STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:3611–3625
Published online 3 September 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3709

Testing for familial aggregation of functional traits

Yixin Fang[1] and Yuanjia Wang[2,*,†]

[1]*Department of Mathematics and Statistics, Georgia State University, U.S.A.*
[2]*Department of Biostatistics, Mailman School of Public Health, Columbia University, U.S.A.*

SUMMARY

In a genetic epidemiology study of a trait, prior to collecting genotype data the foremost task is to test for familial aggregation and examine heritability. Recently, functional traits have drawn attentions from investigators. Here, to test for familial aggregation of a functional trait in the family studies, a test constructed based on the leading functional principal component of heritability, which is a summary measure of temporal genetic variation in a functional trait, is proposed. The p-value of the test can be approximated by a permutation procedure given the family structure. The asymptotic distribution of the test statistic is derived. Simulations are carried out to examine the size and the power of the test. The proposed methods are applied to the total cholesterol data in the Framingham Heart Study. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: permutation; Tracy–Widom distribution; genome-wide association; Framingham Heart Study

1. INTRODUCTION

Disease etiology for complex traits is often complicated, involving genetic and environmental contributions, interaction of the two, and other unobserved factors [1, 2]. Genetic architectures are more complicated for complex functional traits, such as cholesterol level, blood pressure, and HIV load, which change with time [3]. Although cross-sectional designs are widely implemented to map genes influencing quantitative traits, for functional traits, longitudinal designs can provide more information and offer novel insights on how genes affect developmental features of these traits [3].

*Correspondence to: Yuanjia Wang, Department of Biostatistics, Mailman School of Public Health, Columbia University, U.S.A.
†E-mail: yw2016@columbia.edu

Contract/grant sponsor: National Heart, Lung and Blood Institute's Framingham Heart Study; contract/grant number: N01-HC-25195
Contract/grant sponsor: NIH; contract/grant number: AG031113-01A2

Received 26 November 2008
Accepted 13 July 2009

For any trait of interest, prior to collecting genotype data for further genetic studies, the foremost task is to test for familial aggregation and examine heritability. If no strong evidence of familial aggregation is found, there is little need to conduct further genetic linkage or association studies. Despite a rich literature on familial aggregation in scalar traits (e.g. [4]), there has been conspicuously few research on the same topic for functional traits.

A naive method to examine the familial aggregation in a functional trait would be to perform the existing tests derived for scalar traits at each measurement time point. This strategy has several disadvantages. For example, it is subject to the problem of multiple comparisons, it ignores temporal trend in functional traits, and in practice subjects may not be measured at the same time points.

To overcome these problems, we consider tools developed in the field of functional data analysis. When the subjects in the sample are independent, functional principal components analysis (FPCA) can be used to explore sources of temporal variation of functional traits [5]. For family studies, however, without taking the familial structure into account, the standard FPCA describes the total temporal variation of a trait instead of the genetic temporal variation, in which we are primarily interested.

For multivariate traits, a principal components approach based on heritability (PCH) was developed in [6]. Instead of maximizing the total variation of the traits, this approach computes a linear combination of traits that maximize the heritability. Recently, an association test based on PCH for multivariate traits was proposed in [7].

In this manuscript, to test for familial aggregation and examine the temporal heritability of functional traits, a functional principal components approach based on heritability (FPCH) is developed. The test statistic is the leading functional principal component of heritability. The p-value of the test can be approximated by a permutation procedure given the family structure. The asymptotic distribution of the test statistic is also derived. Using asymptotic critical value to assess significance can reduce computational burden significantly, especially for genome-wide association studies. In addition, performance of the test can be improved by incorporating a mixed-effects model.

The manuscript is organized as follows. In Section 2, methods are proposed. In Section 3, simulations are carried out to examine size and power of the test. In Section 4, the proposed methods are applied to the total cholesterol data collected in the Framingham Heart Study. In Section 5, application to genome-wide association study and some possible extensions are discussed.

2. METHODS

2.1. Functional heritability and a direct method

Let $y_{ij}(t)$ denote the observed functional trait of subject j in family i at time t. Let s be the number of families in the sample, n_i be the number of subjects in family i, and n be the total number of subjects. In the absence of genotype information, the total variance of the trait is decomposed into polygenic genetic variance and residual variance. Heritability is then defined as the ratio of genetic variance and total variance. Here, for the sake of illustration, we assume that each family consists of only siblings. The genetic variance and residual variance can be estimated based on the between- and the within-family sum of squares [8]. Specifically, for sibship data the heritability at time t can be estimated by

$$h^2(t) = \sigma_G^2(t)/\sigma_T^2(t) \tag{1}$$

DOI: 10.1002/sim

where $\sigma_G^2(t)=2(\mathrm{MS}_b(t)-\mathrm{MS}_w(t))/n_0$ and $\sigma_T^2(t)=[(s-1)\mathrm{MS}_b(t)+(n-s)\mathrm{MS}_w(t)]/(n-1)$. Here, $\mathrm{MS}_b(t)$ is the between-family mean square, $\sum_i n_i(\bar{y}_{i\cdot}(t)-\bar{y}_{\cdot\cdot}(t))^2/(s-1)$, $\mathrm{MS}_w(t)$ is the within-family mean square, $\sum_{ij}(\bar{y}_{ij}(t)-\bar{y}_{i\cdot}(t))^2/(n-s)$, and n_0 equals to $(n-\sum_i n_i^2/n)/(s-1)$.

In practice, a functional trait is only observed at discrete time points and in most cases these time points are different among subjects. Let $\phi(\cdot)$ denote the K vector of B-spline basis, $(\phi_1(\cdot),\ldots,\phi_K(\cdot))^{\mathrm{T}}$, and assume that the trait can be expressed as

$$y_{ij}(t)=c_{ij}^{\mathrm{T}}\phi(t)+\varepsilon_{ij}(t) \tag{2}$$

where c_{ij} are basis coefficients and $\varepsilon_{ij}(\cdot)$ are stationary white-noise processes. Here, we use B-spline basis system because it is flexible and fast to compute even with very large number of basis functions [9].

The basis coefficients c_{ij} are unknown. A direct method to estimate c_{ij} for subject j in family i based on measurements $(y_{ij}(t_{ij,1}),y_{ij}(t_{ij,2}),\ldots,y_{ij}(t_{ij,T_{ij}}))$ is to use least squares or other smoothing techniques [5]. Using the estimated basis coefficients $\hat{c}_{ij}$, the between- and the within-family mean squares are, respectively,

$$\mathrm{MS}_b(t)=\frac{\phi^{\mathrm{T}}(t)B\phi(t)}{s-1}\quad\text{and}\quad\mathrm{MS}_w(t)=\frac{\phi^{\mathrm{T}}(t)W\phi(t)}{n-s}$$

where

$$B=\sum_i n_i(\bar{\hat{c}}_{i\cdot}-\bar{\hat{c}}_{\cdot\cdot})(\bar{\hat{c}}_{i\cdot}-\bar{\hat{c}}_{\cdot\cdot})^{\mathrm{T}}\quad\text{and}\quad W=\sum_{ij}(\hat{c}_{ij}-\bar{\hat{c}}_{i\cdot})(\hat{c}_{ij}-\bar{\hat{c}}_{i\cdot})^{\mathrm{T}} \tag{3}$$

Then the functional heritability (1) can be computed based on these expressions.

As pointed out in [5], the choice of the number of basis functions used depends on many considerations such as the number of sampling points and the efficiency of the basis functions in reproducing the behavior of the original function. Here, we use a direct method to develop tests for familial aggregation in Sections 2.2–2.4, and then discuss methods of regularizing smoothness in Section 2.5. Although the test based on the direct method in Section 2.3 is robust to model mispecification or model selection, it is inappropriate when the measurement points are sparse. To overcome this, we propose a method based on a mixed-effects model in Section 2.5.

2.2. FPCA based on heritability

The FPCA approach maximizes the total variation in a functional trait collected in independent subjects and therefore does not reflect familial information. As discussed in [6], taking into account the family structure can explore genetic variation instead of total variation in a trait. For a functional trait collected from family studies, we propose FPCH approach to incorporate familial information.

For any weight function $\beta(\cdot)$, the weighted trait of $y_{ij}(\cdot)$, or FPCH score, is defined as $f_{ij}=\int\beta(t)y_{ij}(t)\,\mathrm{d}t$. The FPCH approach attempts to find a weight function $\beta_1(\cdot)$ that maximizes heritability in f_{ij}, which is a measure of the overall heritability of $y_{ij}(\cdot)$ corresponding to $\beta(\cdot)$. It can be shown that the weight function $\beta_1(\cdot)$ also maximizes the ratio of between- and within-family

DOI: 10.1002/sim

variation of f_{ij},

$$\frac{\sum_{i=1}^{s} n_i(\bar{f}_{i\cdot}-\bar{f}_{\cdot\cdot})^2}{\sum_{i=1}^{s}\sum_{j=1}^{n_i}(f_{ij}-\bar{f}_{i\cdot})^2} \tag{4}$$

Similar to $y_{ij}(\cdot)$, $\beta(\cdot)$ is projected onto a K-dimensional B-spline basis system, that is, $\beta(t)=b^{\mathrm{T}}\phi(t)$. It follows that:

$$f_{ij}=\int b^{\mathrm{T}}\phi(t)\phi^{\mathrm{T}}(t)\hat{c}_{ij}\,\mathrm{d}t$$

where the coefficients b and $\hat{c}_{ij}$ are K vectors, and $\hat{c}_{ij}$ are discussed in the previous section. Then, the maximization problem (4) becomes solving

$$\max_{\|b\|=1}\frac{b^{\mathrm{T}}\Gamma B\Gamma b}{b^{\mathrm{T}}\Gamma W\Gamma b} \tag{5}$$

where $\Gamma=\int\phi(t)\phi^{\mathrm{T}}(t)\,\mathrm{d}t$, and B and W are defined in (3). Also note that (5) is equivalent to $\max_{\|b\|=1} b^{\mathrm{T}}Bb/b^{\mathrm{T}}Wb$. It is well known that the solution, say b_1, to this optimization problem is the largest eigenvector of $W^{-1}B$, and then the weight function is $\beta_1(\cdot)=b_1^{\mathrm{T}}\phi(\cdot)$.

2.3. Testing for familial aggregation of functional traits

To test for the null hypothesis that there is no familial aggregation in a functional trait, we develop a test statistic defined as the maxima in (5), which is equivalent to

$$T=\max_{\|b\|=1}\frac{b^{\mathrm{T}}Bb}{b^{\mathrm{T}}Wb} \tag{6}$$

The bigger value of T indicates the stronger evidence of familial aggregation. This test has larger power compared with the one taking average of observations within each subject, because T is based on the optimal combination of observations within each subject. As a special case, authors in [6] examined power of T compared with other methods for multivariate trait including averaging observations on a subject, and it is found that T has the largest power in the most of cases.

Under the null hypothesis of no familial aggregation, functional traits $y_{ij}(\cdot)$ are independent across all subjects. The p-value of the test statistic T can be obtained through a permutation procedure given the family structure. To describe the permutation procedure, re-denote the functional traits as $z_h(\cdot), h=1,\ldots,n$ and their basis coefficients as $\hat{c}_h$ such that $z_h(\cdot)=\hat{c}_h^{\mathrm{T}}\phi(\cdot)$. Since the functional traits are independent under the null hypothesis, we can randomly partition $z_h(\cdot), h=1,\ldots,n$, into s families with the observed family sizes. Denote the curve of the subject j in the family i in a random partition as $y_{ij}^*(\cdot)$ with the corresponding basis coefficient $\hat{c}_{ij}^*$. Then the test statistic calculated from this partition is

$$T^*=\max_{\|b\|=1}\frac{b^{\mathrm{T}}B^*b}{b^{\mathrm{T}}W^*b} \tag{7}$$

DOI: 10.1002/sim

where B^* and W^* are the between- and within-family sum of squares of the basis coefficients $\hat{c}^*_{ij}$, as defined in (3). Repeat the random partition D times and record the test statistic (7) in each partition as T^*_d. Then the one-sided p-value can be approximated by

$$\sum\{d : T^*_d > T\}/D \tag{8}$$

An attractiveness of the permutation test is its robustness to model mispecification of the trait, and to the choice of basis system and the number of bases. The type-I error of the test is always correct regardless of specified model. However, an appropriate model specification can improve power of the test. We will discuss this in Section 2.5.

The algorithm for computing permutation p-value is straightforward to implement. However, the computation may be intensive especially when the sample size is large. We derive the asymptotic distribution of the test statistic in the next section.

2.4. Asymptotic distribution of the test statistic

For any given b, the distribution of the ratio in (6) can be approximated by an F distribution with $(s-1, n-s)$ degrees of freedom. Because of the maximization, however, the distribution of T is more complicated than an F distribution. In recent years, there is a wealth of literature on the asymptotic distribution of the largest eigenvalue of random matrix; see for example [10]. Among these, the following theorem from [11] can be applied to obtain the asymptotic distribution of T.

Theorem 1

Suppose that there are two independent $p \times p$ random matrices W_1 and W_2 following Wishart distributions $W_p(k_1, I_p)$ and $W_p(k_2, I_p)$, respectively. Here, I_p is an $p \times p$ identity matrix. Recall that if p-vectors $z_1, \ldots, z_m$ are i.i.d. as $N_p(0, \Sigma)$, then $\sum_{i=1}^m z_i z_i^{\mathrm{T}} \sim W_p(m, \Sigma)$. Let $k = k_1 + k_2$. Let U_1 be the largest root of equation $\det[U(W_1 + W_2) - W_1]$. Then, with p, $k_1(p)$, and $k_2(p) \to \infty$,

$$P(U_1 \leqslant s) = P(\mu_+ + \sigma_+ \mathrm{TW}_1 \leqslant s) + o(1) \tag{9}$$

where

$$\frac{\gamma_p}{2} = \sin^{-1}\sqrt{\frac{p}{k}}, \quad \frac{\phi_p}{2} = \sin^{-1}\sqrt{\frac{k_1}{k}}, \quad \mu_+ = \cos^2\left(\frac{\pi}{2} - \frac{\phi_p + \gamma_p}{2}\right), \quad \sigma_+^3 = \frac{\sin^4(\phi_p + \gamma_p)}{4k^2 \sin\phi_p \sin\gamma_p}$$

and TW_1 follows the Tracy–Widom F_1 distribution.

See [12] for details about Tracy–Widom F_1 distribution. If p and k_1 are small, use corrections

$$\frac{\gamma_p}{2} = \sin^{-1}\sqrt{\frac{p-0.5}{k-1}}, \quad \frac{\phi_p}{2} = \sin^{-1}\sqrt{\frac{k_1-0.5}{k-1}}, \quad \mu_+ = \cos^2\left(\frac{\pi}{2} - \frac{\phi_p + \gamma_p}{2}\right)$$

and

$$\sigma_+^3 = \frac{\sin^4(\phi_p + \gamma_p)}{4(k-1)^2 \sin\phi_p \sin\gamma_p}$$

Corollary 1

Assume that the functional traits of all subjects are observed at the same time points $t_1, \ldots, t_q$, $q > K$, where K is the number of B-spline bases in the model (2). Also assume that in (2), the residuals

DOI: 10.1002/sim

$\varepsilon_{ij}(t)$ are stationary Gaussian processes. Then under the null hypothesis of no familial aggregation, the distribution of $T/(1+T)$ can be approximated by that of U_1 in Theorem 1 with $p=K$, $k_1=s-1$, and $k_2=n-s$.

Proof of Corollary 1

Let Φ be the matrix $(\phi(t_1),\ldots,\phi(t_q))$ and let σ^2 be the constant variance of $\varepsilon_{ij}(t)$. Let $\Sigma=(\Phi^{\mathrm{T}}\Phi)^{-1}\sigma^2$. Following the arguments in the analysis of variance, it is easy to show that B and W are independent, and under the null hypothesis that of no familial aggregation, $B\sim W_K(s-1,\Sigma)$ and $W\sim W_K(n-s,\Sigma)$ [5]. Then by the transformation $\tilde{c}_{ij}=\Sigma^{-1/2}c_{ij}$, the corollary is proved. □

The condition of all subjects being measured at the same time points in the corollary may not be satisfied. However, if the measurements are not largely different, Corollary 1 can still be applied to obtain approximated critical values for test statistic T. The similar results were obtained in [13] for multivariate traits.

By Corollary 1, we can obtain the approximated critical values for the test statistic T. For example, with 5 per cent significance level, the corresponding percentile of Tracy–Widom F_1 is 0.98 [11]. The approximated critical value of $T/(1+T)$ is $u_*=\mu_+ +0.98\sigma_+$, where μ_+ and σ_+ are defined in (9) and depend on s,n, and K. Then the approximated critical value of T at 5 per cent level is $t_*=1/(1-u_*)-1$.

2.5. Improving power of the proposed test

Although the test proposed in Section 2.3 based on the direct method has correct type-I error no mater what model is specified, more sophisticated methods will improve the power of the test. Beyond the direct method, many other methods are proposed in the field of functional data analysis, such as mixed-effects approach (e.g. [14, 15]), reduced-rank method [16], and FPCA through conditional expectation [17]. Incorporating these methods to the test (6) can improve the power of the test. In this subsection, we use mixed-effects model approach as an example.

Again, under the null hypothesis subjects are independent, so that we can re-denote the traits as $z_h(t), h=1,\ldots,n$. The mixed-effects model has the form

$$z_h(t)=a^{\mathrm{T}}\phi(t)+c_h^{\mathrm{T}}\phi(t)+\varepsilon_h(t),\quad h=1,\ldots,n \tag{10}$$

where $\phi(t)$ is the K-dimensional B-splines, a is the fixed effect, c_h is the subject-specific random effect with covariance matrix Σ_c, and $\varepsilon_h(t)$ is the random error with mean zero and variance σ^2. Let Z_h be the vector consisting of the T_h observed values on the hth subject, let Φ_h be the corresponding T_h by K spline matrix evaluated at the observed time points, and ε_h be the corresponding random error vector with covariance matrix $\sigma^2 I$. The mixed-effects model then becomes

$$Z_h=\Phi_h a+\Phi_h c_h+\varepsilon_h,\quad h=1,\ldots,n \tag{11}$$

A general approach to fitting mixed-effects models of this form uses the EM algorithm to estimate a and Σ_c [18]. Given the estimates, predictions are obtained for basis coefficients c_h using the best linear unbiased prediction (BLUP, [19]). The BLUP for c_h in (11) is

$$\hat{c}_h=(\widehat{\Sigma}_c^{-1}/\hat{\sigma}^2+\Phi_h\Phi_h^{\mathrm{T}})^{-1}\Phi_h(Z_h-\Phi_h\hat{a}) \tag{12}$$

DOI: 10.1002/sim

For a random partition of $z_h(\cdot)$ into s families with the observed family sizes, denote the curve of the subject j in the family i as $y^*_{ij}(\cdot)$ with the corresponding basis coefficient $\hat{c}^*_{ij}$ obtained from (12) and calculate the test statistic (7). Repeating this process many times to get an approximated p-value.

3. SIMULATIONS

Here, we present a simulation study to investigate properties of the proposed methods. The simulation models are partially based on the observed total cholesterol data in the Framingham Heart Study. In each model there is a single disease susceptibility locus with two alleles and the effect of the disease allele is dominant. The functional traits are simulated from

$$y_{ij}(t)=\alpha(t)+g(t)X_{ij}+\varepsilon_{ij}(t)$$

where X_{ij} takes value $\frac{1}{2}$ for carriers of disease allele and value $-\frac{1}{2}$ for non-carriers, $\alpha(t)$ is the population mean trait value, $g(t)$ is the genetic effect, and $\varepsilon_{ij}(t)$ are the residual effects that are stationary white noises with variance σ^2. The age-specific heritability of traits in this model is [20]

$$\frac{2\operatorname{cov}(X_{ij},X_{ij'})g^2(t)}{\operatorname{var}(X_{ij})g^2(t)+\sigma^2} \tag{13}$$

In each of the simulation studies, 100 families each with three children are generated. For each subject, a fully informative marker perfectly linked to the underlying locus is simulated, and 20 observations equally spaced between age 31 and 69 are simulated. The grand mean $\alpha(t)$ is based on the observed cholesterol levels in the Framingham study. The genetic effect $g(t)$ is based on a logarithm function. The logarithm function was also used to simulate the genetic effects in the Genetic Analysis Workshop 13, where the simulations were intended to mimic the Framingham data [21]. The residual effects follow a stationary Gaussian distribution with variance 25.

The first set of simulations is designed to examine type-I error of the permutation test, and to compare the asymptotic critical values with the permutation critical values. In this setting, the genetic effect $g(t)$ is set to zero. To compute permutation p-value, the simulated subjects are randomly partitioned into 100 families with observed family sizes. The random partition is repeated 500 times. The asymptotic critical value is computed by Corollary 1. We also record the type-I error and the critical value from the mixed-effects model-based method. For this method, since the asymptotic distribution may not hold, we only consider the permutation procedure.

Three other simulation settings with varying genetic effects are designed to examine power of the test (7). In setting 1, the genetic effect function, $g(t)$, was $2[1+0.2\log(0.4(t-27))]$, and the maximum heritability over time calculated from (13) is 11 per cent. In setting 2, the genetic effect is $2.2[1+0.25\log(0.4(t-27))]$, and the maximum heritability is 15 per cent. In setting 3, the genetic effect is $2.5[1+0.25\log(0.4(t-27))]$, and the maximum heritability is 19 per cent, see Figure 1 for the three heritability curves computed from these specifications of genetic effect. We generate 500 repetitions for each simulation setting. The number of B-spline basis functions is seven.

Table I summarizes the results for examining type-I errors of the proposed tests. For the direct method, the type-I error of the permutation test is close to the pre-specified size at all levels. The type-I error of the test based on the asymptotic critical value is slightly lower than the pre-specified

DOI: 10.1002/sim

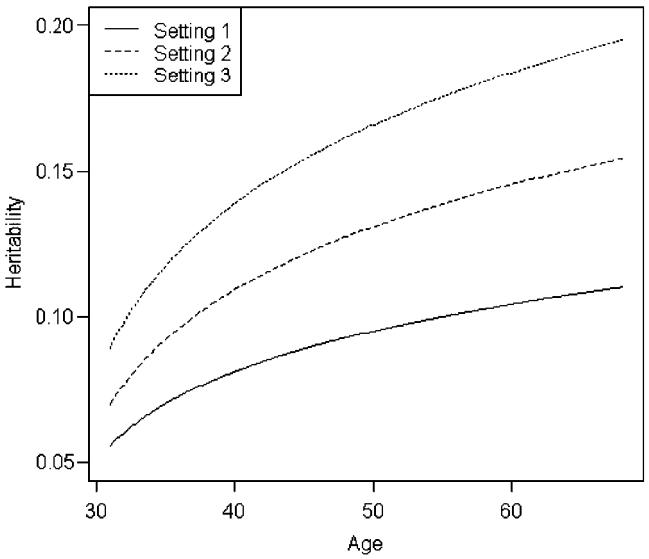

Figure 1. Heritability of traits in the three simulation settings examining power.

Table I. Type-I error and critical values of the tests.

	Type-I error			Critical value*		
α level	Method 1†	Method 2‡	Method 3§	Method 1	Method 2	Method 3
0.01	0.014	0.01	0.006	1.12	1.22	0.77
0.05	0.052	0.03	0.044	1.02	1.12	0.71
0.1	0.09	0.06	0.107	0.98	1.07	0.68

*Averaged over 500 repetitions.
†Method 1: Direct method-based permutation test.
‡Method 2: Direct method-based asymptotic test.
§Method 3: Mixed effects model-based permutation test.

size. Correspondingly, the asymptotic critical values are slightly higher than the empirical critical values of the permutation distribution. This suggests that using the asymptotic critical value in the direct method is conservative. For the mixed-effect model-based method, the permutation test has correct size at all levels.

Table II summarizes the results for examining power of the proposed tests. The tests have good power even when the maximum heritability over time is as small as 11 per cent (setting 1). The test using asymptotic critical value is conservative, and therefore slightly less powerful. The simulations show that the power of both the permutation test and the asymptotic test increases with increasing heritability, and the difference in power between the two decreases with increasing heritability. The mixed-effects model-based method increases power of the test in all settings. For example, the direct method in setting 1 has 87 per cent power for the permutation procedure and 99 per cent power for the mixed-effects model-based approach.

DOI: 10.1002/sim

Table II. Power of the tests.

Setting	Heritability*	Method 1† (per cent)	Method 2‡ (per cent)	Method 3§ (per cent)
1	0.11	87	71	99
2	0.15	98	93	99
3	0.19	99	98	100

*Maximum of the functional heritability as in (3).
†Method 1: Direct method-based permutation test.
‡Method 2: Direct method-based asymptotic test.
§Method 3: Mixed effects model-based permutation test.

Table III. Mean and standard error of the observed FPCH test statistic (6) under the null and the alternative.

	Direct method-based permutation		Mixed-effects model-based permutation	
Setting	Mean(T) (SE)†	Mean(T^0) (SE)‡	Mean(T) (SE)†	Mean(T^0) (SE)‡
Null	0.845 (0.105)	0.844 (0.099)	0.556 (0.091)	0.556 (0.082)
1	1.191 (0.163)	0.819 (0.097)	1.088 (0.159)	0.520 (0.084)
2	1.320 (0.165)	0.803 (0.097)	1.236 (0.156)	0.514 (0.085)
3	1.411 (0.169)	0.788 (0.097)	1.344 (0.165)	0.511 (0.086)

†Mean and standard error of T under the alternative.
‡Mean and standard error of T under the null.

Table III summarizes the mean and the standard error of the permutation test statistics under the null and the alternative. It can be seen that the observed mean test statistic under the alternative increases when the heritability increases. The mean and the standard error of the permutated test statistic under the null hypothesis are similar across all settings.

4. DATA ANALYSIS

Originating in 1948, the Framingham Heart Study [22] is an ongoing prospective study of risk factors for cardiovascular disease (CVD). There were 5209 subjects recruited (2336 men and 2873 women) from the town of Framingham, Massachusetts. Its objective is to identify common risk factors or characteristics that contribute to CVD by following its development over a long period of time in a large group of subjects who had not yet developed overt symptoms of CVD or suffered a heart attack or stroke when they entered the study. The clinical data from Framingham Heart Study includes systolic blood pressure, total cholesterol, fasting HDL cholesterol, triglycerides, blood glucose, and so on. These longitudinal observations are collected from two cohorts. The original Framingham cohort (Cohort 1) was first examined in 1948 and has been examined every 2 years thereafter. Cohort 2, composed primarily of offspring of the original cohort and the spouses of these offspring (5124 subjects, 2483 men, and 2641 women), was examined first in 1971 and has been examined approximately every 4 years [23].

DOI: 10.1002/sim

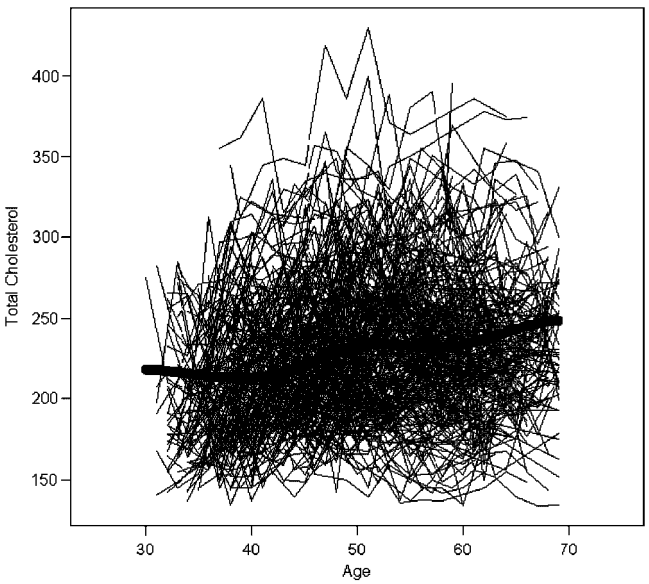

Figure 2. Raw data of cholesterol versus age. Thick solid line: Mean total cholesterol level.

4.1. Functional heritability

We apply the proposed methods to the total cholesterol levels collected in the Framingham Heart Study. We divide large pedigrees into nuclear families and select subjects with more than 10 observations from age 31 to age 69. There are 114 nuclear families and each family has an average of 2.3 offsprings. Only offsprings are included in the analysis. There are 258 subjects and each subject has an average of 12.8 observations. The total cholesterol levels on these subjects are presented in Figure 2.

To provide a description of the polygenic heritability as in (3), we consider two approaches: (1) a discrete heritability by dividing age into 5-year intervals and averaging observations in each interval to estimate a single heritability for that age range and (2) a functional heritability using seven B-spline basis functions with evenly spaced knots.

The long-term average heritability for total cholesterol estimated by Framingham Study investigators was 0.46 (http://www.nhlbi.nih.gov/about/framingham/policies/pageeleven.htm). Figure 3 presents the comparison of the estimated discrete heritability and the estimated functional heritability. The discrete heritability ranges from 0.51 to 0.78, while the functional heritability ranges from 0.41 to 0.77. As expected, the functional heritability serves as a smoothed version of the discrete heritability. The functional heritability provides much more information than the long-term average heritability. It shows the temporal features of the heritability curve over time: the heritability increases from 0.6 at age 30 to 0.77 at age 50, reaches its peak around age 50, and then enters a period of declination after age 50 and reaches the minimum value of 0.41 at age 65. This analysis suggests that for cross-sectional genetic studies of total cholesterol levels, collecting subjects around age 50 would be most beneficial because these subjects exhibit greatest heritability. To account for possible gender difference, we also analyze males and females separately. We find

DOI: 10.1002/sim

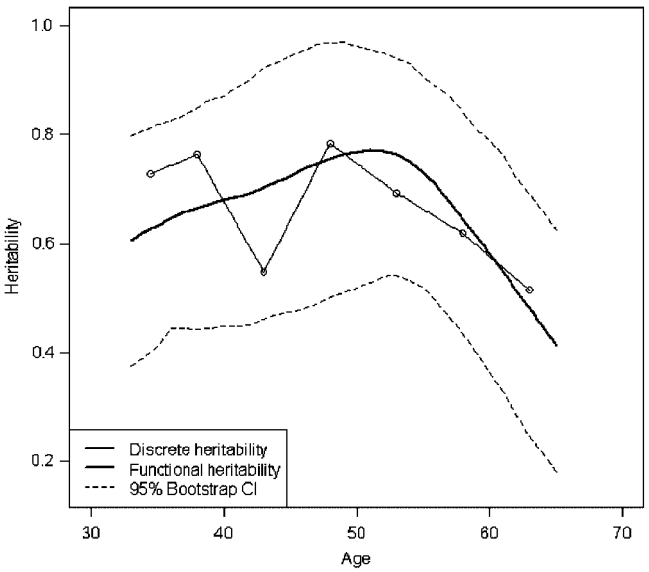

Figure 3. Discrete (solid line) and functional heritability (heavy line) and 95 per cent bootstrap CI for functional heritability (dashed line).

their heritability curves to be similar (results not shown here); therefore, we use the combined sample in the analysis.

4.2. Testing for familial aggregation of total cholesterol

We conduct the test proposed in Section 2.3 for the cholesterol data. The observed value of the test statistic T is 1.7. We then randomly permute the data given the family structure 500 times to assess the p-value. The permutation p-value is zero. The asymptotic critical value for these data is 1.65 at 5 per cent level, which is less than the observed test statistic. The mixed-effects model-based approach gives a higher value of the test statistic ($T=1.89$), and the p-value approximated by the permutation procedure is also zero. Therefore, all methods provided strong evidence that there is familial aggregation of total cholesterol.

4.3. Smoothed FPCA

To explore sources of variation in the data, we carry out the standard FPCA [5]. Since the FPCA requires observations to be independent, we randomly select one person from each family in the Framingham study. We project each subject's observations onto a common basis system considered in the previous sections. To provide smoothed principal component, its roughness is penalized. The objective function of the penalized FPCA is [5]

$$\frac{\operatorname{var}\left(\int \beta(s)\, y_i(s)\, \mathrm{d}s\right)}{\|\beta\|^2+\lambda \int \beta''(s)\, \mathrm{d}s} \tag{14}$$

where λ is the tuning parameter to be determined. The tuning parameter to regularize the smoothness of the leading principal component is selected based on generalized cross-validation (GCV). For

DOI: 10.1002/sim

the Framingham data, the GCV suggests the tuning parameter to be 150. The leading principal component reveals a major cholesterol variation that decreases over time up to age 52, and then slightly increases after that age. The total cholesterol level fluctuates the least around age 50.

The FPCA approach maximizes the total variation in phenotype $y_{ij}(t)$. The pattern of the total variation revealed by the principal component may not be directly related to genetic variation as will be shown in this example. We compare this result to the one based on the FPCH approach in the next section.

4.4. Smoothed functional principal components of heritability

When the test (6) is significant, it may be of interest to describe the temporal variation of the genetic effect by showing the leading functional principal component of heritability. To smooth the weight function $\beta_1(\cdot)$, we penalize its roughness. Similar to the penalized FPCA, the leading FPCH can be smoothed by (see the Appendix)

$$\arg\max_{\|b\|=1} \frac{b^{\mathrm{T}}\Gamma B\Gamma b}{b^{\mathrm{T}}\Gamma W\Gamma b+\lambda b^{\mathrm{T}}Qb} \tag{15}$$

where Q is the roughness penalty matrix of the basis functions, $\int \phi_j''(s)\phi_k''(s)\,\mathrm{d}s$, and λ is the tuning parameter to be determined. The algorithm to solve the above optimization problem and a generalized cross-validation formula to choose λ are discussed in Appendix.

The result of GCV suggests the smoothing parameter to be 1000. The major temporal variation in heritability increases with age, culminates at around age 50 and starts to decrease. This variation echoes the functional heritability estimated in Figure 3, which also increases up to age 50 and then decreases. This trend could correspond to a major gene effect or a synthesis of several gene effects. Note that the pattern revealed by FPCH is completely different from the pattern of FPCA. This example illustrates the advantage of the FPCH approach; the pattern reported by FPCH explores the temporal variation of the genetic effect instead of the temporal trend of the total trait variation.

5. DISCUSSION

The main focus of this work is to develop a test for familial aggregation of a functional trait of interest. The p-value can be approximated by a permutation procedure given family structure or by asymptotic distribution under the null hypothesis. We discuss a mixed-effects model-based approach to improve the power. We also propose to use smoothed functional principal component of heritability to summarize the temporal trend of major genetic variation of the functional trait when the testing result is significant. We choose the smoothing parameter by a GCV method.

If we only conduct a single test, the advantage of the asymptotic approximation over permutation procedure may be moderate. But, when we conduct many tests simultaneously as in a genome-wide association study, the computational advantage of the above approximation is much more beneficial. For example, in a population-based genome-wide association study of multivariate traits, the role of family in the test (6) is replaced by any single SNP along the genome [7, 24]. Testing for genetic effect at each SNP can be done by computing a linear combination of all traits that maximize heritability attributable to the SNP [7]. The p-value of the test statistic can be computed by a similar permutation procedure as discussed in Section 2.3. In [7], a sample splitting procedure was suggested to assess significance. Although the computational burden of the proposed permutation

DOI: 10.1002/sim

procedure is lighter than the methods in [7], in a typical genome-wide association study, the number of SNPs is on the scale of hundreds of thousands and performing such a number of permutation tests are still daunting. Since the test statistic is in a similar form as (6), the asymptotic distribution derived here can be used to assess significance, and the asymptotic approximation is far more computational efficient.

In some cases, it may be of interest to test whether heritability changes with time. We can test this hypothesis by examining the derivative of heritability with respect to time. That is,

$$H_0 : \frac{\mathrm{d}}{\mathrm{d}t} h^2(t) = 0$$

where $h^2(t)$ is defined in (1). The test statistic is $(\mathrm{d}/\mathrm{d}t)\hat{h}^2(t)$. The p-value can be assessed by permuting observations within each subject.

For the sake of illustration, here we describe methods for nuclear-family data where the heritability can be estimated based on the between- and within-family sum of squares. These methods can be extended to arbitrary pedigree using kinship coefficients [8]. For general pedigrees, instead of decomposing the total variation of the traits into between- and within-family sum of squares, covariance structure of the trait will be constructed based on the kinship coefficients. The FPCH is implemented by maximizing the genetic component variation relative to the residual component estimated from this covariance structure. In [25], random regression and splines are used to analyze the longitudinal animal genetics data for extended pedigrees.

The proposed direct method project each subject's observations onto a common basis system. Since the direct method uses only a subjects' own measurements to estimate its basis coefficients, we develop the mixed-effects model-based approach to provide an empirical Bayes-type estimate to pool information from all subjects to estimate individual basis parameters (see also [15]). Our future work is to imbed kinship coefficients into the framework of mixed-effects models to account for general pedigree.

Finally, we should point out that the proposed methods do not account for covariates. One challenge is to remove the effects of important covariates. An approach would be fitting an appropriate regression including covariates and using the residuals as inputs to the principal components of heritability analysis.

ACKNOWLEDGEMENTS

The data was obtained from the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine. This work was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195). Fang (Subcontract PI) and Wang (PI)'s research is supported by NIH AG031113-01A2.

APPENDIX

In this Appendix, we provide an algorithm to solve (15) and a generalized cross-validation formula to determine the tuning parameter. The algorithm is:

1. Expand observed data y_{ij} with respect to the basis functions ϕ to get coefficients c_{ij} and matrix B and W in (3).
2. Perform a Choleski factorization to compute L, where $L^{\mathrm{T}}L = \Gamma W \Gamma + \lambda Q$, and define $S = L^{-1}$.

DOI: 10.1002/sim

3. Solve $Ld_i = \Gamma \bar{c}_{i.}$ to get $d_i = S\Gamma \bar{c}_{i.}$.
4. Carry out a standard PCA on d_i and record eigenvectors u.
5. Solve $L^{\mathrm{T}} b = u$ to find eigenvector b and renormalize the resulting vectors to get $b^{\mathrm{T}}\Gamma b = 1$.
6. Transform back to get principal components weight functions $\beta(s) = b^{\mathrm{T}}\phi(s)$.

To choose the tuning parameter λ, we present a generalized cross-validation formula. In a linear regression analysis with response Y and design matrix X, the GCV formula is

$$\mathrm{GCV}(\lambda) = \frac{\frac{1}{n}\mathrm{SSE}}{\left[\frac{1}{n}\mathrm{trace}(I - P_\lambda)\right]^2}$$

where $\mathrm{SSE} = \sum \|Y_i - \hat{Y}_i\|^2$, $P_\lambda = X(X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}$, and $\hat{Y}_i$ are the fitted values. In the FPCA, the residual sum-of-squares SSE takes the form

$$\mathrm{PCASSE} = \sum_i \int \left[y_i(t) - \sum_k f_i^{(k)} \beta_k(t) \right]^2 \mathrm{d}t$$

where $f_i^{(k)}$ are the principal component scores corresponding to the kth principal component. Note that in the step 4 of the above algorithm, we carry out FPCA on the d_i. Therefore, the GCV formula for the penalized FPCH can be developed base on d_i. Let $u_i(s)$ denote $d_i^{\mathrm{T}}\phi(s)$ and let $\hat{u}_i(s) = \sum_k \bar{f}_{i.}^{(k)} \beta_k(s) = \sum_k \bar{f}_{i.}^{(k)} b_k^{\mathrm{T}} \phi(s)$. The residual sum-of-squares is then

$$\begin{aligned}\mathrm{PCHSSE} &= \sum_i \int [u_i(t) - \hat{u}_i(s)]^2 \mathrm{d}t \\ &= \sum_i (d_i - \beta^{\mathrm{T}} \bar{f}_{i.})^{\mathrm{T}} \Gamma (d_i - \beta^{\mathrm{T}} \bar{f}_{i.})\end{aligned}$$

where β is the basis coefficients matrix $(b_1, \ldots, b_k)^{\mathrm{T}}$, $\bar{f}_{i.}$ is the vector $(\bar{f}_{i.}^{(1)}, \ldots, \bar{f}_{i.}^{(m)})^{\mathrm{T}}$, and m is the number of PCH. The projection matrix mapping u_i to $\hat{u}_i$ is, in this case, $P_\lambda = F(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}$. Here, F is the matrix of principal components scores $(\bar{f}_{1.}, \ldots, \bar{f}_{s.})^{\mathrm{T}}$ computed with a particular λ. The GCV formula for the functional principal components of heritability is therefore

$$\mathrm{PCHGCV}(\lambda) = \frac{\frac{1}{s}\mathrm{PCHSSE}}{\left[\frac{1}{s}\mathrm{trace}(I - P_\lambda)\right]^2}$$

Then find the λ that minimizes the above criterion.

REFERENCES

1. Botto LD, Khoury MJ. Facing commentary: facing the challenge of gene–environment interaction: the two-by-four table and beyond. *American Journal of Epidemiology* 2003; **153**:1016–1020.
2. Smith GD, Ebrahim S, Lewis S, Lansell AL, Palmer LJ, Burton PR. Genetic epidemiology and public health: hope, hype and future prospects. *Lancet* 2005; **366**:1484–1498.

DOI: 10.1002/sim

3. Wu RL, Lin M. Functional mapping—how to map and study the genetics architecture of dynamic complex traits. *Nature Reviews Genetics* 2006; **7**:229–237.
4. Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of Genetic Epidemiology*. Oxford University Press: U.S.A., 1993.
5. Ramsay JO, Silverman BW. *Functional Data Analysis*. Springer: New York, 2005.
6. Ott J, Rabinowitz D. A principal-components approach based on heritability for combining phenotype information. *Human Heredity* 1999; **49**:106–111.
7. Klei L, Luca D, Devlin B, Roeder K. Pleiotropy and principal components of heritability combine to increase power for association analysis. *Genetic Epidemiology* 2008; **32**:9–19.
8. Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates: Massachusetts, 1998.
9. DeBoor C. *A Practical Guide to Splines* (Revised edn). Springer: New York, 2001.
10. Bai Z. Methodologies in spectral analysis of large dimensional random matrices: a review. *Statistica Sinica* 1999; **9**:611–677.
11. Johnstone I. Multivariate analysis and Jacobi ensembles: largest eigenvalue, Tracy Widom limits and rates of convergence. *The Annals of Statistics* 2008; **36**:2638–2716.
12. Johnstone IM. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics* 2001; **29**:295–327.
13. Fang Y. An application of a theorem of Johnstone and Forrester to testing for familial aggregation. *2008 International Conference on BioMedical Engineering and Informatics*, Sanya, Hainan, China, vol. 2, 2008; 773–777.
14. Brumback B, Rice J. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of American Statistical Association* 1998; **93**:961–976.
15. Rice JA, Wu CO. Nonparametric mixed effects models for unequally sampled noisy curves. *Biometrics* 2001; **57**:253–259.
16. James MG, Hastie TJ, Sugar AC. Principal component models for sparse functional data. *Biometrika* 2000; **87**:587–602.
17. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *Journal of American Statistical Association* 2005; **100**:577–590.
18. Laird N, Ware J. Random-effects models for longitudinal data. *Biometrics* 1982; **38**:963–974.
19. Robinson G. That BLUP is a good thing: the estimation of random effects. *Statistical Science* 1991; **6**:15–32.
20. Wang Y, Fang Y, Jin M. A ridge penalized principal-components approach based on heritability for high-dimensional data. *Human Heredity* 2007; **64**:182–191.
21. Daw EW, Morrison J, Zhou X, Thomas D. Genetic analysis workshop 13: simulated longitudinal data on families for a system of oligogenic traits. *BMC Genetics* 2003; **4**(suppl. 1):S3.
22. Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: the Framingham study. *American Journal of Public Health* 1951; **41**:279–286.
23. Cupples LA, Yang Q, Demissie S, Copenhafer D, Levy D. Description of the Framingham heart study data for genetic analysis workshop 13. *BMC Genetics* 2003; **4**(suppl. 1):S2.
24. Falconer DS, Mackay TFC. *Introduction to Quantitative Genetics* (4th edn). Longmans Green: Harlow, Essex, U.K., 1996.
25. DeGroot BF, Keown JF, Van Vleck LD, Kachman SD. Estimates of genetic parameters for Holstein cows for test-day yield traits with a random regression cubic spline model. *Genetics and Molecular Research* 2007; **6**(2):434–444.

DOI: 10.1002/sim