Longitudinal Association Analysis of Quantitative Traits

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Genetic Epidemiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>GenEpi-12-0068</td>
</tr>
<tr>
<td>Wiley - Manuscript type:</td>
<td>Research Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>28-Mar-2012</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>fan, ruzong; NIH, NICHD</td>
</tr>
<tr>
<td>Key Words:</td>
<td>association mapping, quantitative trait loci, longitudinal analysis</td>
</tr>
</tbody>
</table>
Longitudinal Association Analysis of Quantitative Traits

Ruzong Fan\textsuperscript{1,*}, Yiwei Zhang\textsuperscript{2}, Paul S. Albert\textsuperscript{1}, Aiyi Liu\textsuperscript{1},
Yuanjia Wang\textsuperscript{3}, and Momiao Xiong\textsuperscript{4}

\textsuperscript{1}6100 Executive Blvd, Room 7B05, MSC 7510
Biostatistics and Bioinformatics Branch
Division of Epidemiology, Statistics, and Prevention
\textit{Eunice Kennedy Shriver} National Institute
of Child Health and Human Development
National Institutes of Health, Rockville, MD 20852

\textsuperscript{2}Division of Biostatistics, University of Minnesota
A460 Mayo Building, MMC 303, Minneapolis, MN 55455

\textsuperscript{3}Department of Biostatistics, Mailman School of Public Health
Columbia University, New York, NY 10032

\textsuperscript{4}Human Genetics Center, University of Texas - Houston
P.O. Box 20334, Houston, Texas 77225

Short Title: Association Analysis of Quantitative Traits

*Correspondence to: Dr. Ruzong Fan, Tel. 301-496-6813,
Fax 301-402-2084, E-mail: fanr@mail.nih.gov

March 28, 2012
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 incorporates 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 non-parametric models are proposed. The models can be applied to multiple di-allele genetic markers such as single nucleotide polymorphisms and multi-allele markers such as micro-satellites. 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 non-parametric 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.

Key Words: 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.\textsuperscript{1,2} 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.\textsuperscript{3} 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).\textsuperscript{4,5} 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.\textsuperscript{6-8} 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.\textsuperscript{9}

For certain traits of complex diseases, like 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.\textsuperscript{15}

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 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 were developed to analyze the FHS data and the similarly structured simulated data.\textsuperscript{10,11} 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 (2002, 2003).\textsuperscript{12,13} 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 parameter. 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).\textsuperscript{14} 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 non-parametric models are proposed. By utilizing multiple genetic markers which can be either di-allelic or multi-allelic, 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 (2002, 2003)\textsuperscript{12,13} 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 non-parametrically which makes our methods different from the parametric models of Zhang and Zhong (2006).\textsuperscript{14} 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 the robustness, power, and parameter estimation accuracy of the proposed non-parametric 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 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

\[
\begin{align*}
\mu_{11}(t) &= \mu(t) + 2\alpha_1(t) + d_1(t), \\
\mu_{12}(t) &= \mu(t) + \alpha_1(t) + \alpha_2(t) + d_2(t), \\
\mu_{22}(t) &= \mu(t) + 2\alpha_2(t) + d_3(t),
\end{align*}
\]

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 $F(\mu, \alpha_1, \alpha_2) = \sum_{i=1}^{2} \sum_{j=1}^{2} q_i q_j (\mu_{ij}(t_0) - \mu(t_0) - \alpha_i(t_0) - \alpha_j(t_0))^2$, classical theory of quantitative genetics provides the estimates of $\mu(t_0), \alpha_1(t_0), \alpha_2(t_0)$ as \textsuperscript{18,19}

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

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

\[
\begin{align*}
\mu_{11}(t_0) &= \mu_0(t_0) + 2q_2\alpha_Q(t_0) - q_2^2\delta_Q(t_0), \\
\mu_{12}(t_0) &= \mu_0(t_0) + (q_2 - q_1)\alpha_Q(t_0) + q_1q_2\delta_Q(t_0), \\
\mu_{22}(t_0) &= \mu_0(t_0) - 2q_1\alpha_Q(t_0) - q_1^2\delta_Q(t_0).
\end{align*}
\]

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 equations (1) imply that the trait can be expressed as

\[
y(t_0) = \mu_0(t_0) + x_Q\alpha_Q(t_0) + z_Q\delta_Q(t_0) + \epsilon,
\]
where $\epsilon$ is a random error term, and $x_Q$ and $z_Q$ are dummy variables defined by

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

(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 Di-allelic Markers**

In our previous research, population-based regression association models of QTL are constructed in Fan and Xiong (2002) by using multiple di-allelic markers in the analysis. The models are further extended to be variance component models for combined linkage and association mapping of QTL. The temporal models described as follows can be thought as longitudinal extension of the association mapping models in Fan and Xiong (2002).

Assume that $K$ di-allelic markers $M_j$, $j = 1, 2, \cdots, K$ 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 $D_{M_jM_k} = P(M_jM_k) - P_{M_j}P_{M_k}, j, k = 1, 2, \cdots, K$. Here $P(M_jQ_1)$ and $P(M_jM_k)$ are frequencies of haplotypes $M_jQ_1$ and $M_jM_k$, respectively.

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

$$y_i(t) = \mu(t) + w_i(t)^\top \beta(t) + \sum_{j=1}^K x_{ij} \alpha_j(t) + \sum_{j=1}^K z_{ij} \delta_j(t) + U_i(t) + E_i + \epsilon_i.$$  

(4)
The components of the above model are specified as follows. First, $\mu(t)$ is a non-random 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 possible their interaction terms; $\beta(t)$ is a non-random column vector of regression parameters of the covariates $w_i(t)$ with fixed effects. One may want to notice that the covariates can be time invariant like gender or can be time varying such as the BMI. In addition, $x_{ij}$ and $z_{ij}$ are dummy variables defined by

$$x_{ij} = \begin{cases} 2 & \text{if } G_{ij} = M_j M_j \\ 1 & \text{if } G_{ij} = M_j m_j \\ 0 & \text{if } G_{ij} = m_j m_j \end{cases},$$

and $z_{ij}$ = \begin{cases} -P^2_{m_j} & \text{if } G_{ij} = M_j M_j \\ P_{m_j} P_{M_j} & \text{if } G_{ij} = M_j m_j \\ -P^2_{M_j} & \text{if } G_{ij} = m_j m_j \end{cases} \quad (5)$

and $\alpha_j(t)$ and $\delta_j(t)$ are regression coefficients of the dummy variables $x_{ij}$ and $z_{ij}$ at the time $t$.

In model (4), $U_i(t)$ is a random polygenic effect; $E_i$ is a random variation of subject $i$; and $\epsilon_i$ is a random measurement error term. Assume that $U_i(t)$, $E_i$, and $\epsilon_i$ are independent. Moreover, assume that $E_i$ is normal $N(0, \sigma^2_E)$ and $\epsilon_i$ is normal $N(0, \sigma^2)$.

The above formulation is, of course, a general framework. In the context of population data, some of these components may be confounded. In addition, some of the components may need to be removed from the analysis. For instance, assume that the dominance effects are not significantly present. Then model (4) can be simplified to

$$y_i(t) = \mu(t) + w_i(t)\beta(t) + \sum_{j=1}^{K} x_{ij}\alpha_j(t) + U_i(t) + E_i + \epsilon_i. \quad (6)$$

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) (or model (6)) is that we include measured genotype components estimating association with genotyped markers, i.e., the terms involves $x_{ij}$ and $z_{ij}$ (or $x_{ij}$). Let the additive and dominance variance-covariance matrices of the indicator variables defined in (5) be

$$V_A = 2 \begin{pmatrix} P_{M_1} P_{m_1} & D_{M_1 M_2} & \cdots & D_{M_1 M_K} \\ D_{M_1 M_2} & P_{M_2} P_{m_2} & \cdots & D_{M_2 M_K} \\ \vdots & \vdots & \ddots & \vdots \\ D_{M_1 M_K} & D_{M_2 M_K} & \cdots & P_{M_K} P_{m_K} \end{pmatrix}, V_D = \begin{pmatrix} D^2_{M_1 M_1} & D^2_{M_1 M_2} & \cdots & D^2_{M_1 M_K} \\ D^2_{M_2 M_1} & D^2_{M_2 M_2} & \cdots & D^2_{M_2 M_K} \\ \vdots & \vdots & \ddots & \vdots \\ D^2_{M_K M_1} & D^2_{M_K M_2} & \cdots & P^2_{M_K} P^2_{m_K} \end{pmatrix}. \quad (7)$$
Such as equation (5) in Jung, Fan and Jin (2005), the analytical formulas of parameter estimates of models (4) and (6) at the time \( t \) can be obtained as

\[
\begin{pmatrix}
\alpha_1(t) \\
\vdots \\
\alpha_K(t)
\end{pmatrix} = V_A^{-1} \begin{pmatrix}
2D_{M_1Q} \\
\vdots \\
2D_{M_KQ}
\end{pmatrix} \alpha_Q(t) \text{ and } \begin{pmatrix}
\delta_1(t) \\
\vdots \\
\delta_K(t)
\end{pmatrix} = V_D^{-1} \begin{pmatrix}
D^2_{M_1Q} \\
\vdots \\
D^2_{M_KQ}
\end{pmatrix} \delta_Q(t).
\]

(8)

From equations (8), 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., \( \alpha_Q(t) \) and \( \delta_Q(t) \)) are contained in the mean coefficients. Hence, models (4) and (6) 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 models (4) and (6), the markers \( M_j, j = 1, 2, \ldots, K \), 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 don’t provide the details in this article.

**Longitudinal Association Mapping Models Using Multi-allelic Markers**

In a region of the QTL \( Q \), suppose that multiple multi-allelic markers are typed, which may be micro-satellite 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 \( P_{A_i}, i = 1, 2, \ldots, a \). There are \( J_A = a(a + 1)/2 \) possible genotypes, which can be listed as \( A_1A_1, \ldots, A_aA_a, A_1A_2, \ldots, A_1A_a, \ldots, A_{a-1}A_a \). The marker \( B \) has \( b \) alleles denoted by \( B_1, \ldots, B_b \). Let the frequency of allele \( B_k \) be \( P_{B_k}, k = 1, 2, \ldots, b \). There are \( J_B = b(b + 1)/2 \) possible genotypes, which can be listed as \( B_1B_1, \ldots, B_bB_b, B_1B_2, \ldots, B_1B_b, \ldots, B_{b-1}B_b \).

Again, consider a population sample with \( N \) individuals. For the \( i \)-th 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, Jung, and Jin (2006), consider the following “genotype effect model” under normality

\[
y_i(t) = \mu(t) + w_i(t)\beta(t) + \sum_{j=1}^{a-1} x_{Aij}\alpha_{Aj}(t) + \sum_{j=1}^{b-1} x_{Bij}\alpha_{Bj}(t) + \sum_{1\leq j<l\leq a} z_{Aijl}\delta_{Aijl}(t) + \sum_{1\leq j<l\leq b} z_{Bijl}\delta_{Bijl}(t) + U_i(t) + E_i + e_i,
\]

where the dummy variables \(x_{Aij}, z_{Aijl}, x_{Bijl}\) and \(z_{Bijl}\) are defined by

\[
x_{Aij} = \begin{cases} 2 & \text{if } G_{Ai} = A_jA_j \\ 1 & \text{if } G_{Ai} = A_jA_l, l \neq j \\ 0 & \text{else} \end{cases}
\]

\[
z_{Aijl} = \begin{cases} -P_{A_i}^2 & \text{if } G_{Ai} = A_jA_j \\ P_{A_i}P_{A_l} & \text{if } G_{Ai} = A_jA_l, j \neq l \\ -P_{A_i}^2 & \text{if } G_{Ai} = A_lA_l \\ 0 & \text{else} \end{cases}
\]

\[
x_{Bijl} = \begin{cases} 2 & \text{if } G_{Bi} = B_jB_j \\ 1 & \text{if } G_{Bi} = B_jB_l, l \neq j \\ 0 & \text{else} \end{cases}
\]

\[
z_{Bijl} = \begin{cases} -P_{B_i}^2 & \text{if } G_{Bi} = B_jB_j \\ P_{B_i}P_{B_l} & \text{if } G_{Bi} = B_jB_l, j \neq l \\ -P_{B_i}^2 & \text{if } G_{Bi} = B_lB_l \\ 0 & \text{else} \end{cases}
\]

and \(\alpha_{Aj}(t), \alpha_{Bj}(t), \delta_{Aijl}(t), \delta_{Bijl}(t)\) are regression coefficients of the dummy variables at the time \(t\).

The other terms of model (9) are similar as those of model (4). If both markers \(A\) and \(B\) are di-allelic, it can be shown that model (9) is equivalent to model (4). In fact, notice that \(x_{ij} + 2P_M = \begin{cases} 2 & \text{if } G_{ij} = M_JM_J \\ 1 & \text{if } G_{ij} = M_JM_j \text{ since } P_M + P_m = 1, \text{ where } x_{ij} \text{ is defined in (5). Hence, model (9) extends model (4) to use multi-allelic markers in analysis.}
\]

Model (9) takes both additive and dominance effects into account. Such as model (6), model (9) can be modified to an “additive effect model” if only the additive effect is modeled, i.e.,

\[
y_i(t) = \mu(t) + w_i(t)\beta(t) + \sum_{j=1}^{a-1} x_{Aij}\alpha_{Aj}(t) + \sum_{j=1}^{b-1} x_{Bij}\alpha_{Bj}(t) + U_i(t) + E_i + e_i.
\]

In the Appendix, we show that the parameters of LD and gene effects are contained in the regression coefficients. Models (9) and (11) 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. 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 \( \sigma_G^2(t) = \text{Var}(U_i(t)) \) be the
polygenic variance at the time \( t \). For a pair of time points \( t \) and \( s \), the covariance is given by
\[
\sigma_G(t, s) = \text{Cov}(U_i(t), U_i(s)) = \sigma_G(t)\sigma_G(s)\rho_G(s, t),
\]
where \( \rho_G(s, t) \) is correlation of polygenic variations between \( U_i(t) \) and \( U_i(s) \). Let \( \sigma^2_{ga}(t) = 2q_1q_2[\alpha_Q(t)]^2 \)
and \( \sigma^2_{gd}(t) = [q_1q_2\delta_Q(t)]^2 \) 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
\[
\text{Cov}(y_i(t), y_i(s)) = \begin{cases} 
\sigma^2_{ga}(t) + \sigma^2_{gd}(t) + \sigma^2_G(t) + \sigma^2_E + \sigma^2_e & \text{if } t = s \\
\sigma^2_G(t, s) & \text{if } t \neq s 
\end{cases}
\]
In above formulation, the covariance \( \text{Cov}(y_i(t), y_i(s)) \) is assumed to be equal to the covariance of
polygenic variations \( U_i(t) \) and \( U_i(s) \), \( t \neq s \). In practice, the correlation between \( y_i(t) \) and \( y_i(s) \) can be
from the polygenic variations, major additive and dominant effects at the QTL \( Q \), and environmental
factors or their combinations. For the population data, it is impossible to distinguish them. Hence,
we simply put it as the correlation of polygenic variations.

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 polygenic 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 \( \sigma_G(t, s) = \frac{2}{\rho} \exp(-|t - s|/\rho) \) (P400, Ross 1996). In this case, the polygenic 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. 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 less good.

In certain cases, however, the covariance or correlation functions may not be functions of the time range. In this case, the polygenic effect $U_i(t)$ is a non-stationary process. For instance, assume that the polygenic 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 non-stationary Gaussian process is integrated Brownian motion, i.e., $U_i(t) = \theta_1 \int_0^t W_i(s) ds$, 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 $s \leq t$ (pp369-370, Ross 1996). Other examples of Gaussian processes are $U_i(t) = \theta_1 W_i^2(t) - t$ and $U_i(t) = \theta_1 [\exp[\theta_1 W_i(t) - \theta_1^2 t/2] - 1]$ (pp381-382, Ross 1996).

**Penalized Spline Estimations**

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

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

where $\mu_i, i = 0, 1, \cdots, q, q \geq 1$, are fixed effects, and $u_k, k = 1, 2, \cdots, K$, are identically and independently normal distributed random variables, $\kappa_k, k = 1, 2, \cdots, K$, is a pre-assigned sequence of knots, $K$ is the number of knots, and $q$ is the order of the spline. In addition, $(t - \kappa_k)^q_+ = \begin{cases} (t - \kappa_k)^q & \text{if } t - \kappa_k > 0 \\ 0 & \text{else} \end{cases}$. Let $\mathbf{u} = (u_1, \cdots, u_K)^T$. 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 $\alpha_j(t)$ and $\delta_j(t)$ can be approximated by linear penalized spline models. For instance, the $q$-order penalized spline model for $\alpha_1(t)$ is

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

where $\alpha_{1i}, i = 0, 1, \cdots, q$, are fixed effects, and $v_k, k = 1, 2, \cdots, K$, are identically and independently distributed normal random variables.
normal distributed random variables. Let \( \mathbf{v} = (v_1, \ldots, v_K)^T \). 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 (BLUP) criteria to estimate the parameters \( \mu, \alpha_1, \sigma_u^2, \sigma_v^2, \sigma_e^2 \), 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.\(^{32}\)

### 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 two 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.\(^{35}\) The first data we analyzed are from GAW 16, which contains phenotypes from the 3 cohorts and SNP genetic markers. The second data we analyzed are from GAW 13, which contains phenotypes of the first two cohorts of FHS and micro-satellite markers. The detailed description of the GAW 13 and GAW 16 data can be found from these two workshops.\(^{1,2,11}\)

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 number of 11,131 measurements.
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 systolic blood pressure. Of the three regions, two are on chromosome 12, 88.4Mb to 88.7Mb and 110.2Mb to 110.5Mb, and the other on chromosome 11, 16.8Mb to 16.9Mb.

We first fitted models without using any genetic information. When we fitted the mean function \( \mu(t) \) 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 non-parametric models). For quadratic or higher order spline \( \mu(t) = \mu_0 + t \mu_1 + \cdots + t^n \mu_n + \sum_{k=1}^{K} u_k(t - \kappa_k)^n_+ \), the random term \( \sum_{k=1}^{K} u_k(t - \kappa_k)^n_+ \) 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 \( \mu(t) \). We found that the following cubic linear mixed model can fit the data

\[
y_{ij} = \mu_0 + t_{ij} \mu_{age} + t_{ij}^2 \mu_{age^2} + t_{ij}^3 \mu_{age^3} + \text{sex}_i \beta_{sex} + \text{bmi}_{ij} \beta_{bmi} + U_i(t_{ij}) + E_i + e_{ij},
\]

where \( \text{sex}_i \) indicates the gender of the subject \( i \) (\( \text{sex}_i = 1 \) for male, \( \text{sex}_i = 2 \) for female), \( t_{ij} \) = age of subject \( i \) at visit \( j \) – mean of age, \( \text{bmi}_{ij} \) is the BMI for the subject \( i \) at visit \( j \), \( U_i(t_{ij}) \) is the random polygenic variation, \( E_i \) is the random variation of SBP for the subject \( i \), and \( e_{ij} \) is the error term. The variances of \( E_i \) and \( e_{ij} \) are \( \sigma_E^2 \) and \( \sigma_e^2 \), 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-Uhlenbeck Gaussian processes.

In the model (15), 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. By fitting linear mixed effect model in R, it was possible to distinguish \( \text{Var}(e_{ij}) = \sigma_e^2 \) and \( \text{Var}(U_i(t_{ij})) = 2/\rho \). The outputs included the estimations of sum variance \( \text{Var}(U_i(t_{ij}))+\text{Var}(e_{ij}) = 2/\rho + \sigma_e^2 = \sigma_S^2 \), subject variance \( \sigma_E^2 \), and correlation range \( \rho \). The variance estimations of model (15) were \( \hat{\sigma}_S^2 = 8.48^2 \) and \( \hat{\sigma}_S^2 = 12.54^2 \), and correlation range estimation was \( \hat{\rho} = 4.56 \). Thus, the
estimation of $\sigma^2$ is $\hat{\sigma}^2 = 12.54^2 - 2/4.56 = 156.813$. The regression results of model (15) were presented in Table 1.

Next, we performed analysis by using one SNP a time to fit additive model (6). In the region 88.4Mb to 88.7Mb of chromosome 12, three SNPs, rs17249754, rs10858904, and rs17465266, 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 the strong linkage disequilibrium among the 3 SNPs. In the Table 1, we presented the results of the linear mixed additive model with rs17249754 as a di-allelic marker, and the model was

$$y_{ij} = \mu_0 + t_{ij}\mu_{age} + t^2_{ij}\mu_{age^2} + t^3_{ij}\mu_{age^3} + sex_i\beta_{sex} + bmi_{ij}\beta_{bmi} + x_{i1}\alpha_{10} + U_i(t_{ij}) + E_i + e_{ij},$$

(16)

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.

In the region 110.2Mb to 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 to 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 the Table 1, we presented the results of the following linear mixed additive models

$$y_{ij} = \mu_0 + t_{ij}\mu_{age} + t^2_{ij}\mu_{age^2} + t^3_{ij}\mu_{age^3} + sex_i\beta_{sex} + bmi_{ij}\beta_{bmi} + x_{i2}\alpha_{30} + E_i + e_{ij},$$

(17)

$$y_{ij} = \mu_0 + t_{ij}\mu_{age} + t^2_{ij}\mu_{age^2} + t^3_{ij}\mu_{age^3} + sex_i\beta_{sex} + bmi_{ij}\beta_{bmi} + x_{i3}\alpha_{30} + E_i + e_{ij},$$

(18)

where $x_{i2} = \begin{cases} 2 & \text{if } G_i = C/C \\ 1 & \text{if } G_i = C/T \\ 0 & \text{if } G_i = T/T \end{cases}$ is the number of allele $C$ in the genotype $G_i$ of subject $i$ at SNP rs17249754.
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 $C$ in the genotype $G_i$ of subject $i$ at SNP rs11024074.

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

$$y_{ij} = \mu_0 + t_{ij} \mu_{age} + t_{ij}^2 \mu_{age^2} + t_{ij}^3 \mu_{age^3} + \text{sex}_i \beta_{sex} + \text{bmi}_{ij} \beta_{bmi} + x_{i1} \alpha_{10} + x_{i1} t_{ij} \alpha_{1,age} + x_{i2} \alpha_{20} + x_{i3} \alpha_{30} + U_i(t_{ij}) + E_i + e_{ij}.$$  \tag{19}$$

The results of model (19) were presented in the Table 1. We actually fitted $\alpha_i(t)$ by linear spline model (14) and found that the random term $\sum_{k=1}^{K} u_k(t_i - \kappa_k)_+$ had little effect on the model. As the final model (19) 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 $\alpha_{10}, \alpha_{1,age}, \alpha_{20}, \alpha_{30}$ in model (19) was significant at 95% significance level (Table 1). The overall likelihood ratio test of model (19) vs. model (15) 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’s 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.

**Non-Parametric Models.** In model (15), the population mu $\mu(t)$ was fitted by cubic regression without random spline term $\sum_{k=1}^{K} u_k(t_i - \kappa_k)_+^3$. The reason was that the random term was not significant since $\sigma_u^2$ 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

$$y_{ij} = \mu_0 + t_{ij} \mu_{age} + \sum_{k=1}^{K} u_k(t_{ij} - \kappa_k)_+ + \text{sex}_i \beta_{sex} + \text{bmi}_{ij} \beta_{bmi} + U_i(t_{ij}) + E_i + e_{ij}.$$  \tag{20}$$
where the number of knots $K = 20$, and the knots $\kappa_k$ were uniformly chosen on the interval. By adding each of the three SNPs rs17249754, rs3184504, and rs11024074 in the analysis, we found significant results for each of the three models below

\[
y_{ij} = \mu_0 + t_{ij} \mu_{age} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ + \text{sex}_i \beta_{sex} + \text{bmi}_{ij} \beta_{bmi} + x_{i1} \alpha_{10} + U_i(t_{ij}) + E_i + e_{ij}, \tag{21}
\]

\[
y_{ij} = \mu_0 + t_{ij} \mu_{age} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ + \text{sex}_i \beta_{sex} + \text{bmi}_{ij} \beta_{bmi} + x_{i2} \alpha_{20} + U_i(t_{ij}) + E_i + e_{ij}, \tag{22}
\]

\[
y_{ij} = \mu_0 + t_{ij} \mu_{age} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ + \text{sex}_i \beta_{sex} + \text{bmi}_{ij} \beta_{bmi} + x_{i3} \alpha_{30} + U_i(t_{ij}) + E_i + e_{ij}. \tag{23}
\]

The results were presented in Table 2. By adding all three SNPs rs17249754, rs3184504, and rs11024074 in the analysis, we found the final model is

\[
y_{ij} = \mu_0 + t_{ij} \mu_{age} + \sum_{k=1}^{K} u_k (t_{ij} - \kappa_k)_+ + \text{sex}_i \beta_{sex} + \text{bmi}_{ij} \beta_{bmi} + x_{i1} \alpha_{10} + x_{i1} t_{ij} \alpha_{1,age} + x_{i2} \alpha_{20} + x_{i3} \alpha_{30} + U_i(t_{ij}) + E_i + e_{ij}, \tag{24}
\]

and the results were presented in the Table 2. As we did for model (19), we fitted $\alpha_i(t)$ by linear spline model (14) and found that the random term had little effect on the model (24).

Such as parametric model (19), each of $\alpha_{10}, \alpha_{1,age}, \alpha_{20}, \alpha_{30}$ in model (24) was significant at 95% significance level (Table 2). The overall likelihood ratio test of model (24) vs. model (20) 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 (19).

**Comparison of Parametric and Non-parametric Models.** In Figures 2 and 3, we plotted the predicted systolic blood pressure vs age by parametric model (19) and non-parametric model (24). 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 systolic blood pressure vs age by parametric model (19) and non-parametric model (24) are very similar,
although the parametric predictions shown in Figure 2 are more smooth than those of non-parametric predictions shown in Figure 3.

For non-parametric penalized linear models (20), (21), (22), (23), and (24), 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 2). This is somehow expected. Taking model (20) as an example, the coefficient $\mu_{age}$ 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 doesn’t change with time at early ages (between smallest age and the first knot), so we do not expect $\mu_{age}$ to be significant. For cubic parametric linear mixed models (15), (16), (17), (18), and (19), the significant results for $\sigma_u^2$ disappear but the regression coefficients of $age, age^2, age^3$ are significant. Hence, the relation between SBP and age is non-linear.

In short, the non-linear trends in non-parametric 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 non-linearity is reflected by the significant results of regression coefficients of $age, age^2, age^3$.

### Analysis of FHS Data from GAW 13

Levy et al. (2000)\textsuperscript{4} performed a genome-wide linkage analysis of FHS data and identified a multi-allelic locus GATA25A04 (D17S1299) on chromosome 17 which shows strong linkage with a high LOD score 3.8. We concerned about the association between systolic blood pressure and genotypes of the locus GATA25A04 over people’s age. At the locus GATA25A04, there are 8 alleles (184, 188, 192, 196, 200, 204, 208, 212). We investigated the temporal trend of systolic blood pressure related to genetic information, age, and sex. After a thorough model selection by fitting linear mixed model in R (Pinheiro and Bates 2000), we got a final model as

$$y_{ij} = \mu_0 + t_{ij}\mu_{age} + t_{ij}^2\mu_{age^2} + t_{ij}^3\mu_{age^3} + sex_i\beta_{sex}$$
\[ +x_{i,188}t_{ij}^{\alpha_{age}} + x_{i,188}t_{ij}^{2\alpha_{age}} + x_{i,188}t_{ij}^{3\alpha_{age}} + U_i(t_{ij}) + E_i + e_{ij}, \]  

(25)

where \( x_{i,188} = \begin{cases} 2 & \text{if } G_i = 188/188 \\ 1 & \text{if } G_i = 188/\ast \\ 0 & \text{else} \end{cases} \) is the number of allele 188 in the genotype \( G_i \) of subject \( i \), and \( \ast \) represents the alleles other than allele 188.

The variance estimations are \( \hat{\sigma}_E^2 = 11.21^2 \) and \( \hat{\sigma}_S^2 = 12.82^2 \), and correlation range \( \hat{\rho} = 1.84 \). The regression results of model (25) are presented in Table 3. In addition, we presented in Table 3 the regression results of the following model

\[ y_{ij} = \mu_0 + t_{ij}^{\mu_{age}} + t_{ij}^{2\mu_{age}} + t_{ij}^{3\mu_{age}} + \text{sex}_i\beta_{sex} + x_{i,188}\alpha_0 + U_i(t_{ij}) + E_i + e_{ij}, \]  

(26)

which does not include the time-dependent variables \( x_{i,188}t_{ij} \), \( x_{i,188}t_{ij}^{2} \), and \( x_{i,188}t_{ij}^{3} \). From the results of model (26) in Table 3, we can see that allele 188 has no significant effect on the SBP since the high p-value 0.28. However, all three time-dependent variables \( x_{i,188}t_{ij} \), \( x_{i,188}t_{ij}^{2} \), and \( x_{i,188}t_{ij}^{3} \) have significant effects on SBP at a significance level 0.05 for model (25). Therefore, it is important to include the time-dependent genetic variables in the analysis. Ignoring time trends in genetic effects can make the model invalid.

To understand the temporal trend of SBP, Figure 4 provides the predicted SBP vs age for male and female, separately. The predicted SBP of male is 4.95 higher than that of female. Interestingly, the allele 188 at locus GATA25A04 almost has no effect on SBP for age interval of (25, 40). However, it does have positive effect for SBP when one’s age is older than 40. The allele 188 at locus GATA25A04 can be a risk genetic variant since it may lead to higher SBP for middle aged and old people.

In models (25) and (26), we do not include the random spline variables \( u_k \) and \( v_k \) as these of spline models (13) and (14). We did fit the models by using spline models (13) and (14), but none of them significantly improves the fitting. Hence, none of them is included in the final model. We also fitted non-parametric linear penalized spline models, but they failed to detect the genetic effect of allele 188 although the random term \( \sum_{k=1}^{K} u_k(t_{ij} - \kappa_k)_+ \) provides significant result.
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 4. 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 4).

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 \( \mu(t) = -34.2 + 81.7 \log(0.3(t + 21.7)) \) and an exponential function \( \mu(t) = 110 \exp(0.0002(t - 25)^2) \). The curves of the two functions were plotted in Figure 5. The logarithm function \( \mu(t) = -34.2 + 81.7 \log(0.3(t + 21.7)) \) was taken from Daw et al. (2003)\(^{36}\) and Wang et al. (2012)\(^{16}\) 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 \( \sigma^2_{E} \) was 25 and the error variance \( \sigma^2_{e} \) was 10.

We are mainly concerned about the performance of detecting the genetic effect in the model 
\[ y_i(t) = \mu(t) + sex_i \beta_{sex} + x_i \alpha_1 + U_i(t) + E_i + \epsilon_i. \]

In Table 4, 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 non-parametric linear penalized spline model, a correctly specified non-linear function, and misspecified parametric functions of \( \mu(t) \). Each empirical type I error rate in Table 4 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 \( \alpha_1 = 0 \) 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 \( \alpha = 0.05 \) were reported in Table 4 and represented the proportions of the test values calculated for the 2,000 samples, that exceeded the 95-th percentiles of the \( \chi^2_1 \)-distribution.
The empirical power was calculated similarly by assuming $\alpha_1 = 0.5, 1, 2, 3$ 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 non-parametric linear penalized spline model, the correctly specified non-linear 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 5. 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 $\alpha_1$ and then took difference with the true value of $\alpha_1$ as the bias. From the results in Table 4, the non-parametric linear penalized spline model and the correctly specified non-linear 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 non-parametric 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 incorporates temporal variations is important for understanding genetic 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.\textsuperscript{15,37} Moreover, important environmental factors which are asso-
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 paper, 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 di-allelic genetic markers such as single nucleotide polymorphisms (SNPs) and multi-allelic markers such as micro-satellites. 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 non-parametric 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. Another way is to use parametric models in the analysis.

The proposed approaches were applied to analyze GAW 13 and GAW 16 systolic blood pressure (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 non-parametric 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.

To evaluate the robustness of the non-parametric 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 non-parametric 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 specifying regression models and variance-covariance structure. This is different from the method proposed in de Andrade et al. (2002, 2003),\textsuperscript{12,13} 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), (6), (9), and (11); (ii) temporal variance-covariance functions given by equations (12). 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).\textsuperscript{12} 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.\textsuperscript{26,27} 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.\textsuperscript{38} In practice, trait values are observed at a finite set of times $t_1, \cdots, t_m$ 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.\textsuperscript{39,40} An approximation of co-variance matrices by orthogonal polynomials has been also proposed.\textsuperscript{38} Based on the theory of stochastic processes, the character process model was proposed by Pletcher and Geyer (1999).\textsuperscript{27} These methods were summarized and evaluated in Jafferzic and Pletcher (2000),\textsuperscript{26} and used in analyzing empirical data of Drosophila reproduction and mortality and growth in beef cattle. However, the methods do not use genetic
Functional mapping methods were developed to estimate the dynamic changes of QTL effects during a course of ontogenetic growth. 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), (6), (9), and (11) can be used to model the trait means, which take care of the association information. The temporal variance-covariance functions given by equation (12) can be used for one individual’s measurements. For family members, the temporal variance-covariance functions can be constructed in the same way as variance component models presented. 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 5 times. This provides an opportunity to compare the results by using the real data.

In this paper, 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 paper 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.

**Appendix**

Let $x_{Ai}, x_{Bij}, z_{Aij}$ and $z_{Bij}$ be the dummy variables defined by relations (10). Denote $X_A = (x_{A1}, \cdots, x_{A(a-1)})^T$, $X_B = (x_{B11}, \cdots, z_{B1(b-1)})^T$, and $X_{A\cup B} = (X_A^T, X_B^T)^T$. Let us denote the additive variance-covariance matrix of the dummy variables by $V_A = \text{Cov}(X_{A\cup B}, X_{A\cup B}) = E(X_{A\cup B}X_{A\cup B}^T) - E(X_{A\cup B})E(X_{A\cup B}^T)$. Similarly, let $Z_A = (z_{A1}, \cdots, z_{A(a-1)})^T$, $Z_B = (z_{B11}, \cdots, z_{B1(b-1)})^T$, and $Z_{A\cup B} = (Z_A^T, Z_B^T)^T$. Let us denote the dominance variance-covariance matrix of the indicator variables $z_{Aij}, z_{Bij}$ by $V_D = \text{Cov}(Z_{A\cup B}, Z_{A\cup B})$.

For $i = 1, 2, \cdots, a$, let us denote $D_{A,i} = P(Q_1A_i) - q_1P_{A_i}$, which are measures of LD between QTL $Q$ and marker $A$. Here $P(Q_1A_i)$ is the frequency of haplotype $Q_1A_i$. For $k = 1, 2, \cdots, b$, let us denote $D_{B,k} = P(Q_1B_k) - q_1P_{B_k}$, which are measures of LD between QTL $Q$ and marker $B$. Here $P(Q_1B_k)$ is the frequency of haplotype $Q_1B_k$. For $i = 1, 2, \cdots, a, k = 1, \cdots, b$, let us denote $D_{A,iB_k} = P(A_iB_k) - P_{A_i}P_{B_k}$, which are measures of LD between markers $A$ and $B$. Here $P(A_iB_k)$ is frequency of haplotype $A_iB_k$. 

John Wiley & Sons, Inc.
Such as Appendix V, Fan, Jung, and Jin (2006), we can show that the regression coefficients of models (11) and (9) are given by

\[
\begin{pmatrix}
\alpha_{A1}(t) \\
\vdots \\
\alpha_{A(a-1)}(t) \\
\alpha_{B1}(t) \\
\vdots \\
\alpha_{B(b-1)}(t)
\end{pmatrix} = (V_A/2)^{-1} \begin{pmatrix}
D_{A1}Q \\
\vdots \\
D_{A(a-1)Q} \\
D_{B1}Q \\
\vdots \\
D_{B(b-1)Q}
\end{pmatrix} \alpha_Q(t)
\]

\[
\begin{pmatrix}
\delta_{A12}(t) \\
\vdots \\
\delta_{A(a-1)2}(t) \\
\delta_{B12}(t) \\
\vdots \\
\delta_{B(b-1)2}(t)
\end{pmatrix} = V_D^{-1} \begin{pmatrix}
[P_{A1}D_{A1Q} - P_{A1}D_{A2Q}]^2 \\
\vdots \\
[P_{A(a-1)}D_{A(a-1)Q} - P_{A1}D_{A(a-1)Q}]^2 \\
[P_{B1}D_{B1Q} - P_{B1}D_{B2Q}]^2 \\
\vdots \\
[P_{B(b-1)}D_{B(b-1)Q} - P_{B1}D_{B(b-1)Q}]^2
\end{pmatrix} \delta_Q(t).
\]

The elements of matrices $V_A$ and $V_D$ are provided in Appendix V, Fan, Jung, and Jin (2006). Equations (27) show that the parameters of LD (i.e., $D_{A1Q}$ and $D_{B1Q}$) and gene effects (i.e., $\alpha_Q(t)$ and $\delta_Q(t)$) are contained in the regression coefficients. The gene substitution effect $\alpha_Q(t)$ is contained only in $\alpha_{Aj}, \alpha_{Bj}$; and the dominance effect $\delta_Q(t)$ is contained only in $\delta_{Ajl}, \delta_{Bjl}(t)$. Thus, $V_A$ is called additive variance-covariance matrix; and $V_D$ is called dominance variance-covariance matrix. The model (9) orthogonally decomposes genetic effect into summation of additive and dominance effects.

**Acknowledgement**

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.

**Computer Program**

The methods proposed in this paper 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.
References


Table 1: Parametric association results of blood systolic pressure and SNPs $M_1 = rs17249754$, $M_2 = rs11024074$, $M_3 = rs3184504$ for FHS data, GAW 16.

<table>
<thead>
<tr>
<th>Model (15)</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>99.92110</td>
<td>1.116201</td>
<td>89.52</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}$</td>
<td>0.33315</td>
<td>0.019624</td>
<td>16.98</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^2$</td>
<td>0.00521</td>
<td>0.000669</td>
<td>7.79</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^3$</td>
<td>-0.00013</td>
<td>0.000036</td>
<td>-3.43</td>
<td>0.0006</td>
<td></td>
</tr>
<tr>
<td>$\beta_{sex}$</td>
<td>-3.22158</td>
<td>0.409038</td>
<td>-7.88</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\beta_{bmi}$</td>
<td>616.1159</td>
<td>26.939680</td>
<td>22.87</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model (16)</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>100.2625</td>
<td>1.121348</td>
<td>89.41</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}$</td>
<td>0.33286</td>
<td>0.019618</td>
<td>16.97</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^2$</td>
<td>0.00519</td>
<td>0.000669</td>
<td>7.76</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^3$</td>
<td>-0.00012</td>
<td>0.000036</td>
<td>-3.41</td>
<td>0.0006</td>
<td></td>
</tr>
<tr>
<td>$\beta_{sex}$</td>
<td>-3.21477</td>
<td>0.408619</td>
<td>-7.87</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\beta_{bmi}$</td>
<td>617.6547</td>
<td>26.932825</td>
<td>22.94</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{10}$</td>
<td>-1.08444</td>
<td>0.373864</td>
<td>-2.90</td>
<td>0.0037</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model (17)</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>99.41874</td>
<td>1.132271</td>
<td>87.65</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}$</td>
<td>0.33312</td>
<td>0.019622</td>
<td>16.98</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^2$</td>
<td>0.00520</td>
<td>0.000669</td>
<td>7.77</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^3$</td>
<td>-0.00014</td>
<td>0.000036</td>
<td>-3.41</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>$\beta_{sex}$</td>
<td>-3.23002</td>
<td>0.408668</td>
<td>-7.90</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\beta_{bmi}$</td>
<td>616.9971</td>
<td>26.924911</td>
<td>22.92</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{30}$</td>
<td>-0.86881</td>
<td>0.281312</td>
<td>-3.09</td>
<td>0.002</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model (18)</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>99.3735</td>
<td>1.132271</td>
<td>87.65</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}$</td>
<td>0.33312</td>
<td>0.019622</td>
<td>16.98</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^2$</td>
<td>0.00520</td>
<td>0.000669</td>
<td>7.77</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^3$</td>
<td>-0.00014</td>
<td>0.000036</td>
<td>-3.41</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>$\beta_{sex}$</td>
<td>-3.23002</td>
<td>0.408668</td>
<td>-7.90</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\beta_{bmi}$</td>
<td>616.9971</td>
<td>26.924911</td>
<td>22.92</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{30}$</td>
<td>-0.86881</td>
<td>0.281312</td>
<td>-3.09</td>
<td>0.002</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model (19)</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>99.3735</td>
<td>1.132271</td>
<td>87.65</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}$</td>
<td>0.33312</td>
<td>0.019622</td>
<td>16.98</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^2$</td>
<td>0.00520</td>
<td>0.000669</td>
<td>7.77</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{age}^3$</td>
<td>-0.00014</td>
<td>0.000036</td>
<td>-3.41</td>
<td>0.0007</td>
<td></td>
</tr>
<tr>
<td>$\beta_{sex}$</td>
<td>-3.23002</td>
<td>0.408668</td>
<td>-7.90</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\beta_{bmi}$</td>
<td>616.3573</td>
<td>26.885682</td>
<td>22.92</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{10}$</td>
<td>-1.3894</td>
<td>0.373911</td>
<td>-3.05</td>
<td>0.0023</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{age}$</td>
<td>-0.04428</td>
<td>0.019624</td>
<td>-2.38</td>
<td>0.0196</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{30}$</td>
<td>-0.86859</td>
<td>0.280898</td>
<td>-3.09</td>
<td>0.002</td>
<td></td>
</tr>
</tbody>
</table>

The overall likelihood ratio test of model (19) vs. model (15) to test $H_0 : \alpha_{10} = \alpha_{1,age} = \alpha_{20} = \alpha_{30} = 0$ is 28.63, df = 4, p-value < 0.0001.
Table 2: Non-Parametric association results of blood systolic pressure and SNPs $M_1 = rs17249754, M_2 = rs11024074, M_3 = rs3184504$ for FHS data, GAW 16.

<table>
<thead>
<tr>
<th>Model</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model (20)</td>
<td>$\mu_0$</td>
<td>98.6284</td>
<td>6.867915</td>
<td>14.36</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\mu_{age}$</td>
<td>0.0605</td>
<td>0.209519</td>
<td>0.29</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>$\beta_{sex}$</td>
<td>-3.2366</td>
<td>0.408893</td>
<td>-7.92</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{bmi}$</td>
<td>614.1032</td>
<td>26.929997</td>
<td>22.80</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>Model (21)</td>
<td>$\mu_0$</td>
<td>99.0071</td>
<td>6.845307</td>
<td>14.46</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>based on</td>
<td>$\mu_{age}$</td>
<td>0.0629</td>
<td>0.208835</td>
<td>0.30</td>
<td>0.76</td>
</tr>
<tr>
<td>$M_1 = rs17249754$</td>
<td>$\beta_{sex}$</td>
<td>-3.2297</td>
<td>0.408378</td>
<td>-7.91</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{bmi}$</td>
<td>615.6503</td>
<td>26.914968</td>
<td>22.87</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{10}$</td>
<td>-1.0865</td>
<td>0.373734</td>
<td>-2.91</td>
<td>0.0037</td>
</tr>
<tr>
<td>Model (22)</td>
<td>$\mu_0$</td>
<td>99.4904</td>
<td>6.864736</td>
<td>14.49</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>based on</td>
<td>$\mu_{age}$</td>
<td>0.0606</td>
<td>0.209273</td>
<td>0.30</td>
<td>0.76</td>
</tr>
<tr>
<td>$M_2 = rs3184504$</td>
<td>$\beta_{sex}$</td>
<td>-3.2135</td>
<td>0.408378</td>
<td>-7.87</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{bmi}$</td>
<td>613.9346</td>
<td>26.904021</td>
<td>22.82</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{20}$</td>
<td>-0.8690</td>
<td>0.281214</td>
<td>-3.09</td>
<td>0.0020</td>
</tr>
<tr>
<td>Model (23)</td>
<td>$\mu_0$</td>
<td>98.1551</td>
<td>6.846793</td>
<td>14.36</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>based on</td>
<td>$\mu_{age}$</td>
<td>0.0624</td>
<td>0.208559</td>
<td>0.30</td>
<td>0.76</td>
</tr>
<tr>
<td>$M_3 = rs11024074$</td>
<td>$\beta_{sex}$</td>
<td>-3.2449</td>
<td>0.408523</td>
<td>-7.94</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{bmi}$</td>
<td>614.9878</td>
<td>26.915266</td>
<td>22.85</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{30}$</td>
<td>0.8034</td>
<td>0.309029</td>
<td>2.60</td>
<td>0.0109</td>
</tr>
<tr>
<td>Model (24)</td>
<td>$\mu_0$</td>
<td>99.3246</td>
<td>6.846793</td>
<td>14.51</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>based on</td>
<td>$\mu_{age}$</td>
<td>0.0750</td>
<td>0.208707</td>
<td>0.36</td>
<td>0.72</td>
</tr>
<tr>
<td>$M_1 = rs17249754$, $M_2 = rs3184504$, $M_3 = rs11024074$</td>
<td>$\beta_{sex}$</td>
<td>-3.1965</td>
<td>0.407736</td>
<td>-7.90</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\beta_{bmi}$</td>
<td>616.3468</td>
<td>26.875826</td>
<td>22.93</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{10}$</td>
<td>-1.1405</td>
<td>0.373783</td>
<td>-3.05</td>
<td>0.0023</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{20}$</td>
<td>-0.8688</td>
<td>0.280794</td>
<td>-3.09</td>
<td>0.0020</td>
</tr>
<tr>
<td></td>
<td>$\alpha_{30}$</td>
<td>0.7860</td>
<td>0.308430</td>
<td>2.55</td>
<td>0.0109</td>
</tr>
</tbody>
</table>

The overall likelihood ratio test of model (24) vs. model (20) to test $H_0: \alpha_{10} = \alpha_{1,age} = \alpha_{20} = \alpha_{30} = 0$ is 28.52, df = 4, p-value $< 0.0001$. 
Table 3: Association results of blood systolic pressure and marker GATA25A04 (D17S1299) for Framingham Heart Study Data, Genetic Analysis Workshop 13.

<table>
<thead>
<tr>
<th>Model (25)</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>130.12519</td>
<td>0.5845056</td>
<td>222.62</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{\text{age}}$</td>
<td>0.58538</td>
<td>0.0205328</td>
<td>28.51</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{\text{age}^2}$</td>
<td>0.00312</td>
<td>0.0007050</td>
<td>4.43</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{\text{age}^3}$</td>
<td>-0.00015</td>
<td>0.0000312</td>
<td>-4.96</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\beta_{\text{sex}}$</td>
<td>-4.95030</td>
<td>0.7489797</td>
<td>-6.61</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>2.56019</td>
<td>1.4748975</td>
<td>1.74</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{\text{age}}$</td>
<td>0.15237</td>
<td>0.0665988</td>
<td>2.29</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{\text{age}^2}$</td>
<td>-0.00418</td>
<td>0.0021415</td>
<td>-1.95</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>$\alpha_{\text{age}^3}$</td>
<td>-0.00023</td>
<td>0.0000944</td>
<td>-2.45</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model (26)</th>
<th>Coefficient</th>
<th>Estimates</th>
<th>Std Error</th>
<th>t-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_0$</td>
<td>130.21348</td>
<td>0.5829787</td>
<td>223.36</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{\text{age}}$</td>
<td>0.59855</td>
<td>0.0197430</td>
<td>30.32</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{\text{age}^2}$</td>
<td>0.00273</td>
<td>0.0006753</td>
<td>4.04</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\mu_{\text{age}^3}$</td>
<td>-0.00018</td>
<td>0.0000297</td>
<td>-5.97</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\beta_{\text{sex}}$</td>
<td>-4.94591</td>
<td>0.7488726</td>
<td>-6.60</td>
<td>&lt; 0.0001</td>
<td></td>
</tr>
<tr>
<td>$\alpha_0$</td>
<td>1.44832</td>
<td>1.3486853</td>
<td>1.07</td>
<td>0.28</td>
<td></td>
</tr>
</tbody>
</table>

The overall likelihood ratio test of model (25) vs. model (26) to test $H_0: \alpha_{\text{age}} = \alpha_{\text{age}^2} = \alpha_{\text{age}^3} = 0$ is 9.09, df = 3, p-value = 0.0281.
Table 4: Simulation results at 95% significance level of testing additive genetic effect $H_0: \alpha_1 = 0$ in model $y_i(t) = \mu(t) + sex_i \beta_{sex} + x_i \alpha_1 + U_i(t) + E_i + \epsilon_i$ by non-parametric linear penalized spline model, a correctly specified non-linear function, and misspecified parametric functions of $\mu(t)$. $^1$True $\mu(t) = -34.2 + 81.7\log(0.3(t + 21.7))$; $^2$True $\mu(t) = 110\exp(0.0002 \times (t - 25)^2)$; No. of Obs$^3$: the total number of observations; $^4\mu(t)$ is estimated by non-parametric linear penalized spline model; $^5\mu(t)$ is correctly specified; $^6\mu(t)$ is misspecified as $\mu(t) = \mu_0 + \mu_1 t$; $^7\mu(t)$ is misspecified as $\mu(t) = \mu_0 + \mu_1 t + \mu_2 t^2 + \mu_3 t^3$.

<table>
<thead>
<tr>
<th>$\mu(t)$: Mean Function</th>
<th>Type I Error or Power (Bias)</th>
<th>$\alpha_1$</th>
<th>No. of Obs$^3$</th>
<th>Empirical Results: Type I Error or Power (Bias $\alpha_1 - \hat{\alpha}_1$)</th>
<th>Non-parametric Model$^4$</th>
<th>Correctly Specified$^5$</th>
<th>Misspecified: Linear$^6$</th>
<th>Misspecified: Cubic$^7$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logarithm$^1$</td>
<td>Type I Error (Bias)</td>
<td>0</td>
<td>926</td>
<td>0.0535 (-0.011)</td>
<td>0.0555 (0.005)</td>
<td>0.0615 (0.0161)</td>
<td>0.0540 (0.009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>896</td>
<td>0.1620 (-0.068)</td>
<td>0.1595 (-0.054)</td>
<td>0.2430 (-0.254)</td>
<td>0.1605 (0.509)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>926</td>
<td>0.4010 (0.005)</td>
<td>0.4055 (0.003)</td>
<td>0.5695 (-0.291)</td>
<td>0.4015 (1.009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>889</td>
<td>0.9465 (-0.008)</td>
<td>0.9470 (-0.001)</td>
<td>0.9000 (0.137)</td>
<td>0.9475 (2.009)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>902</td>
<td>0.9980 (-0.007)</td>
<td>0.9980 (0.001)</td>
<td>0.9970 (0.124)</td>
<td>0.9980 (3.010)</td>
<td></td>
</tr>
<tr>
<td>Exponential$^2$</td>
<td>Type I Error (Bias)</td>
<td>0</td>
<td>916</td>
<td>0.0525 (-0.015)</td>
<td>0.0525 (-0.018)</td>
<td>0.009 (-0.242)</td>
<td>0.0525 (-0.032)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.5</td>
<td>908</td>
<td>0.1215 (0.003)</td>
<td>0.1240 (0.003)</td>
<td>0.0785 (-0.343)</td>
<td>0.1205 (0.467)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>900</td>
<td>0.3540 (-0.002)</td>
<td>0.3505 (0.001)</td>
<td>0.3120 (-0.522)</td>
<td>0.3430 (0.967)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>906</td>
<td>0.8545 (-0.005)</td>
<td>0.8535 (-0.008)</td>
<td>0.7175 (-0.358)</td>
<td>0.8430 (1.966)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>3</td>
<td>916</td>
<td>0.9990 (-0.015)</td>
<td>0.9990 (-0.018)</td>
<td>0.9910 (-0.242)</td>
<td>0.9990 (2.968)</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1: Time plot of systolic blood pressure and total plasma cholesterol level against age in years.
Figure 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 (19). 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 $M_i, i = 1, 2, 3$. 
Figure 3: 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 non-parametric model (24). 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 $M_i, i = 1, 2, 3$. 
Figure 4: Predicted systolic blood pressure against age in years for Male and Female by marker GATA25A04 (D17S1299) and sex, based on parametric model (25).
Figure 5: The curves of logarithm function $\mu(t) = -34.2 + 81.7 \log(0.3(t + 21.7))$ and an exponential function $\mu(t) = 110 \exp(0.0002 \times (t - 25)^2)$.