Published in final edited form as:

Genet Epidemiol. 2014 November ; 38(7): 622–637. doi:10.1002/gepi.21840.

Generalized Functional Linear Models for Gene-based Case-Control Association Studies

Ruzong Fan^{#1,*}, Yifan Wang^{#1}, James L. Mills², Tonia C. Carter³, Iryna Lobach⁴, Alexander F. Wilson⁵, Joan E. Bailey-Wilson⁵, Daniel E. Weeks⁶, and Momiao Xiong⁷

¹Biostatistics and Bioinformatics Branch, Division of Intramural Population Health Research Eunice Kennedy Shriver National Institute of Child Health and Human Development National Institutes of Health, Rockville, MD 20852

²Epidemiology Branch, Division of Intramural Population Health Research Eunice Kennedy Shriver National Institute of Child Health and Human Development National Institutes of Health, Rockville, MD 20852

³Center for Human Genetics, Marshfield Clinic, Marshfield, WI 54449

⁴Department of Neurology, School of Medicine University of California, San Francisco, CA 94185

⁵Statistical Genetics Section, Computational and Statistical Genomics Branch National Human Genome Research Institute National Institutes of Health, Bethesda, MD 20892

⁶Departments of Human Genetics and Biostatistics, Graduate School of Public Health University of Pittsburgh, Pittsburgh, PA 15261

⁷Human Genetics Center, University of Texas - Houston P.O. Box 20334, Houston, Texas 77225

These authors contributed equally to this work.

Abstract

By using functional data analysis techniques, we developed generalized functional linear models for testing association between a dichotomous trait and multiple genetic variants in a genetic region while adjusting for covariates. Both fixed and mixed effect models are developed and compared. Extensive simulations show that Rao's efficient score tests of the fixed effect models are very conservative since they generate lower type I errors than nominal levels, and global tests of the mixed effect models generate accurate type I errors. Furthermore, we found that the Rao's efficient score test statistics of the fixed effect models have higher power than the sequence kernel association test (SKAT) and its optimal unified version (SKAT-O) in most cases when the causal variants are both rare and common. When the causal variants are all rare (i.e., minor allele frequencies less than 0.03), the Rao's efficient score test statistics and the global tests have similar or slightly lower power than SKAT and SKAT-O. In practice, it is not known whether rare variants or common variants in a gene are disease-related. All we can assume is that a combination of rare and common variants influences disease susceptibility. Thus, the improved performance of

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

Computer Program. The methods proposed in this paper are implemented by using procedure of functional data analysis (fda) in the statistical package R. The R codes for data analysis and simulations are available from the web <http://www.nichd.nih.gov/about/org/diphrr/bbb/software/Pages/default.aspx>

our models when the causal variants are both rare and common shows that the proposed models can be very useful in dissecting complex traits. We compare the performance of our methods with SKAT and SKAT-O on real neural tube defects and Hirschsprung's disease data sets. The Rao's efficient score test statistics and the global tests are more sensitive than SKAT and SKAT-O in the real data analysis. Our methods can be used in either gene-disease genome-wide/exome-wide association studies or candidate gene analyses.

Keywords

rare variants; common variants; case-control association studies; complex diseases; logistic regression; functional data analysis; generalized functional linear models

Introduction

Modern sequencing technologies can assay millions of genetic variants, both common and rare [Bansal et al., 2010; Clarke et al., 2009; Mardis, 2008; Metzker, 2010; Rusk and Kiermer, 2008; Shendure and Ji, 2008]. Common variants have typically been analyzed one by one, which may cause a loss of power due to multiple comparison problems. However, the power of a single variant analysis drops precipitously as the variant becomes rarer, so it is important to develop statistical models that use multiple genetic variants in a unified analysis, instead of one marker at a time. This marker-set analysis approach helps improve power and reduces the multiple comparison problems [Wessel and Schork, 2006]. In addition, the marker-set approach can be thought of as a gene-based approach if the markers selected lie within a gene region [He et al., 2013].

Several methods are available to perform gene-based association studies, such as burden tests and kernel-based test methods. The burden tests are mainly designed to analyze rare variants, which have minor allele frequencies (MAFs) of less than 0.01 — 0.05. Burden tests collapse rare variants in genetic regions into a single variable which is used to test for association with a complex trait and to reduce high dimensionality of genetic data [Gorlov et al., 2008; Han and Pan, 2010; Li and Leal, 2008; Madsen and Browning, 2009; Morgenthaler and Thilly, 2007; Morris and Zeggini, 2010; Neale et al., 2011; Price et al., 2010; Schork et al., 2009]. The kernel-based test methods build a kernel-matrix to aggregate the association between genetic variants and phenotype, and can analyze either rare variants or common variants [Ionita-Laza et al., 2013; Kwee et al., 2008; Lee et al., 2012a, b; Lin and Schaid, 2009; Mukhopadhyay et al., 2010; Wu et al., 2010]. The kernel-based approaches deal with high dimensionality by assuming genetic effects as random variables which have means of zero and constant variances, and association is tested by testing a null hypothesis of zero variances by sequence kernel association test (SKAT). The SKAT and its optimal unified test (SKAT-O) have higher power than the burden tests [Lee et al., 2012a; Wu et al., 2011].

An alternate way to parsimoniously analyze multiple genetic variants in a selected genomic region uses functional data analysis (FDA) techniques, which have been applied successfully in many areas such as engineering, finance, and image analysis [Cardot and Sarda, 2005; Ferraty and Romain, 2010; Goldsmith et al., 2011; Horváth and Kokoszka,

2012; James, 2002; Li and Hsing, 2010; Müller and Stadtmüller, 2005]. In genetics, we have previously developed functional linear models to perform multi-locus association analysis of quantitative traits [Fan et al., 2013; Luo et al., 2011, 2012, and 2013]. Functional linear models assume that the individual realizations (e.g., the genetic variants) in the region are the result of a stochastic process, which in our domain reflects mutation, selection, and the effects of linkage disequilibrium (LD) between the variants. In Fan et al. [2013], F -distributed test statistics of fixed effect functional linear models were built to test for association between multiple genetic variants and a quantitative trait. Simulation studies showed that the F -test statistics not only have accurate type I error rates, but also generally have much higher power than SKAT and SKAT-O. The functional linear models are very flexible since they can analyze rare variants or common variants or a combination of the two [Fan et al., 2013]. The superior performance of the functional linear models is most likely due to their optimal utilization of both similarity between individuals and LD information, while SKAT and SKAT-O model the similarities but do not sufficiently model higher order LD information. Furthermore, not only do our models take LD between the variants into account, but they naturally take into account the physical spacing of the genetic variants.

The key idea of the functional linear models is to treat the discrete genetic data of each individual as a particular realization of an underlying stochastic process, which is summarized as a genetic variant function (GVF) [Luo et al., 2012; Fan et al., 2013]. The genetic markers in the human genome are actually a collection of random variables, and so high-resolution genetic marker data can be treated as dense discrete realizations of an underlying stochastic processes [Ross 1996]. When we treat genetic data functionally, functional data analysis techniques markedly reduce the dimensionality. For example, consider a region with 50 single nucleotide polymorphisms (SNPs). If we use the 50 SNPs in a regression as predictors, we would have 50 regression coefficients which are not easy to deal with due to the large number of genetic variants, collinearity, multiple testing, and variable selection issues. In the functional data analysis framework, each individual's marker data are naturally treated as one function, and so there is no limit on how many genetic variants the FDA method can handle simultaneously. Actually, the accuracy of the GVF estimate increases as the number of genetic variants increases. In addition, collinearity and variable selection are not a problem when using the FDA method. By using B-spline or Fourier or linear spline basis functions, the GVF can be approximated by a fixed number of basis functions. Similarly, the genetic effect of the GVF can be approximated by a fixed number of basis functions. Association is then detected by testing if the genetic effect of the GVF is equal to zero. In this way, the issues associated with high dimensionality are resolved and our models are useful in practice.

In this article, we develop generalized functional linear models (GFLM) to test for association between a dichotomous trait and multiple genetic variants in a genetic region while adjusting for covariates. Fixed and mixed effect models and related test statistics are proposed and compared. We used simulated sequence data to evaluate type I error rates and power of our proposed statistics. Extensive empirical type I error calculation is performed to make sure that false positive rates are properly controlled. Power is primarily compared with that of SKAT and SKAT-O, simply because SKAT and SKAT-O tend to outperform the

burden tests (Wu et al. 2011; Lee et al. 2012a). Our methods were applied to test for association with neural tube defects [Pangilinan et al., 2012] and with Hirschsprung's disease [Carter et al., 2012].

Materials and Methods

Generalized Functional Linear Models

Consider a case-control study with n individuals who are sequenced in a genomic region that has m genetic variants. We assume that the m variants have known ordered physical locations $0 < t_1 < \dots < t_m$. To make the notation simpler, we normalize the region $[t_1, t_m]$ to be $[0, 1]$. For the i th individual, let y_i denote a dichotomous disease trait of interest, $G_i = (g_i(t_1), \dots, g_i(t_m))'$ the genotype of the m variants, and $Z_i = (z_{i1}, \dots, z_{ic})'$ the c covariates. For the disease trait, $y_i = 1$ indicates that the i th individual is an affected case and $y_i = 0$ indicates that the i th individual is a normal control. For the genotypes, we assume that $g_i(t_j)$ ($= 0, 1, 2$) is the number of minor alleles of the i th individual at the j th variant located at the location t_j .

For the i th individual, we denote his/her genetic variant function as $X_i(t)$, $t \in [0, 1]$. By using the genetic variant vector G_i , we may estimate, as shown below, the related genetic variant function $X_i(t)$ [Fan et al., 2013; Luo et al., 2011, 2012, and 2013]. To relate the genetic variant function to the phenotype while adjusting for covariates, we use the following generalized functional logistic regression model

$$\text{logit}(\pi_i) = \alpha_0 + Z_i' \alpha + \int_0^1 X_i(t) \beta(t) dt, \quad (1)$$

where $\pi_i = P(y_i = 1)$ is the disease probability, α_0 is the regression intercept, α is a $c \times 1$ vector of regression coefficients of covariates, and $\beta(t)$ is the genetic effect of genetic variant functions $X_i(t)$ at the location t .

Estimation of Genetic Variant Functions

To estimate the genetic variant functions $X_i(t)$ from the genotypes G_i , we use two methods: (1) an ordinary linear square smoother; (2) a functional principal component analysis (FPCA) technique [Fan et al., 2013; Goldsmith et al., 2011]. The ordinary linear square smoother method assumes that the genetic variant functions are smooth, while no smoothness is assumed by the FPCA technique. In the following, we briefly describe the two approaches.

Using the discrete realizations $G_i = (g_i(t_1), \dots, g_i(t_m))'$, we may estimate the genetic variant function $X_i(t)$ using an ordinary linear square smoother [Ramsay and Silverman, 1996, Chapter 4]. Specifically, let $\phi_k(t)$, $k = 1, \dots, K$, be a series of K basis functions. Let Φ denote the m by K matrix containing the values $\phi_k(t_j)$, where $j \in 1, \dots, m$. Then, $X_i(t)$ is estimated by

$$\hat{X}_i(t) = (g_i(t_1), \dots, g_i(t_m)) \Phi \left[\Phi' \Phi \right]^{-1} \phi(t), \quad (2)$$

where $\varphi(t) = (\varphi_1(t), \dots, \varphi_K(t))'$ is a column vector of basis functions. In this article, we consider two types of basis functions: (1) the B-spline basis: $B_k(t)$, $k = 1, \dots, K$, and (2) the Fourier basis: $\varphi_0(t) = 1$, $\varphi_{2r-1}(t) = \sin(2\pi rt)$, and $\varphi_{2r}(t) = \cos(2\pi rt)$, $r = 1, \dots, (K-1)/2$ (for Fourier basis, K is taken as a positive odd integer) [Ramsay and Silverman, 1996; Ramsay et al., 2009; de Boor, 2001; Ferraty and Romain, 2010; Horváth and Kokoszka, 2012].

To introduce the main idea of FPCA, let $\Sigma_X(s, t)$ be the covariance function of the genetic variant functions. The covariance function $\Sigma_X(s, t)$ can be estimated from the observed genotype data $G_i = (g_i(t_1), \dots, g_i(t_m))'$, $i = 1, 2, \dots, n$ [Ramsay and Silverman, 1996; Horváth and Kokoszka, 2012]. Denote the spectral decomposition of $\Sigma_X(s, t)$ by

$\sum_{k=1}^{\infty} \lambda_k \phi_k(s) \phi_k(t)$, where $\lambda_1 \geq \lambda_2 \geq \dots$ are the non-decreasing eigenvalues and $\phi_k(t)$, $k = 1, 2, \dots$, are the corresponding orthonormal eigenfunctions. An approximation for $X_i(t)$, based on a truncated Karhunen-Loève expansion, is $\hat{X}_{i(t)} = (c_{i1}, \dots, c_{iK})\varphi(t)$, where K is the truncation lag, $\varphi(t) = (\varphi_1(t), \dots, \varphi_K(t))'$, and $c_{ik} = \int_0^1 X_i(t) \phi_k(t) dt$, which can be estimated using the observed genotype data.

Revised Generalized Functional Linear Model

The initial functional logistic regression model (1) is a theoretical model. We need to revise it to be useful for practical data analysis. Basically, we transform the model (1) to be an ordinary logistic regression model

$$\text{logit}(\pi_i) = \alpha_0 + Z_i' \alpha + W_i' \beta. \quad (3)$$

In the following, we propose three approaches to define the term $W_i' \beta$: (1) smoothing both the genetic variant functions $X_i(t)$ and the genetic effect function $\beta(t)$; (2) functional principal component analysis technique; (3) smoothing the genetic effect function $\beta(t)$ only, i.e., beta-smooth only.

Smoothing both the genetic variant functions and the genetic effect function

—As we did with the genetic variant functions $X_i(t)$, we may also expand the genetic effect function $\beta(t)$ using a series of basis functions $\theta_k(t)$, $k = 1, \dots, K_\beta$ as $\beta(t) = (\theta_1(t), \dots, \theta_{K_\beta}(t))(\beta_1, \dots, \beta_{K_\beta})'$, where $\beta = (\beta_1, \dots, \beta_{K_\beta})'$ is a vector of coefficients. Let us denote $\theta(t) = (\theta_1(t), \dots, \theta_{K_\beta}(t))'$. Replacing $X_i(t)$ in the functional logistic regression model (1) by the estimation $\hat{X}_{i(t)}$ in (2) and $\beta(t)$ by the above expansion, we have a revised logistic regression model

$$\text{logit}(\pi_i) = \alpha_0 + Z_i' \alpha + \left[(g_i(t_1), \dots, g_i(t_m)) \Phi \left[\Phi' \Phi \right]^{-1} \int_0^1 \phi(t) \theta'(t) dt \right] \beta.$$

Thus, $W_i' = (g_i(t_1), \dots, g_i(t_m)) \Phi \left[\Phi' \Phi \right]^{-1} \int_0^1 \phi(t) \theta'(t) dt$. The statistical package R and Matlab both contain available codes for doing these calculations [Ramsay et al., 2009].

FPCA approach—In the case of FPCA, the genetic effect $\beta(t)$ is expanded by a linear spline basis as follows. Let $\beta(t) = \beta_1 + \beta_2 t + \sum_{k=3}^{K_\beta} \beta_k (t - \kappa_k)_+$ where κ_k are knots in the interval $[0, 1]$; and $(t - \kappa_k)_+$ indicates if t is larger than κ_k , i.e., $(t - \kappa_k)_+ = 0$ if $t \leq \kappa_k$ and 1 if $t > \kappa_k$. Therefore, the genetic variant function can be expanded as $\beta(t) = \theta'(t)\beta$, where $\beta = (\beta_1, \dots, \beta_{K_\beta})'$, $\theta(t) = (1, t, (t - \kappa_3)_+, \dots, (t - \kappa_{K_\beta})_+)$. Replace $X_i(t)$ in the functional logistic regression model (1) by the truncated Karhunen-Loève expansion $\hat{X}_{i(t)} = (c_{i1}, \dots, c_{iK})\phi(t)$ and $\beta(t)$ by the linear spline basis. Then, the revised logistic regression model is $\text{logit}(\pi_i) = \alpha_0 + Z_i' \alpha + W_i' \beta$, where $W_i = (c_{i1}, \dots, c_{iK}) \int_0^1 \phi(t) \theta'(t) dt$.

The beta-smooth only approach—In the beta-smooth only approach, we use the original genotype data $G_i = (g_i(t_1), \dots, g_i(t_m))'$ directly, making no assumptions about smoothness of the genetic variant functions $X_i(t)$. To achieve this, model (1) is revised as

$$\text{logit}(\pi_i) = \alpha_0 + Z_i' \alpha + \sum_{j=1}^m g_i(t_j) \beta(t_j). \quad (4)$$

Here, the integration term $\int_0^1 X_i(t) \beta(t) dt$ in model (1) is replaced by a summation term $\sum_{j=1}^m g_i(t_j) \beta(t_j)$. The genetic effect function $\beta(t)$ is assumed to be smooth, so one may estimate it by B-spline or Fourier or linear spline basis functions. Replacing $\beta(t)$ by the expansion $\beta(t) = (\theta_1(t), \dots, \theta_{K_\beta}(t))(\beta_1, \dots, \beta_{K_\beta})'$, model (4) can be revised as

$$\text{logit}(\pi_i) = \alpha_0 + Z_i' \alpha + \left[\sum_{j=1}^m g_i(t_j) (\theta_1(t_j), \dots, \theta_{K_\beta}(t_j)) \right] (\beta_1, \dots, \beta_{K_\beta})'. \quad (5)$$

Thus, $W_i' = \sum_{j=1}^m g_i(t_j) (\theta_1(t_j), \dots, \theta_{K_\beta}(t_j))$. The revised model (5) is straightforward and less technical than the two models proposed above.

Smoothness of the genetic variant functions—Note that the first approach above assumes that the genetic variant functions are smooth, while no smoothness is assumed for genetic variant functions in the FPCA and the beta-smooth only approaches.

Rao's efficient Score Test Statistics of Fixed effect Models

We first consider the fixed effect model (3), i.e., we treat the regression coefficients β as known constant parameters. Therefore, the revised regression model (3) is a logistic regression which models the genetic effect of genetic variant functions while adjusting for covariates. To test the association between the m genetic variants and the dichotomous disease trait, the null hypothesis is $H_0: \beta = (\beta_1, \dots, \beta_{K_\beta})' = 0$. By using the standard statistical approach, we may test the null $H_0: \beta = 0$ by a χ^2 -distributed Rao's efficient score statistic with degrees of freedom K_β . In our simulation studies, likelihood ratio test (LRT) statistics were also evaluated.

Global Tests of Mixed effect Model

In the second analysis, we treat the regression coefficients β as a random vector. We assume that each β_k follows a distribution with a mean of zero and a variance τ , and $\beta_1, \dots, \beta_K$ are identically and independently distributed. Therefore, model (3) is treated as a generalized linear mixed effect model with α_0 and α as fixed effect components, and β as a random component. Denote $W = (W_1, \dots, W_n)'$ the model matrix of the regression coefficients β . To test the association between the m genetic variants and the dichotomous disease trait, one may test the null hypothesis $H_0: \tau = 0$. Let $Y = (y_1, \dots, y_n)'$ be a vector of trait values. A variance-component functional kernel score test as follows can be used to test the association

$$S(\hat{\pi}) = (Y - \hat{\pi})' \mathcal{K} (Y - \hat{\pi}), \quad (6)$$

where $\hat{\pi} = (\hat{\pi}_1, \dots, \hat{\pi}_n)'$ is a vector of the estimated disease probabilities of $\pi_i = P(y_i = 1)$ under the null H_0 , and $\mathcal{K} = WW'$ is a kernel matrix. That is, $\text{logit } \hat{\pi}_i = \hat{\alpha}_0 + Z_i' \hat{\alpha}$ and so $\hat{\pi} = \hat{\alpha}_0 + Z \hat{\alpha}$, where $Z = (Z_1, \dots, Z_n)'$ is the covariate matrix, and $\hat{\alpha}_0$ and $\hat{\alpha}$ are estimated under the null model by using the covariate matrix Z without the genetic variant functions. To facilitate the computation of a p -value for significance, one can approximate the distribution of $S(\hat{\pi})$ by a ratio of quadratic forms in normal variables and build test statistics to test the null [Goeman et al., 2011]. In our simulation studies, the global test (gt) proposed in Goeman et al. [2004, 2006, and 2011] was found to perform well.

Real Data Analysis

We tested for the association between neural tube defects, or Hirschsprung's disease, and the SNPs in each related gene by our Rao's efficient score test statistics of fixed effect models using both B-spline basis and Fourier basis, and both the Rao's test and the global test of FPCA model. To make comparisons with the existing methods in the literature, we applied SKAT in R package to test for the association by both SKAT and SKAT-O. In the following, we provide a brief description of the data used in the neural tube defects and Hirschsprung's disease examples [Pangilinan et al., 2012; Carter et al., 2012].

Neural Tube Defects Data—We analyzed data from a study of genetic factors and neural tube defects, which includes a primary dataset of 301 cases and 341 controls [Pangilinan et al., 2012]. The sample set was genotyped for 1339 tagging SNPs in 82 genes related to folate/vitamin B12 metabolism, transport of folate or vitamin B12, or transcriptional or developmental processes implicated in mouse models of neural tube defects. Pangilinan et al. [2012] tested for association between neural tube defects and each single SNP.

Hirschsprung's Disease Data—The second data set we analyzed is from a population based case-control study of Hirschsprung's disease [Carter et al., 2012]. The dataset includes 301 cases and 1,215 controls. A total of 40 tagging SNPs are available in six autosomal gene regions. Carter et al. [2012] tested for association between Hirschsprung's disease and each single SNP, and found that RET proto-oncogene variants were strongly associated Hirschsprung's disease (p -values between 10^{-3} and 10^{-31}).

Simulation Studies

Extensive simulations were performed to evaluate our proposed methods with sample sizes ranging from 200 to 2,000, using the sequence data of Wu et al. [2011] and Lee et al. [2012a, b]. The sequence data are of European ancestry from 10,000 chromosomes covering 1 Mb regions, simulated using the calibrated coalescent model programmed in COSI. Specifically, the sequence data were generated using COSI's calibrated best-fit models, and the generated European haplotypes mimic CEPH Utah individuals with ancestry from northern and western Europe in terms of site frequency spectrum and LD pattern [Figure 4 in Schaffner et al., 2005; The International HapMap Consortium 2007]. Using the same strategy as the simulations in Wu et al. [2011] and Lee et al. [2012a, b], genetic regions of 3 kb length were randomly selected for type I error and power calculations. Two scenarios were considered: (1) the causal variants are all rare; (2) the causal variants are both rare and common. Since the same strategy was used to generate phenotype data as Wu et al. [2011] and Lee et al. [2012a, b] and the same simulated sequence data were used, the comparison with their SKAT and SKAT-O results is valid.

Type I Error Simulations—To evaluate the type I error rates of the proposed statistics, we generated phenotype data sets using the model

$$\text{logit}(\pi_i) = \alpha_0 + 0.5Z_{1i} + 0.5Z_{2i}, \quad (7)$$

where Z_{1i} is a dichotomous covariate taking values 0 and 1 with a probability of 0.5, Z_{2i} is a continuous covariate from a standard normal distribution $N(0, 1)$, and $\alpha_0 = -4.91$ was chosen to create a disease prevalence of around 0.01 under the null hypothesis (e.g., $\pi_i = 0.007$ when $Z_{1i} = Z_{2i} = 0$, and $\pi_i = 0.012$ when $Z_{1i} = Z_{2i} = 1$). Genotypes were selected from variants in 3 kb subregions randomly selected from the 1 Mb region. Notice that the trait values are not related to the genotypes, and so the null hypothesis holds. The sample sizes of the datasets were 200, 250, 300, 350, 500, 1,000, 1,500, and 2,000 when the causal variants are both rare and common, and 500, 1,000, 1,500, and 2,000 when the causal variants are only rare. For each sample size, 10^6 phenotype-genotype datasets were generated to fit the proposed models and to calculate the test statistics and related p -values. In each dataset, 50% cases and 50% controls were simulated under a case-control sampling scheme. Then, an empirical type I error rate was calculated as the proportion of 10^6 p -values which were smaller than a given α level.

Empirical Power Simulations—To evaluate the power of our proposed statistics, we simulated data sets under the alternative hypothesis by randomly selecting 3 kb subregions to obtain causal variants. For each sample dataset, a subset of m causal variants located in the selected 3 kb subregion was then randomly selected, yielding genotypes $(g(t_1), \dots, g(t_m))$. Then, we generated the dichotomous disease traits by

$$\text{logit}(\pi_i) = \alpha_0 + 0.5Z_{1i} + 0.5Z_{2i} + \beta_1 g_i(t_1) + \dots + \beta_m g_i(t_m), \quad (8)$$

where Z_1 , Z_2 , and $\alpha_0 = -4.91$ were the same as in the type I error model (7), $(g_i(t_1), \dots, g_i(t_m))'$ were genotypes of the i th individual at the causal variants, and the β s are additive effects for the causal variants defined as follows. We used $j\beta_j = cj \log_{10}(MAF_j)/2$, where

MAF_j was the MAF of the j th variant. As was done by Wu et al. [2011] and Lee et al. [2012a], three different settings were considered: 10%, 20%, and 50% of variants in the 3 kb subregion are chosen as causal variants. When 10%, 20%, and 50% of the variants were causal, $c = \log(7)$, $\log(5)$, and $\log(2.5)$, respectively. As in Lee et al. [2012a], three different sample sizes were considered: $n = 200$, 500, and 1,000. For each setting, 1,000 datasets were simulated to calculate the empirical power as the proportion of p -values which are smaller than a given α level. In each dataset, 50% cases and 50% controls are simulated. For each dataset, the causal variants are the same for all the individuals in the dataset, but we allow the causal variants to be different from dataset to dataset.

Parameters of Functional Data Analysis

In the data analysis and simulations, we used two functions in the *fda* R package as follows to create basis:

```
basis= create.bspline.basis (norder=order, nbasis=bbasis)
basis= create.fourier.basis (c (0, 1) , nbasis=fbasis)
```

where we set $order = 4$, $bbasis = 10$, $fbasis = 11$. Specifically, the order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. In the data analysis and simulations of FPCA, the number of knots of linear spline basis was taken as $K_\beta = 10$ and the truncation lag $K = 20$.

To make sure that the results are valid and stable, we tried a wide range of parameters: (1) $8 \leq K = K_\beta \leq 21$ for B-spline basis and Fourier basis functions, (2) $8 \leq K_\beta \leq 22$ and $15 \leq K \leq 30$ for FPCA. The results are very similar to those when $order = 4$, $bbasis = 10$, $fbasis = 11$ for B-spline basis and Fourier basis functions, and $K_\beta = 10$ and the truncation lag $K = 20$ for FPCA, which are presented in the **Supplementary Materials I and II**.

Results

Application to The Real SNP Data

In our real data analyses, our tests and SKAT as well as SKAT-O were applied. We use a p -value cutoff of 0.05 to determine which genes are significantly associated with neural tube defects or Hirschsprung's disease.

Neural Tube Defects Results

Table 1 presents the results of the SNP data analysis of nine genes related to neural tube defects. The nine genes, from Pangilinan et al. [2012], were found to show the strongest association signals by single SNP analysis. The SKAT and SKAT-O results confirm the association of two genes, MFTC and CDKN2A, with neural tube defects. The results of the proposed tests confirm the association of six genes, MFTC, CDKN2A, PEMT, CUBN, MTHFD1, and T (Brachyury), with neural tube defects, including the two significant genes by SKAT and SKAT-O. Three genes ADA, GART, and DNMT3A, were not found to be significantly associated with the neural tube defects by either the proposed tests or SKAT or SKAT-O.

In the six genes which are significant at least once by our tests, four are detected by the Rao's efficient score tests of the fixed effect models [MFTC, PEMT, CUBN, and MTHFD1], and of these four genes, SKAT and SKAT-O only detected MFTC, failing to detect the other three. The global test of the FPCA mixed effect model detected four genes [CDKN2A, PEMT, CUBN, and T (Brachyury)], while SKAT detected none and SKAT-O only detected one (CDKN2A). Hence, both the Rao's efficient score tests of the fixed effect models and the global test of the FPCA mixed effect models are more sensitive than SKAT and SKAT-O for neural tube defects data.

Hirschsprung's Disease Results

Table 2 presents the results of data analysis of six genes related to Hirschsprung's disease. The results of our tests confirm the association of the RET proto-oncogene with Hirschsprung's disease with a small p -value range: 2.74×10^{-24} — 8.62×10^{-23} . The SKAT and SKAT-O also show significant association between the RET proto-oncogene and the Hirschsprung's disease, but the p -values are larger than those of our tests (the p -value of SKAT is 2.11×10^{-14} and the p -value of SKAT-O is 1.96×10^{-18} , respectively). In addition, the PROKR1 gene is shown to be significantly associated with the Hirschsprung's disease by our Rao's efficient score test of the fixed effect model with a p -value of 4.23×10^{-2} based on B-spline basis and FPCA approach, while neither SKAT nor SKAT-O shows any association signal.

General Observation of Real SNP Data Analysis

The power simulation results below show that our Rao's efficient score tests perform better than SKAT and SKAT-O when the causal variants are not restricted to rare variants. It is noteworthy that the SNPs of both neural tube defects sample and Hirschsprung's disease sample are common variants, and the good performance of our Rao's efficient score tests for neural tube defects and Hirschsprung's disease data is consistent with the results of the simulation studies.

In the Table 1, the Rao's efficient score test results of beta-smooth only are identical to those of smoothing both the genetic variant functions $X_i(t)$ and the genetic effect function $\beta(t)$ except for gene MFTC. Similarly, except for one gene ASCL1 using B-spline basis in the Table 2, the Rao's efficient score test results of beta-smooth only are identical to those of smoothing both the genetic variant functions and the genetic effect function. Therefore, whether the genetic variant functions are smoothed or not does not have much impact on the results. We observed this for quantitative traits in Fan et al. [2013].

Empirical Type I Error Rates

The empirical type I error rates are reported in Tables 3 and 4 at four nominal significance levels $\alpha = 0.05, 0.01, 0.001$, and 0.0001 . In the Tables, the results of “Basis of both GVF and $\beta(t)$ ” were based on smoothing both the GVF and the genetic effect function $\beta(t)$ by either B-spline or Fourier basis functions, the results of “FPCA Approach” were based on FPCA approach, and the results of “Basis of beta-Smooth Only” were based on smoothing the genetic effect function $\beta(t)$ only approach. In the two cases of “Basis of both GVF and $\beta(t)$ ” and “Basis of beta-Smooth Only”, the results of Rao's efficient score tests of the fixed

effect models were reported for both B-spline basis and Fourier basis functions. In the “FPCA Approach” case, the results of the fixed effect Rao's efficient score test and the mixed effect global test were reported. Therefore, five Rao's efficient score test results and one global test result were reported for each combination of one sample size and one nominal level $\alpha = 0.05, 0.01, 0.001, \text{ and } 0.0001$, respectively.

Tables 3 and 4 clearly shows that the Rao's efficient score test is very conservative, no matter whether the genotype data are smoothed or not and which basis functions are used to smooth the GVF and $\beta(t)$, since the empirical type I error rates are smaller than the corresponding nominal level α for all sample sizes and all cases of “Basis of both GVF and $\beta(t)$ ”, “Basis of beta-Smooth Only”, and “FPCA Approach”. In addition, the results of “Basis of both GVF and $\beta(t)$ ” are very similar to those of “Basis of beta-Smooth Only” in Tables 3 and 4; and actually, many of them are identical for large sample cases $n = 1,000, 1,500, \text{ and } 2,000$.

The empirical type I error rates of the FPCA global test in Table 3 are slightly higher than the nominal levels α when the sample sizes are small and moderate-sized, i.e., $n = 1,000$. When the sample size is big ($n = 1,500 \text{ and } 2,000$), the calculations of our simulation failed for the FPCA global test and so no results are available. When the sample size was $n = 1,000$, we were only able to get the results of FPCA global test for 10^5 phenotype-genotype datasets instead of 10^6 . Hence, the mixed effect global tests are more suitable for small and moderate-sized problems in terms of computational concerns.

In our simulation studies, we also evaluated the LRT statistics of the fixed effect GFLM. Unfortunately, the empirical type I error rates were inflated (data not shown). Thus, we prefer the Rao's efficient score tests and the global tests for further power analysis. We do notice that the empirical type I error rates of the LRT statistics decrease as the sample sizes increase. When the sample size is 2,000, the empirical type I error rates of the LRT statistics are slightly higher than the nominal levels (around 0.07 at a 0.05 nominal level, data not shown).

Statistical Power Evaluation

Based on the simulated sequence data, the power of our tests was compared with SKAT and SKAT-O. Our tests are those considered in the type I error simulations, i.e., the Rao's efficient score test statistics of fixed effect GFLM and the global test of mixed effect FPCA linear models. The results are reported in Figures 1 — 6. In Figures 1, 2, and 3, the causal variants can be both rare and common. In Figures 4, 5, and 6, the causal variants are only rare variants. In Figures 1 and 4, all causal variants have positive effects; when 20%/80% causal variants have negative/positive effects, we present the results in Figures 2 and 5; when 50%/50% causal variants have negative/positive effects, the results are presented in Figures 3 and 6.

In the legend of all the Figures, “GVF&Beta, B-sp” (or “GVF&Beta, F-sp”) means that both genetic variant functions and genetic effect function $\beta(t)$ were smoothed by B-spline (or Fourier) basis functions, “Beta, B-sp” (or “Beta, F-sp”) means that only the genetic effect function $\beta(t)$ was smoothed by B-spline (or Fourier) basis functions (i.e., beta-smooth only),

“B-sp” means B-spline basis was used, and “F-sp” means Fourier basis was used. In addition, both the Rao's efficient score test statistics and global tests are used for “FPCA”, and “gt” is the abbreviation of global test.

Power Comparison—When the causal variants can be both rare and common as shown in Figures 1, 2, and 3, the Rao's efficient score test statistics of the fixed effect GFLM have higher power than that of SKAT and SKAT-O, except that the SKAT-O has slightly higher power for small and moderate sample size cases of $n = 200, 500$ in plots (a3), (b3), and (c3) of a single Figure 1. The power of the global test is generally low in the Figures 1, 2, and 3. However, the power of the global test is the highest in plot (c3) of the Figure 1, and is similar to that of SKAT-O in plots (a3) and (b3) of the Figure 1, when the sample size $n = 200$ is small. In addition, the power of the global test is higher than those of the Rao's efficient score test statistics in plots (a3), (b3), and (c3) of the Figure 1, when the sample size $n = 200$ is small. When the sample size $n = 500$ is moderate-sized in plots (a3), (b3), and (c3) of the Figure 1, the power of the global test is slightly lower than those of the Rao's efficient score test statistics and SKAT-O.

When only rare variants are causal as shown in Figures 4, 5, and 6, the Rao's efficient score test statistics of fixed effect GFLM have similar or lower power than those of SKAT or SKAT-O. Interestingly, the global test of mixed effect generalized functional linear models of FPCA performs well when only rare variants are causal in the Figures 4, 5, and 6. In particular, when the power levels of the Rao's efficient score test statistics of the fixed effect generalized functional linear models are lower than those of SKAT or SKAT-O, the power of the global test of mixed effect models are notably higher than that of the Rao's efficient score test and similar to that of SKAT-O for quite a few plots, e.g., Figures 4 and 5. Therefore, the Rao's efficient score test statistics and the global test of the proposed models are complementary to each other in terms of their power performance and we observe this in plots (a3), (b3), and (c3) of the Figure 1 when the causal variants can be both rare and common for small sample size $n = 200$.

General Observations—In total, we compared the power performance of five Rao's efficient score test statistics of the fixed effect models: two are based on B-spline basis functions, two are based on Fourier basis functions, and one is based on FPCA. In the two Rao's efficient score tests to use B-spline (or Fourier) basis functions, one is to smooth both the genetic variant functions and the genetic effect function $\beta(t)$, and the other is only to smooth the genetic effect function $\beta(t)$ (i.e., beta-smooth only). Generally, the five Rao's test statistics of the fixed effect functional linear models have similar power. The power levels of beta-smooth only are almost identical to those of smoothing both the genetic variant functions and genetic effect function $\beta(t)$ by B-spline basis (or Fourier basis). Therefore, our Rao's efficient score test statistics of the fixed effect models do not strongly depend on whether the genotype data are smoothed or not, or which basis functions are used. Hence, it is very stable in terms of power performance.

LRT Statistics—For the fixed effect generalized functional linear models, we calculated the empirical power levels of the LRT statistics, which provide the highest powers among the Rao's efficient score tests, global tests, and SKAT and SKAT-O (data not shown). Since

the empirical type I error rates are inflated, the LRT statistics are not considered to be valid tests.

Additional Simulations—In the **Supplementary Materials I and II**, we report 6 sets of simulation results in addition to those reported in the main text. In the first simulations presented in the **Supplementary Materials I**, the order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 12$; the number of Fourier basis functions was $K = K_\beta = 13$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K_\beta = 12$ and the truncation lag $K = 25$. The empirical type I error rates are presented in Table A.1, which shows that the type I error rates are properly controlled. The power results based on COSI sequence data are presented in Figures A.1 — A.6. One can notice that the power levels of the Figure 1 are similar to those in the Figure A.1, the power levels of the Figure 2 are similar to those of the Figure A.2, etc. Additionally, the **Supplementary Materials I and II** contain simulation results for a range of the number of basis functions, and they confirm the models are very robust.

Discussion

In this article, we built functional logistic regression models for gene- or region-based association analysis of dichotomous phenotype traits adjusting for covariates. By using functional data analysis techniques, the observed high dimension genetic variant data are used to estimate genetic variant functions based on B-spline or Fourier basis functions or functional principal component decompositions [de Boor, 2001; Ferraty and Romain, 2010; Horváth and Kokoszka, 2012; Ramsay et al., 2009; Ramsay and Silverman, 1996]. Since the genetic data are treated as stochastic functions, the genetic effects are modeled as a function of the genetic distances between the markers. Similarly, the genetic effect function can be estimated by B-spline or Fourier or linear spline basis functions. Then, the estimated genetic variant functions or original genotype data and the estimated genetic effect function are used in fixed or mixed effect generalized functional linear models to test for association between a dichotomous trait and multiple genetic variants in a genetic region. When the original genotype data are used in the model, we simplify the generalized functional logistic regression to be a beta-smooth only model (4), in which only the genetic effect function needs to be estimated.

Extensive simulation analysis demonstrate that the Rao's efficient score tests of our fixed effect models are conservative (since they generate lower type I errors than the nominal levels), and the global tests of the mixed effect models generate accurate type I errors. Our simulation studies found that the Rao's score test statistics of the fixed effect models have higher power than SKAT and SKAT-O in most cases when the causal variants are both rare and common. When the causal variants are all rare, the Rao's efficient score test statistics or the global score tests have similar or slightly lower power than SKAT and SKAT-O.

Our models should be very useful in dissecting complex traits because they have superior performance when the causal variants are both common and rare. This is particularly important because, in practice, it is not known whether only rare variants or only common variants in a disease-related gene have effects. All we can assume is that a mixture of rare

and common variants influences disease susceptibility. When the causal variants are only rare, both our mixed effect model and SKAT can perform well. It should be noted that SKAT is based on a mixed effect model. Therefore, the mixed effect models can be good for analyzing rare variants, while the fixed effect models are better suited when the causal variants are both rare and common. One possibility is that the common variant effects are likely from only one or a few genetic variants which makes the fixed effect models work well. On the other hand, it is likely many genetic variants have effects on the phenotype if the causal variants are only rare and so mixed effect models work well.

One benefit of treating genotype data functionally is that the estimated genetic effect function naturally serves as a weighting function; this function is determined by the data, and takes marker spacing and LD and similarity among individuals into account. This is different from the burden tests and kernel-based approaches which use artificial weighting functions to improve power. Furthermore, for different traits, it is unclear which weighting functions are the best [Ionita-Laza et al., 2013], and so one needs to try different weighting functions to find the one which provides good power. The weighting function may depend on sample sizes and MAFs as described in Ionita-Laza et al. [2013]. For routine data analysis, it can be hard to decide which weighting functions to use. In our generalized functional linear models, the method already incorporates weighting. The genetic effect function $\beta(t)$ is the effect of the genetic variant function at the location t , which can be thought as a weighted effect. We explored using weighted genetic variant functions defined by the MAFs, and found that the power is very similar to the power without weights. Hence, it is not necessary to add weights.

We also evaluated the performance of LRT statistics. It was found that the LRT statistics of the fixed effect models have the highest power among the Rao's efficient score tests, the global tests, SKAT, and SKAT-O. Unfortunately, the LRT statistics inflate the empirical type I error rates (even when the sample size is big, e.g., $n = 2,000$), and hence we do not report the results of the LRT statistics in this paper. The LRT statistics are not useful for testing association for the dichotomous traits and so we had to search for appropriate test statistics which have well-controlled type I error rates and have good power performance, when the sample sizes are less than or equal to 2,000. The Rao's efficient score tests are very conservative and deflated; in addition, their power performance is better than SKAT and SKAT-O when the causal variants are both rare and common, and similar or slightly worse than SKAT and SKAT-O when the causal variants are all rare. The global tests of the mixed effect models proposed in Goeman et al. [2004, 2006, and 2011] have accurate type I error rates; moreover, the global tests have good power levels when the causal variants are rare and when the performance of the Rao's efficient score test statistics is less impressive.

While our methods are promising, they have some limitations. Like other gene-based methods, our methods test at the gene-level, and so have limited ability to determine precisely which variants contribute to the phenotype. In our study, the Rao's efficient score tests and the global tests are found to perform well. It is unclear if other better tests exist, and this is an interesting open area for future investigation. Our simulations are limited to cases in which the sample sizes are less than 2,000. We notice that the type I error rates of LRT decreases as the sample size increases. For large sample problems such as those in

meta-analysis, it could be possible that the LRT could be useful. However, more investigation is needed to determine how large the sample size must be to ensure that the LRT has accurate type I error rates. For quantitative traits, the LRT of functional linear models was found to have inflated type I error rates when sample sizes are smaller than or equal to 1,000 and to have accurate type I error rates when sample sizes are 1,500 or 2,000, while F -tests were found to have accurate type I error rates over all sample sizes examined [Fan et al. 2013]. Similar results regarding type I error rates were found in Kong et al. [2014].

Our research so far focuses on population data. It would be interesting to extend our models to handle pedigree data, which would require properly taking correlation between pedigree members into account [Jakobsdottir and McPeck, 2013; Thornton and McPeck, 2010; Wang and McPeck, 2009].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by the Intramural Research Program of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Ruzong Fan, Yifan Wang, and James L. Mills), by the Intramural Research Program of the National Human Genome Research Institute (Alexander F. Wilson and Joan E. Bailey-Wilson), National Institutes of Health, Bethesda, MD, and by Wei Chen's NIH/NEI grant R01 EY024226 (Daniel E. Weeks, Ruzong Fan is an unpaid collaborator on this grant) and the University of Pittsburgh (Daniel E. Weeks). We thank Dr. Goeman for many e-mail communications about the global tests developed in his group. We thank Dr. Seungeun Lee for sending us his simulation program of SKAT and sequence data generated by Dr. Yun Li using program COSI; and Dr. Yun Li for generation and LD pattern of the sequence data. Two anonymous reviewers and the editor, Dr. Shete, provided very good and insightful comments for us to improve the manuscript.

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD (<http://biowulf.nih.gov>).

References

- Bansal V, Harismendy O, Tewhey R, Murray SS, Schork NJ, Topol EJ, Frazer KA. Accurate detection and genotyping of SNPs utilizing population sequencing data. *Genome Res.* 2010; 20:537–545. [PubMed: 20150320]
- Cardot H, Sarda P. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis.* 2005; 92:24–41.
- Carter TC, Kay DM, Browne ML, Liu AY, Romitti PA, Kuehn D, Conley MR, Caggana M, Druschel CM, Brody LC, et al. Hirschsprungs disease and variants in genes that regulate enteric neural crest cell proliferation, migration, and differentiation. *Journal of Human Genetics.* 2012; 57:485–493. [PubMed: 22648184]
- Clarke J, Wu HC, Jayasinghe L, Patel I A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol.* 2009; 4:265–270. [PubMed: 19350039]
- Cordell HJ, Clayton DG. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: application to HLA in type 1 diabetes. *Am J Hum Genet.* 2002; 70:124–141. [PubMed: 11719900]
- de Boor, C. A Practical Guide to Splines, revised version. Springer; New York: 2001. Applied Mathematical Sciences 27..

- Fan R, Wang Y, Mills JL, Wilson AF, Bailey-Wilson JE, Xiong M. Functional linear models for association analysis of quantitative traits. *Genet Epidemiol*. 2013; 37:726–742. [PubMed: 24130119]
- Ferraty, F.; Romain, Y. *The Oxford Handbook of Functional Data Analysis*. Oxford University Press; New York: 2010.
- Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*. 2004; 20:93–99. [PubMed: 14693814]
- Goeman JJ, van de Geer SA, van Houwelingen HC. Testing against a high-dimensional alternative. *Journal of the Royal Statistical Society Series B Statistical Methodology*. 2006; 68:477–493.
- Goeman JJ, van Houwelingen HC, Finos L. Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*. 2011; 98:381–390.
- Goldsmith J, Bobb J, Crainiceanu CM, Caffo B, Reich D. Penalized functional regression. *J Comput Graph Stat*. 2011; 20:830–851. [PubMed: 22368438]
- Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*. 2008; 82:100–112. [PubMed: 18179889]
- Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered*. 2010; 70:42–54. [PubMed: 20413981]
- He X, Fuller CK, Song Y, Meng Q, Zhang B, Yang X, Li H. Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *Am J Hum Genet*. 2013; 92:667–680. [PubMed: 23643380]
- Horváth, L.; Kokoszka, P. *Inference for Functional Data With Applications*. Springer; New York: 2012.
- The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007; 449:851–861. [PubMed: 17943122]
- Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. *Am J Hum Genet*. 2013; 92:841–853. [PubMed: 23684009]
- Jakobsdottir J, McPeck MS. MASTOR: mixed-model association mapping of quantitative traits in samples with related individuals. *Am J Hum Genet*. 2013; 92:652–666. [PubMed: 23643379]
- James G. Generalized linear models with functional predictor variables. *Journal of the Royal Statistical Society B*. 2002; 64:411–432.
- Kong, D.; Staicu, A.; Maity, A. Classical testing in functional linear models. 2014. <http://www4.stat.ncsu.edu/~staicu/Research.html>
- Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet*. 2008; 82:386–397. [PubMed: 18252219]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012a; 91:224–237. [PubMed: 22863193]
- Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012b; 13:762–775. [PubMed: 22699862]
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008; 83:311–321. [PubMed: 18691683]
- Li Y, Hsing T. Deciding the dimension of effective dimension reduction space for functional and high-dimensional data. *The Annals of Statistics*. 2010; 38:3028–3062.
- Lin WY, Schaid DJ. Power comparisons between similarity-based multilocus association methods, logistic regression, and score tests for haplotypes. *Genet Epidemiol*. 2009; 33:183–197. [PubMed: 18814307]
- Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. *Genome Res*. 2011; 21:1099–1108. [PubMed: 21521787]
- Luo L, Zhu Y, Xiong M. Quantitative trait locus analysis for next-generation sequencing with the functional linear models. *J Med Genet*. 2012; 49:513–524. [PubMed: 22889854]

- Luo L, Zhu Y, Xiong M. Smoothed functional principal component analysis for testing association of the entire allelic spectrum of genetic variation. *Eur J Hum Genet.* 2013; 21:217–224. [PubMed: 22781089]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet.* 2008; 9:387–402. [PubMed: 18576944]
- Metzker ML. Sequencing technologies the next generation. *Nat Rev Genet.* 2010; 11:31–46. [PubMed: 19997069]
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007; 615:28–56. [PubMed: 17101154]
- Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol.* 2010; 34:188–193. [PubMed: 19810025]
- Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol.* 2010; 34:213–221. [PubMed: 19697357]
- Müller H, Stadtmüller U. Generalized functional linear models. *The Annals of Statistics.* 2005; 33:774–805.
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genet.* 2011; 7:e1001322. [PubMed: 21408211]
- Pangilinan F, Molloy AM, Mills JL, Troendle JF, Parle-McDermott A, Signore C, O'Leary VB, Chines P, Seay JM, Geiler-Samerotte K, et al. Evaluation of common genetic variants in 82 candidate genes as risk factors for neural tube defects. *BMC Med Genet.* 2012; 13:62. [PubMed: 22856873]
- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86:832–838. [PubMed: 20471002]
- Ramsay, JO.; Hooker, G.; Graves, S. *Functional Data Analysis With R and Matlab.* Springer; New York: 2009.
- Ramsay, JO.; Silverman, BW. *Functional Data Analysis.* Springer; New York: 1996.
- Ross, SM. *Stochastic Processes.* Second Edition. John Wiley & Sons; New York: 1996.
- Rusk N, Kiermer V. Primer: Sequencing the next generation. *Nat Methods.* 2008; 5:15. [PubMed: 18175411]
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 2005; 15:1576–1583. [PubMed: 16251467]
- Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev.* 2009; 19:212–219. [PubMed: 19481926]
- Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol.* 2008; 26:1135–1145. [PubMed: 18846087]
- Thornton T, McPeck MS. ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am J Hum Genet.* 2010; 86:172–184. [PubMed: 20137780]
- Wang Z, McPeck MS. ATRIUM: testing untyped SNPs in case-control association studies with related individuals. *Am J Hum Genet.* 2009; 85:667–678. [PubMed: 19913122]
- Wessel J, Schork NJ. Generalized genomic distance-based regression methodology for multilocus association analysis. *Am J Hum Genet.* 2006; 79:792–806. [PubMed: 17033957]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010; 86:929–942. [PubMed: 20560208]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89:82–93. [PubMed: 21737059]

Figure 1. The Empirical Power of the Rao's efficient Score Tests, Global Tests, SKAT and SKAT-O, When Causal Variants Were Both Rare and Common, and All Causal Variants Had Positive effects

The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K_\beta = 10$ and the truncation lag $K = 20$.

Figure 2. The Empirical Power of the Rao's efficient Score Tests, Global Tests, SKAT and SKAT-O, When Causal Variants Were both Rare and Common, and 20%/80% Causal Variants Had Negative/Positive effects

The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_{\beta} = 10$; the number of Fourier basis functions was $K = K_{\beta} = 11$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K_{\beta} = 10$ and the truncation lag $K = 20$.

Figure 3. The Empirical Power of the Rao's efficient Score Tests, Global Tests, SKAT and SKAT-O, When Causal Variants Were both Rare and Common, and 50%/50% Causal Variants Had Negative/Positive effects

The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K_\beta = 10$ and the truncation lag $K = 20$.

Figure 4. The Empirical Power of the Rao's efficient Score Tests, Global Tests, SKAT and SKAT-O, When Causal Variants Were only Rare, and All Causal Variants Had Positive effects
 The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K_\beta = 10$ and the truncation lag $K = 20$.

Figure 5. The Empirical Power of the Rao's efficient Score Tests, Global Tests, SKAT and SKAT-O, When Causal Variants Were Only Rare, and 20%/80% Causal Variants Had Negative/Positive effects

The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_\beta = 10$; the number of Fourier basis functions was $K = K_\beta = 11$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K_\beta = 10$ and the truncation lag $K = 20$.

Figure 6. The Empirical Power of the Rao's efficient Score Tests, Global Tests, SKAT and SKAT-O, When Causal Variants Were Only Rare, and 50%/50% Causal Variants Had Negative/Positive effects

The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K_{\beta} = 10$; the number of Fourier basis functions was $K = K_{\beta} = 11$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K_{\beta} = 10$ and the truncation lag $K = 20$.

Table 1

Association Analysis of Neural Tube Defects Data.

The Name of Gene	The Number of SNPs	P-values								
		Basis of both GVF and $\beta(t)$		FPCA Approach		Basis of beta-Smooth Only			Kernel Based Tests	
		Rao's Efficient Score Tests		Test Statistics		Rao's Efficient Score Tests			SKAT	SKAT-O
		B-sp Basis	Fourier Basis	Rao	gt	B-sp Basis	Fourier Basis			
MFTC	9	4.89×10^{-4}	1.95×10^{-2}	5.40×10^{-3}	0.375	2.10×10^{-3}	8.28×10^{-4}	2.57×10^{-3}	5.02×10^{-3}	
CDKN2A	18	0.081	0.063	0.081	2.92×10^{-2}	0.081	0.063	0.061	1.08×10^{-2}	
ADA	16	0.857	0.646	0.977	0.608	0.857	0.646	0.903	1.000	
PENT	8	1.71×10^{-2}	9.78×10^{-3}	1.71×10^{-2}	4.69×10^{-2}	1.71×10^{-2}	9.78×10^{-3}	0.122	0.055	
CUBN	134	1.15×10^{-2}	0.172	2.95×10^{-2}	1.73×10^{-2}	1.15×10^{-2}	0.172	0.457	0.267	
GART	8	0.099	0.071	0.099	0.768	0.099	0.071	0.922	1.000	
DNMT3A	23	0.492	0.177	0.166	0.604	0.492	0.177	0.946	1.000	
MTHFD1	18	1.59×10^{-2}	1.57×10^{-2}	1.99×10^{-2}	0.730	1.59×10^{-2}	1.57×10^{-2}	0.587	0.740	
T(Brachyury)	18	0.135	0.237	0.250	6.44×10^{-3}	0.135	0.237	0.628	0.366	

The results of “Basis of both GVF and $\beta(t)$ ” were based on smoothing both the GVF and the genetic effect function $\beta(t)$, the results of “FPCA Approach” were based on FPCA approach, and the results of “Basis of beta-Smooth Only” were based on smoothing the genetic effect function $\beta(t)$ only approach. Abbreviation: gt = global test. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K\beta = 10$; the number of Fourier basis functions was $K = K\beta = 11$. In the case of FPCA, the number of knots of linear spline basis was taken as $K\beta = 10$ and the truncation lag $K = 20$. The p -values which are smaller than 0.05 are marked by boldface.

Table 2

Association Analysis of Hirschsprung's Disease Data.

The Name of Gene	The Number of SNPs	P-values							
		Basis of both GVF and $\beta(t)$		FPCA Approach		Basis of beta-Smooth Only		Kernel Based Tests	
		Rao's Efficient Score Tests		Test Statistics		Rao's Efficient Score Tests		SKAT	SKAT-O
		B-sp Basis	Fourier Basis	Rao	gt	B-sp Basis	Fourier Basis		
PROK1	10	0.283	0.553	0.283	0.159	0.283	0.553	0.245	0.262
PROKR1	7	4.23×10^{-2}	0.191	4.23×10^{-2}	0.386	4.23×10^{-2}	0.191	0.298	0.365
PHOX2B	5	0.063	0.069	0.063	0.755	0.063	0.069	0.257	0.265
ASCL1	5	0.882	0.551	0.668	0.170	0.668	0.551	0.565	0.493
HOXB5	6	0.552	0.596	0.552	0.680	0.552	0.596	0.586	0.684
RET	7	8.62×10^{-23}	4.81×10^{-23}	8.62×10^{-23}	2.74×10^{-24}	8.62×10^{-23}	4.81×10^{-23}	2.11×10^{-14}	1.96×10^{-18}

The results of "Basis of both GVF and $\beta(t)$ " were based on smoothing both the GVF and the genetic effect function $\beta(t)$, the results of "FPCA Approach" were based on FPCA approach, and the results of "Basis of beta-Smooth Only" were based on smoothing the genetic effect function $\beta(t)$ only approach. Abbreviation: gt = global test. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K\beta = 10$; the number of Fourier basis functions was $K = K\beta = 11$. In the case of FPCA, the number of knots of linear spline basis was taken as $K\beta = 10$ and the truncation lag $K = 20$. The p -values which are smaller than 0.05 are marked by boldface.

Table 3

Empirical Type I Error Rates of the Proposed Rao's Efficient Score Tests of the Fixed Effect Models and the Global Test of the FPCA Mixed Effect Models, When the Causal Variants Are Both Rare and Common.

Nominal Level α	Sample Size n	Basis of both GVF and $\beta(t)$		FPCA Approach		Basis of beta-Smooth Only	
		Rao's Efficient Score Tests		Test Statistics		Rao's Efficient Score Tests	
		B-sp Basis	Fourier Basis	Rao	gt	B-sp Basis	Fourier Basis
0.05	200	0.035814	0.034715	0.036819	0.050922	0.035685	0.034648
	250	0.036998	0.035438	0.038135	0.050471	0.036903	0.035374
	300	0.037574	0.036152	0.039339	0.050355	0.037536	0.036138
	350	0.038173	0.037027	0.040051	0.050517	0.038127	0.037012
	500	0.039692	0.039133	0.042021	0.050101	0.039688	0.039120
	1,000	0.043553	0.043674	0.045964	0.050440 [#]	0.043550	0.043674
	1,500	0.045111	0.045315	0.046896	X	0.045115	0.045315
	2,000	0.045571	0.046309	0.047011	X	0.045571	0.046309
0.01	200	0.005242	0.004918	0.005251	0.010580	0.005214	0.004901
	250	0.005545	0.005194	0.005618	0.010435	0.005512	0.005169
	300	0.005613	0.005241	0.005931	0.010370	0.005594	0.005213
	350	0.005887	0.005491	0.006156	0.010386	0.005870	0.005483
	500	0.006290	0.006090	0.006767	0.010324	0.006288	0.006090
	1,000	0.007329	0.007304	0.007995	0.010150 [#]	0.007328	0.007304
	1,500	0.008000	0.008061	0.008610	X	0.008000	0.008061
	2,000	0.008135	0.008399	0.008536	X	0.008136	0.008399
0.001	200	0.000380	0.000373	0.000341	0.001168	0.000389	0.000370
	250	0.000398	0.000354	0.000360	0.001135	0.000401	0.000357
	300	0.000379	0.000356	0.000405	0.001101	0.000380	0.000354
	350	0.000400	0.000371	0.000398	0.001123	0.000397	0.000368
	500	0.000439	0.000417	0.000498	0.001073	0.000436	0.000418
	1,000	0.000550	0.000582	0.000662	0.001060 [#]	0.000550	0.000583
	1,500	0.000639	0.000633	0.000706	X	0.000640	0.000633
	2,000	0.000671	0.000673	0.000715	X	0.000671	0.000673
0.0001	200	0.000052	0.000049	0.000038	0.000118	0.000051	0.000047
	250	0.000050	0.000038	0.000033	0.000123	0.000048	0.000038
	300	0.000035	0.000037	0.000030	0.000120	0.000035	0.000037
	350	0.000036	0.000036	0.000023	0.000118	0.000035	0.000036
	500	0.000039	0.000029	0.000041	0.000114	0.000039	0.000029
	1,000	0.000033	0.000040	0.000037	0.000090 [#]	0.000033	0.000040
	1,500	0.000041	0.000036	0.000049	X	0.000041	0.000036
	2,000	0.000054	0.000045	0.000047	X	0.000054	0.000045

The results of "Basis of both GVF and $\beta(t)$ " were based on smoothing both the GVF and the genetic effect function $\beta(t)$, the results of "FPCA Approach" were based on FPCA approach, and the results of "Basis of beta-Smooth Only" were based on smoothing the genetic effect function $\beta(t)$ only approach.

X : result is not available. Abbreviation: gt = global test. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K\beta = 10$; the number of Fourier basis functions was $K = K\beta = 11$. In the simulations of FPCA, the number of knots of linear spline basis was taken as $K\beta = 10$ and the truncation lag $K = 20$.

[#] results were based on 10^5 phenotype-genotype datasets.

Table 4

Empirical Type I Error Rates of the Proposed Rao's Efficient Score Tests of the Fixed Effect Models, When the Causal Variants Are Only Rare.

Nominal Level α	Sample Size n	Rao's Efficient Score Tests			
		Basis of both GVF and $\beta(t)$		Basis of beta-Smooth Only	
		B-sp Basis	Fourier Basis	B-sp Basis	Fourier Basis
0.05	500	0.030993	0.029429	0.030960	0.029382
	1,000	0.037398	0.036902	0.037400	0.036898
	1,500	0.040507	0.040476	0.040507	0.040477
	2,000	0.042458	0.042855	0.042457	0.042855
0.01	500	0.003911	0.003566	0.003907	0.003563
	1,000	0.005443	0.005181	0.005440	0.005180
	1,500	0.006444	0.006247	0.006446	0.006247
	2,000	0.006924	0.006896	0.006925	0.006896
0.001	500	0.000227	0.000188	0.000226	0.000188
	1,000	0.000343	0.000292	0.000343	0.000293
	1,500	0.000469	0.000425	0.000469	0.000425
	2,000	0.000469	0.000467	0.000469	0.000467
0.0001	500	0.000025	0.000027	0.000026	0.000027
	1,000	0.000018	0.000013	0.000018	0.000013
	1,500	0.000028	0.000034	0.000028	0.000034
	2,000	0.000029	0.000032	0.000029	0.000032

The results of "Basis of both GVF and $\beta(t)$ " were based on smoothing both the GVF and the genetic effect function $\beta(t)$, and the results of "Basis of beta-Smooth Only" were based on smoothing the genetic effect function $\beta(t)$ only approach. The order of B-spline basis was 4, and the number of basis functions of B-spline was $K = K\beta = 10$; the number of Fourier basis functions was $K = K\beta = 11$.