HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2017 April 01.

Published in final edited form as:

Genet Epidemiol. 2016 April ; 40(3): 210–221. doi:10.1002/gepi.21955.

Uncovering local trends in genetic effects of multiple phenotypes via functional linear models

Olga A. Vsevolozhskaya[#1,*], **Dmitri V. Zaykin**[#2], **David A. Barondess**[3], **Xiaoren Tong**[3], **Sneha Jadhav**[4], and **Qing Lu**[3,*]

[1] Department of Biostatistics, University of Kentucky, Lexington, USA

[2] National Institute of Environmental Health Sciences, National Institutes of Health, USA

[3] Department of Epidemiology and Biostatistics, Michigan State University, East Lansing, USA

[4] Department of Statistics, Michigan State University, East Lansing, USA

[#] These authors contributed equally to this work.

Abstract

Recent technological advances equipped researchers with capabilities that go beyond traditional genotyping of loci known to be polymorphic in a general population. Genetic sequences of study participants can now be assessed directly. This capability removed technology-driven bias toward scoring predominantly common polymorphisms and let researchers reveal a wealth of rare and sample-specific variants. While the relative contributions of rare and common polymorphisms to trait variation are being debated, researchers are faced with the need for new statistical tools for simultaneous evaluation of all variants within a region. Several research groups demonstrated flexibility and good statistical power of the functional linear model approach. In this work we extend previous developments to allow inclusion of multiple traits and adjustment for additional covariates. Our functional approach is unique in that it provides a nuanced depiction of effects and interactions for the variables in the model by representing them as curves varying over a genetic region. We demonstrate flexibility and competitive power of our approach by contrasting its performance with commonly used statistical tools and illustrate its potential for discovery and characterization of genetic architecture of complex traits using sequencing data from the Dallas Heart Study.

Keywords

multivariate analysis; pleiotropy; genome-wide association studies; sequencing studies; quantitative traits; qualitative traits; functional analysis

[*] Corresponding Author: Qing Lu, Department of Epidemiology and Biostatistics, College of Human Medicine, Michigan State University, 909 Fee Road, East Lansing, MI 48824-1030, qlu@msu.edu. [*] Co-corresponding Author: Olga A. Vsevolozhskaya, Department of Biostatistics, College of Public Health, University of Kentucky, 725 Rose Street, Lexington, KY 40536-0082, vsevolozhskaya@uky.edu.

Conflict of Interests

We have no conflicts of interest to declare.

1 Introduction

Genome-wide association studies (GWAS) have identified numerous risk loci for common complex diseases, and next-generation sequencing based association strategies are now emerging to characterize the contribution of rare genetic variants to human genetic disorders. Analysis of the 'rare variant - common complex disease' hypothesis requires tailored statistical methods, as single marker tests fail to uncover these rare variants [Carvajal-Carmona, 2010]. An entirely new powerful class of statistical methods based on non-parametric functions was recently developed for genetic association testing that can accommodate both rare and common variants, or the combination of the two [Fan et al., 2014, 2013; Lee et al., 2014; Luo et al., 2011, 2012a,b; Svishcheva et al., 2015; Vsevolozhskaya et al., 2014; Wang et al., 2015; Zhu and Xiong, 2012]. A comprehensive comparison of non-parametric functional-based methods (FBMs) via simulation studies and real data applications have repeatedly shown that FBMs have a valid type-I error rate and a substantially higher power to detect an association compared with alternative approaches. Additionally, FMBs were proven to be a powerful approach for genetic association studies with longitudinal data [Reimherr et al., 2014], or for the analysis of gene expression data [Storey et al., 2005].

Recently, our research group has demonstrated that within FBMs, functional analysis of variance (FANOVA) attains higher power to detect an association between a genetic region and a dichotomous trait compared to methods based on functional linear models (FLM) [Vsevolozhskaya et al., 2014]. Specifically, we have shown that FANOVA outperforms FLM for small to moderate effect sizes of the variants within a genetic region. Nonetheless, from a practical point of view, FANOVA had a notable limitation in that it was not able to accommodate quantitative traits or adjust for continuous covariates.

In light of these shortcomings, our aim was to extend the existing FANOVA method to association analyses of multiple quantitative and qualitative traits and to accommodate situations in which (1) a gene influences more than one trait (i.e., pleiotropy), (2) where there are confounding/mediation effects (due to population substructure or other sources), and (3) where the effect of disease risk can be modified by a trait or an exposure – a phenomena that we refer hereafter as "Treatment by Trait" (T×T) interaction.

To conceptualize T×T interaction, consider a study of genetic risk factors of substance abuse disorder. It is well known that personality traits like impulsivity and sensation-seeking are highly prevalent in drug-dependent individuals (e.g., [De Wit, 2009]). It is also known that personality traits are substantially influenced by genes (e.g., [Bouchard Jr and Loehlin, 2001]). Suppose there are genetic risk factors that contribute to the increased risk of developing drug addiction among individuals with high trait-impulsivity. Suppose, further, that a different genetic disposition might be involved in the increased risk of developing drug addiction among individuals with low trait-impulsively. Hence, risk alleles for drug-dependence (i.e., 'treatment') might vary by the level of personality traits, which will be modeled as T×T interaction in our generalized FANOVA approach – more on this later.

A distinctive contribution of the approach presented here to the emerging field of FBMs for genetic association studies is the introduction of an efficient way to estimate the effects of phenotypes, confounding factors and T×T interactions using continuous curves smoothly varying over genetic loci. Previously proposed functional methods for genetic association studies (e.g., [Fan et al., 2013; Luo et al., 2011, 2012a]) and other methods that combine information across multiple variants within a gene (e.g., [Liu and Leal, 2010; Wu et al., 2011]) aggregate across both the association signals of genetic variants as well as over covariate effects. We exploit the flexibility of the functional approach to unveil a more nuanced blueprint of how covariate and interaction effects vary within a genetic region by estimating partial regression coefficient curves that change over variant positions.

Unlike traditional statistical models that treat a disease phenotype as an outcome (i.e., on the left-hand side of the equation), our model puts non-genetic variables on the right-hand side, including traits, environmental exposures, and confounders. The response function in our model is an allelic dosage curve, fitted through genetic variants within a region. If we start our modeling by including a binary trait such as drug dependence as a single predictor, the continuous regression coefficient will be the difference between average allelic dosages over multiple variants of the two groups. That is, a continuous intercept curve will estimate smoothed average allelic dosage among non drug dependent controls, and a continuous regression coefficient will estimate a deviation from this baseline allelic dosage over multiple variants among drug-dependent cases. Further, if we include personality trait as a covariate, the regression coefficient curve for drug-addiction will be adjusted for personality trait. Finally, if we include a T×T interaction between drug-dependence status and a personality trait, the deviation from the baseline allelic dosage among drug-dependent cases will vary by the level of a personality trait.

Functional models where genetic predictor (X) and the outcome (Y) are swapped in the regression equation are reminiscent of the reverse regression approach [Maddala, 1992]. In general, coefficients of the direct and the reverse regressions are not the same, however the test statistic for the X (adjusted for any covariates) as well as the corresponding partial correlation coefficient remain the same after the swapping. For example, adjustment for confounding or mediation is unaffected and remains valid in the reverse regression approach.

To estimate continuous coefficient curves, our new generalized FANOVA approach utilizes a connection between penalized spline regression and best linear unbiased predictors (BLUPs), enabling a straightforward practical implementation using standard linear mixed models statistical software. A connection between BLUPs and penalized functional regression has been explored in statistical and machine learning literature [Brumback et al., 1999; Crainiceanu et al., 2005; Crainiceanu and Goldsmith, 2010; Eilers and Marx, 1996; Goldsmith et al., 2011; Ivanescu et al., 2014; Lian, 2007; Nosedal-Sanchez et al., 2012; Pearce and Wand, 2006; Ruppert et al., 2003; Wand and Ormerod, 2008; Wang, 1998]. However, this connection has largely been ignored in functional method approaches for genetic association studies.

We provide an illustration of our method using data from the Dallas Heart Study [Romeo et al., 2007], by characterizing associations of sequence variants with plasma triglyceride

levels, modified by race and adjusted for sex. In addition to identifying the originally reported association between triglyceride levels and the *ANGPTL4* gene, our new FANOVA approach identified specific sub-regions of the *ANGPTL4* gene associated with plasma triglyceride levels among European Americans, African Americans, and Hispanics.

2 Methods

2.1 Genotypic functions: a brief overview

In brief, our method is an extension of the previously proposed FANOVA methodology, which seeks to quantify the relationship between scalar phenotypes $X_1, X_2, \ldots, X_k$ and smooth genotypic functions $G(t)$'s, with t indexing a genetic variant's position over a genetic region, $t \in [0, \tau]$ [Vsevolozhskaya et al., 2014]. By using the term 'genotypic functions,' we refer to nonparametric functions fitted with a basis expansion method [Ramsay and Silverman, 2005; Ruppert et al., 2003; Wood, 2006]. Thus, for each subject, the genetic data is not of a discrete (i.e., counted) nature, such as would be the case for genotype frequencies, but rather a single nonparametric genotypic function, $G(t)$, of a continuous nature.

A genotypic function is obtained by either (i) a cubic B-spline basis expansion over a dense set of knots, $\kappa_1, \ldots, \kappa_K$, over the range of the variant's genomic positions t_i's (in the one-base coordinate system) or (ii) penalized spline smoothing that avoids the knot selection problem completely (e.g., [Luo et al., 2012a; Vsevolozhskaya et al., 2014]). Earlier investigations of functional linear models designed for genetic association testing include comprehensive coverage of the estimation procedure for the genotypic functions $G(t)$'s [Fan et al., 2014, 2013; Lee et al., 2014; Luo et al., 2011, 2012a,b; Svishcheva et al., 2015; Vsevolozhskaya et al., 2014; Wang et al., 2015; Zhu and Xiong, 2012].

If we let $G_1(t), \ldots, G_N(t)$, $t \in [0, \tau]$ denote the functional genotypic data for N individuals, and we let $X_{1i}, \ldots, X_{Pi}$, $i = 1, \ldots, N$ denote a set of P variables that consists of covariates and traits (either quantitative or qualitative) that may contribute to a disease, our model for each individual's genotypic function is:

$$G_i(t) = \beta_0(t) + \beta_1(t) X_{1i} + \cdots + \beta_P(t) X_{Pi} + \epsilon_i(t), \quad (1)$$

where $\beta_l(t)$'s are continuous regression coefficients that describe an association between a scalar trait and a set of variants in a genetic region $t \in [0, \tau]$, and where $\varepsilon(t)$ is a residual function. Unlike traditional models where the outcome is regressed on a set of predictors, this model treats genetic information as an outcome. Outside of the functional approach, utility of such "reverse regressions" has been explored previously for analysis of genetic associations [Feng, 2014; Kwan et al., 2011]. While coefficient estimates change, in general, due to swapping of variables between two sides of a regression equation, the partial correlations as well as the test statistics and P-values for the coefficients remain invariant: this follows simply from expressing these quantities in terms of the entries of the inverse of the correlation matrix between all variables including the outcome. Thus, testing for effects or for validity of regression adjustments are preserved under the reversal. There is also convenience in having the same type of outcome (i.e., genetic information) and thus the same type of a link function regardless of the type and the number of other variables in the

model. Additionally, within the functional approach, exploration of $\hat{\beta}(t)$'s may allow researchers to determine sub-regions of $[0, \tau]$ that harbor causal genetic variants (i.e., sub-regions over which $\hat{\beta}(t) \neq 0$).

To estimate $\hat{\beta}(t)$'s, we place a function-on-scalar regression in Eq. 1 into the context of a mixed-effects model or, more generally, embed the penalized splines problem into the class of reproducing kernel methods. To introduce the method, we first present a case of a single curve estimation, and conclude with the general case that allows us to estimate continuous coefficients of multiple traits, construct their confidence intervals and test for an association. We finally note that in the context of this paper, the word "kernel" should not be confused with a weight function as in the local regression (or local smoothing), which is also called a kernel [Hastie et al., 2009].

2.2 Estimating a single curve

To draw connections between smoothing splines and reproducing kernels, first consider a simpler problem of estimating a single curve from the observed y_i's and t_i's, $i = 1, \ldots, n$. One possible approach to estimating a nonparametric function $f(t)$ from discrete data is to invoke penalized spline smoothing (e.g., [Wahba, 1990]). This smooth interpolation of the data is achieved by minimizing least squares fits with an additional roughness penalty (i.e., penalized sums of squares) as follows:

$$min\left\{n^{-1}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda\int_0^1[Ly(t)]^2dt\right\}. \quad (2)$$

Here, the roughness of a function is quantified by the square of a linear differential operator $Ly(t)$ (a typical choice is $Ly(t) = f''(t)$ that corresponds to penalizing curvature of the function). The constant term, λ, referred to as a smoothing or a tuning parameter, should be either specified by a user or determined through the generalized cross-validation (GCV) [Wood, 2006].

The above minimization problem is analogous to a corresponding regularization problem within the machine learning domain:

$$\min_{f\in\mathscr{H}}\left\{\sum_{i=1}^{n}L(y_i, f(t_i)) + \lambda\|Pf\|^2_{\mathscr{H}}\right\}, \quad (3)$$

where $L(y_i, f(t_i))$ is a loss function, $\|Pf\|^2$ penalizes f in terms of the variability of its function values, and $\mathscr{H}$ is the reproducing kernel Hilbert space (RKHS) of real functions f. The theory of RKHS was developed by Aronszajn [1950] and Saitoh [1988], with good overviews provided by Smola and Schölkopf [1998]; Wahba [1990] and Rasmussen and Williams [2006]. Briefly, a RKHS on $\mathbb{R}^d$ is a Hilbert space of real-valued functions generated by a bivariate symmetric, positive definite kernel $k(\cdot, \cdot)$ with the following properties: (i) for every $\mathbf{t}$ in $\mathbb{R}^d$, $k(\mathbf{t}, \mathbf{t}')$ is a function of $\mathbf{t}'$ in $\mathscr{H}$ and (ii) k has the reproducing property $\langle k(\cdot, t_i), f\rangle_{\mathscr{H}} = f(t_i)$, where $\langle\cdot, \cdot\rangle$ denotes the inner product. To conceptualize penalized splines in Eq. (2) as BLUPs in a mixed model framework, we explore the solution

to the regularization problem in Eq. (3) from the machine learning theory. Based on the results of the *representer theorem* [Kimeldorf and Wahba, 1971], it can be shown that each minimizer $f \in \mathscr{H}$ of Eq. (3) can be written as a linear combination of kernel functions, as follows:

$$f(t) = \sum_{i=1}^{n} \alpha_i k(t, t_i). \quad (4)$$

The solution for $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_n]'$ can be obtained as $\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}$, in which $\mathbf{K}$ is then $n \times n$ matrix with ijth entry of $k(t_i, t_j)$, $\mathbf{I}$ is the $n \times n$ identity matrix, nad $\mathbf{y}$ is the $n \times 1$ vector of observed y_i's [Hastie et al., 2009; Rasmussen and Williams, 2006]. Further, the vector of n fitted values is given by $\hat{\mathbf{f}} = \boldsymbol{K}\hat{\boldsymbol{\alpha}}$. This solution looks very similar to that from a linear regression model (i.e., $\hat{\mathbf{y}} = \mathbf{T}\hat{\boldsymbol{\beta}}$ since we used t_i instead x_i in Eqs. (2-3)). Regrettably, this reproducing kernel transformation of t_i's does not simply move our non-linear problem into the 'friendly' linear model domain, because the solution for $\boldsymbol{\alpha}$ depends on λ.

A slight variation to the *representer theorem* can be achieved by decomposing $\mathscr{H}$ into $\mathscr{H}_0 \oplus \mathscr{H}_1$, where $\mathscr{H}_0$ is a finite-dimensional null space containing terms which will not be penalized, and $\mathscr{H}_1$ is it's orthogonal complement (i.e., penalized terms). For example, for $\|Pf\|^2$ defined by differential operators of the form $Ly(t) = f^{(m)}(t)$, the null space $\mathscr{H}_0$ is spanned by polynomials of degree up to $m - 1$. More specifically, if $m = 2$, then constant and linear functions are in the null space, because they are not penalized for 'curvature.' With the decomposition of $\mathscr{H}$, the minimizer f of the regularization function in Eq. (3) now has the form:

$$f(t) = \sum_{j=1}^{m} d_j \phi_j(t) + \sum_{i=1}^{n} c_i k_1(t, t_i), \quad (5)$$

where $\varphi_1(t), \ldots, \varphi_m(t)$ form the basis of $\mathscr{H}_0$ and $k_1(\cdot, \cdot)$ is a reproducing kernel that generates $\mathscr{H}_1$. If $m = 2$ as in the example above, then $\varphi_1(t) = 1$ and $\varphi_2(t) = t$ span the null space of unpenalized functions.

There are relatively few published recommendations in the statistical literature on how to construct $k_1(\cdot, \cdot)$. For example, Lian [2007] writes "[...] *the construction of* k_1 *in general is difficult* and a search of the literature does not seem to provide us with any clues about how to construct a positive definite kernel in general." Nonetheless, if we shift our attention to the machine learning literature, we see that $k_1(t, t_i) = G(t, t_i)$, where $G(t, t_i)$ is a Green's function of the linear differential operator $Ly(t)$ [Fasshauer, 2012; Fasshauer and Ye, 2013; Poggio and Girosi, 1990; Rasmussen and Williams, 2006]. Note that the Green's function also depends on the boundary conditions. A 'natural' choice is the "Natural Boundary Condition" $f^{(j)}(a) = f^{(j)}(b) = 0$, $j = 1, \ldots, m$; where a and b are the boundaries of the functional domain [Green and Silverman, 1993].

How can we estimate the fitted values of the coefficients $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ in Eq. (5) for a specific problem? If we re-write Eq. (5) using linear algebra notations as:

$$\hat{\mathbf{f}}=\mathbf{\Phi}\hat{\mathbf{d}}+\mathbf{K_1}\hat{\mathbf{c}}, \quad (6)$$

it becomes evident that Eq. (6) represents a solution to the linear mixed-effects model with design matrices $\mathbf{\Phi}$ and $\mathbf{K}_1$, and $\hat{\mathbf{d}}$ and $\hat{\mathbf{c}}$ estimated as best linear unbiased predictors (BLUPs) from this model [Speed, 1991]. In addition, the BLUP solution for the coefficients is independent of the smoothing parameters λ, which is equal to the ratio of the variances of the residuals and random effects. For numerical stability reasons, the design matrices are specified for a sequence of knots $k_1, \ldots, k_\kappa$ places at sample quantiles over the range of t_i's [Ruppert, 2002] as:

$$\mathbf{\Phi}=\begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix} \quad \text{and} \quad \mathbf{K_1}=\begin{bmatrix} (t_1-k_1)_+ & \cdots & (t_1-k_\kappa)_+ \\ \vdots & \ddots & \vdots \\ (t_n-k_1)_+ & \cdots & (t_n-t_\kappa)_+ \end{bmatrix}, \quad (7)$$

where $G(t_i, t_j) = (t_i - t_j)_+$ is the Green's function of the linear differential operator $f^{(2)}(t)$, and $x_+ = \max\{0, x\}$. This specification of the design matrices corresponds to a truncated lines series basis expansion $\hat{f}(t) = \hat{d}_0 + \hat{d}_1 t + \sum_{i=1}^{\kappa} \hat{c}_i (t-k_i)_+$. Other choices of basis functions can also be used with corresponding changes to penalized terms. Possible choices include, but are not limited to, (a) truncated power basis $(t-k_i)_+^p$, (b) O'sullivan splines [Wand and Ormerod, 2008], (c) thin plate splines [Ivanescu et al., 2014], or (d) the Gaussian kernel [Lian, 2007].

Some readers might wonder whether the mixed model formulation for penalized splines bear the same parameter interpretation as in a typical application to nested hierarchical data. We should clarify that the functional representation in Eq. (6) is just a convenient way of shifting a non-linear problem into a linear domain, while simultaneously estimating a smoothing parameter. Similarly, the random effects in c, are just a convenience device to model the curvature in **f** and should not be interpreted as random effects, per se.

2.3 Estimating $\beta(t)$'s

With respect to the conceptual model in Eq. (1), continuous regression coefficients can be estimated as follows. For each subject, the genotypic function is evaluated on the grid of genomic positions $t_1, \ldots, t_n$, i.e, $\hat{G}_i(t) = \hat{G}_i(t_1), \ldots, \hat{G}_i(t_n)$. For the sequence of knots $k_1, \ldots, k_\kappa$, each functional regression coefficient is expanded in terms of the linear combination of φ's and k_1's. This expansion yields the following mixed-model representation of Eq. (1):

$$\begin{aligned}\hat{G}_i(t) &= \hat{\beta}_0(t) + \hat{\beta}_1(t) X_{1i} + \cdots + \hat{\beta}_P(t) X_{Pi} \\ &= \underbrace{\left(\hat{d}_1 + \hat{d}_2 t + \sum_{i=1}^{\kappa} k_1(t, k_i)\, \hat{c}_i\right)}_{\hat{\beta}_0(t)} \\ &+ \underbrace{\left(\hat{d}_1^* + \hat{d}_2^* t + \sum_{i=1}^{\kappa} k_1(t, k_i)\, \hat{c}_i^*\right)}_{\hat{\beta}_1(t)} X_{1i} + \cdots \\ &+ \underbrace{\left(\hat{d}'_1 + \hat{d}'_2 t + \sum_{i=1}^{\kappa} k(t, k_i)\, \hat{c}'_i\right)}_{\hat{\beta}_P(t)} X_{Pi}.\end{aligned} \quad (8)$$

Conceptually, the generalized FANOVA-based regression coefficients, $\beta(t)$'s, are similar to the genetic effect coefficients in the recently published paper by Wang et al. [2015]. Specifically, Wang et al. [2015] also proposed to estimate regression coefficients, $\beta_j(t)$'s, smoothly varying over the genetic position t. However, unlike the methodology proposed in the present study, their approach can not simultaneously handle quantitative and qualitative traits, adjust coefficients for confounders/mediators over a continuum $[0, \tau]$ or modify effects by the level of another trait. With our approach, this adjustments can be easily incorporated into the model.

Suppose we want to adjust the effect of a risk factor X_1 by trait X_2 over all t. The model will be written as:

$$\hat{G}_i(t) = \hat{\beta}_0(t) + \hat{\beta}_1(t) X_{1i} + \hat{\beta}_2(t) X_{2i}.$$

Suppose, further, we want to modify the effect of a risk factor X_1 by the level of trait X_2, i.e., model a T×T interaction (for simplicity, assume that X_2 has only two levels). The model can be expressed as:

$$\hat{G}_i(t) = \hat{\beta}_0(t) + \hat{\beta}_1(t) X_{1i} + \hat{\beta}_2(t) X_{2i} + \hat{\beta}_{12}(t) X_{1i} X_{2i}.$$

Then, for the first level of X_2, dummy coded as 0, the association between a gene and X_1 will be estimated by $\hat{\beta}_1(t)$:

$$\hat{G}_i(t) = \hat{\beta}_0(t) + \hat{\beta}_1(t) X_{1i},$$

and for the second level of X_2, dummy coded as 1, the association between a gene and X_1 will be modified as:

$$\hat{G}_i(t) = \left(\hat{\beta}_0(t) + \hat{\beta}_2(t)\right) + \left(\hat{\beta}_1(t) + \hat{\beta}_{12}(t)\right) X_{1i}.$$

To facilitate the data analysis using mixed-effects software, an input response should be a vectorized matrix of genotype functions for N subjects evaluated on the grid of genomic

positions, $vec\left(\hat{\mathbf{G}}\right)=\left[\hat{G}_1\left(t_1\right),\ldots,\hat{G}_1\left(t_n\right),\ldots,\hat{G}_N\left(t_1\right),\ldots,\hat{G}_N\left(t_n\right)\right]$. Input predictors should be $N \cdot n \times 1$ vectors $\mathbf{X}_1, \ldots, \mathbf{X}_P$ which are generated by repeating each phenotype observation n times and stacking them on top of one another. The fixed and the random effects design matrices, $\mathbf{\Phi}$ and $\mathbf{K}_1$, are then constructed as follows:

$$\begin{aligned}\mathbf{\Phi} &= \begin{bmatrix} 1 & t_1 & X_{11} & t_1X_{11} & \cdots & X_{P1} & t_1X_{P1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & t_n & X_{11} & t_nX_{11} & \cdots & X_{P1} & t_nX_{P1} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & t_1 & X_{1N} & t_1X_{1N} & \cdots & X_{PN} & t_1X_{PN} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & t_n & X_{1N} & t_nX_{1N} & \cdots & X_{PN} & t_nX_{PN} \end{bmatrix} \\ &= \left[\mathbf{1}_N \otimes \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \quad \mathbf{X}_1 \otimes \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}, \ldots, \mathbf{X}_P \otimes \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{bmatrix}\right],\end{aligned}$$

and $\mathbf{K}_1=[\mathbf{1}_N \otimes \mathbf{K},\ \mathbf{X}_1 \otimes \mathbf{K}, \ldots, \mathbf{X}_P \otimes \mathbf{K}]$, where $\mathbf{1}_N$ is $N \times 1$ vector of 1's, $\otimes$ is the Kronecker product, and $\mathbf{K}$ is the $n \times \kappa$ matrix with the ijth entry of $k_1(t_i, k_j)$ calculated over the sequence of knots $k_1, \ldots, k_\kappa$.

2.4 Confidence interval for $\hat{\beta}(t)$

Since the conceptual model in Eq. (1) can be expressed as a mixed-effects model in Eq. (8), the typical inferential machinery for mixed-effects models can be used to obtain the variance-covariance estimates of the model parameters [Ruppert et al., 2003]. An explicit formulation for the estimated standard deviation of $\hat{\beta}(t)$ is:

$$\widehat{st.dev}\left(\hat{\beta}(t)\right)=\hat{\sigma}_\epsilon \sqrt{\mathbf{C}\left(\mathbf{C}^\top \mathbf{C}+\hat{\lambda}\mathbf{D}\right)^{-1}\mathbf{C}^\top \mathbf{C}\left(\mathbf{C}^\top \mathbf{C}+\hat{\lambda}\mathbf{D}\right)^{-1}\mathbf{C}^\top}, \quad (9)$$

where $\hat{\sigma}_\epsilon$ is a REML estimate of $\sigma_\mathcal{E}$, $\mathbf{C} = [\mathbf{\Phi}\ \mathbf{K}_1]$ is formed by two design matrices described in Eq. (7), $\hat{\lambda}=\hat{\sigma}^2_\epsilon/\hat{\sigma}^2_c$ is the estimated smoothing parameter, and $\mathbf{D}$ is formed as follows:

$$\mathbf{D}=\begin{bmatrix} \mathbf{0}_{m\times m} & \mathbf{0}_{m\times\kappa} \\ \mathbf{0}_{\kappa\times m} & \mathbf{I}_{\kappa\times\kappa} \end{bmatrix},$$

where m is the number of 'fixed effects' and κ is the number of 'random effects.' An approximate point-wise 100%(1 – α) confidence interval is $\hat{\beta}(t) \pm z_{(1-\alpha/2)}\widehat{st.dev}\left(\hat{\beta}(t)\right)$ Alternatively, Bayesian credible intervals can be obtained by realizing a connection between Gaussian processes and spline construction [Crainiceanu et al., 2005; Rasmussen and Williams, 2006], or "subject re-sampling" bootstrap error bars can be obtained to construct the confidence intervals [Wu and Yu, 2002].

In the application of point-wise bands to functional genotype data, the issue of bias-variance trade-o associated with the selection of the degree of smoothing might deserve more careful attention. Specifically, in the context of the mixed-effects model in Eq. (8), the response variable is a fitted genotypic function $\hat{G}(t)$. If the fitted function is somewhat wiggly, this 'noise' will account for the increased width of the point-wise standard error bands for $\hat{\beta}(t)$. We previously proposed the 'flipping algorithm' for genotype re-labeling that decreases the number of noisy oscillations for smoothed genotype data and showed that this approach results in a substantial increase of statistical power to detect a genetic association [Vsevolozhskaya et al., 2014]. Nonetheless, too smooth genotype functions might result in narrow standard error bands for $\hat{\beta}(t)$ and thus estimate a biased version of a true function with great reliability. Further research is needed on the issue of optimal choice of a smoothing parameter in the context of genotype function fitting.

2.5 Testing for an association

In this section we turn our attention to a test statistic used for evaluating an association between a genetic region and one or more phenotypes. Whereas different types of point-wise confidence intervals for the coefficient curves can be constructed, the hypothesis testing problem of distinguishing an optimal sub-model of $\beta(t)$'s is still of interest. To address this issue, we will use the function $\mathscr{F}$ statistic [Shen and Faraway, 2004] as previously used in our FANOVA methodology [Vsevolozhskaya et al., 2014]. Specifically, suppose we want to test the nullity of a single predictor:

$$H_0: \beta_i(t) = 0, \quad i = 1, \ldots, P.$$

By using Theorem 2 in Shen and Faraway [2004], a test statistic to determine if $\beta(t)$ is equivalent to the zero function can be constructed as follows:

$$\mathscr{F} = \frac{(N-P)\int \hat{\beta}_i^2(t)\,dt}{rss_1(\mathbf{X}^\top \mathbf{X})^{-1}_{ii}}, \quad (10)$$

where $\mathbf{X} = (\mathbf{1}\ \mathbf{X}M_1 \ldots \mathbf{X}_P)$ is a design matrix for the full model containing all phenotypic variables, and $rss_1 = \sum_{i=1}^{N} \int \left(\hat{G}_i(t) - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^2 dt$ is the residual sum of squares for the full model. Under the null hypothesis, it can be easily shown (e.g. [Reimherr et al., 2014; Shen and Faraway, 2004; Zhang, 2013]) that the distribution of $\mathscr{F}$ can be approximated by an F-distribution as:

$$\mathscr{F} \sim F_{\hat{d},(N-P)\hat{d}},$$

where $\hat{d} = \frac{\left(\sum_{i=1}^{n} r_i\right)^2}{\sum_{i=1}^{n} r_i^2}$, with n being the number of genetic variants, and r_i is the ith order eigenvalue of the empirical variance-covariance matrix under the full model, $\hat{\Sigma}^1$.

Alternatively, if we want to test the nullity of K predictors simultaneously, that is, to compare the full model:

$$\hat{G}_i(t) = \sum_{j=1}^{P} \hat{\beta}_j(t) X_{ij},$$

to the reduced model:

$$\hat{G}_i(t) = \sum_{j=1}^{(P-K)} \hat{\beta}_j(t) X_{ij},$$

the test statistic $\mathscr{F}$ can be defined in terms of the reduction in the sums of squared errors, as follows:

$$\mathscr{F} = \frac{(rss_0 - rss_1)/K}{rss_1/(N-P)} \approx \frac{trace\left(\hat{\Sigma}^0 - \hat{\Sigma}^1\right)/K}{trace\left(\hat{\Sigma}^1\right)/(N-P)}, \quad (11)$$

where rss_0 is the residual sum of squares for the reduced model, and $\hat{\Sigma}^0$ is the empirical variance-covariance matrix under the reduced model. Under the null hypothesis, the distribution of $\mathscr{F}$ is approximated by $F_{Kd,(N-P)d}$.

We note that the test statistic in Eq. (11) is computationally more complex than the one in Eq. (10). That is, if the goal is to test the nullity of only one predictor at a time, the test statistic in Eq. (10) can be calculated directly by fitting only the full model, and thus omitting fitting the reduced model. Further details and comparisons of the two formulas can be found in Shen and Faraway [2004].

3 Simulation Study

3.1 Design

The flexibility of our method allows us to accommodate various analysis settings and types of variables, including multiple, possibly correlated or pleiotropic phenotypes, and T×T interactions. One way to analyzing multiple traits is to test for an association one trait at a time. For a proper control of the experiment-wise false-discovery error rate, this ‘one at a time’ testing approach requires accounting for the number of tests performed and correcting for each individual trait's *P*-value. This individual correction typically leads to an inflation in the observed *P*-values. However, our method provides an efficient way of testing multiple traits simultaneously, with no *P*-value correction required, and thus naturally provides superior performance in terms of statistical power to detect an association. Moreover, to handle T×T interactions, or to assess modification of genetic susceptibility to disease by trait, our model requires a test of nullity for an interaction term. Previously, we investigated the power of FANOVA to detect an association with a single predictor [Vsevolozhskaya et

al., 2014]. Simulation studies presented here reflect the extension of our previous basic model with the addition of mediation/confounding scenarios

Figure 1 aids in conceptualization of our data simulation process. We focused on a three variable system and hypothesized that there is a genetic predisposition (G) to continuous phenotypes (Z) and (X). We also assumed a relationship between (Z) and (X) and were interested in testing for an association between (G) and (X), while adjusting for the third variable (Z). Clearly, data generated under this scenario fits the mediation analysis framework, but MacKinnon et al. [2000] point out that the label of (Z) (i.e., either as a mediator or a confounder) depends on the framework used to conceptualize the phenomenon. From a statistical modeling point of view, directionality and the causality are indistinguishable, making these seemingly different concepts of mediation and confounding statistically equivalent. Therefore, data generated under our design can be used to check for both a mediator and a confounding control.

3.2 Data generation

We generated genetic data (G) using the 1,000 genome project [Durbin et al., 2010] to mimic the real sequencing data structure (e.g., linkage disequilibrium patterns, allele frequencies, and randomly missing genotype data). Specifically, at each simulation iteration, a random 30 kb section of genetic region was drawn. Within this 30 kb region, each simulated data contained an average of 300 variants with minor allele frequencies (MAF) ranging from less than 0.001 to almost 0.5. The complete distribution of MAF for all variants across simulations is provided in the left panel of Figure 2.

Next, a continuous trait (Z) was simulated as:

$$Z_i=\sum_{j=1}^{n} G_i^{\chi}(t_j)\times\gamma(t_j)+\epsilon_i,\quad i=1,\ldots,N,\quad j=1,\ldots,n, \tag{12}$$

where N is the number of subjects, n is the number of variants, t_j indexes the position of variants, $\gamma(t_j)$ is the effect of the variant in t_j's, $\varepsilon_i \sim N(0, 1)$, and "$\chi$" indicates a subset of genetic variants harboring causal alleles. For example, if χ = 10%, then a random sample of 10% of all variants for subject i were causal, and the effect of each causal variant j, $\gamma(t_j)$, was drawn from an $N\left(\mu_\gamma, \sigma_\gamma^2\right)$ distribution (the rest of $\gamma(t)$'s, corresponding to non-causal variants, were zero). If μ_γ = 0, the effect of a given causal variant was either protective or deleterious . If $\mu_\gamma > 0$, then the majority of causal variants had the same direction of the effects (i.e., deleterious), and the magnitude of the effect size varied by manipulating σ_γ^2. The middle panel of Figure 2 illustrates simulated effects by MAF for the choice μ_γ = 0 and $\sigma_\gamma^2=1$; the right panel for μ_γ = 0.25 and $\sigma_\gamma^2=1$. The reader should note that under our simulation scenario, the causal variants can be both rare and common. Alternative situations with only rare or common causal variants were previously investigated by our group and showed favorable performance by FANOVA [Vsevolozhskaya et al., 2014].

Another continuous trait (X) was simulated as:

$$X_i = \sum_{j=1}^{n} G_i^{\chi}(t_j) \times \alpha(t_j) + \beta \times Z_i + \epsilon_i. \quad (13)$$

Similar to $\gamma(t_j)$, $\alpha(t_j) \sim N\left(\mu_\alpha, \sigma_\alpha^2\right)$ represents the effect of a causal variant *j* on the trait (X), and $\beta \sim N(3,1)$ represents the effect of the third variable (Z) on the trait (X).

3.3 Type I error results

For empirical type I error simulations, we set the genetic effect on the continuous trait (X) to zero, i.e., $\alpha(t_j) = 0$ for all *j*, and tested for an association between (G) and (X), while adjusting for (Z). The percentage of risk variants for the association between (G) and (Z) in Eq.(12) was set to $\chi = 30\%$ and γ_j's were simulated from an $N(\mu_\gamma = 0, \sigma_\gamma = 3)$ distribution. For the different sample sizes, we compared the generalized FANOVA approach to the SKAT methodology [Wu et al., 2011]. The results are summarized in Table 1. For both methods, all empirical type I error rates are around the nominal α levels with the exception of SKAT for a small sample size. To further contrast the differences between FANOVA and SKAT, we proceeded to a comparison of power simulations.

3.4 Statistical power results

For the statistical power comparison, both traits (Z) and (X) shared the same percentage, but a random set of risk variants. The percent of risk variants were set to 5%, 10%, 30%, 50%, 70%, 90%, and 100%. The sample size values were $N = 50$, 500, 2500, and 5000. The execution time of a single iteration of the simulations (the statistical power is presented based on at least 1,000 iterations) on a single core (2.5Ghz Intel Xeon E5-2670v2) of high-performance computing center (HPCC: https://wiki.hpcc.msu.edu/) ranged from 20 seconds for $N = 50$ up to an hour for $N = 5000$. The allocated memory for $N = 5000$ subjects was 64GB.

Figure 3 summarizes empirical power results for the scenario with risk variants having either positive or negative effects (i.e., $\mu_\gamma = \mu_\alpha = 0$) for the different number of subjects. In Figure 4 the majority of risk variants had deleterious effects for both traits (i.e., $\mu_\gamma > 0$ and $\mu_\alpha > 0$). In each figure, the generalized FANOVA statistical power to detect an association between (G) and (X), while adjusting for (Z), is represented by a solid line, and the power of SKAT is represented by a dashed line.

In general, the proposed FANOVA approach attained higher power than SKAT, especially for small sample sizes, small effect sizes, and when the percentage of risk variants is small. The empirical power of the two approaches become comparable if the effect sizes and the proportion of risk variants were large.

4 Application to real data: *ANGPRL4* association with triglyceride

To further illustrate the utility of our generalized FANOVA approach, we turn to the issue of association testing between sequence variations in *ANGPTL4* gene and lipid metabolism. In mice, the involvement of *ANGPTL4* in lipid metabolism was shown by intravenous injection

of recombinant *ANGPTL4*, resulting in an increase in plasma triglycerides (TG) levels [Yoshida et al., 2002]. In humans, the involvement of *ANGPTL4* in lipid metabolism is probable and may be associated with a higher risk of cardiovascular disorder [Kathiresan et al., 2009; Muendlein et al., 2014; Romeo et al., 2007]. However, each individual *ANGPTL4* variant confers a modest effect [Kathiresan et al., 2009], suggesting an improved statistical power for methods like generalized FANOVA that perform a joint gene-based association analysis.

We conducted an analysis of 93 sequence variations in *ANGPTL4* that were identified among 3,551 participants in the Dallas Heart Study [Romeo et al., 2007]. To examine an increase in plasma TG levels, we binned individuals into the 'low-triglyceride' group (660 individuals with plasma triglyceride level ≤25th percentile) and into the 'high-triglyceride' group (679 individuals with plasma triglyceride level ≥75th percentile). The resulting sample included 443 individuals of mixed European descent, 651 African Americans, and 245 Hispanics.

As discussed elsewhere (e.g., [Svishcheva et al., 2015; Vsevolozhskaya et al., 2014]), statistical power of functional methods may depend on the quality of genotype data smoothing. To obtain smooth genotypic functions, we first coded allelic dosage based on the minor allele counts (i.e., either 0, 1 or 2) and applied the "flipping algorithm" [Vsevolozhskaya et al., 2014] to minimize the number of 0-2 (or 2-0) patterns in every two subsequent variant positions. However, because the majority of 93 sequenced variants were rare [Romeo et al., 2007], the coding based on minor allele counts was concluded to be optimal and no re-coding of allelic dosage was necessary.

To examine an effect of increase in TG levels, modified by race and adjusted for sex, we built the following model:

$$\begin{aligned}\hat{G}_i(t_j) = \; & \beta_0(t_j) + \beta_1(t_j) X_{\mathbf{TG}i} + \\ & \beta_2(t_j) X_{\mathbf{AA}i} + \beta_3(t_j) X_{\mathbf{H}i} + \beta_{12}(t_j) X_{\mathbf{TG}i} X_{\mathbf{AA}i} + \beta_{13}(t_j) X_{\mathbf{TG}i}(t_j) X_{\mathbf{H}i} + \\ & \beta_4(t_j) X_{\mathbf{Sex}i} + \epsilon_i(t_j),\end{aligned}$$

where $\beta_0(t_j)$ is the smoothed baseline allelic dosage $j = 1, \ldots, 93$; $\beta_1(t_j)$ is the effect of TG-increase on allelic dosage. The next four terms are added to examine T×T interaction or whether the effect of TG increase varies among European Americans ($\beta_1(t_j)$), African Americans ($\beta_1(t_j) + \beta_{12}(t_j)$), and Hispanics ($\beta_1(t_j) + \beta_{13}(t_j)$). Finally, $\beta_4(t_j)$ is the adjustment for sex.

To determine the most parsimonious model, we first performed a test for T×T interaction, i.e., $H_0 : \beta_{12}(t_j) = \beta_{13}(t_j) = 0$ for all t_j, and found statistically significant differences in TG-increasing effect among individuals of different racial descent (*P*-value=0.0028). We note that the magnitude of this *P*-value remained the same for different choices of kernels and as such, we proceeded to explore specific sub-regions of the *ANGPTL4* gene that may harbor causal variants for the different racial groups.

Each panel of Figure 5 illustrates the estimated TG-increasing effect among different racial groups and across 93 variants of the *ANGPTL4* gene. Further, the positions of the recently identified variants E40K and T266M [Romeo et al., 2007; Talmud et al., 2008] are added as vertical lines to each panel. The left panel of Figure (5) shows $\hat{\beta}_1(t)$ or the estimated effect of TG increase among European Americans. From this panel we can infer that the region around the E40K variant has the top contribution among European Americans, since it is the region over which $\hat{\beta}_1(t)$ deviates the most from the zero line. Additionally, the direction of $\hat{\beta}_1(t)$ around E40K is negative, indicating that TG increase is associated with a lower dosage of E40K variant, which implies that European American E40K carriers can be expected to have lower TG levels. However, the confidence bands for $\hat{\beta}_1(t)$ include zero and indicate lack of statistical significance.

The right panel of Figure 5 shows $\hat{\beta}_1(t)+\hat{\beta}_{13}(t)$ or the estimated effect of TG increase among Hispanics. Once again, the effect has the top magnitude around E40K region, but it's direction is reversed, indicating that Hispanic E40K carriers tend to have higher TG levels. Additionally, among Hispanics, E40K region association with TG-increase reached statistical significance.

The middle panel of Figure 5 shows $\hat{\beta}_1(t)+\hat{\beta}_{12}(t)$ or the estimated effect of TG increase among African Americans. Unlike European Americans and Hispanics, the contribution of E40K variant does not appear to be appreciably associated with TG increase. Also, no contribution of T266M variant to either TG increase (or decrease) was found among any racial groups.

Finally, to compare our T×T interaction results to SKAT, we performed a subgroup analysis on data from European Americans, African Americans, and Hispanics. The *P*-values, adjusted for sex, for the test of an association between TG-levels and variants in the *ANGPTL4* gene were as follows: among European Americans $P_{SKAT} = 0.0006$, $P_{FANOVA} = 0.0262$; among Hispanics $P_{SKAT} = 0.1738$, $P_{FANOVA} = 0.0001$; among African Americans $P_{SKAT} = 0.2321$, $P_{FANOVA} = 0.9447$. Accordingly, both methods concluded an association between *ANGPTL4* variants and plasma triglycerides levels among European Americans, no association among African Americans, and discordant results among Hispanics. The reader should not be surprised by seemingly disagreeing FANOVA conclusions for European Americans summarized via the confidence bends in Figure 5 and via the *P*-value for an association test. It has been noted multiple times, including by our research group [Vsevolozhskaya et al., 2015], that a combination of multiple “marginally significant” outcomes across different variants may result in the overall significance for a genetic region.

5 Discussion

By generalizing previously proposed FANOVA methodology, we offer a novel approach not previously explored in FLM-based association studies for estimating multiple phenotype-specific effects smoothly varying over genetic variants. Furthermore, by treating genetic information as the response variable and all traits as predictors (qualitative or quantitative), the generalized FANOVA provides a straightforward way to account for hidden population

stratification, confounders, mediators and T×T interactions. The established connection between penalized least squares and best linear unbiased predictors allows for a straightforward implementation of the proposed methodology using standard mixed linear model software.

The introduced notion of T×T interaction deserves additional clarification. We are not necessarily putting emphasis on the interaction itself or the value of its coefficient. Rather, the inclusion of this term gives a simple way of detecting possible effects of various combinations of treatment and trait values that may go beyond what is captured by the sum of their individual effects.

How well do our generalized FANOVA regression coefficient estimates replicate what others have found in prior studies of *ANGPTL4*? Studies of Romeo et al. [2007] and Talmud et al. [2008] revealed that among European Americans E40K carriers have significantly lower TG levels. Talmud et al. [2008] also showed TG-lowering effect of T266M variant, but only among E40K carriers (i.e., whenever E40K men were excluded from the reanalysis, there was no longer a significant association between T266M and TG levels). T266M is more prevalent than E40K and in our sample out of 620 T266M carriers only 16 were also carriers of E40K, which may be a reason behind lack of association. Furthermore, no studies presented conclusive findings over TG-lowering effect and mutations in *ANGPTL4*, so a replication of the reported association is required.

Our generalized FANOVA model is a functional model analogue of “reverse regression” (e.g., [Maddala, 1992]), where genetic information, X, becomes the response while phenotypes, Y, are treated as predictors. Regression coefficients are not invariant to swapping of predictor and response variables. However, partial correlations, as well as test statistics and the corresponding P-values remain the same after swapping. Thus, effects of adjustments for covariates in a direct model are properly preserved when testing for association in a reverse model. With multiple correlated predictors at an arbitrary variant's position t_j, the test statistic for the regression coefficient β_i can be re-expressed based on the partial correlation between Y and X_i, which is not affected by swapping of variables, and the test statistic (and therefore the P-value) is also invariant under the reversal in a functional model. One limitation of this approach is that for the direct and reverse tests to be equivalent, X_i cannot enter any interaction terms with other variables.

The generalized FANOVA is an extension of the previously proposed FANOVA approach and thus inherits some of its features. For example, generalized FANOVA fully utilizes variants’ position information and linkage-disequilibrium structure when computing the test statistic $\mathcal{F}$. However, unlike the previously proposed FANOVA, our current approach allows inclusion of multiple traits and adjustment for additional covariates. Moreover, our new functional approach provides a unique way of graphically depicting phenotypic effects and interactions by representing them as continuous curves varying over a genetic region. We also hypothesize that the functional approach may hold increased robustness to genotyping errors. This may be due to the fact that the estimated genotype functions, $\hat{G}(t)$, are used for the analysis in place of allele frequencies of the marked locus. It is noted that genotyping errors can have severe consequences for the analysis of low frequency alleles (e.g.,

[Abecasis et al., 2001]). Although genotype functions are estimated via allele counts, they incorporate a certain degree of smoothing, therefore the fitted functions are expected to be less prone to genotyping errors.

In terms of the application of the generalized FANOVA methodology, practitioners can use standard mixed-effects software to estimate continuous regression coefficients as illustrated in the Methods section of this article. Previous research in penalized regression models [Scheipl and Greven, 2012] suggests that a penalty with a small null space should be preferred (a typical choice for the number of "fixed effects" is 2) and a 'rule of thumb' for the number of "random effects" is $\kappa = 35$. However, the specific number of kernel functions is unimportant as long as the fitted genotype functions are not too smooth.

Acknowledgments

Funding

This work was supported by a National Institute of Drug Abuse T32 research training program award (NIDA; T32DA021129) for OAV's postdoctoral fellowship, DVZ's Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS), DAB's research award (NIDA; R01DA016558), and QL's Mentored Research Scientist Development Award (NIDA; K01DA033346).

References

Abecasis GR, Cherny SS, Cardon LR. The impact of genotyping error on family-based analysis of quantitative traits. Eur J Hum Genet. 2001; 9(2):130–134. [PubMed: 11313746]

Aronszajn N. Theory of reproducing kernels. Trans Amer Math Soc. 1950:337–404.

Bouchard TJ Jr, Loehlin JC. Genes, evolution, and personality. Behav Genet. 2001; 31(3):243–273. [PubMed: 11699599]

Brumback BA, Ruppert D, Wand MP. Comment. J Am Stat Assoc. 1999; 94(447):794–797.

Carvajal-Carmona LG. Challenges in the identification and use of rare disease-associated predisposition variants. Curr Opin Genet Dev. 2010; 20(3):277–281. [PubMed: 20564784]

Crainiceanu C, Ruppert D, Wand MP. Bayesian analysis for penalized spline regression using winbugs. Journal of Statistical Software. 2005; 14(14):1–24.

Crainiceanu CM, Goldsmith AJ. Bayesian functional data analysis using winbugs. Journal of Statistical Software. 2010; 32(11)

De Wit H. Impulsivity as a determinant and consequence of drug use: a review of underlying processes. Addict Biol. 2009; 14(1):22–31. [PubMed: 18855805]

Durbin RM, Altshuler DL, Durbin RM, Abecasis GAR, Bentley DR, Chakravarti A, Clark AG, Collins FS, et al. A map of human genome variation from population-scale sequencing. Nature. 2010; 467(7319):1061–1073. [PubMed: 20981092]

Eilers PH, Marx BD. Flexible smoothing with b-splines and penalties. Stat Sci. 1996:89–102.

Fan R, Wang Y, Mills JL, Carter TC, Lobach I, Wilson AF, Bailey-Wilson JE, Weeks DE, Xiong M. Generalized functional linear models for gene-based case-control association studies. Genet Epidemiol. 2014; 38(7):622–637. [PubMed: 25203683]

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(7):726–742. [PubMed: 24130119]

Fasshauer, GE. Approximation Theory XIII. Springer; San Antonio: 2012. Greens functions: Taking another look at kernel approximation, radial basis functions, and splines.; p. 37-63.2010

Fasshauer GE, Ye Q. Reproducing kernels of sobolev spaces via a green kernel approach with differential operators and boundary operators. Adv Comput Math. 2013; 38(4):891–921.

Feng Z. A generalized quasi-likelihood scoring approach for simultaneously testing the genetic association of multiple traits. J Roy Stat Soc C-App. 2014; 63(3):483–498.

Goldsmith J, Bobb J, Crainiceanu CM, Ca o B, Reich D. Penalized functional regression. J Comput Graph Stat. 2011; 20(4)

Green, PJ.; Silverman, BW. Nonparametric regression and generalized linear models: a roughness penalty approach. CRC Press; 1993.

Hastie, T.; Tibshirani, R.; Friedman, J.; Hastie, T.; Friedman, J.; Tibshirani, R. The elements of statistical learning. Springer; 2009.

Ivanescu AE, Staicu A-M, Scheipl F, Greven S. Penalized function-on-function regression. Computation Stat. 2014:1–30.

Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet. 2009; 41(1):56–65. [PubMed: 19060906]

Kimeldorf G, Wahba G. Some results on tchebycheffian spline functions. J Mathe Anal Appl. 1971; 33(1):82–95.

Kwan JS, Kung AW, Sham PC. A simple bias correction in linear regression for quantitative trait association under two-tail extreme selection. Behavior genetics. 2011; 41(5):776–779. [PubMed: 21626281]

Lee D-Y, Hanis C, Bell G, Aguilar D, Redline S, Below J, Xiong M. Genetic studies of physiological traits with their application to sleep apnea. 2014:1410, 7363. arXiv preprint arXiv.

Lian H. Nonlinear functional models for functional responses in reproducing kernel hilbert spaces. Can J Stat. 2007; 35(4):597–606.

Liu DJ, Leal SM. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. PLoS Genetics. 2010; 6(10):e1001156. [PubMed: 20976247]

Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. Genome Res. 2011; 21(7):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. 2012a; 49(8):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. 2012b; 21(2):217–224. [PubMed: 22781089]

MacKinnon DP, Krull JL, Lockwood CM. Equivalence of the mediation, confounding and suppression effect. Prevention Science. 2000; 1(4):173–181. [PubMed: 11523746]

Maddala, GS. Introduction to econometrics. Vol. 2. Macmillan; New York: 1992.

Muendlein A, Saely CH, Leiherer A, Fraunberger P, Kinz E, Rein P, Vonbank A, Zanolin D, Malin C, Drexel H. Angiopoietin-like protein 4 significantly predicts future cardiovascular events in coronary patients. Atherosclerosis. 2014; 237(2):632–638. [PubMed: 25463098]

Nosedal-Sanchez A, Storlie CB, Lee TC, Christensen R. Reproducing kernel hilbert spaces for penalized regression: A tutorial. Am Stat. 2012; 66(1):50–60.

Pearce ND, Wand MP. Penalized splines and reproducing kernel methods. Am Stat. 2006; 60(3)

Poggio T, Girosi F. Networks for approximation and learning. Proceedings of the IEEE. 1990; 78(9): 1481–1497.

Ramsay, J.; Silverman, B. Functional Data Analysis. second edition. Springer; 2005.

Rasmussen CE, Williams CKI. Gaussian processes for machine learning. Citeseer. 2006

Reimherr M, Nicolae D, et al. A functional data analysis approach for genetic association studies. Ann Appl Stat. 2014; 8(1):406–429.

Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen JC. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase hdl. Nat Genet. 2007; 39(4):513–516. [PubMed: 17322881]

Ruppert D. Selecting the number of knots for penalized splines. J Comput Graph Stat. 2002; 11(4)

Ruppert, D.; Wand, MP.; Carroll, RJ. Semiparametric regression. Cambridge university press; 2003.

Saitoh S. Theory of reproducing kernels and its applications. Longman Scientific & Technical Harlow. 1988

Scheipl, F.; Greven, S. Technical Report 125. University of Munich - Department of Statistics; 2012. Identifiability in penalized function-on-function regression models..

Shen Q, Faraway J. An f test for linear models with functional responses. Statistica Sinica. 2004; 14(4):1239–1258.

Smola AJ, Schölkopf B. Learning with kernels. Citeseer. 1998

Speed T. [that blup is a good thing: The estimation of random effects]: Comment. Statistical Science. 1991:42–44.

Storey JD, Xiao W, Leek JT, Tompkins RG, Davis RW. Significance analysis of time course microarray experiments. Proc Natl Acad Sci U S A. 2005; 102(36):12837–12842. [PubMed: 16141318]

Svishcheva GR, Belonogova NM, Axenovich TI. Region-based association test for familial data under functional linear models. PloS One. 2015; 10(6):e0128999. [PubMed: 26111046]

Talmud PJ, Smart M, Presswood E, Cooper JA, Nicaud V, Drenos F, Palmen J, Marmot MG, Boekholdt SM, Wareham NJ, et al. Angptl4 e40k and t266m effects on plasma triglyceride and hdl levels, postprandial responses, and chd risk. Arterioscler Thromb Vasc Biol. 2008; 28(12):2319–2325. [PubMed: 18974381]

Vsevolozhskaya OA, Greenwood MC, Powell SL, Zaykin DV. Resampling-based multiple comparison procedure with application to point-wise testing with functional data. Environmental and Ecological Statistics. 2015; 22(1):45–59.

Vsevolozhskaya OA, Zaykin DV, Greenwood MC, Wei C, Lu Q. Functional analysis of variance for association studies. PloS One. 2014; 9(9):e105074. [PubMed: 25244256]

Wahba G. Spline models for observational data. Siam. 1990

Wand M, Ormerod J. On semiparametric regression with o'sullivan penalized splines. Australian & New Zealand Journal of Statistics. 2008; 50(2):179–198.

Wang Y. Smoothing spline models with correlated random errors. J Am Stat Assoc. 1998; 93(441): 341–348.

Wang Y, Liu A, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, Xiong M, Wu CO, Fan R. Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models. Genet Epidemiol. 2015

Wood S. Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC Texts in Statistical Science. 2006

Wu CO, Yu KF. Nonparametric varying-coefficient models for the analysis of longitudinal data. Int Stat Rev. 2002; 70(3):373–393.

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(1):82–93. [PubMed: 21737059]

Yoshida K, Shimizugawa T, Ono M, Furukawa H. Angiopoietin-like protein 4 is a potent hyperlipidemia-inducing factor in mice and inhibitor of lipoprotein lipase. J Lipid Res. 2002; 43(11):1770–1772. [PubMed: 12401877]

Zhang, J-T. Analysis of variance for functional data. CRC Press; 2013.

Zhu Y, Xiong M. Family-based association studies for next-generation sequencing. Am J Hum Genet. 2012; 90(6):1028–1045. [PubMed: 22682329]

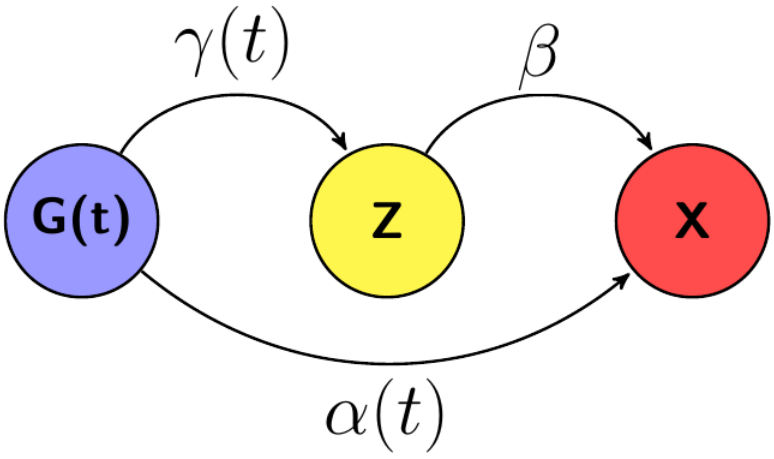

Figure 1.
The genetic information ($G(t)$) is directly associated with the outcome of interest (X) and indirectly through the third variable (Z).

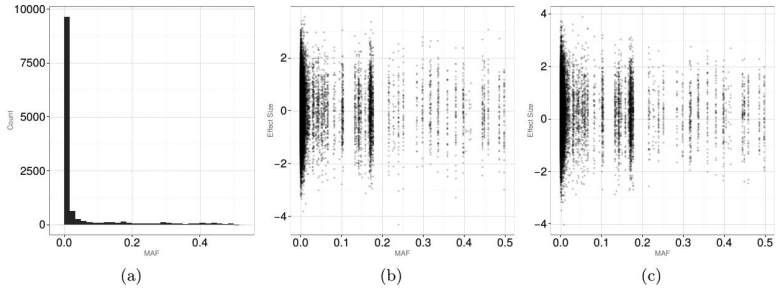

Figure 2.

Panel (a): The range and the distribution of MAF for all variants. Panels (b)-(c): MAF distribution of causal variants by the effect size.

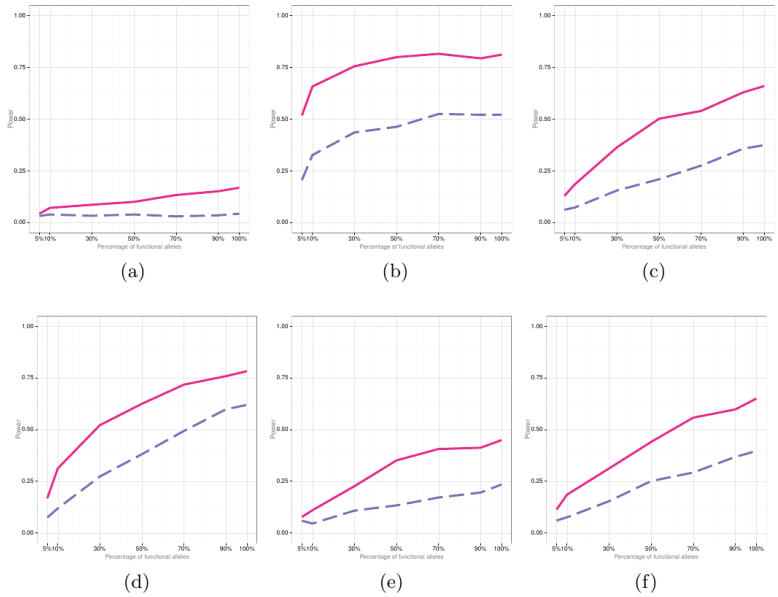

Figure 3.

Empirical power of FANOVA (solid line) and SKAT (dashed line) when the variants can have either protective or deleterious effects (i.e., $\mu_\gamma = \mu_\alpha = 0$). Panel (a): $N = 50$ $\sigma_\gamma = \sigma_\alpha = 0.05$; (b): $N = 50$ $\sigma_\gamma = \sigma_\alpha = 1$; (c): $N = 500$ $\sigma_\gamma = \sigma_\alpha = 0.05$; (d): $N = 1000$ $\sigma_\gamma = \sigma_\alpha = 0.05$; (e) $N = 2500$ $\sigma_\gamma = \sigma_\alpha = 0.015$; (f) $N = 5000$ $\sigma_\gamma = \sigma_\alpha = 0.015$.

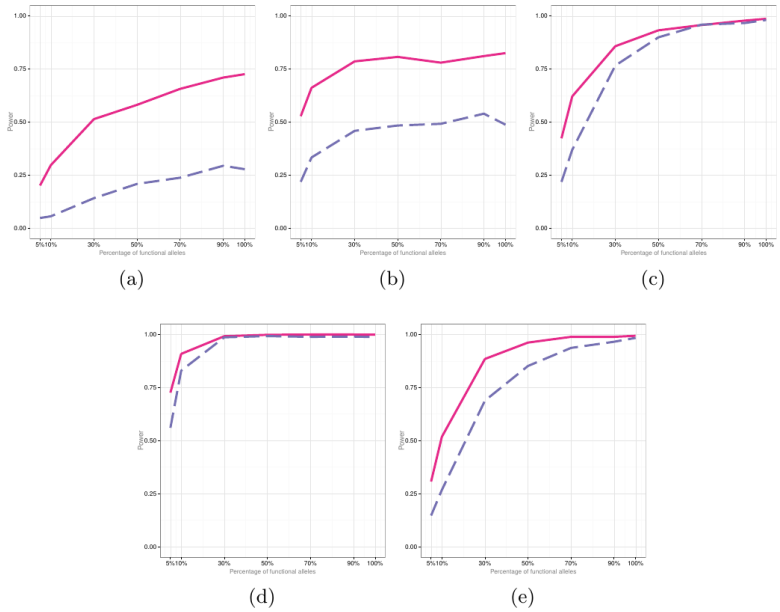

Figure 4.

Empirical power of FANOVA (solid line) and SKAT (dashed line) when the majority of variants have deleterious effect (i.e, $\mu_\gamma > 0$ $\mu_a > 0$) . Panel (a): $N = 50$ $\mu_\gamma = \mu_a = 0.05$ $\sigma_\gamma = \sigma_a = 0.25$; (b): $N = 50$ $\mu_\gamma = \mu_a = 0.05$ $\sigma_\gamma = \sigma_a = 1$; (c): $N = 500$ $\mu_\gamma = \mu_a = 0.05$ $\sigma_\gamma = \sigma_a = 0.15$; (d): $N = 500$ $\mu_\gamma = \mu_a = 0.25$ $\sigma_\gamma = \sigma_a = 0.15$; (e): $N = 1000$ $\mu_\gamma = \mu_a = 0.05$ $\sigma_\gamma = \sigma_a = 0.05$

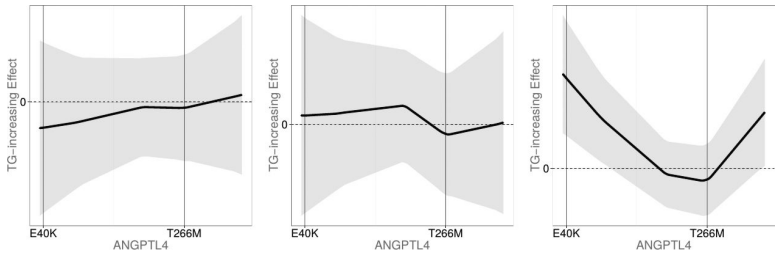

Figure 5.

TG-increasing effect among European Americans (left panel), African Americans (middle panel), and Hispanics (right panel) with the 95% confidence bands (shaded regions).

Table 1

Empirical type I error rates for the association tests between (G) and (X), while adjusting for (Z).

Sample size	Nominal level α	FANOVA	SKAT
50	0.05	0.04319	0.04164
	0.01	0.01037	0.00845
	0.001	0.00191	0.00018
	0.0001	0.00036	0.00000
500	0.05	0.04346	0.04854
	0.01	0.00941	0.01002
	0.001	0.00108	0.00123
	0.0001	0.00023	0.00000