RESEARCH ARTICLE

Genetic Epidemiology
OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Functional Logistic Regression Approach to Detecting Gene by Longitudinal Environmental Exposure Interaction in a Case-Control Study

Peng Wei,[1]* Hongwei Tang,[2] and Donghui Li[2]

[1]*Division of Biostatistics and Human Genetics Center, The University of Texas School of Public Health, Houston, Texas, United States of America;* [2]*Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America*

Received 18 January 2014; Revised 29 May 2014; accepted revised manuscript 29 July 2014.
Published online 0 xxxx 0000 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21852

ABSTRACT: Most complex human diseases are likely the consequence of the joint actions of genetic and environmental factors. Identification of gene-environment (G × E) interactions not only contributes to a better understanding of the disease mechanisms, but also improves disease risk prediction and targeted intervention. In contrast to the large number of genetic susceptibility loci discovered by genome-wide association studies, there have been very few successes in identifying G × E interactions, which may be partly due to limited statistical power and inaccurately measured exposures. Although existing statistical methods only consider interactions between genes and static environmental exposures, many environmental/lifestyle factors, such as air pollution and diet, change over time, and cannot be accurately captured at one measurement time point or by simply categorizing into static exposure categories. There is a dearth of statistical methods for detecting gene by time-varying environmental exposure interactions. Here, we propose a powerful functional logistic regression (FLR) approach to model the time-varying effect of longitudinal environmental exposure and its interaction with genetic factors on disease risk. Capitalizing on the powerful functional data analysis framework, our proposed FLR model is capable of accommodating longitudinal exposures measured at irregular time points and contaminated by measurement errors, commonly encountered in observational studies. We use extensive simulations to show that the proposed method can control the Type I error and is more powerful than alternative ad hoc methods. We demonstrate the utility of this new method using data from a case-control study of pancreatic cancer to identify the windows of vulnerability of lifetime body mass index on the risk of pancreatic cancer as well as genes that may modify this association.

Genet Epidemiol 00:1–14, 2014. © 2014 Wiley Periodicals, Inc.

KEY WORDS: gene-environment interaction; GWAS; functional data analysis; longitudinal exposure; measurement error

Introduction

Most complex human diseases are likely caused by the interplay between genetic and environmental risk factors [Thomas, 2010]. With the rapidly decreasing cost of high-throughput genotyping and next-generation sequencing technology, many large-scale genome-wide association studies (GWAS) and meta-analysis efforts have identified thousands of genetic loci associated with hundreds of human diseases and complex traits [Hindorff et al., 2009]. These findings have provided novel insights into the biological mechanisms of complex disease [Visscher et al., 2012]. Despite the great success of GWAS, a large proportion of the heritability of complex traits remains unexplained, which is often referred to as "missing heritability" [Manolio, 2013; Zuk et al., 2012]. The missing heritability has been attributed to yet to be identified rare genetic variants, gene-gene (G × G) and gene-environment (G × E) interactions. Identification of G × E interactions may contribute to finding the "missing heritability," provide novel insights into the underlying disease mechanism [Wu et al., 2012], improve disease risk prediction [Garcia-Closas et al., 2013], and help develop disease prevention strategies because environmental exposures, which are unlike genetic factors, are often modifiable.

Despite the great promise of G × E interactions, in contrast to the large number of GWAS-identified loci, only a few robustly replicated G × E interactions have been reported in the literature [Hutter et al., 2013; Wu et al., 2012]. The limited success can be attributed to a few challenges facing the G × E investigations. First, although typical GWAS are designed to ensure sufficient power for detecting genetic main effects, straightforward genome-wide scanning of G × E interactions has very limited power [Mukherjee et al., 2012a; Murcray et al., 2011]. In addition to ongoing efforts to increase sample sizes through study consortia collaborations, many new and powerful statistical methods and analysis strategies have been recently proposed to increase the statistical power for detecting genome-wide G × E interactions,

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

*Correspondence to: Peng Wei, Division of Biostatistics and Human Genetics Center, The University of Texas School of Public Health, 1200 Pressler Street, RAS E817, Houston, TX 77030, USA. E-mail: Peng.Wei@uth.tmc.edu including the empirical Bayes method [Mukherjee and Chatterjee, 2008], two-stage methods [Kooperberg and Leblanc, 2008; Murcray et al., 2009], hybrid methods [Gauderman et al., 2013; Hsu et al., 2012; Murcray et al., 2011], and biological pathway based methods [Jiao et al., 2013; Tang et al., 2014a,b], among others [Hutter et al., 2013]. Second, environmental exposures are often not accurately measured. For example, dietary intake data are usually collected from food frequency questionnaires, which can be contaminated by measurement errors [Aschard et al., 2012]. As large collaborative consortia have become the mainstream in genetic epidemiology studies, environmental variables may be measured differently in individual cohorts and thus impose challenges for data harmonization and replication studies [Li et al., 2014]. Third, many environmental exposures are longitudinal and time-varying, for example, air pollution, toxic exposure, and dietary intake, while time-varying exposures in current G × E analysis are often simply treated as static and measured at some arbitrary time point, which may lead to possible bias and loss of power. Although the importance of taking into account time-varying environmental exposures in G × E analysis has been recognized [Bookman et al., 2011; Hutter et al., 2013; Mechanic et al., 2012], almost all statistical methods for G × E analysis, including the aforementioned works, only consider static environmental exposures. A few exceptions include Ko et al. [2013] and Mukherjee et al. [2012b], who proposed novel statistical methods to detect time-varying gene by cumulative environmental exposure interactions for repeatedly measured quantitative traits in longitudinal cohort studies, i.e., G × E × time interactions.

On the other hand, case-control studies are often embedded in large cohort studies, for example, the Framingham Heart Study [Splansky et al., 2007] and Kaiser Permanente cohort [Hoffmann et al., 2011], where the past longitudinal exposure data for both cases and controls may be available. Less commonly, as in the case-control study of pancreatic cancer to be analyzed here, lifetime body weight, thus body mass index (BMI), history was collected retrospectively by personal interview. Although considerable efforts have been devoted to modeling the association between longitudinal exposure and disease risk [Bhadra et al., 2012; Pepe et al., 1999; Sanchez et al., 2011], there is a lack of statistical methods to model and detect the interaction between gene and longitudinal environmental exposure in a case-control study.

To help alleviate this gap, we propose a novel statistical method for testing gene by longitudinal environmental exposure in the functional data analysis (FDA) framework. FDA is a powerful tool to extract informative features from high-dimensional longitudinal data, often contaminated by measurement errors, and achieve substantial dimension reduction [Ramsay and Silverman, 2005; Yao et al., 2005]. This technique has been successfully employed in image analysis [Sorensen et al., 2013], time-course gene expression analysis [Leng and Muller, 2006], and statistical genetics to model the association between rare genetic variants and disease phenotype [Luo et al., 2011]. In our proposed new method, we first reduce the dimension of the longitudinal environmental exposure by the functional principle component analysis (FPCA), taking into account measurement errors, and then model gene by longitudinal environmental exposure interaction by the functional logistic regression (FLR) model [Müller and Stadtmüller, 2005]. The new method is demonstrated by using a pancreatic cancer candidate gene-based case-control study to detect interactions between gene and BMI across an individual's lifespan. Our real data based simulation study shows that the proposed method can satisfactorily control the Type I error and is more powerful than alternative methods.

Materials and Methods

Notations

We consider a case-control study with a total sample size n including n_1 cases and n_2 controls ($n = n_1 + n_2$). Let D_i denote the binary disease status of individual i: 0 for controls and 1 for cases ($i = 1, \ldots, n$). Let Z_i denote the covariate vector, including, for example, sex, age, and leading principle components capturing population substructure. Given a single nucleotide polymorphism (SNP) to be tested for G × E interaction, let G_i denote the genotype of the SNP in subject i, equal to 0, 1, and 2 for major allele homozygotes, heterozygotes, and minor allele homozygotes, respectively. In contrast to traditional G × E analysis in which a static environmental exposure is considered, here for each individual i, we consider time-varying environmental exposure, denoted by $E_i(t_{ij})$, where t_{ij} is the time of the individual i's jth measurement ($j = 1, \ldots, J_i$) and J_i is the individual i's total number of measurements. t_{ij} takes values in the time interval $\mathcal{T} = [T_1, T_2]$, which can be rescaled to the unit interval [0, 1]. For example, in our pancreatic cancer case-control study example to be detailed later on, t_{ij} is the jth age period (in 10-year intervals from age 14 to 19 years until 1 year prior to diagnosis for cases or recruitment into the study for frequency-matched controls) in which the individual i's BMI information was collected (see Bhadra et al. [2012] for a more complete treatment of the time index t_{ij}). Other examples of longitudinal exposures include air pollutant intensity over time, dietary intake over a follow-up period, or lifestyle/host factors such as sex hormone levels and lifetime BMI in the application here. Note that E_i's can be measured at either the same or different time points across individuals, the latter of which is more typical in observational studies.

Standard Logistic Regression for Gene by Longitudinal Environmental Exposure Interaction

Because the standard logistic regression model can only test the interaction between gene and environmental exposures one at a time, it requires the longitudinal environmental exposures to be measured at the same time points across all n individuals, that is, $J_1 = J_2 = \ldots = J_n = J$ and $t_{1j} = t_{2j} = \ldots = t_n$ for all $j = 1, \ldots, J$. Should this not hold, measurement time points need to be grouped into intervals in an ad hoc way, for example, into 5-year intervals [Sanchez et al., 2011]. Measurements in the same interval are then averaged, leading to a single environmental exposure value for each individual. We rewrite $E_i\left(t_{ij}\right)$ as E_{ij} to denote the individual *i*'s exposure measure at the *j*th aligned time point. We fit the following logistic regression model, for $j = 1, \ldots, J$:

$$\operatorname{logit}\left[\Pr\left(D_i = 1|G_i, E_{ij}\right)\right] = \alpha_0 + \beta_Z Z_i + \beta_G G_i + \beta_{E,j} E_{ij} + \beta_{GE,j} G_i \times E_{ij}, \quad (1)$$

under the additive genetic model. Other alternative genetic models, such as the dominant or recessive model, can also be assumed. We test the null hypothesis $H_0: \beta_{GE,1} = \beta_{GE,2} = \ldots = \beta_{GE,J} = 0$ against the alternative H_1 : at least one of $\beta_{GE,j}$ is not 0. This can be done by calculating a *P*-value for each $\beta_{GE,j}$ and taking the minimum of the J *P*-values, denoted as min*P*. We then compare min*P* with the Bonferroni correction significance threshold $0.05/J$. Because the J longitudinal exposures E_{ij}'s are likely to be correlated, the Bonferroni correction, which assumes independent multiple tests, tend to be conservative, leading to reduced statistical power. Alternatively, we can resort to the parametric bootstrap method to obtain the null distribution of min*P* taking into account correlated measurements [Buzkova et al., 2011].

In the standard logistic regression framework, we can also model the longitudinal exposures jointly via

$$\operatorname{logit}\left[\Pr\left(D_i = 1|G_i, E_{ij}s\right)\right] = \alpha_0 + \beta_Z Z_i + \beta_G G_i + \sum_{j=1}^{J} \beta_{E,j} E_{ij} + \sum_{j=1}^{J} \beta_{GE,j} G_i \times E_{ij}. \quad (2)$$

We test the null hypothesis $H_0: \beta_{GE,1} = \beta_{GE,2} = \ldots = \beta_{GE,J} = 0$ against the alternative H_1 : at least one of $\beta_{GE,j}$ is not 0. This seems to be the same as that in Model (1); however, the key difference is that $\beta_{E,j}$ and $\beta_{GE,j}$ in Equation (1) are estimated for each environmental exposure measurement separately, whereas those in Equation (2) are estimated jointly for all measurements. We can employ a J degrees-of-freedom (df) likelihood ratio test (LRT), or its asymptotically equivalent score and Wald tests, to test the above hypothesis, which may suffer from loss of power if the number of measurements J is large [Pan et al., 2011]. Another caveat of Model (2) is that correlated longitudinal exposures may lead to unstable numerical solutions due to multicollinearity.

New Method: FLR and FPCA

The FDA, including the FPCA and functional linear/generalized linear models, has emerged as a powerful approach to modeling noisy and irregularly measured longitudinal data in association with a scalar response variable, for example, disease outcome [Li et al., 2010; Müller and Stadtmüller, 2005]. Here, we propose to model the longitudinal environmental exposure in the FDA framework. First, we decompose the longitudinal exposure trajectory into a few uncorrelated components using the FPCA, taking into account possible measurement errors, and then model gene by longitudinal exposure interaction using the FLR model.

FPCA

We model the *n* individuals' exposure trajectories as independent realizations from a square integrable stochastic process $\{E(t), t \in \mathcal{T}\}$ with mean function $\mu(t)$ and covariance function $R(s, t) = \operatorname{cov}\{E(s), E(t)\}$ in time domain $s, t \in \mathcal{T} = [T_1, T_2]$. By *Mercer's* Theorem [Leng and Muller, 2006], we have eigendecomposition $R(s, t) = \sum_{k=1}^{\infty} \lambda_k \rho_k(s) \rho_k(t)$, where ρ_k and λ_k are eigenfunctions and eigenvalues ordered by size $\lambda_1 \geq \lambda_2 \geq \ldots$. The orthonormal eigenfunctions satisfy $\int_{\mathcal{T}} \rho_j(t) \rho_l(t)\, dt = \delta_{jl}$, which is 1 if $j = l$ and 0 otherwise. By the *Karhunen-Loève* decomposition [Yao et al., 2005], a random curve $E_i(t)$ from the population can be represented by

$$E_i(t) = \mu(t) + \sum_{k=1}^{\infty} FPC_{ik} \rho_k(t), \quad (3)$$

where $FPC_{ik} = \int_{\mathcal{T}} (E_i(t) - \mu(t)) \rho_k(t)\, dt$ is the *k*th FPC score for the *i*th subject. In addition, FPC_{ik} satisfies $E(FPC_{ik}) = 0$ and $Var(FPC_{ik}) = \lambda_k$. The value of FPC_{ik} measures the similarity between the deviation of individual curve $E_i(t)$ from the population mean and the *k*th eigenfunction (FPC) $\rho_k(t)$. The above FPCA framework for functional data is analogous to the representation of random vectors in multivariate analysis by principle components: a random vector can be represented as linear combination of the orthonormal basis defined by the eigenvectors of its covariance matrix, which is the finite-dimensional equivalent of the *Karhunen-Loève* decomposition.

We further assume that we observe the *i*th individual's exposure trajectory $\left\{E_i\left(t_{ij}\right), j = 1, \ldots, J_i\right\}$, contaminated by measurement errors, at J_i time points for $i = 1, \ldots, n$:

$$E_{ij} = E_i\left(t_{ij}\right) + \varepsilon_{ij} = \mu(t) + \sum_{k=1}^{\infty} FPC_{ik} \rho_k(t) + \varepsilon_{ij}, \quad (4)$$

where the measurement error ε_{ij}, independent of $E_i\left(t_{ij}\right)$, has mean 0 and variance σ^2 following the classical measurement error model [Carroll et al., 2006]. The observation time points t_{ij}'s can be either the same across individuals (regular time intervals) or irregular and sparse, the latter of which is often encountered in observational studies.

FPCA by Principle Analysis via Conditional Expectation (PACE)

We propose to employ the PACE method [Yao et al., 2005] to estimate the mean function $\mu(t)$, covariance function $R(s, t)$, eigenfunctions $\rho_k(t)$, and FPC scores FPC_{ik} from the entire observed data $\left\{E_{ij}, i = 1, \ldots, n \text{ and } j = 1, \ldots, J_i\right\}$, including both cases and controls. The PACE method has been shown to be versatile and powerful when applied to sparse and irregularly measured longitudinal data contaminated with measurement errors, as well as regularly measured longitudinal data with possible missing values [Müller, 2009]. Briefly, the PACE method carries out the FPCA as follows. First, all available measurements $\{(t_{ij}, E_{ij}), i = 1, \ldots, n$ and $j = 1, \ldots, J_i\}$ are pooled to form a scatter plot and the estimate $\hat{\mu}(t)$ of the mean function $\mu(t)$ is obtained by a one-dimensional kernel smoother. Second, the estimated covariance function $\hat{R}(s, t)$, for $s, t \in \mathcal{T} = [T_1, T_2]$, is obtained by a two-dimensional kernel smoother with all pairwise products $\left\{E_{ij} - \hat{\mu}\left(t_{ij}\right)\right\}\{E_{il} - \hat{\mu}(t_{il})\}$, for $j \neq l$, as the responses and $\left(t_{ij}, t_{il}\right)$ as the predictors. Third, estimated eigenfunctions $\hat{\rho}_k(t)$ and eigenvalues $\hat{\lambda}_k$ are obtained by applying spectral decomposition to the smoothed covariance surface $\hat{R}(s, t)$ after discretization. Because the smoothed $\hat{R}(s, t)$ may no longer be positive definite, only positive eigenvalues are retained [Yao et al., 2005]. Fourth, in contrast to estimating the FPC scores by numerical integration $FPC_{ik} = \int_{\mathcal{T}}(E_i(t) - \mu(t))\rho_k(t)\,dt$, the PACE method takes measurement errors and sparse measurements into account by assuming FPC_{ik} and ε_{ij} to be jointly normal and predicting the random effects $\widehat{FPC}_{ik}$ based on its conditional expectation: $\widetilde{FPC}_{ik} = \mathrm{E}\left(FPC_{ik}|\tilde{E}_i = \left(E_{i1,\ldots,}E_{iJ_i}\right)'\right)$ (see Yao et al. [2005] for details). Predictions $\widehat{FPC}_{ik}$'s are then obtained by plugging in estimates of the parameters from the entire dataset, borrowing information from all subjects. Finally, it is often the case that the infinite-dimensional stochastic process $\{E(t), t \in \mathcal{T}\}$ can be well approximated by the function space spanned by the leading K eigenfunctions, leading to a truncated version of the *Karhunen-Loève* representation,

$$\hat{E}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{K} \widehat{FPC}_{ik}\hat{\rho}_k(t) \text{ for } t \in [T_1, T_2]. \qquad (5)$$

The choice of K can be based on either the fraction of variance explained or some model selection criteria, such as modified Akaike information criterion (AIC) and Bayesian information criterion (BIC) [Li et al., 2010; Yao et al., 2005]. As demonstrated in Equation (5), the infinite-dimensional trajectory $E_i(t)$ is reduced to K FPC scores $\widehat{FPC}_{ik}$ for each individual. The PACE method is implemented in the Matlab toolbox and R package "PACE."

FLR

We first consider modeling the association between the disease status and longitudinal environmental exposure (BMI in the application here) and then introduce the gene by longitudinal exposure interaction model.

Model the Main Effect of Longitudinal Exposure on the Disease Risk

We assume that the disease status is dependent on the longitudinal exposure via the following FLR model [Müller and Stadtmüller, 2005]:

$$\text{logit}\left[Pr(D_i = 1)\right] = \alpha_0 + \beta_Z Z_i + \int_{\mathcal{T}} \beta(t) E_i(t)\,dt, \qquad (6)$$

where $\beta(t)$ is a time-varying coefficient, implying that a constant unit increase in the longitudinal exposure $E_i(t)$ from time t_1 to t_2 will increase the odds of disease by $\int_{t_1}^{t_2} \beta(t)\,dt$, given other covariates Z_i are fixed. By the *Karhunen-Loève* representation, both $\beta(t)$ and $E_i(t)$ can be expanded by the eigenfunctions ρ_k : $E_i(t) = \mu(t) + \sum_{k=1}^{\infty} FPC_{ik}\rho_k(t)$ and $\beta(t) = \sum_{k=1}^{\infty} \beta_k\rho_k(t)$. Without loss of generality, we assume $\mu(t) \equiv 0$, as $\mu(t)$ is not subject specific and can be absorbed to the intercept α_0 in Model (6). Furthermore, we assume a truncated version of the *Karhunen-Loève* representation involving only the leading K FPCs as in Equation (5). From the orthonormality of the eigenfunctions ρ_k, it follows that Model (6) is equivalent to

$$\text{logit}\left[Pr(D_i = 1)\right] = \alpha_0 + \beta_Z Z_i + \sum_{k=1}^{K} \beta_k FPC_{ik}, \qquad (7)$$

that is, the outcome is dependent only on the leading K FPC scores, which are uncorrelated random effects by construction. We thus propose to first apply the PACE method to obtain the estimated $\widehat{\rho_k(t)}$ and $\widehat{FPC}_{ik}$; second, plug in the estimated FPC scores $\widehat{FPC}_{ik}$ in Model (7) and estimate the regression coefficients β_k as usual logistic regression; and third, obtain the estimate of the time-varying coefficient $\beta(t)$ by $\widehat{\beta(t)} = \sum_{k=1}^{K} \widehat{\beta}_k\widehat{\rho_k(t)}$. We can perform a global test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_K = 0$, via, for example, a K-df LRT, to investigate if there is an overall association between the longitudinal exposure and disease risk. To take into account the uncertainty in the estimated FPC scores $\widehat{FPC}_{ik}$ in Model (7), we can employ the nonparametric bootstrap procedure to obtain the standard error (SE) for the regression coefficients β_k's by resampling the paired observations $\left(D_i, Z_i, E_{i1}, \ldots, E_{iJ_i}\right)'$ with replacement and repeating the PACE procedure and FLR [Li et al., 2010].

Model the Interaction Between Gene and Longitudinal Exposure

To model the interaction between an SNP and the longitudinal exposure, we propose the following FLR model:

$$\text{logit}\left[Pr(D_i = 1)\right] = \alpha_0 + \beta_Z Z_i + \beta_G G_i + \int_{\mathcal{T}} \beta(t) E_i(t)\,dt + \int_{\mathcal{T}} \beta_{GE}(t) E_i(t) G_i dt, \qquad (8)$$

where $\beta(t)$ and $\beta_{GE}(t)$ are time-varying coefficients for the longitudinal exposure main effect and its interaction effect with the SNP G. Similarly to Model (6), we can decompose $\beta(t)$, $\beta_{GE}(t)$, and $E_i(t)$ using the leading K orthonormal eigenfunctions ρ_k and rewrite Model (8) as

$$\text{logit}\left[Pr(D_i = 1)\right] = \alpha_0 + \beta_Z Z_i + \beta_G G_i + \sum_{k=1}^{K} \beta_k FPC_{ik} + \sum_{k=1}^{K} \beta_{Gk} G_i * FPC_{ik}. \qquad (9)$$

Testing the null hypothesis that there is no gene by longitudinal exposure interaction amounts to testing the time-varying interaction coefficient in Model (8) to be 0, that is, $H_0 : \beta_{GE}(t) \equiv 0$ for any $t \in [T_1, T_2]$, which is equivalent to testing $H_0 : \beta_{G1} = \beta_{G2} = \ldots = \beta_{GK} = 0$ in Model (9). The latter can be tested by a K-df LRT or its asymptotically equivalent score or Wald test. However, the LRT may be less powerful if K is large and/or the interaction effect sizes β_{Gk} are moderate. Here, we propose to employ the sum of squared score (SSU) test [Pan, 2009; Pan et al., 2011] to test the global null hypothesis $H_0 : \beta_{G1} = \beta_{G2} = \ldots = \beta_{GK} = 0$. Let $U_{GE} = (U_{G1}, \ldots, U_{GK})'$ denote the efficient score vector for the K interaction terms in Model (9), where $U_{Gk} = \sum_{i=1}^{n} U_{i,Gk} = \sum_{i=1}^{n} (D_i - \hat{p}_i)\left(G_i * FPC_{ik} - \hat{\mu}_{i,Gk}\right)$, $\hat{p}_i$ is the fitted probability of $D_i = 1$ from the main-effect logistic regression model, that is, Model (9) without the interaction terms, and $\hat{\mu}_{i,Gk} = \hat{E}(G_i * FPC_{ik})$ is the fitted response value from the linear regression model $E(G_i * FPC_{ik}) = \alpha_0 + \beta_Z Z_i + \beta_G G_i + \sum_{k=1}^{K} \beta_k FPC_{ik}$. The SSU test statistic is defined as the sum of squared elements of the efficient score vector U_{GE}, that is, $T_{SSU} = U'_{GE} U_{GE}$, which has an asymptotic null distribution of a mixture of χ_1^2's and can be well approximated by a scaled and shifted χ^2 distribution [Pan, 2009]. The SSU test has been shown to be equivalent to the permutation version of Goeman's variance component score test for a random-effects logistic regression model [Goeman et al., 2006] and is closely related to kernel machine regression for SNP-set association test [Pan, 2011; Wu et al., 2010]. We adapt the SSU to test pairwise interaction between an SNP and multiple FPCs, in comparison with testing the SNP-set association or interaction between multiple SNPs and an environmental exposure as originally proposed [Pan, 2009; Pan et al., 2011]. As to be shown in the numerical results, the SSU test was found to outperform other alternative tests.

Data Application: A Case-Control Study of Pancreatic Cancer

Pancreatic cancer is the fourth leading cause of cancer-related deaths for both men and women in the United States with a 5-year survival rate of 6% [American Cancer Society, 2013]. Known risk factors for pancreatic cancer include cigarette smoking, long-term Type 2 diabetes, heavy alcohol consumption, and family history. There is increasing evidence that risk of pancreatic cancer is elevated among individuals who are obese or have high BMI (weight in kilograms divided by height in meters squared). Specifically, we have previously shown that excess BMI in young adulthood confers a higher risk of pancreatic cancer than weight gain at a later age in a case-control study conducted at The University of Texas MD Anderson Cancer Center during 2004–2009 [Li et al., 2009]. Cases were patients with pathologically confirmed pancreatic adenocarcinoma and controls were healthy individuals frequency matched to cases by age, race, and sex. To assess the lifetime BMI influence on pancreatic cancer risk, we collected, by personal interview, height and body weight history of each study participant starting at ages 14–19 years and over 10-year intervals progressing to the year prior to recruitment in the study. The BMI for each individual was then calculated as weight in kilograms at each age period divided by adult body height in meters squared. It is of interest to discover genes that modify the association between changing BMI at different age periods and risk of pancreatic cancer, that is, gene by longitudinal BMI interaction. As pancreatic cancer is a late-onset malignancy with a median age of diagnosis at 71 years [American Cancer Society, 2013], we performed gene by longitudinal BMI interaction analysis on 553 cases and 580 controls who were older than 50 at recruitment and had complete BMI information from age 14 to 19 progressing in 10-year intervals to age 50–59. These individuals were also genotyped for SNPs in susceptibility genes identified in previous GWAS of pancreatic cancer [Amundadottir et al., 2009; Petersen et al., 2010], including *ABO*, *NR5A2*, and *CLTPM1L-TERT*, as well as *FTO*, an obesity-associated gene [Berndt et al., 2013]. Genotyping was performed on genomic DNA from peripheral blood samples using the Taqman method. The study was approved by the institutional review board of The University of Texas MD Anderson Cancer Center.

To analyze this real dataset, we applied the univariate and multiple logistic regression Models (1) and (2) as well as the proposed FLR Model (9), in addition to some simple ad hoc methods previously proposed in the literature [Bhadra et al., 2012]. Specifically, we summarized the longitudinal BMI as a scalar via the following functions and then used the summary exposure in a standard logistic regression interaction model: (a) taking average over time to create average exposure $\overline{E_{i.}} = \frac{1}{J_i} \sum_{j=1}^{J_i} E_{ij}$ (called "aveBMI"), (b) using the maximum exposure over time $E_{i,max} = \max_{1 \leq j \leq J_i} E_{ij}$ (called "maxBMI"), (c) using the area under the longitudinal exposure curve (AUC) $AUC_i = \int_{\mathcal{T}} E_i(t)\, dt$ (called "aucBMI"), which can be approximated by the trapezoidal rule (see p. 122 of Fitzmaurice et al. [2011]), and (4) the first principle component from multivariate PCA of the longitudinal exposures (called "PC1"). We also considered the leading few PCs explaining at least 95% of the total variation in the observed exposures (called "PC95"). Compared with the proposed FLR Models (8) and (9), the ad hoc models (a) and (c) essentially assumed constant coefficient of the exposure and its interaction with the genetic factor over time, while model (b) assumed that the maximum exposure interacted with the genetic factor in modifying the disease risk regardless when the maximum occurred. These assumptions may not hold in practice, as to be shown in the analysis of the pancreatic cancer example. Also of note, multivariate PCA, including "PC1" and "PC95," can only be applied to longitudinal exposures measured at regular time points and cannot accommodate missing values as the FPCA.

Results

FPCA of Longitudinal BMI

We applied the FPCA to the BMI measures from the following five age intervals with cases and controls pooled

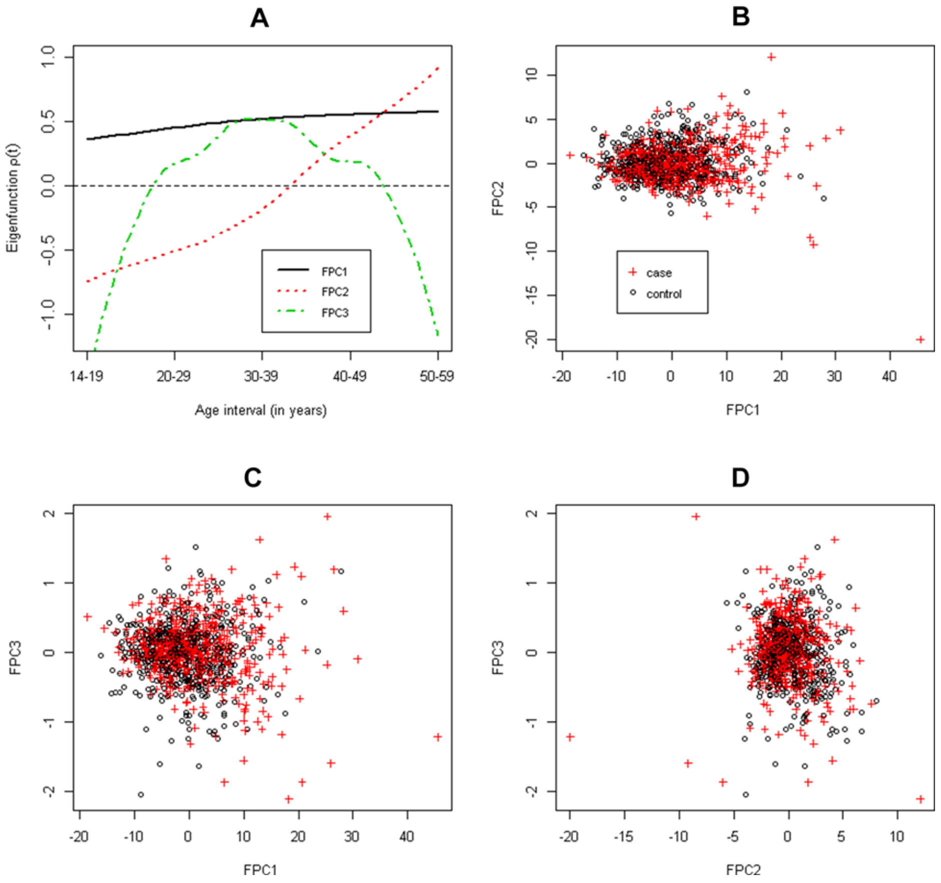

Figure 1. Functional principle components (FPCs) and FPC scores. Panel (A) displays the leading three FPCs; panels (B)–(D) display pairwise FPC scores with circles and crosses corresponding to controls and cases, respectively.

together: 14–19, 20–29, 30–39, 40–49, and 50–59 years. We denote the time domain as $\mathcal{T} = [T_1, T_2] = [1, 5]$, with $t = 1,\ 2,\ 3,\ 4$, and 5 corresponding to each of the five age intervals. Therefore, the BMI was measured on a regular grid for all individuals: $J_i \equiv 5$ for all $i = 1, \ldots, \mathrm{n}$, and $t_{1j} = t_{2j} = \cdots = t_{nj}$ for all $j = 1, \ldots, 5$. As shown in supplementary Figure S1, the raw BMI profiles exhibited substantial variations in both cases and controls. Taking into account possible measurement errors in the recalled BMI, we employed the PACE method to perform the FPCA of the longitudinal BMI pooled from all cases and controls. The kernel smoothed mean function $\hat{\mu}(t)$ and covariance function $\hat{R}(s, t)$ are shown in supplementary Figures S2 and S3. The mean function captured the overall increasing trend of BMI with aging, while the covariance function indicated that the covariance between two BMI measures decreased as they became farther away. The PACE method estimated the measure error σ^2 to be 1.2^2, which suggests that the BMI measure was contaminated by a mean zero random error with a standard deviation of 1.2. In addition, the leading three FPCs, that is, $K = 3$, were selected by a modified BIC, which is the default model selection method in the PACE package. The top three FPCs explained, respectively, 89%, 9.6%, and 1% of the total variation in the observed BMI and explained 99.6% of the total variation collectively (supplementary Fig. S4). Figure 1 shows the first three FPCs along with scatter plots for pairwise FPC scores. The first FPC represented a relative constant vertical shift from the pooled mean curve. The second FPC captured the pattern of initially underweight and eventually overweight in later adulthood, while the third FPC represented overweight in the middle three age intervals (20–29, 30–39, and 40–49 years). Although the majority of the cases and controls were not obviously separated by the first three FPC scores, there were a number of cases with extreme scores, suggesting that the FPCs might extract useful information regarding the disease risk. We also plotted a few randomly selected individuals' observed vs. PACE predicted (Equation (5)) BMI trajectories, as well as for individuals with extreme FPC scores (Fig. 2 and supplementary Fig. S5). We can see that the PACE method predicted the BMI trajectories very well in general and shrunk extreme BMI values toward the population mean for more stable predictions, especially for those individuals with extreme FPC scores (Fig. 2(C) and (D) and supplementary Fig. S5(C) and (D)). On the other hand, as demonstrated in Figure 2(C) and (D), the two individuals with the largest FPC3 and FPC2 score, respectively, matched the corresponding FPCs in Figure 1(A) very well.

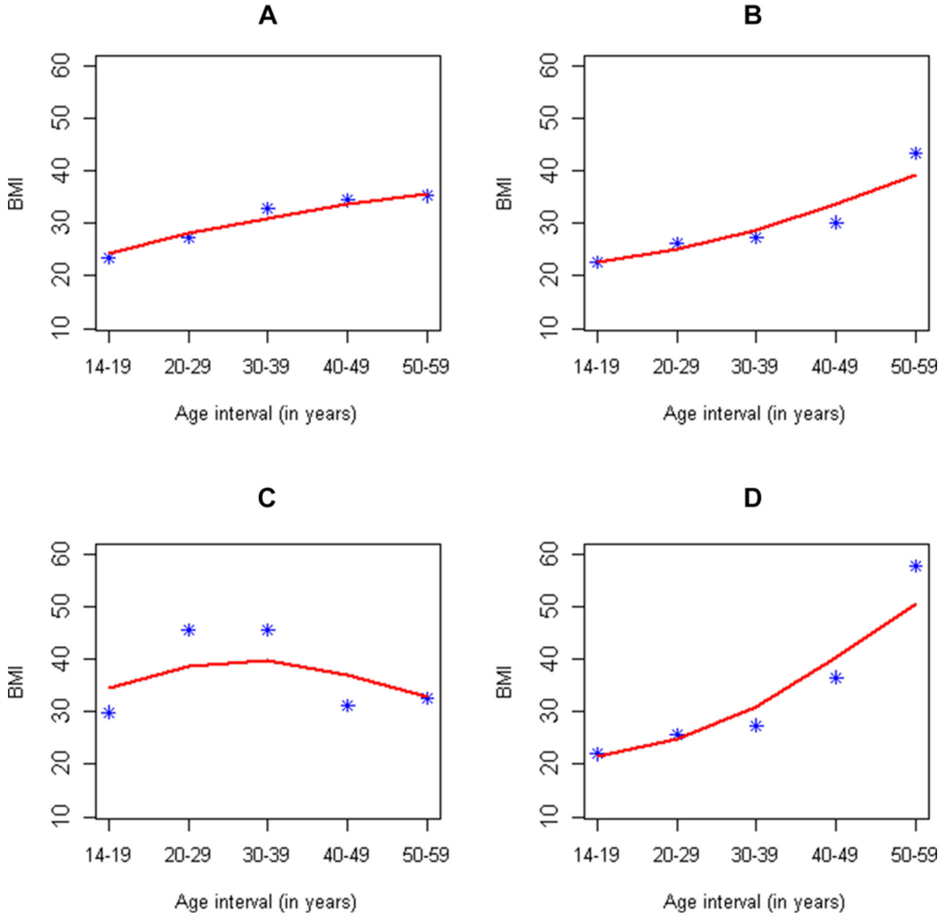

Figure 2. Observed vs. PACE-predicted BMI profiles. Panels (A) and (B) display the BMI profiles of two randomly selected individuals with asterisks and solid lines corresponding to observed and PACE-predicted BMI, respectively. Panel (C) shows an individual with the largest FPC3 score and the third smallest FPC2 score. Panel (D) shows an individual with the largest FPC2 score and the smallest FPC3 score.

Modeling the Main Effect of Longitudinal BMI on Disease Risk by FLR

We first investigated the main effect of lifetime BMI on the pancreatic cancer risk via the FLR (Model (6)). By resorting to the FPCA procedure as described above, we fitted Model (7) as a usual logistic regression with the predicted FPC scores as covariates adjusted for age at recruitment and gender. Supplementary Table S1 shows the estimated regression coefficients, their model-based SEs and 95% confidence intervals (CIs), as well as bootstrap-based SEs and 95% CIs taking into account the uncertainty in predicting the random FPC scores. It is noted that the model- and bootstrap-based SEs were almost identical, as were the 95% CIs, suggesting that the estimation error was likely ignorable in this data example. The first and third FPCs were both significantly and positively associated with the disease risk (P-value = 1.39 × 10^{-9} and 0.048, respectively). The P-value for the LRT of the global null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ was 1.55 × 10^{-9}, indicating that the longitudinal BMI was highly significantly associated with the pancreatic cancer risk. Figure 3 shows the time-varying coefficient for the longitudinal BMI obtained

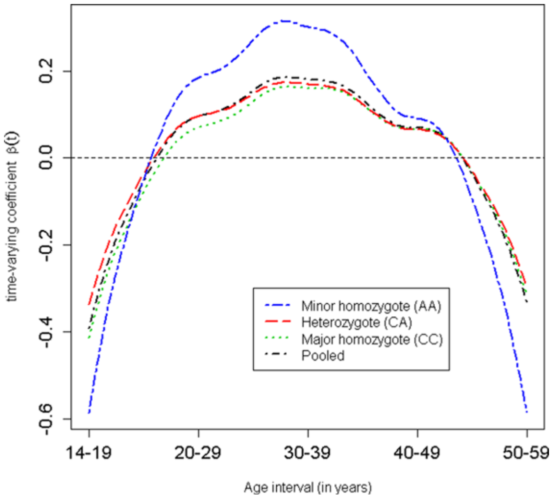

Figure 3. Genotype-specific time-varying coefficient for the longitudinal BMI. The individual curves were obtained by fitting Model (7) stratified by the genotype of *FTO* SNP rs8050136 adjusted for age and gender, while the “pooled” curve was obtained by fitting Model (7) to all the individuals adjusted for age and gender.

by $\widehat{\beta(t)} = \sum_{k=1}^{3} \widehat{\beta_k}\widehat{\rho_k(t)}$, confirming our previous finding that overweight in the age intervals 20–29 and 30–39 years conferred a higher risk of pancreatic cancer than did weight gain at later ages [Li et al., 2009] (the "pooled" curve corresponds to the BMI main effect; other curves to be discussed in the next section). We can estimate the increase in disease odds for a constant unit increase in the BMI from time t_1 to t_2 by exponentiating the AUC $\widehat{\beta(t)}$, that is, $\exp\left(\int_{t_1}^{t_2} \widehat{\beta(t)}dt\right)$. For example, if an individual's BMI increased by 1 unit from age 25 to 35 years given the BMI during other ages was fixed, the odds of developing pancreatic cancer would increase by approximately 12%. We have demonstrated here that the FLR coupled with FPCA is a powerful means to identify the windows of vulnerability of longitudinal exposure on the disease risk [Hutter et al., 2013].

Detecting Gene by Longitudinal BMI Interaction by FLR

To identify genes that may modify the association between longitudinal BMI and pancreatic cancer risk, that is, gene by longitudinal BMI interaction, we applied the FLR interaction Model (9) to SNPs from the candidate gene study of pancreatic cancer. We performed a variety of interaction tests, including FLR-based tests and SNP by age-specific BMI interaction tests, as specified in Models (1) and (2). As shown in Table 1, SNP rs8050136 in *FTO* was found to be nominally interacting with the longitudinal BMI (*P*-value for the FLR-based SSU test = 0.02). Age-specific BMI interaction tests showed that the age interval 20–29 years had the smallest *P*-value = 0.02; however, after adjusting for multiple comparisons, that is, testing five age-specific BMI interactions in total, the SNP by BMI interaction was no longer significant with the parametric bootstrap based min*P* *P*-value = 0.14. Noticeably, the interaction test for rs8050136 by BMI for the last age interval, that is, 50–59 years, was the least significant interval (*P*-value = 0.12), underscoring the importance of considering longitudinal/lifespan environmental exposure in G × E analysis. SNP rs8050136 together with other SNPs in the *FTO* gene has been robustly associated with BMI/obesity in previous GWAS [Berndt et al., 2013]. Supplementary Table S2 shows the regression coefficient estimates in the FLR Model (9) for rs805013; only the first FPC appeared to interact with the SNP (*P*-value = 0.02). The FLR-based SSU test for the global null hypothesis $H_0 : \beta_{G1} = \beta_{G2} = \beta_{G3} = 0$ detected this significant interaction (*P*-value = 0.02), while the FLR-based score test had a nonsignificant *P*-value = 0.10 (Table 1). Our simulation study showed that the SSU test was more powerful than the generic LRT and score test while maintaining the Type I error at the nominal level (to be discussed in the next section). In addition, in a previous study of the same population with a larger sample size of 1,000 case-control pairs of pancreatic cancer, we identified a highly significant interaction between rs8050136 and dichotomized adult BMI (age interval 30–39 years), thereby supporting our findings reported here [Tang et al., 2011]. Similar to the main-effect FLR for the longitudinal BMI, the model-based SEs and 95% CIs in the FLR-based interaction

Table 1. *P*-values for gene by longitudinal BMI interaction

SNP	Position[b]	Gene	BMI[c] (14–19 years)	BMI (20–29 years)	BMI (30–39 years)	BMI (40–49 years)	BMI (50–59 years)	BMI[d] (Bonferroni)	BMI[e] (MinP; parametric bootstrap)	ave BMI[f]	max BMI[f]	auc BMI[f]	PC1[g]	PC95[g]	BMI[h] (Score)	FLR[i] (Score)	FLR[i] (SSU)
rs505922	chr9: 136149229	ABO	0.63[a]	0.28	0.87	0.82	0.67	1.0	0.83	0.84	0.66	0.80	0.79	0.53	0.42	0.61	0.77
rs1558902	chr16: 53803574	FTO	0.11	0.14	0.72	0.33	0.99	0.55	0.46	0.40	0.97	0.38	0.35	0.38	0.06	0.39	0.36
rs8050136	chr16: 53816275	**FTO**	**0.04**	**0.02**	**0.04**	**0.03**	0.12	0.12	0.14	**0.03**	0.09	**0.03**	**0.03**	0.11	0.17	0.10	**0.02**
rs12029406	chr1: 199905828	NR5A2	0.69	0.98	0.15	0.09	0.19	0.44	0.38	0.24	0.12	0.23	0.26	0.38	0.23	0.39	0.18
rs3790844	chr1: 200007432	NR5A2	0.62	0.73	0.35	0.09	0.11	0.43	0.34	0.21	0.12	0.22	0.24	0.41	0.63	0.43	0.23
rs3790843	chr1: 200010824	NR5A2	0.58	0.96	0.22	0.13	0.21	0.65	0.54	0.26	0.17	0.27	0.29	0.61	0.70	0.65	0.32
rs401681	chr5: 1322087	CLPTM1L -TERT	0.42	0.99	0.94	0.95	0.98	1.0	0.94	0.92	0.83	0.97	0.92	0.64	0.88	0.63	1.0

[a] All *P*-values were adjusted for age and gender. Tests with *P*-values less than 0.05 are in bold.
[b] Chromosome number and position (human reference genome hg19).
[c] Gene by age-specific BMI interaction by Model (1).
[d] Adjusted minimum *P*-value (min*P*) for age-specific interactions based on the Bonferroni procedure.
[e] min*P* based on parametric bootstrap.
[f] aveBM, average BMI; maxBMI, maximum BMI; aucBMI, area under the BMI curve.
[g] PC1, the first PC from multivariate PCA; PC95, leading PCs explaining at least 95% of total variation.
[h] Score test *P*-value for gene by longitudinal BMI interaction via multiple logistic regression Model (2).
[i] Functional logistic regression (FLR) based score test and sum of squared score (SSU) test.

model were very close to those based on bootstrapping the "pair" of observations, that is, the pair of disease status and covariates vector, confirming that the estimation error in the PACE-based FPCA was ignorable in this data example. To further investigate the interaction, we fitted the longitudinal BMI main-effect FLR Model (7) stratified by the genotype of rs8050136 and obtained genotype-specific time-varying coefficients for BMI. As shown in Figure 3, the gene by longitudinal BMI interaction mainly occurred in early adulthood, that is, ages 20–29 and 30–39 years, consistent with the results from the age-specific BMI interaction analysis (Table 1). For example, for the same 1 unit increase in BMI from age 25 to 35 years, homozygous minor allele carriers of rs8050136 (AA genotype) had approximately 7% higher odds of developing pancreatic cancer than homozygous major allele carriers (CC genotype).

To evaluate possible confounding effect of cancer-related weight loss in patients, we performed a sensitivity analysis and set the BMI information in the age interval 50–59 years to missing should patients be diagnosed during their 50s (n = 167 cases). Because the PACE-FPCA framework allows missing values, we were able to fit the FLR interaction Model (9) as previously described for complete BMI information. The rs8050136 by longitudinal BMI interaction remained nominally significant by the FLR-based SSU test with a P-value = 0.016, suggesting that our analyses based on complete BMI were unlikely to be confounded by the reverse causality derived from pancreatic cancer associated weight loss.

Also shown in Table 1, two ad hoc summaries of the longitudinal BMI, including "aveBMI" and "aucBMI," were able to detect the rs8050136 by BMI interaction (both P-values = 0.03), while "maxBMI" did not identify it (P-value = 0.09). As the time points were equally spaced and each individual had the same number of measurements in this data example, it is not difficult to see that "aveBMI" was approximately a rescaled version of "aucBMI," which explained why they gave similar results across the SNPs. A closer look revealed that "aveBMI" and "aucBMI" were highly correlated with BMI in the 30s, while "maxBMI" had the highest correlation with BMI in the 50s, consistent with the age-specific BMI interaction analysis results. The first PC from multivariate PCA, explaining 91% of the total variation, had roughly the same loadings across the five BMI measures, and thus, was highly correlated with "aveBMI." Not surprisingly, it was able to identify the rs8050136 by BMI interaction (P-value = 0.03), while the first three PCs, explaining 96% of the total variation, had an insignificant score test P-value = 0.11 due to the extra df. Although the multivariate PCA and FPCA performed similarly in detecting the rs8050136 by BMI interaction, we would like to emphasize that the FPCA is more flexible and powerful in that it can accommodate missing values, both regularly and irregularly measured longitudinal exposures, as well as measurement errors. More importantly, the FPCA coupled with the FLR can not only test for interaction effects, but also estimate the time-varying interaction effects as exemplified by Figure 3.

We also identified significant gender by longitudinal BMI interaction using the proposed FLR model with an SSU test P-value = 0.037 (supplementary Tables S3 and S4). Further analysis of the gender-specific time-varying coefficient for the longitudinal BMI revealed that the interaction was the strongest in the age interval 40–49 years (supplementary Fig. S6), in contrast to age intervals 20–29 and 30–39 years for SNP rs8050136 in the *FTO* gene. Our analyses here exemplify the need to look at the entire range of longitudinal exposure in the G × E analysis [Hutter et al., 2013] and the proposed FLR model coupled with the FPCA thereby provides a powerful tool for such investigations.

Simulation Studies

Simulation I: Type I Error and Power

To further evaluate the proposed method's properties, we performed a simulation study that resembled the pancreatic cancer real-data example. We used the parameter values estimated from the real data when generating the simulated genotype, longitudinal BMI, and disease status. Specifically, following Leng and Muller [2006], the disease status of each subject was simulated based on the following model: $\mathrm{logit}\left[\Pr\left(D_i = 1\right)\right] = \beta_0 + \beta_G G_i + \sum_{k=1}^{3} \beta_k FPC_{ik} + \sum_{k=1}^{3} \beta_{Gk} G_i \times FPC_{ik}$, where G_i was simulated from $\mathrm{Binomial}\left(2, P = 0.4\right)$ to resemble the *FTO* SNP rs8050136 with a minor allele frequency of 40%, and the three FPC scores were independently simulated from normal distributions $FPC_{i1} \sim N\left(0, 7.1^2\right)$, $FPC_{i2} \sim N\left(0, 2.3^2\right)$, and $FPC_{i3} \sim N\left(0, 0.76^2\right)$. We set β_0 to be –4.6, corresponding to a baseline 1% disease prevalence, and other parameters estimated from the pancreatic cancer dataset: $(\beta_G, \beta_1, \beta_2, \beta_3, \beta_{G1}, \beta_{G2}, \beta_{G3})' = (0.04, 0.048, –0.068, 0.25, 0, 0, 0)'$ and $(0.04, 0.02, –0.02, 0.28, 0.03, – 0.03, 0.02)'$ to evaluate the Type I error and power, respectively. The latter parameter values were based on the rs8050136 by longitudinal BMI interaction analysis unadjusted for age and gender (supplementary Table S5). We generated the observed BMI profiles at five equally spaced time points that corresponded to the five age intervals, according to the following model without measurement error, $BMI_{ij} = \hat{\mu}\left(t_{ij}\right) + \sum_{k=1}^{3} FPC_{ik}\hat{\rho}_k\left(t_{ij}\right)$, or the classical measurement error model $BMI_{ij} = \hat{\mu}\left(t_{ij}\right) + \sum_{k=1}^{3} FPC_{ik}\hat{\rho}_k\left(t_{ij}\right) + \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N\left(0, \hat{\sigma}^2\right)$ or $\varepsilon_{ij} \sim N\left(0, (2\hat{\sigma})^2\right)$, $\hat{\sigma} = 1.2$ was the standard deviation for the measurement error estimated from the real data, and $\hat{\mu}\left(t\right)$ and $\hat{\rho}_k\left(t\right)$ were the estimated mean function and FPCs. For each simulation replicate, we generated a large homogeneous study population and then randomly sampled 1,000 cases and 1,000 controls to form a simulated dataset. We applied the proposed FLR coupled with PACE-based FPCA (Model (9)) and age-specific BMI interaction analysis methods (Models (1) and (2)) to the simulated datasets and evaluated the empirical Type I error and power under significance level $\alpha = 0.05$ based on 2,000 replications. To further evaluate the Type I error at a lower α level, for example, 0.001, we also increased the replications to 20,000.

Results from Simulation I

As shown in Table 2, all methods under consideration maintained the Type I error at the nominal level except the unadjusted minimum P-value (min$P$) of the age-specific BMI interaction analysis, which had substantial Type I error inflation. On the other hand, the Bonferroni adjusted min$P$ appeared to have conservative Type I errors, while the parametric bootstrap adjusted min$P$ was less conservative. Increased variances of the measurement error did not affect the Type I error control except for the unadjusted min$P$. The same conclusions held under lower α levels (0.01, 0.001, and 0.0005) based on 20,000 replications (supplementary Table S6).

Table 3 shows the empirical statistical powers at $\alpha = 0.05$. The FLR-based SSU test was the most powerful across different simulation scenarios. Although age-specific BMI interaction analysis in certain age intervals might be more powerful than the FLR-based SSU, the min$P$ tests (Bonferroni-adjusted and parametric bootstrap adjusted to take into account multiple testing) were always dominated by the SSU test. Noticeably, even though the unadjusted min$P$ appeared to have the highest power, it could not control the Type I error (Table 2) and should be excluded from the comparison. As the variance of the measurement error increased, most tests' performance deteriorated; however, the FLR-based SSU test was quite robust to increased measurement error. The FLR-based LRT and score tests had comparable statistical power as expected. Although they were less powerful than the score test in the multiple BMI based logistic regression Model (2) when there was no measurement error in the observed BMI, their power loss was less severe than the latter in the presence of increased measurement errors, suggesting that the proposed FLR model was more robust to contaminated longitudinal exposure measurements. Nevertheless, the FLR-based LRT and score tests were less powerful than the SSU test, which we recommend to use coupled with the proposed FLR model.

Simulation II: Effect of the Number of Longitudinal Measurements

We conducted additional simulations to evaluate the effect of the number of longitudinal measurements on the Type I error and power of the proposed FLR method. Specifically, we followed the setup in Simulation I, but, instead of five time points, we generated the observed BMI at 3, 4, 5, 7, 10, and 20 equally spaced time points between the two endpoints, 1 and 5, and always included these two points. For example, when there were three measurements, each individual's BMI was observed at time points 1, 3, and 5 with measurement error standard deviation $\hat{\sigma} = 1.2$. We used the modified BIC method (the default in PACE) to select the optimal number of FPCs. The empirical Type I error rate and power were based on 1,000 and 200 replications, respectively. As shown in supplementary Table S7, both FLR-SSU and multiple logistic regression fitting of the observed BMI (Model (2)) were able to control the Type I error rate at the nominal level regardless of the number of longitudinal measurements. There was a

Table 2. Empirical Type I error at significance level $\alpha = 0.05$

Measurement error in FPCA simulation model[a]	BMI[c] (14–19 years)	BMI (20–29 years)	BMI (30–39 years)	BMI (40–49 years)	BMI (50–59 years)	BMI MinP[d] (unadjusted)	BMI Min$P$[e] (Bonferroni)	BMI Min$P$[f] (parametric bootstrap)	BMI[g] (Score)	FLR[h] (LRT)	FLR[h] (Score)	FLR[h] (SSU)
Without error	0.051[b]	0.049	0.048	0.047	0.054	**0.103**	0.020	0.025	0.049	0.052	0.052	0.048
With error (σ)	0.047	0.049	0.047	0.047	0.049	**0.121**	0.026	0.031	0.044	0.050	0.050	0.050
With error (2σ)	0.041	0.047	0.046	0.049	0.047	**0.140**	0.031	0.038	0.043	0.052	0.051	0.044

[a] Whether observed BMI was contaminated by measurement error in the simulation model; σ is the estimated standard deviation of measurement error from the real data.
[b] All empirical Type I errors were based on 2,000 simulation replications with 1,000 cases and 1,000 controls in each replicate.
[c] Gene by age-specific BMI interaction by Model (1).
[d] Unadjusted minimum P-value (min$P$) for age-specific interactions. Tests with inflated Type I errors are in bold.
[e] Adjusted min$P$ for age-specific interactions based on the Bonferroni procedure.
[f] Adjusted min$P$ for age-specific interactions based on parametric bootstrap.
[g] Score test for gene by longitudinal BMI interaction via multiple logistic regression Model (2).
[h] Functional logistic regression (FLR) based likelihood ratio test (LRT), score test, and sum of squared score (SSU) test.

Table 3. Empirical statistical power at significance level $\alpha = 0.05$

Measurement error in FPCA simulation model[a]	BMI[c] (14–19 years)	BMI (20–29 years)	BMI (30–39 years)	BMI (40–49 years)	BMI (50–59 years)	BMI Min*P*[d] (unadjusted)	BMI Min*P*[e] (Bonferroni)	BMI Min*P*[f] (parametric bootstrap)	BMI[g] (Score)	FLR[h] (LRT)	FLR[h] (Score)	FLR[h] (SSU)
Without error	0.768[b]	0.842	0.837	0.762	0.579	0.873[i]	0.740	0.754	0.794	0.705	0.708	**0.811**
With error (σ)	0.727	0.815	0.816	0.734	0.570	0.893[i]	0.724	0.736	0.593	0.690	0.690	**0.810**
With error (2σ)	0.633	0.739	0.754	0.658	0.509	0.894[i]	0.711	0.719	0.564	0.650	0.655	**0.792**

[a] Whether observed BMI was contaminated by measurement error in the simulation model; σ is estimated standard deviation of measurement error from the real data.
[b] All empirical powers were based on 2,000 simulation replications with 1,000 cases and 1,000 controls in each replicate.
[c] Gene by age-specific BMI interaction by Model (1).
[d] Unadjusted minimum *P*-value (min*P*) for age-specific interactions.
[e] Adjusted min*P* for age-specific interactions based on the Bonferroni procedure.
[f] Adjusted min*P* for age-specific interactions based on parametric bootstrap.
[g] Score test for gene by longitudinal BMI interaction via multiple logistic regression Model (2).
[h] Functional logistic regression (FLR) based likelihood ratio test (LRT), score test, and sum of squared score (SSU) test. Tests with the highest power are in bold.
[i] Type I error cannot be controlled as shown in Table 2.

3% increase in power when the number of measurements increased from three to four for the proposed FLR-SSU test, while the power remained almost constant with further increased numbers of measurements beyond four. On the other hand, the multiple logistic regression's power severely deteriorated as the number of measurements increased beyond four, because of the multicollinearity among the multiple longitudinal measurements. This simulation study supported that we were able to reliably test for interaction effects with five BMI measurements in the pancreatic cancer dataset using the FLR-SSU test.

Simulation III: Model Robustness

We also performed simulations to evaluate how robust the proposed FLR method was with respect to different model misspecifications, including (1) incorrect number of FPCs, (2) misspecified exposure main effects, and (3) longitudinal BMI not simulated from the FPCA model.

First, we evaluated the Type I error and power when the disease status was dependent on three FPCs under Simulation I, but only the first one or two were selected in the FPCA step. This incorrect number of FPCs was possible due to noise in the data and model selection uncertainty in the first FPCA step. As shown in supplementary Table S8, the FLR model coupled with the Wald test (for one FPC) or SSU test (for two or three FPCs) was able to control the Type I error rate at the nominal level even with incorrect number of FPCs, that is, one or two FPCs; in addition, the power was around 81% regardless of the number of FPCs included, likely because the first FPC had the largest interaction effect size among the three under Simulation I resembling the real data (supplementary Table S5).

Second, we evaluated the effect of misspecified exposure main effects on the Type I error and power. It has been reported in the literature that model misspecifications, especially the exposure main effects, can lead to inflated Type I error in G × E tests [Cornelis et al., 2012; Tchetgen Tchetgen and Kraft, 2011]. We first tested whether a quadratic term of age-specific BMI main effect was needed in Model (1) in the pancreatic cancer dataset, and found no significant quadratic term for any of the five age intervals. We also fitted a generalized spline regression to age-specific BMI. As shown in supplementary Figure S7, BMI in the 30–39 years age interval (denoted by BMI30) appeared to be linearly associated with the logit of disease risk. Therefore, a linear term of age-specific BMI main effect turned out to be adequate in the real-data application. We further carried out simulations to evaluate the robustness of Models (1) and (2) and the proposed FLR model. We simulated longitudinal BMI from the FPCA model as in Simulation I, but the disease status was only dependent on BMI30: $\text{logit}\left[\Pr\left(D_i = 1\right)\right] = \beta_0 + \beta_G G_i + \beta_E BMI30_i + \beta_{E2} BMI30_i^2 + \beta_{GE} G_i \times BMI30_i + \beta_{GE2} G_i \times BMI30_i^2$. The simulation details and results are described in Supplemental Text Section 3.1. In summary, we confirmed the previous finding reported in the literature that misspecified main effect of BMI30 led to inflated Type I error rate; in addition, we found that, for age-specific BMI other than BMI30, even the interaction Model (1) with quadratic term of BMI main effect had inflated Type I error. In contrast, the proposed FLR model controlled the Type I error satisfactorily and remained high power without explicitly incorporating quadratic main or interaction effects. Although the FLR model appeared to be robust to nonlinear interaction effects, we would like to point out that the FLR model was developed to estimate and test for time-varying main and interaction effects, that is, for datasets with longitudinal exposure within the same individual, rather than for modeling nonlinear effects of a static exposure, for example, BMI30, across individuals. Nevertheless, because multiple FPCs were included in the FLR model, some nonlinear effects of the exposure at a given time point might be captured, leading to the appealing model robustness.

Finally, we simulated the longitudinal BMI and disease status from non-FPCA/FLR model to further evaluate the model robustness. As detailed in Supplemental Text Section 3.2, we resampled the observed rs8050136 genotype and longitudinal BMI pairs with replacement from the pooled cases and controls in the pancreatic cancer dataset, and generated the disease status based on the real data fitted multiple logistic regression Model (2). As shown in supplementary Table S10, the comparative performance of the different tests was qualitatively the same as that in Simulation I, and the FLR-SSU test remained the most powerful one.

Discussion

We have proposed a novel FLR-based statistical framework for detecting gene by longitudinal environmental exposure interactions. The proposed two-stage approach first summarizes the longitudinal exposure into a few FPCs via the PACE method that takes into account possible measurement errors, and then tests for interaction between an SNP and FPCs in an FLR model. Using data from a case-control study of pancreatic cancer and real data based simulations, we demonstrated that the SSU test nested in the FLR model was more powerful than alternative methods. In addition, the SSU test was found to be robust to measurement errors in the longitudinal environmental exposure. Although in our real-data example the environmental exposure was measured on a regular time grid, the proposed method can be equally applied to irregularly measured time-varying exposures as commonly seen in observational studies. Because the first-stage PACE-based FPCA only needs to be applied once for all SNPs, the proposed method can be easily scaled up to genome-wide scale interaction scan. Therefore, we recommend the use of the FLR model coupled with the SSU test to detect gene by longitudinal exposure interactions.

It has been well recognized that early-life exposures, for example, maternal, childhood, or adolescence exposures, may be critical to disease occurrence later in life [Hutter et al., 2013; Sutcliffe and Colditz, 2013]. Therefore, it is important to consider long-term exposure history beyond the immediate short window prior to disease onset to identify the critical window of disease development. Powerful statistical methods have also recently been proposed to detect time-varying gene by cumulative environmental exposure interactions for repeated-measured quantitative phenotypes, that is, G × E × time interactions [Ko et al., 2013; Mukherjee et al., 2012b]. To the best of our knowledge, the proposed FLR method described here is the first to detect gene by longitudinal environmental exposure interactions in a case-control study. It would be of interest to extend the current framework to quantitative and longitudinal phenotypes, for example, via the functional linear model for functional response variable [Müller, 2009].

Recently it has been demonstrated that robustly identified G × E interactions can improve disease risk prediction and help develop intervention strategies [Garcia-Closas et al., 2013]. It remains a great challenge to develop effective prevention and intervention strategies for pancreatic cancer in the general population due to its low incidence rate and a poor understanding of the disease etiology. However, the identified *FTO* by early adulthood obesity interaction, once replicated in independent studies, may hold the promise to be incorporated into and improve the existing risk prediction model for pancreatic cancer [Klein et al., 2013]. The risk model could then be used to identify high-risk individuals, for example, those who are smokers and have family history of pancreatic cancer, minor allele homozygote of rs8050136, and obesity in early adulthood, for targeted interventions, such as participation in weight loss and smoking cessation programs.

To summarize the longitudinal exposure, we employed the FPCA that is conceptually similar to the multivariate PCA. Both methods are used for dimension reduction and have been applied in genetic studies. For example, the PCA has been used to control population stratification and summarize multiple SNPs in a region in SNP-set association tests [Wei et al., 2012], while the FPCA has been employed in rare variant association tests to model rare mutation profiles [Luo et al., 2011]. The PCA and FPCA, however, differ in the many aspects. First, the FPCA models functional data with a time or spatial domain, and, thus, measurements at different time points, unlike in PCA, are not exchangeable. Second, the FPCA is often coupled with some smoothing procedures to take into account measurement errors by borrowing information from neighboring measurements, for example, either smoothing the mean and covariance functions as in the PACE method, or smoothing the raw longitudinal trajectories as advocated by others [Ramsay and Silverman, 2005]. Third, the FPCA can be applied to longitudinal data measured at either regular or irregular time points and can accommodate missing data, while the PCA can only be applied to regularly measured data without missing values.

In the FPCA-FLR framework, we propose to perform the FPCA on the combined case-control samples rather than only control samples for two main reasons. First, ignoring disease status in the FPCA step avoids differential estimation errors between cases and controls as well as using the disease status twice, and thus ensures that the Type I error rate is maintained at the nominal level in the subsequent FLR association test step. This is in line with the "EG2" two-step G × E test of Murcray et al. [2009] and multivariate PCA-based SNP-set association test of case-control data [Wei et al., 2012]. Second, it is necessary to perform the FPCA on the pooled samples in order to extract longitudinal exposure patterns, that is, the FPCs, which may be only present in the cases, but nevertheless are relevant with the disease association and interaction with the genetic factor. Another related concern is that, given the FPCA-FLR method was only applied to cohort data before [Müller and Stadtmüller, 2005; Yao et al., 2005], whether we can apply it to outcome-dependent sampling schemes, such as the case-control study design. Intuitively, if we consider the FPCA as a means of applying dimension reduction to the covariates prior to association testing, it should not matter whether we have cohort or case-control data. As for effect size estimation, it has been well known that valid estimation of the odds ratio and its asymptotic variance can be obtained by applying the logistic regression model to case-control data as if the sampling were perspective [Prentice and Pyke, 1979]. Considering this result is established for simple exposure variables and the FPC scores in the FLR are nonlinear transformations of the original longitudinal exposures, we resorted to simulations to address the above concern. As detailed in Supplemental Text Section 3.3, we simulated cohort samples with disease prevalence around 50%, each consisting of roughly 1,000 cases and 1,000 controls, as well as case-control studies of 1,000 case-control pairs with baseline disease prevalence at 1%. We found that, between the cohort and case-control designs, the FLR model gave very similar estimates of the regression coefficients in terms of bias and variance (supplementary Table S11), supporting the validity of applying the FLR model to case-control data.

There are some potential limitations in our proposed statistical framework. First, the proposed method is two-stage: the first-stage FPCA is only applied to the pooled longitudinal exposures without considering the disease status and genetic data. Although this approach has the advantages of maintaining the Type I error rate in the second-stage interaction test, generating data-adaptive and interpretable basis functions, that is, the FPCs, and facilitating genome-wide interaction analysis, it is possible that the extracted FPCs contain irrelevant information about the disease status or G × E interactions, leading to reduced statistical power. It would be of interest to develop computationally efficient one-step methods for detecting gene by longitudinal exposure interactions. Second, the longitudinal exposure data in our real-data example is recalled BMI information collected by personal interviews, which, however, may be contaminated by recall bias and other measurement errors. Although our proposed method can accommodate the classical measurement errors [Carroll et al., 2006] and previous studies have found a high level of accuracy of self-reported past body weights compared with measured weights [Casey et al., 1991; Stevens et al., 1990], there is a lack of validation data for this study population and possible recall bias inherently associated with the case-control design cannot be excluded. On the other hand, in nested case-control studies within prospective longitudinal cohorts, such as the Framingham Heart Study [Splansky et al., 2007] and Kaiser Permanente cohort of nearly 100,000 individuals with both GWAS and longitudinal electronic health records [Hoffmann et al., 2011], exposure variables are more accurately measured and unlikely prone to recall bias. A recent study was able to investigate the association between childhood heights from age 8 to 13 years and adult prostate cancer risk in over 125,000 individuals by linking the Copenhagen School Health Records Register prospective cohort and the Danish Cancer Register [Cook et al., 2013]. Finally, as the research community increasingly appreciates the importance of exposure history across one's lifespan on the risk of complex disease and its interaction with genetic factors [Hutter et al., 2013; Sutcliffe and Colditz, 2013], more and more studies, such as the National Children's Study, are collecting accurately measured time-varying exposure information, thereby providing unprecedented opportunities for investigation into the complex interplay between genes and longitudinal environmental exposures.

R programs implementing the proposed FLR method will be posted on our website at: https://sites.google.com/site/utpengwei/

Acknowledgments

This research was supported by NIH grants R01CA169122, R01HL116720, and R01HL106034 (to P.W.); and a supplemental grant to R01CA98380 and the Sheikh Ahmed Center for Pancreatic Cancer Research Funds (to D.L.). We thank Dr. Wei Pan for helpful discussions and Ms. Sara Barton for editorial assistance. We are grateful to two anonymous reviewers for their many helpful and constructive comments that improved the presentation of the paper. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper.

The authors declare no conflict of interest.

References

American Cancer Society. 2013. *Cancer Facts & Figures 2013.* Atlanta: American Cancer Society.

Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, Bueno-de-Mesquita HB, Gross M, Helzlsouer K, Jacobs EJ and others. 2009. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet* 41(9):986–990.

Aschard H, Lutz S, Maus B, Duell EJ, Fingerlin TE, Chatterjee N, Kraft P, Van Steen K. 2012. Challenges and opportunities in genome-wide environmental interaction (GWEI) studies. *Hum Genet* 131(10):1591–1613.

Berndt SI, Gustafsson S, Magi R, Ganna A, Wheeler E, Feitosa MF, Justice AE, Monda KL, Croteau-Chonka DC, Day FR and others. 2013. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet* 45(5):501–512.

Bhadra D, Daniels MJ, Kim S, Ghosh M, Mukherjee B. 2012. A Bayesian semiparametric approach for incorporating longitudinal information on exposure history for inference in case-control studies. *Biometrics* 68(2):361–370.

Bookman EB, McAllister K, Gillanders E, Wanke K, Balshaw D, Rutter J, Reedy J, Shaughnessy D, Agurs-Collins T, Paltoo D and others. 2011. Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop. *Genet Epidemiol* 35(4):217–225.

Buzkova P, Lumley T, Rice K. 2011. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Ann Hum Genet* 75(1):36–45.

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu C. 2006. *Measurement Error in Nonlinear Models: A Modern Perspective.* New York: Chapman & Hall.

Casey VA, Dwyer JT, Berkey CS, Coleman KA, Gardner J, Valadian I. 1991. Long-term memory of body weight and past weight satisfaction: a longitudinal follow-up study. *Am J Clin Nutr* 53(6):1493–1498.

Cook MB, Gamborg M, Aarestrup J, Sorensen TI, Baker JL. 2013. Childhood height and birth weight in relation to future prostate cancer risk: a cohort study based on the copenhagen school health records register. *Cancer Epidemiol Biomarkers Prev* 22(12):2232–2240.

Cornelis MC, Tchetgen EJ, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. 2012. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. *Am J Epidemiol* 175(3):191–202.

Fitzmaurice GM, Laird NM, Ware JH. 2011. *Applied Longitudinal Analysis*. Hoboken, NJ: Wiley.

Garcia-Closas M, Rothman N, Figueroa JD, Prokunina-Olsson L, Han SS, Baris D, Jacobs EJ, Malats N, De Vivo I, Albanes D and others. 2013. Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer Res* 73(7):2211–2220.

Gauderman WJ, Zhang P, Morrison JL, Lewinger JP. 2013. Finding novel genes by testing G x E interactions in a genome-wide association study. *Genet Epidemiol* 37(6):603–13.

Goeman JJ, van de Geer S, van Houwelingen HC. 2006. Testing against a high dimensional alternative. *J R Stat Soc B* 68:477–493.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106(23):9362–9367.

Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, Cawley S, Chung E, Connell S, Eshragh J and others 2011. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* 98(2):79–89.

Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. 2012. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet Epidemiol* 36(3):183–194.

Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM. 2013. Gene-Environment Interactions in cancer epidemiology: A National Cancer Institute Think Tank Report. *Genet Epidemiol* 37(7):643–657.

Jiao S, Hsu L, Bezieau S, Brenner H, Chan AT, Chang-Claude J, Le Marchand L, Lemire M, Newcomb PA, Slattery ML and others. 2013. SBERIA: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genet Epidemiol* 37(5):452–464.

Klein AP, Lindstrom S, Mendelsohn JB, Steplowski E, Arslan AA, Bueno-de-Mesquita HB, Fuchs CS, Gallinger S, Gross M, Helzlsouer K and others. 2013. An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. *PLoS ONE* 8(9):e72311.

Ko YA, Saha-Chaudhuri P, Park SK, Vokonas PS, Mukherjee B. 2013. Novel likelihood ratio tests for screening gene-gene and gene-environment interactions with unbalanced repeated-measures data. *Genet Epidemiol* 37(6):581–591.

Kooperberg C, Leblanc M. 2008. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol* 32(3):255–263.

Leng X, Muller HG. 2006. Classification using functional data analysis for temporal gene expression data. *Bioinformatics* 22(1):68–76.

Li D, Morris JS, Liu J, Hassan MM, Day RS, Bondy ML, Abbruzzese JL. 2009. Body mass index and risk, age of onset, and survival in patients with pancreatic cancer. *JAMA* 301(24):2553–2562.

Li Y, Wang N, Carroll RJ. 2010. Generalized functional linear models with semiparametric single-Index interactions. *J Am Stat Assoc* 105(490):621–633.

Li S, Mukherjee B, Taylor JM, Rice KM, Wen X, Rice JD, Stringham HM, Boehnke M. 2014. The role of environmental heterogeneity in meta-analysis of gene-environment interactions with quantitative traits. *Genet Epidemiol* 38(5):416–429.

Luo L, Boerwinkle E, Xiong M. 2011. Association studies for next-generation sequencing. *Genome Res* 21(7):1099–1108.

Manolio TA. 2013. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* 14(8):549–558.

Mechanic LE, Chen HS, Amos CI, Chatterjee N, Cox NJ, Divi RL, Fan R, Harris EL, Jacobs K, Kraft P and others. 2012. Next generation analytic tools for large scale genetic epidemiology studies of complex diseases. *Genet Epidemiol* 36: 22–35.

Mukherjee B, Chatterjee N. 2008. Exploiting gene-environment independence for analysis of case-control studies: an empirical Bayes-type shrinkage estimator to trade-off between bias and efficiency. *Biometrics* 64(3):685–694.

Mukherjee B, Ahn J, Gruber SB, Chatterjee N. 2012a. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am J Epidemiol* 175(3):177–190.

Mukherjee B, Ko YA, Vanderweele T, Roy A, Park SK, Chen J. 2012b. Principal interactions analysis for repeated measures data: application to gene-gene and gene-environment interactions. *Statist Med* 31(22):2531–2551.

Müller H-G. 2009. Functional Modeling of Longitudinal Data. In: Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, editors. *Longitudinal Data Analysis (Handbooks of Modern Statistical Methods)*. New York: Wiley, pp. 223–252.

Müller H-G, Stadtmüller, U. 2005. Generalized functional linear models. *Ann Statist* 33(2):774–805.

Murcray CE, Lewinger JP, Gauderman WJ. 2009. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol* 169(2):219–226.

Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. 2011. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol* 35(3):201–210.

Pan W. 2009. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 33(6):497–507.

Pan W. 2011. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet Epidemiol* 35:211–216.

Pan W, Basu S, Shen X. 2011. Adaptive tests for detecting gene-gene and gene-environment interactions. *Hum Hered* 72(2):98–109.

Pepe MS, Heagerty P, Whitaker R. 1999. Prediction using partly conditional time-varying coefficients regression models. *Biometrics* 55(3):944–950.

Petersen GM, Amundadottir L, Fuchs CS, Kraft P, Stolzenberg-Solomon RZ, Jacobs KB, Arslan AA, Bueno-de-Mesquita HB, Gallinger S, Gross M and others. 2010. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet* 42(3):224–228.

Prentice RL, Pyke R. 1979. Logistic disease incidence models and case-control studies. *Biometrika* 66(3):403–411.

Ramsay JO, Silverman BW. 2005. *Functional Data Analysis*. New York: Springer-Verlag.

Sanchez BN, Hu H, Litman HJ, Tellez-Rojo MM. 2011. Statistical methods to study timing of vulnerability with sparsely sampled data on environmental toxicants. *Environ Health Persp* 119(3):409–415.

Sorensen H, Goldsmith J, Sangalli LM. 2013. An introduction with medical applications to functional data analysis. *Stat Med* 32(30):5222–5240.

Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, D'Agostino RB, Sr., Fox CS, Larson MG, Murabito JM and others. 2007. The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* 165(11):1328–1335.

Stevens J, Keil JE, Waid LR, Gazes PC. 1990. Accuracy of current, 4-year, and 28-year self-reported body weight in an elderly population. *Am J Epidemiol* 132(6):1156–1163.

Sutcliffe S, Colditz GA. 2013. Prostate cancer: is it time to expand the research focus to early-life exposures? *Nat Rev Cancer* 13(3):208–518.

Tang H, Dong X, Hassan M, Abbruzzese JL, Li D. 2011. Body mass index and obesity- and diabetes-associated genotypes and risk for pancreatic cancer. *Cancer Epidemiol Biomarkers Prev* 20(5):779–792.

Tang H, Wei P, Duell EJ, Risch HA, Olson SH, Bueno-de-Mesquita HB, Gallinger S, Holly EA, Petersen GM, Bracci PM and others. 2014a. Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene and pathway-based interaction analysis of GWAS data. *Carcinogenesis* 35:1039–1045.

Tang H, Wei P, Duell EJ, Risch HA, Olson SH, Bueno-de-Mesquita HB, Gallinger S, Holly EA, Petersen GM, Bracci PM and others. 2014b. Genes-environment interactions in obesity- and diabetes-associated pancreatic cancer: a GWAS data analysis. *Cancer Epidemiol Biomarkers Prev* 23(1):98–106.

Tchetgen Tchetgen EJ, Kraft P. 2011. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. *Epidemiology* 22(2):257–261.

Thomas D. 2010. Gene–environment-wide association studies: emerging approaches. *Nat Rev Genet* 11(4):259–272.

Visscher PM, Brown MA, McCarthy MI, Yang J. 2012. Five years of GWAS discovery. *Am J Hum Genet* 90(1):7–24.

Wei P, Tang H, Li D. 2012. Insights into pancreatic cancer etiology from pathway analysis of genome-wide association study data. *PLoS ONE* 7(10):e46887.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 86(6):929–942.

Wu C, Kraft P, Zhai K, Chang J, Wang Z, Li Y, Hu Z, He Z, Jia W, Abnet CC and others. 2012. Genome-wide association analyses of esophageal squamous cell carcinoma in Chinese identify multiple susceptibility loci and gene-environment interactions. *Nat Genet* 44(10):1090–1097.

Yao F, Müller HG, Wang JL. 2005. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc* 100:577–590.

Zuk O, Hechter E, Sunyaev SR, Lander ES. 2012. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci USA* 109(4):1193–1198.