0)+(q_2-q_1)\alpha_Q(t_0)+q_1q_2\delta_Q(t_0),\\ \mu_{22}(t_0)&=\mu_0(t_0)-2q_1\alpha_Q(t_0)-q_1^2\delta_Q(t_0).\end{aligned}\quad (1)$$

Here, $\alpha_Q(t_0) = q_1\mu_{11}(t_0) + (q_2 - q_1)\mu_{12}(t_0) - q_2\mu_{22}(t_0)$ is the average effect of gene substitution, and $\delta_Q(t_0)=2\mu_{12}(t_0) - \mu_{11}(t_0) - \mu_{22}(t_0)$ is the dominance deviation. For an individual of a population, let the trait value be $y(t)$ at the time t. Let G be the genotype of the individual at the QTL Q. Under an assumption of normality, the equation (1) imply that the trait can be expressed as

$$y(t_0)=\mu_0(t_0)+x_Q\alpha_Q(t_0)+z_Q\delta_Q(t_0)+\varepsilon,\quad (2)$$

where ε is a random error term, and x_Q and z_Q are dummy variables defined by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$x_Q=\begin{cases} 2q_2 \\ \text{if } G=Q_1Q_1 \\ q_2-q_1 \\ \text{if } G=Q_1Q_2 \\ -2q_1 \\ \text{if } G=Q_2Q_2 \end{cases}, \quad z_Q=\begin{cases} -q_2^2 \\ \text{if } G=Q_1Q_1 \\ q_2q_1 \\ \text{if } G=Q_1Q_2 \\ -q_1^2 \\ \text{if } G=Q_2Q_2. \end{cases} \quad (3)$$

Assume that the trait locus Q is known, and the trait alleles Q_1 and Q_2 are correctly typed. Then the trait can be fully described by expression (2). In practice, information about trait locus Q is unknown, but the information of marker loci is available. This motivates us to develop appropriate models based on marker information to map QTL. In the following, we introduce population-based longitudinal LD mapping models.

LONGITUDINAL ASSOCIATION MAPPING MODELS USING DIALLELIC MARKERS

In our previous research, population-based regression association models of QTL are constructed in Fan and Xiong [2002] by using multiple diallelic markers in the analysis. The models are further extended to be variance component models for combined linkage and association mapping of QTL [Fan and Xiong, 2003; Jung et al., 2005]. The additive temporal models described as follows can be thought as longitudinal extension of the association mapping models in Fan and Xiong [2002]. In Supplementary Materials, we present models which model both additive and dominant effect.

Assume that I diallelic markers M_j, $j = 1, 2,\ldots, I$ are typed in a region of the trait locus Q. For marker M_j, the two alleles are denoted by M_j and m_j with frequencies P_{M_j} and P_{m_j}, respectively (note here the notation M_j can be either marker or allele, whichever applies). Suppose that markers M_j are in Hardy-Weinberg equilibrium (HWE). However, they may be in LD. Denote the measure of LD between trait locus Q and marker M_j by $D_{M_jQ}=P(M_jQ_1)-q_1P_{M_j}$, and the measure of LD between marker M_j and marker M_k by $D_{M_jM_k}=P(M_jM_k)-P_{M_j}P_{M_k}, j, k=1, 2, \ldots, I$ [Hartl and Clark, 1989; Hedrick, 1987; Lewontin, 1988]. Here, $P(M_j\, Q_1)$ and $P(M_j\, M_k)$ are frequencies of haplotypes $M_j\, Q_1$ and $M_j\, M_k$, respectively.

Consider a population sample with N individuals. For the ith individual, let y_i be his/her quantitative trait value and let G_{ij} be his/her genotype at the marker M_j. An additive temporal LD regression mixed model extending (2) at the time t can be defined as

$$y_i(t)=\mu(t)+w_i(t)\beta(t)+\sum_{j=1}^{I}x_{ij}\alpha_j(t)+U_i(t)+E_i+\varepsilon_i. \quad (4)$$

The components of the above model are specified as follows. First, $\mu(t)$ is a nonrandom overall mean at time t and $\mu(t)$ is unspecified; $w_i(t)$ is a row vector of covariates such as gender, BMI at the time t, and possibly their interaction terms; $\beta(t)$ is a nonrandom column vector of regression parameters of the covariates $w_i(t)$ with fixed effects. One may want to

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

notice that the covariates can be time invariant like gender or can be time varying such as the BMI. In addition, x_{ij} are dummy variables defined by

$$x_{ij}=\left\{\begin{matrix} 2 \\ \text{if } G_{ij}=M_jM_j \\ 1 \\ \text{if } G_{ij}=M_jm_j \\ 0 \\ \text{if } G_{ij}=m_jm_j \end{matrix}\right. ,$$

and $\alpha_j(t)$ are regression coefficients of the dummy variables x_{ij} at the time t. In model (4), $U_i(t)$ is the correlation effect among repeated measurements due to both genetic and environmental factors of an individual, E_i is a random variation of subject i, and ε_i is a random measurement error term. Assume that $U_i(t)$, E_i, and ε_i are independent. Moreover, assume that E_i is normal $N(0, \sigma_E^2)$ and ε_i is normal $N(0, \sigma_e^2)$

A similar character process model was developed by Pletcher and Geyer [1999] and Jafferzic and Pletcher [2000], which does not measure effects from specific genes and uses no marker information. The novel part of model (4) is that we include measured genotype components estimating association with genotyped markers, i.e., the terms involves x_{ij} [Fan and Jung, 2003; Fan et al., 2005; Fan and Xiong, 2002, 2003; Jung et al., 2005]. Let the variance-covariance matrices of the indicator variables x_{ij} be

$$V_A=2\left\{\begin{matrix} P_{M_1}P_{m_1} \\ D_{M_1M_2} \\ \cdots \\ D_{M_1M_I} \\ D_{M_1M_2} \\ P_{M_2}P_{m_2} \\ \cdots \\ D_{M_2M_I} \\ \vdots \\ \vdots \\ \cdots \\ \vdots \\ D_{M_1M_I} \\ D_{M_2M_I} \\ \cdots \\ P_{M_I}P_{m_I} \end{matrix}\right\}.$$

Such as equation (5) in Jung et al. [2005], the analytical formulas of parameter estimates of model (4) at the time t can be obtained as

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$$\left\{\begin{array}{c}\alpha_1(t)\\ \vdots\\ \alpha_I(t)\end{array}\right\}=V_A^{-1}\left\{\begin{array}{c}2D_{M_1Q}\\ \vdots\\ 2D_{M_IQ}\end{array}\right\}\alpha_Q(t).$$

Thus, it is clear that the parameters of LD (i.e., D_{M_jQ} and $D_{M_jM_k}$) and gene effects at the time t (i.e., $\alpha_Q(t)$) are contained in the mean coefficients. Model (4) simultaneously take care of the LD and the effects of the putative trait locus Q. Moreover, the interaction between the genetic effects and time or age is modeled.

In the model (4), the markers M_j, $j = 1, 2,\ldots, I$, are assumed to be located in a region of a single trait locus Q. This assumption can be removed, i.e., the markers can be from different regions of one chromosome or even from different chromosomes. In one region, there can be one or more trait loci. Thus, the multiple trait loci jointly affect the phenotype. For most interest genetic traits, this is a realistic assumption. Similar arguments as above can be done to justify the models, but notations and formulations can be more complex and we do not provide the details in this article.

LONGITUDINAL ASSOCIATION MAPPING MODELS USING MULTIALLELIC MARKERS

In a region of the QTL Q, suppose that multiple multiallelic markers are typed, which may be microsatellite markers. For simplicity, we use two marker A and B in our analysis, but the models and methods can be easily generalized to use multiple markers. Suppose that the markers A and B are in HWE. Let us denote the alleles of marker A by $A_1,\ldots, A_a$, where a is the number of alleles. Let the frequency of A_i be P_{A_i}, i=1, 2,…, a. There are $J_A = a(a + 1)/2$ possible genotypes, which can be listed as $A_1 A_1, \ldots, A_a A_a, A_1 A_2, \ldots, A_1 A_a, \ldots, A_{a-1} A_a$. The marker B has b alleles denoted by $B_1,\ldots, B_b$. Let the frequency of allele B_k be P_{B_k}, k = 1, 2,…, b. There are $J_B = b(b + 1)/2$ possible genotypes, which can be listed as $B_1B_1,\ldots, B_b B_b, B_1B_2,\ldots, B_1B_b,\ldots, B_{b-1}B_b$.

Again, consider a population sample with N individuals. For the ith individual, let y_i be his/her quantitative trait value with genotype G_{Ai} at marker A and genotype G_{Bi} at marker B. Following Fan et al. [2006], consider the following “additive effect model” under normality

$$\begin{aligned}y_i(t)=\mu(t)+w_i(t)\beta(t)+\sum_{j=1}^{a-1}x_{Aij}\alpha_{Aj}(t)\\ +\sum_{j=1}^{b-1}x_{Bij}\alpha_{Bj}(t)+U_i(t)+E_i+e_i,\end{aligned} \tag{5}$$

where the dummy variables x_{Aij} and x_{Bijl} are defined by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\begin{array}{l} x_{A_{ij}}=\begin{cases} 2 & \text{if } G_{A_i}=A_jA_j \\ 1 & \text{if } G_{A_i}=A_jA_l, l \neq j \\ 0 & \text{else} \end{cases}, \\ x_{B_{ij}}=\begin{cases} 2 & \text{if } G_{B_i}=B_jB_j \\ 1 & \text{if } G_{B_i}=B_jB_l, l \neq j \\ 0 & \text{else} \end{cases}, \end{array} \quad (6)$$

and $\alpha_{Aj}(t)$, $\alpha_{Bj}(t)$ are regression coefficients of the dummy variables at the time t. The other terms of model (5) are similar as those of model (4).

In the Supporting Information, we extend additive effect model (5) to "genotype effect model" which takes both additive and dominance effects into account [Fan et al., 2006]. Moreover, we show that the parameters of LD and gene effects are contained in the regression coefficients. The models take care of both the LD and the effects of the trait locus Q. They are valid temporal models to fit association between genetic markers and the trait.

VARIANCE-COVARIANCE STRUCTURE

The variance-covariance structure of stochastic processes $y_i(t)$ depends on the time or age, and can account for heterogeneity between subjects [Soler and Blangero, 2003]. Therefore, appropriate specification of the variance-covariance structure is a key step to successfully model the phenotype trait. For unrelated individuals in the sample, it is reasonable to assume that the quantitative traits are independent. For the same individual, however, the quantitative traits at different times/ages depend on each other. Hence, it is necessary to consider the variance-covariance structure carefully. Let $\sigma_G^2(t)=\mathrm{Var}(U_i(t))$ be the variance of $U_i(t)$ at the time t. For a pair of time points t and s, the covariance is given by

$$\sigma_G(t,s)=\mathrm{Cov}(U_i(t),U_i(s))=\sigma_G(t)\sigma_G(s)\rho_G\,(s,t)\,,$$

where $\rho_G(s, t)$ is correlation between $U_i(t)$ and $U_i(s)$. Let $\sigma_{ga}^2(t)=2q_1q_2[\alpha_Q(t)]^2$ and $\sigma_{gd}^2(t)=\left[q_1q_2\delta_Q(t)\right]^2$ be the major additive and dominant variances at the time t, respectively. The variance-covariance structure of stochastic process $y_i(t)$ is characterized by

$$\mathrm{Cov}(y_i(t),y_i(s)) =\begin{cases} \sigma_{ga}^2(t)+\sigma_{gd}^2(t)+\sigma_G^2(t)+\sigma_E^2+\sigma_e^2 & \text{if } t=s \\ \sigma_G(t,s) & \text{if } t \neq s \end{cases}. \quad (7)$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In above formulation, the covariance 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 covariance function $\sigma_G(t, s)$ (or correlation functions $\rho_G(s, t)$) is a function of $t-s$, i.e., they are functions of the time range. For instance, assume that the correlation effect is an Ornstein-Uhlenbeck Gaussian process $U_i(t)=exp(-t/\rho)W_i(\frac{2}{\rho}e^{2t/\rho})$, $\rho > 0$, where $W_i(t)$ is a standard Brownian motion. Then clearly, $U_i(t)$ has zero mean at all times t. Moreover, the covariance function is $\sigma_G(t,s)=\frac{2}{\rho}\exp(-\left|t-s\right|/\rho)$ [Ross 1996, p. 400]. In this case, the correlation effect $U_i(t)$ is a stationary Gaussian process. In our analysis, 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 [Pinheiro and Bates, 2000]. Third, we fit models by assuming linear correlation in our data 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.

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 nonstationary 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 $\sigma_G(t,s)=\theta_1^2\min(t,s)$.

Another example of nonstationary Gaussian process is integrated Brownian motion, i.e., $U_i(t)=\theta_1\int_0^\tau W_i(s)ds$ where $W_i(t)$ is a standard Brownian motion. Then $U_i(t)$ has zero mean at all times t. The covariance function is $\sigma_G(t,s)=\theta_1^2 s^2(t/2-s/6)$, if $s \leq t$ [Ross 1996, pp. 369–370]. Other examples of Gaussian processes are $U_i(t)=\theta_1[W_i^2(t)-t]$ and $U_i(t)=\theta_1[\exp[\theta_1 W_i(t)-\theta_1^2 t/2]-1]$ [Ross 1996, pp. 381–382].

PENALIZED SPLINE ESTIMATIONS

To estimate the mean function $\mu(t)$ and genetic regression coefficients $\alpha_j(t)$, we may approximate them by linear combinations of penalized spline functions [Wang, 2011]. For instance, the q-order penalized spline model for $\mu(t)$ is

$$\mu(t)=\mu_0+t\mu_1+\ldots+t^q\mu_q+\sum_{k=1}^{K}u_k(t-\kappa_k)_+^q, \quad (8)$$

where μ_i, $i = 0, 1,\ldots, q$, $q \geq 1$, are fixed effects, and u_k, k=1, 2,…, K, are identically and independently normal distributed random variables, κ_k, $k = 1, 2,\ldots, K$, is a preassigned sequence of knots, K is the number of knots, and q is the order of the spline. In addition,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$(t-\kappa_k)_+^q=\begin{cases} (t-\kappa_k)^q \\ \text{if}\, t-\kappa_k>0 \\ 0 \\ \text{else} \end{cases}$$

Let **u** = $(u_1,\ldots, u_K)^\tau$. Assume that $\text{Cov}(\mathbf{u})=\sigma_u^2 I_K$, where I_K is the identity matrix of rank K. Similarly, the regression coefficients $\alpha_j(t)$ can be approximated by linear penalized spline models. For instance, the q-order penalized spline model for $\alpha_1(t)$ is

$$\alpha_1(t)=\alpha_{10}+t\alpha_{11}+\cdots+t^q\alpha_{1_q}+\sum_{k=1}^{K} v_k(t-\kappa_k)_+^q, \quad (9)$$

where α_{1i}, $i = 0, 1,\ldots, q$, are fixed effects, and v_k, k 1, 2,…, K, are identically and independently normal distributed random variables. Let **v** $(v_1,\ldots, v_K)^\tau$. Assume that $\text{Cov}(\mathbf{v})=\sigma_v^2 I_K$

Putting all these together, the final model is a linear mixed model and it can be fitted in R. Usually, one may use the best linear unbiased prediction criteria to estimate the parameters μ_i, α_{1i}, σ_u^2, σ_v^2, σ_e^2, etc. The details of the parameter procedure can be found in literature, and we omit them here. At the first glance, the model seems to be complicated, but in reality very convenient package is available for data analysis in R [Pinheiro and Bates, 2000].

RESULTS

EXAMPLE

We applied the proposed methods to analyze the FHS genetic data. 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, Massachusetts. Since 1948, the subjects have continued to return to the study every 2 years. In 1971, the study enrolled a second-generation group to participate in similar examinations, i.e., Cohort 2. Between 2002 and 2005, the study enrolled the third generation of the FHS—4,095 offspring of the second generation [Splansky et al., 2007]. The first data we analyzed are from GAW 16, which contains phenotypes from the three cohorts and single-nucleotide polymorphism (SNP) genetic markers. The second data we analyzed are from GAW 13, which contains phenotypes of the first two cohorts of FHS and microsatellite markers. The detailed description of the GAW 13 and GAW 16 data can be found from these two workshops [Almasy et al., 2003b; Cupples et al., 2003; MacCluer et al., 2009].

In our analysis, we only use the information of unrelated individuals of GAW 13 or GAW 16 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. For GAW 16 data, a total of 4,156 individuals are eligible in our analysis with a total number of 11,136 measurements. For the GAW 13 data, the number of eligible individuals is 1,129 with a total

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

number of 11,131 measurements. In the following, we present the results of GAW 16, while the results of GAW 13 are presented in the Supporting Information.

ANALYSIS OF FHS DATA FROM GAW 16

To analyze the GAW 16 data, we focused on three candidate regions reported by Levy et al. (2009) for the trait of SBP. Of the three regions, two are on chromosome 12, 88.4Mb–88.7Mb and 110.2Mb–110.5Mb, and the other on chromosome 11, 16.8Mb–16.9Mb.

We first fitted models without using any genetic information. When we fitted the mean function μ(*t*) by linear penalized spline $\mu(t)=\mu_0+t\mu_1+\sum_{k=1}^{K}u_k(t-\kappa_k)_+$, the random term $\sum_{k=1}^{K}u_k(t-\kappa_k)_+$ was significant (the details are presented below as results of nonparametric models). For quadratic or higher order spline $\mu(t)=\mu_0+t\mu_1+\cdots+t^q\mu_q+\sum_{k=1}^{K}u_k(t-\kappa_k)_+^q$ the random term $\sum_{k=1}^{K}u_k(t-\kappa_k)_+^q$ was not significant, and the results of cubic approximation are presented as parametric models below.

Parametric Models—In the following, we present the results of parametric polynomial approximation of μ(*t*). We found that the following cubic linear mixed model can fit the data

$$\begin{aligned}y_{ij}=\mu_0+t_{ij}\mu_{age}+t_{ij}^2\mu_{age^2}+t_{ij}^3\mu_{age^3}+sex_i\beta_{sex}+bmi_{ij}\beta_{bmi}\\+U_i(t_{ij})+E_i+e_{ij},\end{aligned} \qquad (10)$$

where sex_i indicates the gender of the subject *i* (sex_i = 1 for male, sex_i = 2 for female), t_{ij} = *age of subject i at visit j—mean of age,* bmi_{ij} is the BMI for the subject *i* at visit *j*, $U_i(t_{ij})$ is the random correlation effect, E_i is the random variation of SBP for the subject *i*, and e_{ij} is the error term. The variances of *E* and e_{ij} are σ_E^2 and σ_e^2 respectively. In addition, we assumed that y_{ij} and y_{ik} are correlated to each other with an exponential correlation exp $(-|t_{ij} - t_{ik}|/\rho)$, i.e., $U_i(t_{ij})$ are Ornstein-Uhlenbck Gaussian processes.

In the model (10), we used the difference between age and its mean as time variable *t*. This was for the convenience of computational consideration, and we took the difference as time *t* in a hope to avoid big number multiplications and to achieve numeric stability. We also fitted linear correlation for the trait values, but it led to higher AIC and BIC values. Hence, we preferred the exponential correlation [Pinheiro and Bates, 2000]. By fitting linear mixed effect model in R, it was possible to distinguish $\mathrm{Var}(e_{ij})=\sigma_e^2$ and Var($U_i(t_{ij})$) = 2/ρ. The outputs included the estimations of sum variance $\mathrm{Var}(U_i(t_{ij}))+\mathrm{Var}(e_{ij})=2/\rho+\sigma_e^2=\sigma_s^2$, subject variance σ_E^2, and correlation range ρ. The variance estimations of model (10) were $\hat{\sigma}_E^2=8.48^2$ and $\hat{\sigma}_S^2=12.54^2$, and correlation range estimation $\hat{\rho}=4.56$. Thus, the estimation of σ_e^2 is $\hat{\sigma}_e^2=12.54^2-2/4.56=156.813$ The regression results of model (10) are presented in Table I.

Next, we performed analysis by using one SNP a time to fit additive model (4). In the region 88.4Mb–88.7Mb of chromosome 12, three SNPs, rs17249754, rs10858904, and rs17465266,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

were found to have association signal with the SBP at a significance level of 0.05. However, rs17249754 was the only SNP which showed significant association when we added one of rs10858904 or rs17465266 or both to the model in addition to rs17249754. This must be due to the strong LD among the three SNPs. In the region 110.2Mb–110.5Mb of chromosome 12, two SNPs, rs3184504 and rs2301658, were found to have association signal with the SBP at a significance level of 0.05. However, rs3184504 was the only SNP which showed significant association when we added rs2301658 to the model in addition to rs3184504. In the region 16.8Mb–16.9Mb of chromosome 11, three SNPs, rs414219, rs11024074, and rs2041236, were found to have association signal with the SBP at a significance level of 0.05. However, rs11024074 was the only SNP which showed significant association.

In Table I, we present the results of the linear mixed additive models with each of rs17249754, rs3184504, and rs11024074 as a diallelic marker, and the models were

$$\begin{gathered} y_{ij}=\mu_0+t_{ij}\mu_{age}+t_{ij}^2\mu_{age^2}+t_{ij}^2\mu_{age^3}+sex_i\beta_{sex} \\ +bmi_{ij}\beta_{bmi}+x_{ik}\alpha_{k0}+U_i(t_{ij})+E_i+e_{ij}, \end{gathered} \tag{11}$$

where

$$x_{i1}=\begin{cases} 2 & \text{if } G_i=A/A \\ 1 & \text{if } G_i=A/G \\ 0 & \text{if } G_i=G/G \end{cases}$$

is the number of allele *A* in the genotype G_i of subject *i* at SNP rs17249754,

$$x_{i2}=\begin{cases} 2 & \text{if } G_i=C/C \\ 1 & \text{if } G_i=C/T \\ 0 & \text{if } G_i=T/T \end{cases}$$

is the number of allele *C* at SNP rs3184504, and similarly

$$x_{i3}=\begin{cases} 2 & \text{if } G_i=C/C \\ 1 & \text{if } G_i=C/T \\ 0 & \text{if } G_i=T/T \end{cases}$$

is the number of allele *C* at SNP rs11024074.

Finally, we used all three SNPs, rs17249754, rs3184504, and rs11024074, in the analysis, and we found the final model is

$$\begin{gathered} y_{ij}=\mu_0+t_{ij}\mu_{age}+t_{ij}^2\mu_{age^2}+t_{ij}^3\mu_{age^3}+sex_i\beta_{sex} \\ +bmi_{ij}\beta_{bmi}+x_{i1}\alpha_{10}+x_{i1}t_{ij}\alpha_{1,age}+x_{i2}\alpha_{20}+x_{i3}\alpha_{30} \\ +U_i(t_{ij})+E_i+e_{ij}. \end{gathered} \tag{12}$$

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

The results of model (12) are presented in Table I. We actually fitted $\alpha_i(t)$ by linear spline model (9) and found that the random term $\sum_{k=1}^{K} v_k (t - \kappa_k)_+$ had little effect on the model. As the final model (12) shows, $\alpha_1(t) = \alpha_{10} + t\alpha_{1,age}$ can be fitted as a linear relation, and $\alpha_2(t)$ and $\alpha_3(t)$ do not depend on the time t. Each of α_{10}, $\alpha_{1,age}$, α_{20}, α_{30} in model (12) was significant at 95% significance level (Table I). The overall likelihood ratio test of model (12) vs. model (10) to test H_0: $\alpha_{10} = \alpha_{1,age} = \alpha_{20} = \alpha_{30} = 0$ was 28.63, df = 4, P-value < 0.0001.

Therefore, the three SNPs, rs17249754, rs3184504, and rs11024074, are independently associated with the SBP in addition to the effects of age, sex, and BMI. The impact of allele *A* of SNP rs17249754 on the SBP decreases as the time *t* or age increases; the allele *C* of the SNP rs3184504 has negative impact on SBP and the allele *C* of the SNP rs11024074 has positive impact, but both have no significant time- or age-dependent impact.

Nonparametric Models—In model (10), the population mean $\mu(t)$ was fitted by cubic regression without random spline term $\sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+^3$. The reason was that the random term was not significant since σ_u^2 was not significantly larger than 0 at a significance level 0.05, and the same story happened for the quadratic case (data not shown). For linear case, however, the random term $\sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+$was significant (the likelihood ratio test is 62.64 with a P-value < 0.0001 to test the null $H_0{:}\sigma_u^2{=}0$). Therefore, we started with the linear spline model as follows

$$\begin{aligned} y_{ij} = \mu_0 + t_{ij}\mu_{age} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ + sex_i\beta_{sex} + bmi_{ij}\beta_{bmi} \\ + U_i(t_{ij}) + E_i + e_{ij}, \end{aligned} \quad (13)$$

where the number of knots K = 20, and the knots κ_k were uniformly chosen on the interval. By adding each of the three SNPs, rs17249754, rs3184504, and rs11024074, in the analysis, we found significant result for the model below

$$\begin{aligned} y_{ij} = \mu_0 + t_{ij}\mu_{age} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ + sex_i\beta_{sex} + bmi_{ij}\beta_{bmi} \\ + x_{ik}\alpha_{k0} + U_i(t_{ij}) + E_i + e_{ij}, \end{aligned} \quad (14)$$

where x_{ik} are defined above using SNPs, rs17249754, rs3184504, and rs11024074, and k = 1, 2, 3. The results are presented in Table II. By adding all the three SNPs in the analysis, the final model is

$$\begin{aligned} y_{ij} = \mu_0 + t_{ij}\mu_{age} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ + sex_i\beta_{sex} + bmi_{ij}\beta_{bmi} \\ + x_{i1}\alpha_{10} + x_{i1}t_{ij}\alpha_{1,age} + x_{i2}\alpha_{20} + x_{i3}\alpha_{30} + U_i(t_{ij}) \\ + E_i + e_{ij}, \end{aligned} \quad (15)$$

and the results are presented in Table II. As we did for model (12), we fitted $\alpha_i(t)$ by linear spline model (9) and found that the random term had little effect on the model (15).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Like parametric model (12), each of α_{10}, $\alpha_{1,age}$, α_{20}, α_{30} in model (15) was significant at 95% significance level (Table II). The overall likelihood ratio test of model (15) vs. model (13) to test H_0: $\alpha_{10} = \alpha_{1,age} = \alpha_{20} = \alpha_{30} = 0$ was 28.52, df = 4, P-value < 0.0001, which is similar to that of the parametric model (12).

Comparison of Parametric and Nonparametric Models—In Figures 2 and 3, we plotted the predicted SBP vs. age by parametric model (12) and nonparametric model (15). The plots captured the temporal trends of SBP. Before age 40, the SBP was relatively stable and after age 40, it increased. The predicted SBP of female was about 3.2 lower than that of male. Before age 30, the genetic effects were relatively small. After that, the genetic effects gradually got larger. Interestingly, Figures 2 and 3 showed that the temporal trends of the predicted SBP vs. age by parametric model (12) and nonparametric model (15) are very similar, although the parametric predictions shown in Figure 2 are more smooth that those of nonparametric predictions shown in Figure 3.

For nonparametric penalized linear models (13), (14), and (15), the random term $\sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+$ is significant since the null hypothesis $H_0{:}\sigma_u^2{=}0$ is rejected with extremely small P-values. However the regression coefficient of age is not significant, i.e., the null hypothesis H_0: $\mu_{age}=0$ is not rejected due to big P-values (Table II). This is somehow expected. Taking model (13) as an example, the coefficient μ_{age} is the coefficient for the time trend between smallest age and the first knot. The coefficient for the time trend between the first knot and the second knot is $\mu_{age}+u_1$, and the coefficient between the second and the third+ knot is $\mu_{age} + u_1 + u_2$, and so on. Since the SBP does not change with time at early ages (between smallest age and the first knot), so we do not expect μ_{age} to be significant. For cubic parametric linear mixed models (10), (11), and (12), the significant results for σ_u^2 disappear but the regression coefficients of *age*, *age*^{2}, *age*^{3} are significant. Hence, the relation between SBP and age is nonlinear.

In short, the nonlinear trends in nonparametric linear penalized spline models were absorbed into the random component $\sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+$. In cubic parametric linear mixed models, the nonlinearity is reflected by the significant results of regression coefficients of *age*, *age*^{2}, *age*^{3}.

SIMULATION STUDY

To evaluate the performance of the proposed models, simulation studies were carried out to calculate empirical type I error rates, power, and bias of parameter estimation. The results are presented in Table III. We simulated 200 individuals with an age range from 20 to 65 years. For each individual, the number of observations ranged from 3 to 6 and each individual was examined every 2 or 4 years. Due to the random nature of each simulation, the number of total observations were slightly different from each other in the simulation settings, which ranged from 889 to 926 (column 4 of Table III). In the simulation, we assumed that the phenotype was affected by gender that male people's trait was bigger than that of females by 5. One SNP marker was simulated with additive effect and a minor allele frequency of 0.25. For the mean function, we used one logarithm function $\mu(t) = -34.2 +$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

81.7 log(0.3(t + 21.7)) and an exponential function $\mu(t) = 110 \exp(0.0002(t - 25)^2)$. The curves of the two functions are plotted in Figure 4. The logarithm function $\mu(t) = -34.2 + 81.7 \log(0.3(t + 21.7))$ was taken from Daw et al. [2003] and Wang et al. [2012] whose estimates were from the FHS cholesterol data, and the exponential function $\mu(t) = 110 \exp(0.0002(t - 25)^2)$ was used to mimic the SBP data of FHS. For the variance components, the subject variance σ_E^2 was 25 and the error variance σ_e^2 was 10.

We are mainly concerned about the performance of detecting the genetic effect in the model $y_i(t) = \mu(t) + sex_i\beta_{sex} + x_{i1}\alpha_1 + U_i(t) + E_i + \varepsilon_i$. In Table III, the empirical results of type I error rates and power at 95% significance level were reported to test additive genetic effect H_0: $\alpha_1 = 0$ by nonparametric linear penalized spline model, a correctly specified nonlinear function, and misspecified parametric functions of $\mu(t)$. Each empirical type I error rate in Table III was calculated based on 2,000 simulations. That is, we simulated 2,000 random samples. To calculate the type I error rates, we assumed that α_1=0 in our simulation, i.e., the trait was independent of genetic factor. We calculated an empirical test value of likelihood ratio test for each sample. The empirical type I error rates at nominal level $\alpha = 0.05$ are reported in Table III and represented the proportions of the test values calculated for the 2,000 samples, that exceeded the 95th percentiles of the χ_1^2-distribution. The empirical power was calculated similarly by assuming α_1 = 0.5, 1, 2, 3 based on 2,000 simulations.

Encouragingly, the empirical type I error rates were all around the nominal level 0.05 except for the linear misspecified case. The empirical power was very close for each of the three cases: the nonparametric linear penalized spline model, the correctly specified nonlinear function, and the cubic misspecified case. In the linear misspecified case, the empirical type I error rate of 0.009 was unbelievably low for the exponential function and it was 0.0615 which is relatively high for the logarithm case. Thus, the linear misspecified case can give unstable results. The power of linear misspecified case was different from the other cases. Most likely, this was because linear function was far away from the true logarithm and exponential functions, which can be seen from Figure 4. Relatively, the cubic function performs better than the linear function.

To get an understanding about the parameter estimation, we calculated the average of the 2,000 estimates of the coefficient α_1 and then took difference with the true value of α_1 as the bias. From the results in Table III, the nonparametric linear penalized spline model and the correctly specified nonlinear function gave very small bias values. However, both linear and cubic misspecified cases generated big biases. In practice, it is almost impossible to correctly specify the true mean function. Thus, the nonparametric linear penalized spline model is the best choice to analyze data. In addition, high-order parametric methods such as cubic polynomial function can give reasonable results for power and type I errors, but they can generate large biases in parameter estimations.

DISCUSSION

Longitudinal genetic studies provide a very valuable resource for exploring key genetic and environmental factors that affect complex traits over time. Genetic analysis of longitudinal phenotypic data that incorporate temporal variations is important for understanding genetic

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

architecture and biological variations of common complex diseases. It may provide a powerful tool to identify genetic determinants of complex diseases, and to understand at which stage of human development that the genetic determinants are important [Friedlander et al., 1997; Lasky-Su et al., 2008]. Moreover, important environmental factors that are associated with the complex diseases, such as diet, familial income, and smoking status, can be identified. Thereafter, the interactions of genetic determinants and environmental factors, i.e., gene-gene and gene-environment interactions, can be investigated in the presence of temporal trends of phenotypic traits. Although they are important, there is a paucity of statistical methods to analyze longitudinal human genetic data.

In this article, we develop association models to analyze longitudinal data of human genetic studies. Population-based association mapping models are proposed on the basis of temporal population genetic models. The models can be applied to multiple diallelic genetic markers such as SNPs and multiallelic markers such as microsatellites. Theoretical arguments are provided to justify the approaches. The variance-covariance structure is constructed to analyze multiple measurements per individual based on the theory of stochastic processes. To estimate time-dependent mean functions and genetic regression coefficients, we propose approximations by nonparametric penalized spline models. Similar approximations are used in association mapping of nuclear family data and pedigree linkage analysis of longitudinal traits in which only one marker is used in the analysis [Wang 2012; Wang and Huang, 2012]. Another way is to use parametric models in the analysis.

The proposed approaches were applied to analyze GAW 13 and GAW 16 SBP data from FHS. We focused on three candidate regions detected by Levy et al. [2009] for GAW 16 data and an important marker GATA25A04 (D17S1299) for GAW 13 data detected by Levy et al. [2000]. One may want to notice that no temporal trends were studied for the SBP in Levy et al. [2000, 2009] since sample average of each individual’s measurements was used in the analysis. Both parametric and nonparametric models were fitted to identify the important SNPs for GAW 16 data and important allele for GAW 13 at marker locus GATA25A04. We tried to obtain the temporal relations and genetic effect on SBP for these data. When markers are in high LD, collinearity may affect model fitting and selection. In our analysis, collinearity does not cause problem in our analysis of GAW 16 data. However, this does not mean it is not a problem for the other data analysis. In practice, one can calculate the variance inflation factor to make sure that collinearity does not appear to be a potential problem. For markers in strong LD, one may want to use the most significant SNP in the analysis to report the results.

To evaluate the robustness of the nonparametric penalized spline models and parametric models, simulation studies were carried out to calculate and to compare empirical type I error rates and power. In order to understand the accuracy of the parameter estimation, we calculated the biases for parameter to model the genetic effect. The nonparametric penalized spline models are found to perform well in terms of reasonable type I error rates, power, and parameter estimation accuracy.

One merit of the proposed models is that the number of parameters does not depend on the number of multiple measurements. The number of parameters is fixed after carefully

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

specifying regression models and variance-covariance structure. This is different from the method proposed in de Andrade et al. [2002] and de Andrade and Olswold [2003], in which the number of variance-covariance terms to be estimated depends on the number of multiple measurements and grows rapidly when the number of measurements increases. In our proposed models, the parameters are specified through two components based on the theory of stochastic processes: (i) temporal regression models (4) and (5); (ii) temporal variance-covariance functions given by equation (7). If spline functions are used, some parameters can be specified by spline models. In theory, more measurements will lead to more accurate estimation of the parameters. On the one hand, the number of parameters in the proposed models can be significantly smaller than that of de Andrade et al. [2002]. On the other hand, the structure of variance and covariance matrix and mean coefficients of the proposed models is very flexible. These features can be crucial in successful modeling.

In the literature, the phenotypes of longitudinal data can be characterized as function-valued traits [Jaffrezic and Pletcher, 2000; Pletcher and Geyer, 1999]. Specifically, a function-valued trait is a function $y(t)$, where t is a continuous variable, such as age or time. These traits are also called infinite-dimensional traits since the traits can take values at an infinite number of ages [Kirpatrick and Heckman, 1989]. In practice, trait values are observed at a finite set of times $t_1,\dots, t_m$ for an individual, i.e., longitudinal data. Based on the observed data, different methods are proposed to analyze the function-valued traits. In animal breeding, random regression models have been used for the longitudinal data [Diggle et al., 1994; Jamrozik et al., 1997]. An approximation of covariance matrices by orthogonal polynomials has been also proposed [Kirpatrick and Heckman, 1989]. Based on the theory of stochastic processes, the character process model was proposed by Pletcher and Geyer [1999]. These methods were summarized and evaluated in Jafferzic and Pletcher [2000], and used in analyzing empirical data of Drosophila reproduction and mortality and growth in beef cattle. However, the methods do not use any genetic marker data. Functional mapping methods were developed to estimate the dynamic changes of QTL effects during a course of ontogenetic growth [Ma et al., 2002; Wu et al., 2003]. The general concepts and theory of functional data analysis can be found in Ramsay and Silverman [1996]. In our analysis, we adopted some ideas of the character process to build variance-covariance structure, and we use polynomials to approximate the temporal mean function and regression coefficients.

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 population and family data. For genetics community, there is no handy software and statistical models for longitudinal phenotypic traits. For instance, there is no combined linkage and association analysis of the FHS data. The reason is that there are no longitudinal statistical models, methods, and software for a joint linkage and association study of temporal quantitative traits of complex diseases. The proposed methods, in theory, can be extended to analyze the family data or combinations of population and family data. To achieve the goals, temporal variance component models can be built as follows.

The temporal regression models (4) and (5) can be used to model the trait means, which take care of the association information. The temporal variance-covariance functions given by equation (7) can be used for one individual’s measurements. For family members, the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

temporal variance-covariance functions can be constructed in the same way as variance component models presented [Lange, 2002]. The linkage information then is incorporated into the variance-covariance matrix function of pedigree data. If the number of measurements or the size of the pedigrees is large, the dimension of the variance-covariance matrix is large but it should be manageable. If a moderate number of measurements is taken for each individual, it will be interesting to compare the results of our models with those of de Andrade et al. [2002]. For instance, for GAW 13 Cohort 2 of Problem 1 of the FHS, each individual's phenotypes and covariates are measured five times. This provides an opportunity to compare the results by using the real data.

In this article, we propose temporal association mapping models for longitudinal quantitative traits. The models are basically linear mixed effect models. It is interesting in developing temporal models to analyze qualitative genetic traits. To deal with the discrete longitudinal traits, one may use generalized linear mixed models. As the first step, one may start with population data, and then extend to family data or combinations of family data and population data.

We do not deal with various issues such as missing data, population stratification, and heterogeneity in the current study. Surely, these are important topics. For instance, it is unclear how the models performs in the presence of missing data, population stratification, and heterogeneity. All these issues deserve more investigation for future studies.

In summary, the research in this article sheds light on the important area of longitudinal genetic analysis, and it provides a basis for future methodological investigations and practical applications. Many important issues need more insight investigations in the future studies. In our analysis, we use R for our data analysis and simulations. In a long run, user-friendly software and algorithms are needed for genetic public to facilitate data analysis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the Intramural Research Program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development, National Institutes of Health, Maryland.

References

Almasy L, Amos C, Bailey-Wilson JE, Cantor RM, Janquish CE, Martinez M, Neuman RJ, Olson JM, Palmer LJ, Rich SS, Spence MA, MacCluer JW. Genetic Analysis Workshop 13: analysis of longitudinal family data for complex diseases and related risk factors. BMC Genet. 2003a; 4(Suppl 1):S1. [PubMed: 14975069]

Almasy L, Cupples LA, Daw EW, Levy D, Thomas D, Rice JP, Santangelo S, MacCluer JW. Genetic Analysis Workshop 13: introduction to workshop summaries. Genet Epidemiol. 2003b; 25(Suppl 1):S1–S4. [PubMed: 14635163]

Cupples LA, Yang Q, Demissie S, Copenhafer D, Levy D. Description of the Framingham Heart Study data for Genetic Analysis Workshop 13. BMC Genet. 2003; 4(Suppl 1):S2. [PubMed: 14975070]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Daw EW, Morrison J, Zhou XJ, Thomas DC. Genetic Analysis Workshop 13: simulated longitudinal data on families for a system of oligogenic traits. BMC Genet. 2003; 4(Suppl 1):S3. [PubMed: 14975071]

de Andrade M, Gueguen R, Visvikis S, Sass C, Siest G, Amos CI. Extension of variance components approach to incorporate temporal trends and longitudinal pedigree data analysis. Genet Epidemiol. 2002; 22:221–232. [PubMed: 11921082]

de Andrade M, Olswold C. Comparison of longitudinal variance components and regression-based approach for linkage detection on chromosome 17 for systolic blood pressure. BMC Genet. 2003; 4(Suppl 1):S17. [PubMed: 14975085]

Diggle, PJ.; Liang, KY.; Zeger, SL. Analysis of Longitudinal Data. Oxford: Oxford Science Publications; 1994.

Fan RZ, Albert PS, Schisterman EF. A discussion of gene-gene and gene-environment interactions and longitudinal genetic analysis of complex traits. Stat Med. 2012; 33(22)

Fan RZ, Jung JS. High resolution joint linkage disequilibrium and linkage mapping of quantitative trait loci based on sibship data. Hum Hered. 2003; 56:166–187. [PubMed: 15031619]

Fan RZ, Jung JS, Jin L. High resolution association mapping of quantitative trait loci, a population based approach. Genetics. 2006; 172:663–686. [PubMed: 16172503]

Fan RZ, Spinka C, Jin L, Jung JS. Pedigree linkage disequilibrium mapping of quantitative trait loci. Eur J Hum Genet. 2005; 13:216–231. [PubMed: 15483647]

Fan RZ, Xiong MM. High resolution mapping of quantitative trait loci by linkage disequilibrium analysis. Eur J Hum Genet. 2002; 10:607–615. [PubMed: 12357331]

Fan RZ, Xiong MM. Combined high resolution linkage and association mapping of quantitative trait loci. Eur J Hum Genet. 2003; 11:125–137. [PubMed: 12634860]

Friedlander Y, Austin MA, Newman B, Edwards K, Mayer-Davis EI, King MC. Heritability of longitudinal changes in coronary-heart-disease risk factors in women twins. Am J Hum Genet. 1997; 60:1502–1512. [PubMed: 9199573]

Hartl, DL.; Clark, AG. Principles of Population Genetics. 2. Sunderland, MA: Sinauer Associates, Inc; 1989.

Hedrick PW. Gametic disequilibrium measures: proceed with caution. Genetics. 1987; 117:331–341. [PubMed: 3666445]

Jacquard, A. The Genetic Structure of Populations. New York: Springer-Verlag; 1974.

Jaffrezic F, Pletcher SD. Statistical models for estimating the genetic basis of repeated measures and other function-valued traits. Genetics. 2000; 156:913–922. [PubMed: 11014836]

Jamrozik J, Schaeffer LR, Dekkers JCM. Genetic evaluation of dairy cattle using test day yields and random regression model. J Dairy Sci. 1997; 80:1217–1226. [PubMed: 9201594]

Jung JS, Fan RZ, Jin L. Combined linkage and association mapping of quantitative trait loci by multiple markers. Genetics. 2005; 170:881–898. [PubMed: 15802526]

Kirpatrick M, Heckman N. A quantitative genetic model for growth, shape, reaction norms, other infinite-dimensional characters. J Math Biol. 1989; 27:429–450. [PubMed: 2769086]

Lange, K. Mathematical and Statistical Methods for Genetic Analysis. 2. Springer; 2002.

Lasky-Su J, Lyon HN, Emilsson V, Heid IM, Molony C, Raby BA, Lazarus R, Klanderman B, Soto-Quiros ME, Avila L, Silverman EK, Thorleifsson G, Thorsteinsdottir U, Kronenberg F, Vollmert C, Illig T, Fox CS, Levy D, Laird N, Ding X, McQueen MB, Butler J, Ardlie K, Papoutsakis C, Dedoussis G, O’Donnell CJ, Wichmann HE, Celedón JC, Schadt E, Hirschhorn J, Weiss ST, Stefansson K, Lange C. On the replication of genetic associations: timing can be everything! Am J Hum Genet. 2008; 82:848–858.

Levy D, DeStefano AL, Larson MG, O’Donnell CJ, Lifton RP, Gavras H, Cupples LA, Myers RH. 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. 2000; 36:477–483. [PubMed: 11040222]

Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, Aulchenko Y, Lumley T, Köttgen A, Vasan RS, Rivadeneira F, Eiriksdottir G, Guo X, Arking DE, Mitchell GF, Mattace-Raso FU, Smith AV, Taylor K, Scharpf RB, Hwang SJ, Sijbrands EJ, Bis J, Harris TB, Ganesh SK, O’Donnell CJ, Hofman A, Rotter JI, Coresh J,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Benjamin EJ, Uitterlinden AG, Heiss G, Fox CS, Witteman JC, Boerwinkle E, Wang TJ, Gudnason V, Larson MG, Chakravarti A, Psaty BM, van Duijn CM. Genome-wide association study of blood pressure and hypertension. Nat Genet. 2009; 41:677–687. [PubMed: 19430479]

Lewontin RC. On measures of gametic disequilibrium. Genetics. 1988; 120:849–852. [PubMed: 3224810]

Ma CX, Casella G, Wu RL. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. Genetics. 2002; 161:1751–1762. [PubMed: 12196415]

MacCluer JW, Amos CI, Gregersen PK, Heard-Costa N, Lee M, Kraja AT, Borecki IB, Cupples LA, Almasy L. Genetic Analysis Workshop 16: introduction to workshop summaries. Genet Epidemiol. 2009; 33(Suppl 1):S1–S9. [PubMed: 19924709]

Mountz JD, Van Zant GE, Zhang HG, Grizzle WE, Ahmed R, Williams RW, Hsu HC. Genetic dissection of age-related changes of immune function in mice. Scan J Immunol. 2001; 54:10–20.

Mukherjee B, Ko Y, Vanderweele T, Roy A, Park SK, Chen JB. Principal interactions analysis for repeated measures data: application to gene-gene, gene-environment interactions. Stat Med. 2012; 33(22)

Pinheiro, JC.; Bates, DM. Mixed-Effects Models in S and S-PLUS. Springer; 2000.

Pletcher SD, Geyer CJ. The genetic analysis of age-dependent traits: modeling the character process. Genetics. 1999; 151:825–835. [PubMed: 10610347]

Ramsay, JO.; Silverman, BW. Functional Data Analysis. New York: Springer; 1996.

Ross, SM. Stochastic Processes. 2. New York: Wiley; 1996.

Shi G, Rao DC. Ignoring temporal trends in genetic effects substantially reduces power of quantitative trait linkage analysis. Genet Epidemiol. 2008; 32:61–72. [PubMed: 17703462]

Soler JMP, Blangero J. Longitudinal familial analysis of blood pressure involving parametric (co)variance functions. BMC Genet. 2003; 4(Suppl 1):S87. [PubMed: 14975155]

Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, D'Agostino RB Sr, Fox CS, Larson MG, Murabito JM, O'Donnell CJ, Vasan RS, Wolf PA, Levy D. The third generation cohort of the National Heart, Lung, Blood Institute's Framingham Heart Study: design, recruitment, initial examination. Am J Epidemiology. 2007; 165:1328–1335.

Wang, Y. Smoothing Splines, Methods Applications. Boca Raton, FL: CRC Press, A Chapman & Hall Book; 2011.

Wang Y, Huang C. Semiparametric variance components models for genetic studies with longitudinal phenotypes. Biostatistics. 2012; 13:482–496. [PubMed: 21933778]

Wang Y, Huang C, Fang Y, Yang Q, Li R. Flexible semiparametric analysis of longitudinal genetic studies by reduced rank smoothing. Appl Stat-J R Stat Soc Ser C. 2012; 61:1–24.

Wu RL, Ma CX, Zhao W, Casella G. Functional mapping for quantitative trait loci governing growth rates: a parametric model. Physiol Genomics. 2003; 14:241–249. [PubMed: 12923301]

Zhang HP, Zhong XY. Linkage analysis of longitudinal data and design consideration. BMC Genet. 2006; 7:37. [PubMed: 16768806]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

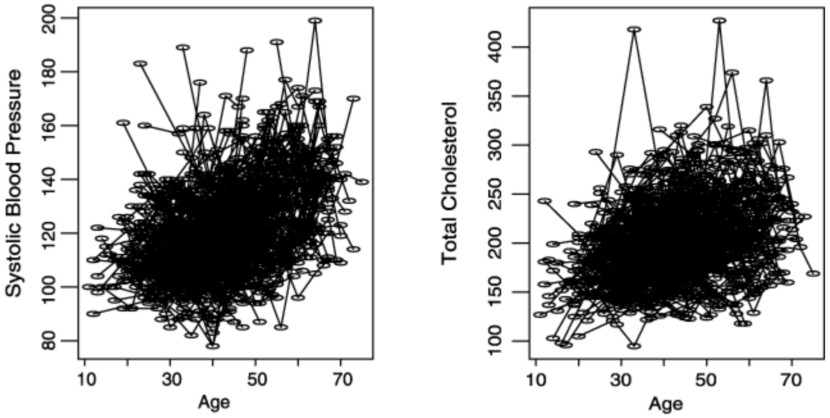

Fig. 1.

Time plot of systolic blood pressure and total plasma cholesterol level against age in years.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

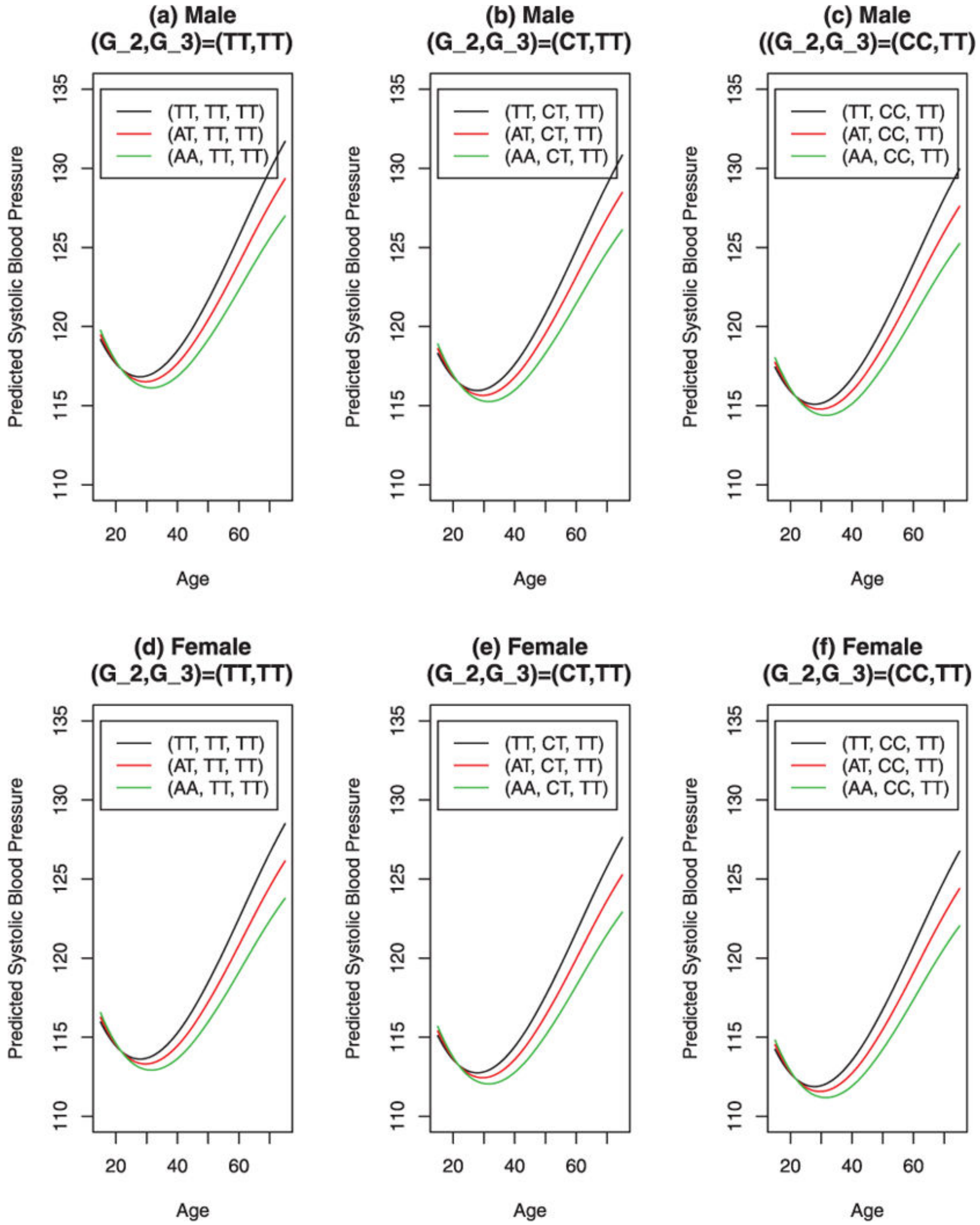

Fig. 2.

Predicted systolic blood pressure against age in years for male and female by SNPs $M_1 = rs17249754$, $M_2 = rs11024074$, $M_3 = rs3184504$, and sex, based on parametric model (12). In the graphs, the legends give the genotypes of the three SNPs; for instance, (AT, CT, TT) means $G_1 = AT$, $G_2 = CT$, and $G_3 = TT$. G_i is the genotype of M_i, $i = 1, 2, 3$.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

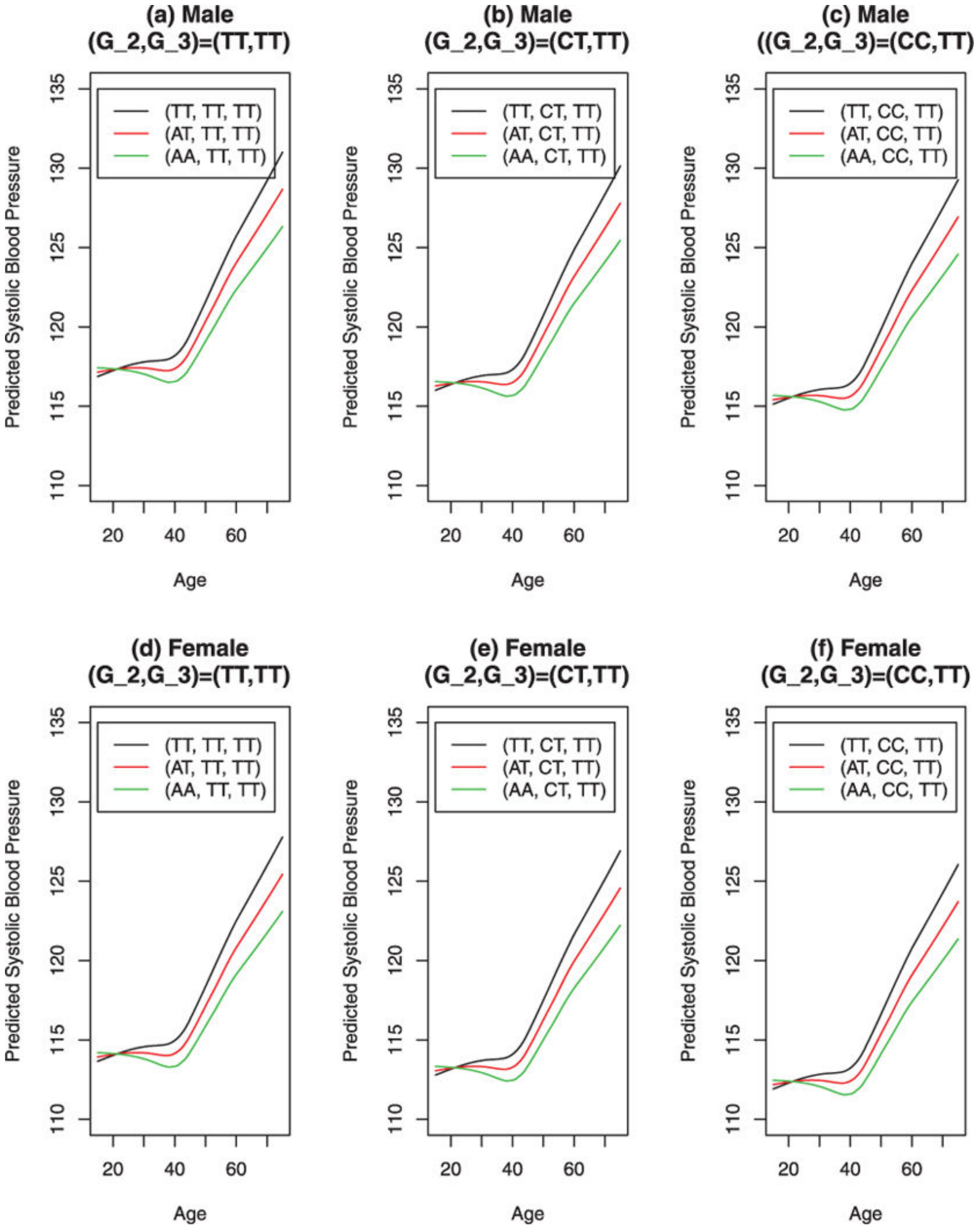

Fig. 3.

Predicted systolic blood pressure against age in years for male and female by SNPs M_1 = *rs*17249754, M_2 = *rs*11024074, M_3 = *rs*3184504, and sex, based on nonparametric model (15). In the graphs, the legends give the genotypes of the three SNPs; for instance, (*AT*,*CT*, *TT*) means G_1 = *AT*,G_2 = *CT*, and G_3 = *TT*. G_i is the genotype of M_i, i = 1, 2, 3.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

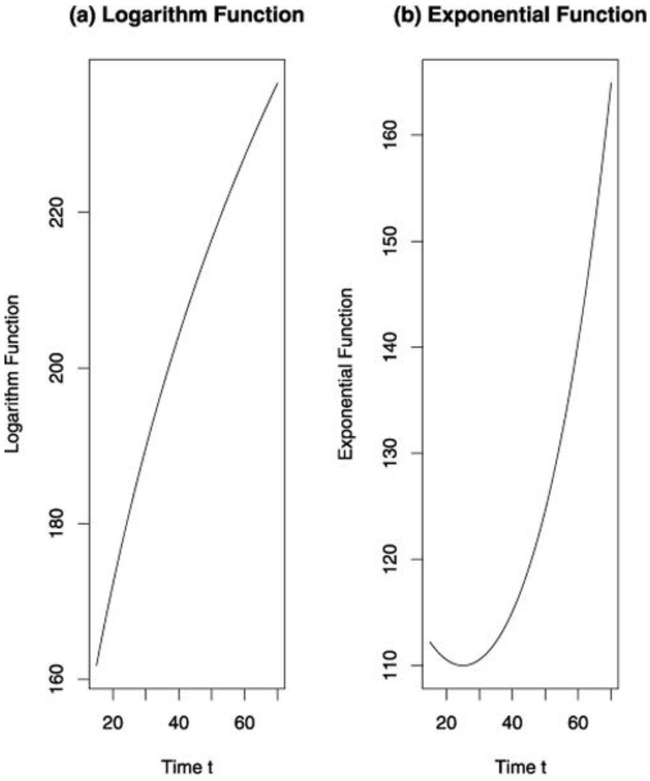

Fig. 4.

The curves of (a) logarithm function μ(t) = −34.2 + 81.7 log(0.3(t + 21.7)) and an (b) exponential function μ(t) = 110 exp(0.0002 × (t − 25)^2)).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE I

Parametric association results of blood systolic pressure and SNPs M_1 = rs17249754, M_2 = rs11024074, M_3 = rs3184504 for FHS data, GAW 16

Model	Coefficient	Estimates	Std error	*t*-value	*P*-value
Model (10)	μ_0	99.92110	1.116201	89.52	<0.0001
	μ_{age}	0.33315	0.019624	16.98	<0.0001
	μ_{age^2}	0.00521	0.000669	7.79	<0.0001
	μ_{age^3}	−0.00013	0.000036	−3.43	0.0006
	β_{sex}	−3.22158	0.409038	−7.88	<0.0001
	β_{bmi}	616.1159	26.939680	22.87	<0.0001
Model (11) based on M_1 = rs17249754	μ_0	100.2625	1.121348	89.41	<0.0001
	μ_{age}	0.33286	0.019618	16.97	<0.0001
	μ_{age^2}	0.00519	0.000669	7.76	<0.0001
	μ_{age^3}	−0.00012	0.000036	−3.41	0.0006
	β_{sex}	−3.21477	0.408619	−7.87	<0.0001
	β_{bmi}	617.6547	26.924825	22.94	<0.0001
	α_{10}	−1.08444	0.373864	−2.90	0.0037
Model (11) based on M_2 = rs3184504	μ_0	100.7971	1.150034	87.65	<0.0001
	μ_{age}	0.33362	0.019624	17.00	<0.0001
	μ_{age^2}	0.00523	0.000669	7.82	<0.0001
	μ_{age^3}	−0.00015	0.000036	−3.41	0.0006
	β_{sex}	−3.19846	0.408522	−7.83	<0.0001
	β_{bmi}	615.9503	26.913737	22.89	<0.0001
	α_{20}	−0.86881	0.281312	−3.09	0.002
Model (11) based on M_3 = rs11024074	μ_0	99.41874	1.132271	87.81	<0.0001
	μ_{age}	0.33312	0.019622	16.98	<0.0001
	μ_{age2}	0.00520	0.000669	7.77	<0.0001
	μ_{age3}	−0.00014	0.000036	−3.41	0.0007

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Model	Coefficient	Estimates	Std error	*t*-value	*P*-value
	β_{sex}	−3.23002	0.408668	−7.90	<0.0001
	β_{bmi}	616.9971	26.924911	22.92	<0.0001
	α_{30}	0.80590	0.309139	2.61	0.0092
Model (12) based on $M_1 = rs17249754$, $M_2 = rs3184504$, $M_3 = rs11024074$	μ_0	100.66600	1.171086	85.96	<0.0001
	μ_{age}	0.35035	0.021427	16.35	<0.0001
	μ_{age^2}	0.00522	0.000668	7.79	<0.0001
	μ_{age^3}	−0.00013	0.000036	−3.47	<0.0001
	β_{sex}	−3.20479	0.407887	−7.86	<0.0001
	β_{bmi}	618.3573	26.885682	23.00	<0.0001
	α_{10}	−1.13894	0.373911	−3.05	0.0023
	$\alpha_{1,age}$	−0.04428	0.022312	−1.99	0.0472
	α_{20}	−0.86859	0.280898	−3.09	0.0020
	α_{30}	0.78825	0.308546	2.55	0.0107

The overall likelihood ratio test of model (12) vs. model (10) to test H_0: $\alpha_{10} = \alpha_{1,age} = \alpha_{20} = \alpha_{30} = 0$ is 28.63, df = 4, *P*-value < 0.0001.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

TABLE II

Nonparametric association results of blood systolic pressure and SNPs M_1 = *rs*17249754, M_2 = *rs*11024074, M_3 = *rs*3184504 for FHS data, GAW 16

Model	**Coefficient**	**Estimates**	**Std error**	***t*-value**	***P*-value**
Model (13)	μ_0	98.6284	6.867915	14.36	<0.0001
	μ_{age}	0.0605	0.209519	0.29	0.77
	β_{sex}	−3.2366	0.408893	−7.92	<0.0001
	β_{bmi}	614.1032	26.929997	22.80	<0.0001
Model (14) based on M_1 = *rs*17249754	μ_0	99.0071	6.845307	14.46	<0.0001
	μ_{age}	0.0629	0.208835	0.30	0.76
	β_{sex}	−3.2297	0.408470	−7.91	<0.0001
	β_{bmi}	615.6503	26.914968	22.87	<0.0001
	α_{10}	−1.0865	0.373734	−2.91	0.0037
Model (14) based on M_2 = *rs*3184504	μ_0	99.4904	6.864736	14.49	<0.0001
	μ_{age}	0.0606	0.209273	0.29	0.77
	β_{sex}	−3.2135	0.408378	−7.87	<0.0001
	β_{bmi}	613.9346	26.904021	22.82	<0.0001
	α_{20}	−0.8690	0.281214	−3.09	0.0020
Model (14) based on M_3 = *rs*11024074	μ_0	98.1551	6.836669	14.36	<0.0001
	μ_{age}	0.0624	0.208559	0.30	0.76
	β_{sex}	−3.2449	0.408523	−7.94	<0.0001
	β_{bmi}	614.9878	26.915266	22.85	<0.0001
	α_{30}	0.8034	0.309029	2.60	0.0094
Model (15) based on M_1 = *rs*17249754, M_2 = *rs*3184504,M_3 = *rs*11024074	μ_0	99.3246	6.846793	14.51	<0.0001
	μ_{age}	0.0750	0.208707	0.36	0.72
	β_{sex}	−3.2196	0.407736	−7.90	<0.0001
	β_{bmi}	616.3468	26.875826	22.93	<0.0001
	α_{10}	−1.1405	0.373783	−3.05	0.0023
	$\alpha_{1,age}$	−0.0436	0.022303	−1.96	0.05
	α_{20}	−0.8688	0.280794	−3.09	0.0020

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Model	**Coefficient**	**Estimates**	**Std error**	***t*-value**	***P*-value**
	α_{30}	0.7860	0.308430	2.55	0.0109

The overall likelihood ratio test of model (15) vs. model (13) to test H_0: $\alpha_{10} = \alpha_{1,age} = \alpha_{20} = \alpha_{30} = 0$ is 28.52, df = 4, P-value < 0.0001.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE III

Simulation results at 95% significance level of testing additive genetic effect *H0:* $\alpha_1 = 0$ in model $y_i(t) = \mu(t) + sex_i\beta_{sex} + x_{i1}\alpha_1 + U_i(t) + E_i + \varepsilon_i$ by nonparametric linear penalized spline model, a correctly specified nonlinear function, and misspecified parametric functions of $\mu(t)$

μ(*t*): Mean function	**Type I error or power (bias)**	α_1	**No. of obs**[c]	**Empirical results: type I error or power (bias $\alpha_1 - \hat{\alpha}_1$)**			
				Nonparametric model[d]	**Correctly specified**[e]	**Misspecified: Linear**[f]	**Misspecified: Cubic**[g]
Logarithm[a]	Type I error (bias)	0	926	0.0535 (0.011)	0.0555 (0.005)	0.0615 (0.0161)	0.0540 (0.009)
	Power (bias)	0.5	896	0.1620 (−0.068)	0.1595 (−0.054)	0.2430 (−0.254)	0.1605 (0.509)
		1	926	0.4010 (0.005)	0.4055 (0.003)	0.5695 (−0.291)	0.4015 (1.009)
		2	889	0.9465 (−0.008)	0.9470 (−0.001)	0.9000 (0.137)	0.9475 (2.009)
		3	902	0.9980 (−0.007)	0.9980 (0.001)	0.9970 (0.124)	0.9980 (3.010)
Exponential[b]	Type I error (bias)	0	916	0.0525 (−0.015)	0.0525 (−0.018)	0.009 (−0.242)	0.0525 (−0.032)
	Power (bias)	0.5	908	0.1215 (0.003)	0.1240 (0.003)	0.0785 (−0.343)	0.1205 (0.467)
		1	900	0.3540 (−0.002)	0.3505 (0.001)	0.3120 (−0.522)	0.3430 (0.967)
		2	906	0.8545 (−0.005)	0.8535 (−0.008)	0.7175 (−0.358)	0.8430 (1.966)
		3	916	0.9990 (−0.015)	0.9990 (−0.018)	0.9910 (−0.242)	0.9990 (2.968)

[a] True $\mu(t) = -34.2 + 81.7\log(0.3(t + 21.7))$.

[b] True $\mu(t) = 110\exp(0.0002\times(t - 25)^2)$.

[c] The total number of observations.

[d] $\mu(t)$ is estimated by nonparametric linear penalized spline model.

[e] $\mu(t)$ is correctly specified.

[f] $\mu(t)$ is misspecified as $\mu(t) = \mu_0 + \mu_1 t$.

[g] $\mu(t)$ is misspecified as $\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2 + \mu_3 t^3$.