HHS Public Access

Author manuscript

J Am Stat Assoc. Author manuscript; available in PMC 2018 May 16.

Published in final edited form as:

J Am Stat Assoc. 2017 ; 112(519): 966–978. doi:10.1080/01621459.2016.1252266.

Set-Based Tests for the Gene–Environment Interaction in Longitudinal Studies

Zihuai He[1], **Min Zhang**[1], **Seunggeun Lee**[1], **Jennifer A. Smith**[2], **Sharon L.R. Kardia**[2], **Ana V. Diez Roux**[3], and **Bhramar Mukherjee**[1]

[1]Department of Biostatistics, University of Michigan

[2]Department of Epidemiology, University of Michigan

[3]Department of Epidemiology, Drexel University

Abstract

We propose a generalized score type test for set-based inference for gene-environment interaction with longitudinally measured quantitative traits. The test is robust to misspecification of within subject correlation structure and has enhanced power compared to existing alternatives. Unlike tests for marginal genetic association, set-based tests for gene-environment interaction face the challenges of a potentially misspecified and high-dimensional main effect model under the null hypothesis. We show that our proposed test is robust to main effect misspecification of environmental exposure and genetic factors under the gene-environment independence condition. When genetic and environmental factors are dependent, the method of sieves is further proposed to eliminate potential bias due to a misspecified main effect of a continuous environmental exposure. A weighted principal component analysis approach is developed to perform dimension reduction when the number of genetic variants in the set is large relative to the sample size. The methods are motivated by an example from the Multi-Ethnic Study of Atherosclerosis (MESA), investigating interaction between measures of neighborhood environment and genetic regions on longitudinal measures of blood pressure over a study period of about seven years with 4 exams.

Keywords

Gene-environment independence; Generalized score test; MESA neighborhood study; Model misspecification; Robustness

1 Introduction

Most complex traits have a multifactorial etiology involving the dynamic interplay of genes and environmental exposures over the life course. Studies of gene-environment interaction (GEI) often suffer from single one time measurement of exposure or a crude proxy thereof, without proper characterization of lifetime history of cumulative exposure. Longitudinal studies with time varying measures of outcome and exposure data help with characterizing the temporal features of exposure and outcomes, handling exposure measurement error and often enhance power when compared to a cross-sectional analysis. While environmental factors considered in an epidemiological analysis are often behavioral factors like diet, physical activity, use of tobacco or alcohol, in recent years, there has been an increasing

interest in measuring the neighborhood environment that the individual lives in. For example, the MESA neighborhood Study, an ancillary study to the Multi-Ethnic Study of Atherosclerosis (MESA), includes a set of novel time varying measures of healthy food availability and access to recreational facilities. Previous studies have shown that individuals living in neighborhoods with better food and physical activity environments are less likely to develop hypertension (Kaiser et al. (2015)). In the present analysis, we are primarily interested in investigating whether a set of single nucleotide polymorphisms (SNPs) measured in a genome-wide association study modifies the effect of neighborhood exposures on longitudinal measures of blood pressure.

Gene-environment interaction is often statistically assessed by fitting a regression model for the quantitative outcome (Y) by including the main effects and a product between a genetic variant (G) and an environmental exposure (E), adjusting for covariates (X). A typical genome-wide interaction search repeats the test for interaction under this model for millions of SNPs, adjusting for multiple comparison. Although numerous single SNP based analyses for gene-environment interaction have been conducted, relatively few of the findings have been replicated because of various reasons such as: limited statistical power due to the burden of multiple comparison; measurement error and misclassification of exposure; detection of spurious interactions due to not properly adjusting for main effect of E and G (for example due to missing a non-linear terms in a continuous exposure E) (Thomas (2010); Tchetgen Tchetgen and Kraft (2011); Mukherjee et al. (2012); Cornelis et al. (2012); Boonstra et al. (2016)).

To improve power and to reduce the burden of multiple comparison, many genetic association studies have now considered an alternate or supplementary analytic approach towards jointly testing the effect of all SNPs in a biologically defined set, such as a gene, pathway or specific genomic region as opposed to a one-at-a-time single SNP analysis. Aggregation of SNPs is particularly critical for studies of rare variants (Derkach et al. (2014); Basu and Pan (2011)). A number of methods have gained popularity including kernel machine regression methods (Wu et al. (2011)), similarity regression (Tzeng et al. (2011)), sum of squared score test (Pan (2009)) and genetic random field model (He et al. (2014); He et al. (2015)). In the context of testing gene-gene/gene-environment interaction for cross-sectional studies, Tzeng et al. (2011), Li et al. (2012), Lin et al. (2013), Chen et al. (2014), Marceau et al. (2015) and Lin et al. (2016) extended the set-based tests for marginal associations to testing interactions. These papers demonstrated superior power of set-based tests for gene-environment interaction by aggregating signals across multiple SNPs. However, no set-based test for gene-environment interaction has been proposed for longitudinal studies where improved power regarding gene-environment interaction is possible by using longitudinally varying outcome and exposure trajectories.

Most GEI studies consider a linear main effect of E. A growing body of literature has shown that a misspecified main effect of E can lead to type I error inflation in tests for gene-environment interaction, and gene-environment independence in the underlying population plays an important role in reducing the detection of spurious gene-environment interactions (Tchetgen Tchetgen and Kraft (2011); Voorman et al. (2011); Cornelis et al. (2012)). However, the theoretical justification for this result has not been established (VanderWeele et

al. (2013)). Also, there is no method proposed for handling misspecified E effect when G and E are dependent, particularly for set-based analysis. For the main effect of G, Lin et al. (2013) pointed out that single SNP analyses for gene-environment interaction can be biased due to ignoring SNPs in the same region that are in linkage disequilibrium (LD) with the tested SNP. Set-based analysis can serve as a potential remedy to this issue, but one practical challenge that is new to deriving set-based tests for GEI is that the null model contains main effects of multiple SNPs and fitting the null model could potentially be problematic when the number of SNPs in a region is large relative to the sample size. The tests can suffer from type I error inflation as the asymptotic distributional properties of the reference test statistic may not hold under such situations.

In this article, we propose a new statistical approach to test for gene-environment interactions with a set of genetic variants and longitudinally measured outcome and exposure data. The test is robust to misspecification of within subject correlation and is substantially more powerful than an analysis that uses subject-specific averages/summaries of outcome and exposure data. We show that the proposed test is robust to the misspecification of E and G main effects under the gene-environment independence condition. We further propose using the method of sieves to flexibly model the main effect of E for improved type I error control when the gene-environment independence condition does not hold, and for better power. We also proposed a weighted principal component analysis (PCA) to remedy the curse of dimensionality when the number of SNPs in the tested set is close to or larger than the sample size. We illustrate the proposed methods by both an analysis of targeted GEI (restricted to genetic regions defined around previous GWAS hits) and an agnostic genome-wide gene-based GEI search, with novel time-varying neighborhood features of the environment as exposure, and blood pressure as the longitudinally measured outcome in MESA. Extensive simulation studies, designed to mimic the data structure of MESA are conducted to assess the operating characteristics of the different methods.

2 Application: Multi-Ethnic Study of Atherosclerosis

MESA was initiated in the year 2000 with the goal of investigating the prevalence, correlates and progression of subclinical cardiovascular disease (Bild et al. (2002)). A total of 6360 MESA subjects who consented to genetic analyses, including 2526 European Americans (EUR), 1611 African Americans (AFA), 1448 Hispanics (HIS) and 775 Asian of Chinese descent (CHN), were included in the current analysis. From 2000 to 2007, four examinations were conducted at approximately 1.5–2 year intervals for participants residing at six study sites: New York, New York; Baltimore, Maryland; Forsyth County, North Carolina; Chicago, Illinois; St Paul, Minnesota; and Los Angeles, California. Blood pressure measurements were available at each MESA exam. An ancillary study of MESA, the MESA neighborhood study, collected longitudinal information on neighborhood characteristics in the four examinations, including four time varying measures of healthy food availability and physical activity resources (Moore et al. (2008); Christine et al. (2015)). These neighborhood environments may influence individual diet and exercise levels, and therefore influence risk factors for chronic diseases, e.g. systolic/diastolic blood pressure (Mujahid et al. (2008)).

The four neighborhood measures include two geographic information system (GIS) based measures and two survey based measures: 1. Density of favorable food stores (GIS-based); 2. Density of recreational facilities (GIS-based); 3. Perceived healthy foods availability (survey-based); 4. Perceived walkability (survey-based). The GIS measures were constructed using the National Establishment Time Series (NETS) database from Wall and Associates for 2000 to 2007 on food stores and commercially-available recreational facilities for every ZIP code within a 5 miles radius of MESA participant households. The survey based measures of healthy food availability and walkability were obtained from questionnaires administered to MESA participants and supplementary sample of other community residents. The detailed description of these neighborhood features can be found in section 4.1 the Supplementary Materials. A growing body of literature has suggested that altering these neighborhood environments may foster behavioral changes and may aid in prevention of chronic diseases (Papas et al. (2007); Sallis et al. (2012); Christine et al. (2015)). Our interest lies in understanding whether an individual's genomic profile modifies the effect of neighborhood features on blood pressure.

We conducted both a targeted GEI analysis and a gene-based genome-wide GEI analysis. Our targeted GEI analysis studied 29 candidate genomic regions which were selected around 29 index SNPs that are significantly associated with blood pressures (p-value $< 10^{-9}$) by the International Consortium for Blood Pressure Genome-Wide Association Studies, ICBP (2011). The criteria of determining each genomic region is same as He et al. (2015): when the index SNP fell within a gene, we selected all SNPs within the gene +/− 5kb and adopted the gene's name to label the region. When the index SNP fell outside of a gene, we selected the index SNP plus all SNPs +/− 50kb and name the region after the index SNP. Number of SNPs in these regions ranges from 10 to 840 SNPs. Our genome-wide gene-based analysis studied 24743 protein coding genes +/− 5kb defined by the UCSC genome browser (Karolchik et al. (2003)). The SNPs in the regions were directly genotyped using the Affymetrix Genome-Wide Human SNP Array 6.0 or imputed as per MESA protocol. Imputation was performed using the IMPUTE 2.1.0 program by Marchini et al. (2007) in conjunction with HapMap Phase I and II reference panels (CEU+YRI+CHB+JPT, release 22 - NCBI Build 36 for African-, Chinese- and Hispanic-American participants; CEU, release 24 - NCBI Build 36 for European Americans). All common and rare variants are included in our analysis without any minor allele frequency filters.

3 Model and Inference

Consider a study population with m independent subjects where the i-th subject has n_i longitudinal observations, $n = \sum_{i=1}^{m} n_i$. When $n_i = 1$ for all $1 \le i \le m$, this corresponds to a cross-sectional study. Let $Y_{i,j}$ be the quantitative outcome value, $\boldsymbol{X}_{i,j} = (X_{i,j}^1, \ldots, X_{i,j}^p)^T$ be the p covariates which can include age, gender, education, etc., $E_{i,j}$ be the environmental exposures for the j-th observations on the i-th subject measured at time $t_{i,j}$; $\bar{\boldsymbol{G}}_i = (G_i^1, \ldots, G_i^q)^T$ be the q time-invariant genetic variants in the target region, where $G_i^k \in \{0, 1, 2\}$. We define $\boldsymbol{Y}_i = (Y_{i,1}, \ldots, Y_{i,n_i})^T$ as a vector of all observations and $\boldsymbol{G}_i = (\bar{\boldsymbol{G}}_i, \ldots, \bar{\boldsymbol{G}}_i)^T$ as an $n_i \times q$ matrix of genetic variants where $\bar{\boldsymbol{G}}_i$ is repeated n_i times; $\boldsymbol{X}_i$, $\boldsymbol{E}_i$ are defined as the matrix forms of

covariates and environmental exposure similarly. We are interested in the statistical interaction between $E_{i,j}$ and $\bar{\boldsymbol{G}}_i$ on outcome $Y_{i,j}$, adjusting for $\boldsymbol{X}_{i,j}$ in addition to the main effect of $E_{i,j}$ and $\bar{\boldsymbol{G}}_i$. $E_{i,j}$ can be a summary statistic of measures of environmental exposure prior or up to exam j if the investigators believe the outcome not only depends on the current values of exposure but also the previous exposure history. For example, $E_{i,j}$ can be the cumulative average of repeated exposure measures up to exam j. The statistical interaction between the environmental exposure and the k-th genetic variant is characterized by $E_{i,j}G_i^k$. We define $E_{i,j} * \bar{\boldsymbol{G}}_i = (E_{i,j}G_i^1, \ldots, E_{i,j}G_i^q)^T$ and its matrix form is denoted by $\boldsymbol{E}_i * \boldsymbol{G}_i$, an $n \times q$ matrix.

One popular approach for analyzing longitudinal genetic data is a single SNP analysis, repeated for each of the G_i^k separately, $k = 1, \ldots, q$, based on a generalized estimating equation (GEE) approach,

$$E(Y_{i,j} \mid \boldsymbol{X}_i, \boldsymbol{E}_i, G_i^k) = \boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_X + E_{i,j}\beta_E + G_i^k \beta_{G,k} + E_{i,j}G_i^k \gamma_k,$$

where $\boldsymbol{\beta}_X = (\beta_{X,1}, \ldots, \beta_{X,p})^T$, β_E and $\beta_{G,k}$ are the coefficients for covariates, main effect of exposure and the k-th SNP respectively; γ_k is the gene-environment interaction parameter of interest. Both the main effects ($\boldsymbol{\beta}_X, \beta_E, \beta_{G,k}$) and the interaction effect γ_k are modeled as fixed effects. The null hypothesis is $H_0 : \gamma_k = 0$. To extend it to a set-based analysis, a natural multivariate model includes all SNPs in the same region simultaneously,

$$\mu_{i,j} = E(Y_{i,j} \mid \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_X + E_{i,j}\beta_E + \bar{\boldsymbol{G}}_i^T \boldsymbol{\beta}_G + (E_{i,j} * \bar{\boldsymbol{G}}_i)^T \boldsymbol{\gamma}, \quad (1)$$

where $\boldsymbol{\beta}_G = (\beta_{G,1}, \ldots, \beta_{G,q})^T$; $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)^T$. The null hypothesis jointly tests the entire interaction vector of length q, namely, $H_0 : \boldsymbol{\gamma} = 0$. The working covariance matrix of $\boldsymbol{Y}_i$ is denoted as $\boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})$, which is of size $n_i \times n_i$ and depends on a vector of parameters $\boldsymbol{\zeta}$. For cross-sectional studies, Lin et al. (2013) considered $\boldsymbol{\beta}_G$ as fixed effects and assumed that each coefficient γ_k follows i.i.d $N(0, \tau^2)$ and proposed a variance component score test for $H_0 : \tau^2 = 0$. Instead of the mixed effect model, we propose a GEE approach based on the unified fixed effect model (1), where the parameters have a more natural interpretation.

The classical approach for testing $H_0 : \boldsymbol{\gamma} = 0$ is a q-degree of freedom likelihood ratio/wald/score test. However, Goeman et al. (2006) showed the power of such tests tend to diminish rapidly when the dimensionality q is large, which is common when the region considered consists of hundreds of variants. To address this, we develop a generalized score type test that can exploit the LD among the SNPs to reduce the test degrees of freedom under model (1). The score vector from model (1) with respect to $\boldsymbol{\gamma}$ is:

$$S_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\gamma}) = \sum_{i=1}^{m} S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta},\boldsymbol{\zeta},\boldsymbol{\gamma}) = \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i),$$

where $\boldsymbol{\mu}_i = (\mu_{i,1}, \ldots, \mu_{i,n_i})^T$. By M-estimation theory, the score statistic $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)$ asymptotically follows a multivariate normal distribution with mean zero and covariance $\boldsymbol{\Sigma}$ under H_0, where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\zeta}}$ are the estimators under $H_0 : \boldsymbol{\gamma} = 0$ obtained by using the usual GEE proposed by Liang and Zeger (1986). Each element $\frac{1}{\sqrt{m}} S^k_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)$ follows an asymptotic normal distribution with mean zero. The classical score test summarizes the vector $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)$ into a scalar by considering $\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)^T \widehat{\boldsymbol{\Sigma}}^{-1} S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)$ where $\hat{\boldsymbol{\Sigma}}$ is an estimator of $\boldsymbol{\Sigma}$. In this case, the test statistic follows a chi-square distribution with q degrees of freedom, i.e., a sum of q squared independent normal random variables. This approach involves the inversion of $\hat{\boldsymbol{\Sigma}}$, which is not stable when q is large relative to m, and cannot be applied to scenarios when $q > m$. To address this, we define a test statistic Q for testing $H_0 : \boldsymbol{\gamma} = 0$ by aggregating the score statistics in a different way,

$$Q = \frac{1}{m} S^T_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0) S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0) = \frac{1}{m} \sum_{k=1}^{q} \{S^k_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)\}^2,$$

where $S^k_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)$ corresponds to the k-th interaction term. The statistic can be understood as the overall deviation from 0 of all score statistics where each of them measures the strength of a specific interaction effect. Let $S\boldsymbol{\beta}(\boldsymbol{\beta}, \boldsymbol{\zeta}, \boldsymbol{\gamma})$ denote the score vector with respect to $\boldsymbol{\beta}$.

Result 3.1—Under model (1) and $H_0 : \boldsymbol{\gamma} = 0$, if q is fixed and $m \to \infty$, Q is asymptotically distributed as

$$\sum_{k=1}^{q} \lambda_k \chi^2_k \quad (2)$$

where χ^2_ks are i.i.d. Chi-square distributions with degree of freedom one; $\lambda_1 \geq \ldots \geq \lambda_q$ are the eigen-values of $\boldsymbol{\Sigma}$ and can be estimated by $\{\bar{\lambda}_k\}_{1 \leq k \leq q}$,

$$\max_{1 \leq k \leq q} |\hat{\lambda}_k - \lambda_k| = o_p(1), \quad m \to \infty;$$

$\bar{\lambda}_1 \geq \ldots \geq \hat{\lambda}_q$ are the ordered eigen-values of $\hat{\boldsymbol{\Sigma}}$. Specifically, $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{A}}\hat{\boldsymbol{D}}\hat{\boldsymbol{A}}^T$,

$$\hat{\boldsymbol{A}} = \{\boldsymbol{I}_q, -[\sum_{i=1}^{m}(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\xi}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)][\sum_{i=1}^{m}(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\hat{\boldsymbol{\xi}})(\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i)]^{-1}\},$$

$$\hat{\boldsymbol{D}} = \frac{1}{m-p-q-1}\sum_{i=1}^{m} S_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0) S_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)^T, S_i(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0) = [S_{\boldsymbol{\gamma}, i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)^T, S_{\boldsymbol{\beta}, i}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0)^T]^T.$$

Result 3.1 shows the asymptotic behavior of the test statistic Q as m goes to infinity. The proof is given in the supplemental materials. The variance component test proposed by Lin et al. (2013) also follows a similar weighted summation of chi-square distributions, but their weights are estimated using a model based inference. Instead, we estimate the weights using the “sandwich estimators”. The empirical estimated weights make the test robust against misspecification of within-subject correlation, which is a desirable property in longitudinal studies with repeated measurements. This sandwich estimation also plays a role in reducing spurious gene-environment interactions caused by potential main effect misspecification of E when G and E are independent, as observed by Voorman et al. (2011) and Cornelis et al. (2012). The rigorous result that explains these observations will be left to the next section.

The proposed test statistic belongs to the class of quadratic test statistics of the form $Q = \boldsymbol{S}^T\boldsymbol{A}\boldsymbol{S}$ as described in Derkach et al. (2014), where $\boldsymbol{S}$ is the score vector. Other examples of test statistics which belong to this class include the ones used in the methods rareGE (Chen et al. (2014)), iSKAT (Lin et al. (2013); Lin et al. (2016)) and the classical q d.f. score test. For our proposed test, rareGE and iSKAT, $\boldsymbol{A}$ equals $\boldsymbol{I}$. For the classical score test, $\boldsymbol{A}$ equals $\hat{\boldsymbol{\Sigma}}^{-1}$ where $\hat{\boldsymbol{\Sigma}}$ is the estimated covariance matrix of $\boldsymbol{S}$. We describe this comparison in detail in section 2 of the Supplementary Materials. Since SNPs in a region can be strongly correlated due to linkage disequilibrium, many eigen-values of $\boldsymbol{\Sigma}$ are close to 0, and the effective test degrees of freedom is less than q. Therefore the proposed test implicitly reduces the test degrees of freedom compared to the classical score test. It is worth noting that the power of a test not only depends on the test degrees of freedom, but also the non-centrality parameter. Since both the effective test degrees of freedom and non-centrality parameter may change across various scenarios, there is no theoretical result for a uniformly optimal choice for constructing a test statistic achieving the highest power in the class of quadratic test statistics. However, many empirical studies have demonstrated the tests with $\boldsymbol{A} = \boldsymbol{I}$, such as the proposed test, has superior power than classical score test in genetic association studies (Wu et al. (2010); Tzeng et al. (2011); He et al. (2014)). Basu and Pan (2011) also pointed out that these tests can be regarded as modified score test by ignoring the non-diagonal elements of $\boldsymbol{A}$, which is known to be advantageous for high-dimensional data.

4 Main Effect Adjustment

So far, we have discussed inference under a correctly specified main effect model under H_0. Unlike set-based tests for genetic association, set-based tests for gene-environment interaction face the unique challenge of having a potentially misspecified and high-dimensional null model. In this section, we consider potential strategies when the main effect of E may be misspecified and the dimension of G, namely q, is large relative to m. A

key step for implementing the proposed generalized score type test is fitting the following main effect model under the null hypothesis

$$\mu_{i,j} = E_{H_0}(Y_{i,j} \mid \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_X + E_{i,j}\beta_E + \bar{\boldsymbol{G}}_i^T \boldsymbol{\beta}_G.$$

There are two challenges with respect to this step. First, a misspecified main effect of $E_{i,j}$ can lead to a biased score ($E_{H_0}[S_{\boldsymbol{\gamma},l}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] \neq 0$) and severe type I error inflation. This may happen when the underlying main effect of the environmental exposure is nonlinear but a linear model is specified. Second, the dimension of $\bar{\boldsymbol{G}}_i$ can be large relative to the sample size, such as the *MECOM* region in MESA which includes 821 SNPs but the Chinese Americans only have 775 subjects. The estimates of $\boldsymbol{\beta} = (\boldsymbol{\beta}_X^T, \beta_E, \boldsymbol{\beta}_G^T)^T$ are not consistent and the approximation to the asymptotic distribution of Q as presented in Result 3.1 does not hold anymore. To address these challenges, we first ensure the robustness of the proposed test to main effect misspecification by exploiting the gene-environment independence condition, then develop methods to handle the main effect misspecification of E and high-dimensionality of G when the gene-environment independence condition does not hold.

4.1 Gene-environment independence condition

Gene-environment independence plays a crucial role in the main effect adjustment. We show in Result 4.1 that the test proposed in Section 3 will be robust to main effect mis-specification under the gene-environment independence condition, by centering $\boldsymbol{E}_i$ and $\boldsymbol{G}_i$ using weighted average as described in section 1.2 of the Supplementary Materials.

Result 4.1—If the following two assumptions hold:

C1. $\boldsymbol{X}_i$ can be separated as $(\boldsymbol{X}_i^E, \boldsymbol{X}_i^G)$ where $(\boldsymbol{X}_i^E, \boldsymbol{E}_i)$ is independent of $\boldsymbol{G}_i$ and $(\boldsymbol{X}_i^G, \boldsymbol{G}_i)$ is independent of $\boldsymbol{E}_i$,

C2. $\mathrm{cov}(\boldsymbol{X}_{i,l}^G, G_i)$ is time invariant,

then the expectation of the score vector equals zero, i.e., $E_{H_0}[S_{\boldsymbol{\gamma},l}(\boldsymbol{\beta}, \boldsymbol{\zeta}, 0)] = 0$, regardless of the main effect model of E and G when $\boldsymbol{E}_i$ and $\boldsymbol{G}_i$ are centered appropriately.

Condition C1 can be seen as the more commonly used condition of gene-environment independence with additional requirement on the covariates $\boldsymbol{X}_i$. For instance, time and age are likely to be correlated with the time varying environmental exposure but independent of the time invariant SNPs. It reduces to the gene-environment independence condition in the special case of no covariates. Condition C2 is specifically for longitudinal studies, and it always holds for cross-sectional studies. It is also satisfied in the special case when $\boldsymbol{X}_i^G$ is time invariant, which is common in a genetic study, e.g. when $\boldsymbol{X}_i^G$ consists of the leading principal components to control for population stratification. The weighted average used to center E and G are proposed to take into account of the within-subject correlation among

observations on the same subject (see section 1.2 of the Supplementary Materials). For cross-sectional studies, this approach reduces to simply centering E and G by the usual average.

Under C1 and C2, the proposed test is robust to a misspecified main effect model, if $\mathbf{\Sigma}$ is estimated using the sandwich covariance estimator and E and G are weighted and centered. This is because the score statistic $\frac{1}{\sqrt{m}}S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}},0)$ will asymptotically follow a mean zero multivariate normal distribution, whose covariance matrix is empirically estimated by sandwich estimators. Therefore the asymptotic distribution of Q, as a function of $\frac{1}{\sqrt{m}}S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\xi}},0)$, can be correctly estimated. Under C1 and C2, this result shows using a linear model for E is sufficient for controlling type I error rate regardless of the true functional form of the main effect of E. The problem of inconsistency due to high-dimensionality of G can be simply solved by excluding the main effects of all SNPs in the model. However, these strategies are not adequate, especially when C1 and C2 are violated. We further develop methods for main effect adjustment of E and G in the subsequent sections that are appropriate under violations of C1 and C2.

This result also explains the findings in Voorman et al. (2011) and Cornelis et al. (2012), where the authors showed that using sandwich estimators can reduce the detection of spurious gene-environment interactions in cross-sectional studies. Specifically, the simulation studies conducted by Voorman et al. (2011) did not observe any type I error inflation under misspecification of main effect of E when a sandwich estimator was used, because no association between G and E was simulated; The genome-wide analysis for gene-environment interactions conducted by Cornelis et al. (2012) used QQ-plots to show that using a sandwich estimator can reduce the type I error inflation. This is likely due to the fact that a vast majority of the SNPs are usually not correlated with the environmental exposure. Using sandwich estimators for variance will eliminate the inflation for these SNPs as gene-environment independence is effectively true in these situations.

4.2 Main effect misspecification of *E*

Most GEI studies consider a linear main effects model as described in (1). When C1 and C2 do not hold, ignoring a nonlinear main effect can result in a biased score function and lead to severe type I error inflation. Even if C1 and C2 hold and type I error is not a concern, a misspecified main effect model for E can significantly reduce power for testing interaction. Examples include the cases when the main effect of E has a quadratic effect, or E is a log-transformed exposure but the true effect is on the original scale. In this subsection, we make further effort to control the bias in the scores due to a misspecified main effect of E when C1 and C2 do not hold, and improve the power. Since the true main effect $h_E(\cdot)$ is unknown, we propose to approximate it non-parametrically by the method of “sieves”: expand $h_E(\cdot)$ by a sequence of finite dimensional models $\Phi_U(\text{sieves})$, then allow the model complexity U to grow slowly with the sample size (Grenander (1981)). Numerous sieve estimators have been proposed such as the polynomial sieves and the spline sieves:

$$\Phi_U^P = \{h_{E,U} : h_{E,U}(x, \boldsymbol{\beta}_E) = \sum_{u=1}^{U} x^u \beta_{E,u}\}; \quad \Phi_U^S = \{h_{E,U} : h_{E,U}(x, \boldsymbol{\beta}_E) = \sum_{u=1}^{U} B_u^U(x) \beta_{E,u}\},$$

where $B_u^U(\cdot)$ is the u-th spline basis function. So the function $h_E(\cdot)$ can be approximated by a series of sieves. The uniform convergence rate of $h_{E,U}(x, \hat{\boldsymbol{\beta}}_E)$ as $m \to \infty$ depends on the smoothness of $h_E(x)$. The details of asymptotic results can be found in Newey (1997).

The main effect model based on the sieve representation can be written as

$$\mu_{i,j} = E_{H_0}(Y_{i,j} \mid \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_X + h_{E,U}(E_{i,j}; \boldsymbol{\beta}_E) + \bar{\boldsymbol{G}}_i^T \boldsymbol{\beta}_G, \quad (3)$$

where $h_{E,U}(\cdot)$ is a finite dimensional model using spline/polynomial sieves. Result 4.2 shows that, under C1 and C2, a test for gene-environment interaction based on a main effect model (3) will be asymptotically equivalent to using the true model. Thus the test not only has correct type I error rate, but also is as powerful as using the true model.

Result 4.2—If C1 and C2 hold and $h_{E,U}(x; \hat{\boldsymbol{\beta}}_E)$ uniformly converges to $h_E(x)$ for $\forall\ x$ as $m \to \infty$,

$$\frac{1}{\sqrt{m}} S_{\boldsymbol{\gamma}}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\xi}}, 0) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})(\boldsymbol{Y}_i - \boldsymbol{\mu}_i^0) + o_p(1).$$

where $\boldsymbol{\mu}_i^0$ is the stacked vector of conditional means in (3) with true main effect $h_E(\cdot)$. This includes a scenario where U is larger than the underlying model complexity. For example, if the underlying main effect of E is linear but we model it using cubic-spline sieves with $U > 1$, the test will be asymptotically equivalent to a linear model and will not be less powerful under C1 and C2. When C1 and C2 do not hold, introducing unnecessary model complexity can reduce the power. However, we note that the proportion of total variation of an exposure explained by a single genomic region is usually not expected to be very large. With this weak dependency, our simulation studies demonstrate that type I error inflation due to main effect misspecification of E can be severe, but the power loss due to using more complex model is negligible (Table 1). In summary, flexibly modeling the main effect of E does not substaintially hurt power for tests of gene-environment interaction, and greatly helps in controlling type I error rate. This is a very important observation for practice. However, we note that this is different from using more flexible models for the GEI terms in the alternative hypothesis, which certainly entail substantial loss of power.

Result 4.2 also helps to choose the model complexity U, which plays a crucial role in the method of sieves. The common criteria include cross-validation that minimizes the integrated mean square error, the Mallows criterion by Mallows (1973), the Akaike

information criterion described in Akaike (1998) and the Bayesian information criterion by Schwarz (1978). Although these methods are still reasonable, Result 4.2 indicates that the ideal criteria for main effect adjustment can be different, because the primary focus is to test another set of variables (the interactions terms). Based on Result 4.2, a larger U allowed by sample size is recommended for better controlling type I error rate and will not hurt power. In this paper, we specifically illustrate the proposed test with a sufficiently rich main effect model for E with $U = m^{\frac{1}{2}}$. The choice of U is driven by existing results that ensure the asymptotic estimation of the coefficients by GEE is reliable. The detailed discussion can be found in Wang (2011).

4.3 High-dimensionality of *G*

When a large genomic region is considered in a set-based analysis, the number of parameters can be large relative to the sample size. The top panel in Supplementary Figure 1 shows an example in MESA where region *MECOM* includes 821 SNPs but the Chinese Americans only have 775 subjects. When C1 and C2 do not hold, the main effect of G cannot be ignored because its confounding effect can lead to bias and type I error inflation. Lin et al. (2013) proposed to use ridge regression for handling the main effect of G, but their test is still based on the assumption that q is fixed and $m \to \infty$, same as the method presented in Section 3. These methods work well when the dimension of G is moderate, but suffer from severe type I error inflation when the number of SNPs is close to or larger than the sample size (Table 2; Supplementary Table 1). This is a curse of dimensionality and some form of dimension reduction in the G space is needed. In this subsection, we make further effort to deal with both the high-dimensionality and the confounding effect of G.

To deal with the high-dimensionality of G, one natural choice is taking advantage of the LD structure in genetic regions, and use some form of PCA. The first panel in Supplementary Figure 1 shows a typical genome region that contains several LD blocks and SNPs within each block are correlated. Therefore eigen-values corresponding to the principal components (PC) decrease to zero very quickly as a function of the leading number of components (Supplementary Figure 1). This enables us to use a small number of PCs to explain most variation in G. A standard PCA results in orthogonal components $\{P_i^s\}_{1 \le s \le q}$ ranked by the corresponding eigen-values $\kappa_1 \ge \ldots \ge \kappa_q$, $E(P_i^s) = 0$, $var(P_i^s) = \kappa_s$. Each component is a linear combination of $\{G_i^k\}_{1 \le k \le q}$. We usually fit the leading PCs:

$$\mu_{i,j} = E_{H_0}(Y_{i,j} \mid \boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{G}_i) = \boldsymbol{X}_{i,j}^T \boldsymbol{\beta}_X + E_{i,j}\beta_E + \sum_{s=1}^{S} P_i^s \beta_{P,s}, \quad (4)$$

where $1 \le S \le q$. The PCA approach with a well chosen S is a plausible remedy for the curse of dimensionality, but not ideal for adjusting the confounding effect of G because there can be low-rank PCs that has non-zero effect on the outcome. When C1 does not hold, it is subject to bias because the model ignores the missed set of $q - S$ PCs so that, now, the main

effect of G is misspecified. Let $\boldsymbol{P}_i^s = (P_i^s, \ldots, P_i^s)^T$ be the stack of PCs corresponding to subject i. Result 4.3 explicitly gives the bias expression due to missing $q - S$ PCs.

Result 4.3—The bias due to fitting model (4) is given by

$$E_{H_0}[S_{\boldsymbol{\gamma},i}(\boldsymbol{\beta},\boldsymbol{\zeta},0)] = \sum_{s=S+1}^{q} \{E[(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{P}_i^s - \boldsymbol{\phi}^s\}\beta_{P,s}^0,$$

where $\beta_{P,s}^0$ is the coefficient in the full model where all PCs are included;

$$\begin{aligned}
\boldsymbol{\phi}^s &= E\{(\boldsymbol{E}_i * \boldsymbol{G}_i)^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]\}\boldsymbol{A}^{-1}\boldsymbol{b}^s \\
\boldsymbol{A} &= E\{[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]\} \\
\boldsymbol{b}^s &= E\{[\boldsymbol{X}_i, \boldsymbol{E}_i, \boldsymbol{P}_i^1, \ldots, \boldsymbol{P}_i^S]^T \boldsymbol{V}_i^{-1}(\boldsymbol{\zeta})\boldsymbol{P}_i^s\}.
\end{aligned}$$

The result shows that the bias due to a PC that was not included is proportional to its association with the outcome conditional on $(\boldsymbol{X}_i, \boldsymbol{E}_i)$. This is also closely related to the definition of confounders discussed by VanderWeele and Shpitser (2013).

To reduce the bias due to the confounding effect of G, a better approach should consider the correlation between the outcome and the PCs in addition to the eigenvalues. A well-known method that takes this correlation into account is the partial least squares regression (PLS). PLS generates orthogonal components by sequentially optimizing their correlation with the outcome and correlation with G (Boulesteix and Strimmer (2007)). However, when the sample size is small and the region is large, PLS components are constructed by overfitting an outcome regression model, which makes the test for the interaction terms less powerful (Supplementary Table 2). Instead, we propose to use the components $\{P_i^s\}_{1 \le s \le q}$ from PCA but rank them by

$$\text{corr}(Y_{i,j}, P_i^s \mid \boldsymbol{X}_i, \boldsymbol{E}_i)^2 var(P_i^s) = R_s^2 \kappa_s,$$

where R_s^2 stands for the variation of $Y_{i,j}$ explained by P_i^s conditional on $(\boldsymbol{X}_i, \boldsymbol{E}_i)$. It is reasonable to assume R_s^2 is not likely to vary across visits j under model (1) because G is time invariant and we do not consider the situation that the association between G and Y may vary by visit j under the null hypothesis. This weighted PCA approach uses a criterion that is close to the objective function of PLS, but the selected PCA components are not constructed by fitting an outcome regression model. Similar approach of using correlation-selected PCs was also successfully used in GWAS to find PCs for population stratification adjustment (Lee et al. (2011)). To adjust for the effect of the exposure and covariates, we

first regress $\boldsymbol{Y}_i$ on $(\boldsymbol{X}_i, \boldsymbol{E}_i)$, then use the residuals to estimate R^2 for each principal component.

To reduce the dimension of the fitted model (4), we again suggest to use $S = m^{\frac{1}{2}}$ in practice to have reliable asymptotic estimation of $\boldsymbol{\beta}$ and illustrate it using extensive simulation studies (Wang (2011)).

5 Numerical Studies

We evaluated type I error rate and power of the proposed test using simulation studies for both cross-sectional and longitudinal data, and compared our method with existing choices: 1. set based tests for GEI using a single average or baseline outcome and exposure measure: iSKAT with $\rho = 0$ and rareGE assuming a random main effect of G (Lin et al. (2013); Lin et al. (2016); Chen et al. (2014)); 2. a single SNP based test for longitudinal outcomes and exposures: the minimum p-value test (MinP) using GEE. For each simulated dataset, we directly sampled SNPs from gene regions in MESA and then conditionally simulated the phenotype and environmental exposure. When there are repeated measurements, we first simulated the complete data, and then applied a missingness indicator with 4% fixed dropout rate at each exam assuming data missing completely at random. The coefficients in the simulation studies were chosen such that each variable explains a reasonable variation in the outcome as in real data scenarios. For example, the variation in the outcome explained by the main effect of E or G (a set of SNPs) ranges from 5% to 15% in the longitudinal settings. We simulated top four principal components as covariates directly from MESA genome-wide data to retain its correlation with the target region, and their coefficients were elicited based on the analysis of the corresponding ethnic group. The simulation studies are structured into three scenarios where each part empirically evaluates both type I error and power based on 1000 replicates.

Scenario 1: Role of main effect specification of *E*

In the first simulation setting, we evaluated the proposed method when the main effect of E is linear/nonlinear in both cross-sectional and longitudinal settings. We focused on cubic-spline sieves generated by knots at equally spaced quantiles of all observations. We used all SNPs from region indexed by *rs*10850411 (190 SNPs) in European Americans (2526 subjects), and simulated one environmental exposure independent/dependent of the SNPs. To focus on the effect of E, this region was chosen such that the sample size is sufficiently large relative to the number of SNPs. The true model is of the form:

$$E_{i,j} = \alpha_{E,0} t_{i,j} + \alpha_{E,1} X_i + \sum_{k=1}^{5} \alpha_{E,2} G_i^k + b_{E,i} + \varepsilon_{E,i,j}, \quad j = 1, \ldots, d,$$

$$Y_{i,j} = \sum_{s=1}^{4} \alpha_{PC,s} PC_i^s + \alpha_0 t_{i,j} + \alpha_1 X_i + \alpha_2 h_M(E_{i,j}) + \sum_{k=1}^{5} \alpha_3 G_i^k + \alpha_4 E_{i,j} \sum_{k=1}^{5} G_i^k + b_i + \varepsilon_{i,j},$$

where $d = 1$ is for cross-sectional data and $d = 4$ is for longitudinal data; $t_{i,j} = j - 1$ (0, 1, 2, 3 standing for visits); $X_i \sim N(0, 1)$ is a time-invariant covariate; PC_i^s is the s-th principal component of subject i directly from the MESA genome-wide data; five out of the 190 SNPs

(2.6%) are causal and G_i^k is the genotype of subject i for the k-th randomly selected causal SNP; $(\alpha_{PC,1}, \alpha_{PC,2}, \alpha_{PC,3}, \alpha_{PC,4}) = (-4.7, -0.9, 13.1, 1.3)$; $\alpha_{E,0} = \alpha_{E,1} = \alpha_0 = \alpha_1 = 1$, $\alpha_2 = 0.5$, $\alpha_3 = 2$; $\alpha_{E,2}$ measure the association between E and G. $\alpha_{E,2} = 0$ when E is independent of G and $\alpha_{E,2} = 0.5$ when E is dependent of G (e.g., ~ 3% variation in E is explained by G in the longitudinal setting); $\alpha_4 = 0.10/0.05$ for evaluating cross-sectional/longitudinal power and $\alpha_4 = 0$ for evaluating type I error rate; $b_{E,i} \sim N(0, 4)$, $\varepsilon_{E,i,j} \sim N(0, 4)$, $b_i \sim N(0, 9)$, $\varepsilon_{i,j} \sim N(0, 9)$ and they are all independent. h_M is the main effect function specified as "E", "$0.3E_2$", "$E + 0.2E_2$" or "exp($0.4E$)" for cross-sectional data, and "$0.8E$", "$0.2E_2$", "$0.5E + 0.1E_2$" or "exp($0.3E$)" for longitudinal data. The functions were scaled such that they explain similar variation of $Y_{i,l}$ as compared to the linear model (e.g., ~10% in the longitudinal setting). Table 1 presents the results.

Type I error rate—Even when C1 holds, iSKAT using a model based inference has inflated type I error rate (e.g., 0.172, 0.113 and 0.185 where the true models are E_2, $E+E_2$ and exp(E) respectively, cross-sectional setting). rareGE has inflated type I error rate when the main effect of E is nonlinear, similar to iSKAT. However, the proposed method using the sandwich estimator is robust regardless of the main effect misspecification; When C1 does not hold, only assuming a linear main effect does have type I error inflation even if sandwich estimation is used (e.g., 0.906, 0.729 and 0.869 for E_2, $E + E_2$ and exp(E), longitudinal setting). However, the proposed method using the method of sieves still has robust type I error rate.

Power—When C1 holds, the proposed method using the method of sieves always has similar power as the method based on the true model, even if the true effect is linear and additional model complexity was assumed for the main effects (e.g., 0.786 vs. 0.789, longitudinal setting). When C1 is violated, the method of sieves results in slightly lower power than using the true model (e.g., 0.774 vs. 0.796, longitudinal setting), but the power difference is small. Moreover, the method of sieves often leads to improved power compared with the method assuming a linear main effect when the true effect is nonlinear (e.g., 0.786 vs. 0.606 when the true main effect is E_2, longitudinal setting).

Scenario 2: Role of main effect specification of G

In the second simulation setting, we evaluated the proposed method for the main effect adjustment of G in both cross-sectional and longitudinal settings. We varied the number of SNPs (400 – 700) simulated from genotype region *MECOM* (821 SNPs) in Chinese Americans (775 subjects), and simulated one environmental exposure independent/dependent of the SNPs. The region was chosen to reflect a scenario where the number of SNPs is large relative to the sample size. The model is same as that in Scenario 1 with a linear main effect of E, so we omit the detailed equations and only present the parameters that are different from Scenario 1. In this scenario, five out of the 400/700 SNPs (1.3%/0.7%) are causal; $(\alpha_{PC,1}, \alpha_{PC,2}, \alpha_{PC,3}, \alpha_{PC,4}) = (-2.3, -24.9, 5.6, -13.3)$; $\alpha_{E,2} = 0$ when E is independent of G and $\alpha_{E,2} = 2$ when E is dependent of G (e.g., ~25% variation in E is explained by G in the longitudinal setting). We chose a large $\alpha_{E,2}$ to observe the type I error inflation due to main effect misspecification of G. $\alpha_4 = 0.2/0.1$ for evaluating cross-

sectional/longitudinal power and $\alpha_4 = 0$ for evaluating type I error rate. The results are summarized in Table 2.

Type I error rate—The MinP test based on single SNP analyses has inflated type I error rate when C1 does not hold (e.g., 0.088/0.108 for 400/700 SNPs, longitudinal setting). This result is consistent with the results in Lin et al. (2013). iSKAT using a ridge regression has type I error inflation when the number of SNPs is large, especially when the number is close to the sample size (0.089 for 700 SNPs, cross-sectional setting). Supplementary Table 1 further shows an example where iSKAT has type I error rate close to one when the number of SNPs exceed the sample size. rareGE has slightly inflated type I error rate when the number of SNPs is greater than the sample size (0.072 for 700 SNPs, cross-sectional setting, Supplementary Table 1). The proposed method has well controlled type I error rate for all scenarios considered in this stimulation setting. We further evaluated PCA, PLS and weighted PCA as other possible approaches to reduce dimension of G and summarized the results in Supplementary Table 2. When C1 holds and the number of adjusted components is five, type I error rates of PLS and weighted PCA are well controlled, but that of PCA is inflated (e.g., 0.033/0.057/0.094 for PLS/weighted PCA/PCA, 700 SNPs, longitudinal setting). When the number of components increases to $m^{\frac{1}{2}}$, all three have well controlled type I error rate. The proposed methods tend to be slightly conservative due to the use of sandwich estimator as in regular GEE, even if a correct mean model is used.

Power—The proposed method has similar power as using the true model and it is more powerful than the MinP test (e.g., 0.588 vs. 0.472, 400 SNPs, longitudinal setting when E and G are independent). We also evaluated the proposed test using a model based inference for estimating $\boldsymbol{\Sigma}$ that is typically used for cross-sectional data. It has slightly higher power than using the sandwich estimation (e.g., 0.626 vs. 0.588, 400 SNPs, longitudinal setting when C1 holds). The power of rareGE is comparable to our proposed test using a model based inference in situations when there are no Type 1 error inflation (for example, in Table 2, both are equal (0.561) when C1 holds and the number of SNPs is 700). Moreover, Supplementary Table 2 shows that PLS has lower power than the proposed method when the number of SNPs is close to the sample size (e.g., 0.381 vs. 0.483, 700 SNPs, cross-sectional setting when C1 holds), although its type I error rate is well controlled.

Scenario 3: Role of longitudinal data

In the third simulation setting, we aimed to illustrate that the proposed method is robust to misspecification of within-subject correlation when there are repeated measurements, and show the advantage of using full trajectory of the longitudinal outcome and exposure. When more than one repeated measures are involved, we compare our method with iSKAT using the average/baseline value of the repeated measurements on both Y and E. We used all SNPs from genotype region indexed by *rs*10850411 (190 SNPs) in European Americans (2526 subjects), and simulated one environmental exposure independent of the SNPs. The model is same as the longitudinal setting in Scenario 1 with an linear main effect of E, so we omit the detailed equations and only present the parameters that are different from Scenario 1. In this scenario, $\alpha_4 = 0.05$ for evaluating power and $\alpha_4 = 0$ for evaluating type I error rate; $b_{E,i} \sim$

$N(0, 0.25)$, $\varepsilon_{E,i,j} \sim N(0, 4)$, $b_i \sim N(0, 9/16)$, $\varepsilon_{i,j} \sim N(0, 9)$ and they are all independent. We note that we simulated a large magnitude of within-subject variation to show the type I error inflation due to using the average value of repeated measures. The relative power difference remains the same when a smaller within-subject variation is simulated. Table 3 presents the results.

Type I error rate—The proposed method using the first order autoregressive correlation structure still has valid type I error rate, when the true correlation structure is compound symmetric. iSKAT and rareGE using the average value of repeated measurements has inflated type I error rate because of their model based inference and the heterogeneous variance due to unbalanced data structure (e.g., 0.092 for iSKAT and 0.078 for rareGE, $d = 4$).

Power—The tests using the full trajectory of longitudinal outcome and exposure have much higher power than using the average values, as the number of repeated measurements increases (e.g., 0.805 vs. 0.414, $d = 4$). This is because averaging the environmental exposure reduces its variance and therefore decreases the power of testing gene-environment interaction. The results demonstrate the advantage of using the longitudinal information.

6 Data Analysis

We illustrate the proposed set-based test using data from the Multi-Ethnic Study of Atherosclerosis (MESA) to test the interaction between each neighborhood variable and each SNP set on blood pressure (systolic and diastolic blood pressure) for the four ethnic groups separately, followed by a meta-analysis. Supplementary Tables 3 – 8 present the summary statistics and the marginal association analysis of E/G in MESA. Density of favorable food stores, density of recreational facilities and perceived healthy foods availability are marginally significantly associated with systolic blood pressure (p-value = 3.65×10^{-3}, 4.69×10^{-4} and 8.74×10^{-7} respectively); Perceived healthy foods availability is also associated with diastolic blood pressure (p-value = 4.21×10^{-3}). The marginal effects of the environmental exposures appear to be mostly linear. We conducted both a targeted GEI analysis for the 29 candidate regions and a set-based genome-wide GEI analysis as described in section 2. We adjusted for age, gender, body mass index (BMI), a socioeconomic status variable (SES) and top four ethnicity-specific principal components (PCs) to correct for potential within-ethnicity stratification. BMI was calculated from direct measurements of weight (kg) and height (meters) available for all MESA exams. The socioeconomic status variable was obtained by performing a principal component analysis on a set of housing, residential stability, education, employment, occupation and income variables. We adjusted for the first leading component which is more highly weighted on education, occupation and income. We adjusted the measured blood pressures for participants taking anti-hypertension medication using the standard procedure of adding 10 mmHg to systolic blood pressure and 5 mmHg to diastolic blood pressure as in Cui et al. (2003). Based on the p-values of the ethnicity-stratified analysis, a meta-analysis was done by Fisher's combined probability test (Fisher (1925)). The vast majority of SNPs in our dataset are common variants with MAF greater than 1%, therefore the genetic principal

components calculated for each region in the weighted PCA approach mostly capture the genetic variation in common variants.

Targeted GEI analysis

We conducted a set-based analysis for the 29 candidate genomic regions and compared our method with GEE-based MinP test and iSKAT using either the average or baseline value of repeated measurements. This set-based analysis led to 29 sets $\times$ 4 exposures = 116 tests. We also conducted a single SNP analysis for all SNPs in the 29 regions that led to 5622 SNPs $\times$ 4 exposures = 22488 tests and present the results using the locus-zoom plots (Supplementary Figure 2) (Pruim et al. (2010)).

Set-based analysis—Table 4 presents the most significant region identified by the set-based analysis. The proposed methods exhibit highly suggestive p-value (0.0005 using a linear main effect of E, 0.0009 using the natural cubic-spline) for the interaction between perceived healthy food availability and the region indexed by rs10850411 on systolic blood pressure in European Americans. These p-values are very close to the Bonferroni threshold ($0.05/(4\times29) = 0.00043$). MinP test also results in a suggestive p-value (0.0047) but iSKAT and rareGE using average/baseline value fail to identify this interaction (p-value = 0.8205 and 0.4331 for iSKAT, 0.7542 and 0.5336 for rareGE respectively). This interaction is also suggestive for its GIS counterpart (density of favorable food stores) by using the proposed method (p-value = 0.0427 using a linear main effect of E, 0.0570 using the natural cubic-spline). The most significant SNP in this region is the index SNP rs10850411 (p-value = 4.08×10^{-5}). The locus-zoom plot (the left panel in Supplementary Figure 2) shows there are multiple other SNPs with small p-values uniformly distributed in the region and they are in linkage disequilibrium with the index SNP. We conducted sensitivity analysis additionally adjusting for site and present the results for this region in Supplementary Table 9. The results with and without adjusting for study site are qualitatively similar with some small numerical differences. We also conducted additional analyses to compare strategies using different forms of longitudinal exposures and present results in Supplementary Table 10. The results show that using repeated longitudinal measures appear to be a better strategy in general.

Single-SNP analysis—The most significant SNP identified by the single-SNP analysis is the interaction between density of recreational facilities and a SNP in region *CACNB2*, namely rs7085587, on systolic blood pressure in Hispanic Americans (p-value = 2.17×10^{-6}). This p-value is still significant after the Bonferroni correction (Bonferroni threshold: $0.05/22488 = 2.22\times10^{-6}$). The locus-zoom plot (the right panel in Supplementary Figure 2) shows the signals are concentrated in a small area around rs7085587. This is also a situation where the MinP test results in a smaller p-value (0.0012) than the proposed test (GE-linear p-value = 0.0753) in the corresponding set-based analysis of *CACNB2* (Supplementary Table 11).

In summary, the set-based test performs better in the first example where the signals are dispersed across many SNPs in the region, while the single SNP based test performs better in the second example where the signals are concentrated. For the significant interactions

noted with systolic blood pressure as outcome in Table 4, we also observed suggestive p-values (0.0844 using a linear main effect of E, 0.0537 using the natural cubic-spline for the main effect of E) for diastolic blood pressure (Supplementary Table 12). These interactions that appear to be noteworthy in the European Americans in the MESA analysis are ethnicity specific and are not significant in the other three ethnic groups. The present finding will require replication in other cohorts and need to be followed up in future studies. The genes nearest to rs10850411 are two members of the phylogenetically conserved T-box family of genes, *TBX*3 and *TBX*5. Proteins encoded by T-box family genes act as transcription factors, and have been shown to play a role in development of the heart and limbs (McKusick (1998); OMIM 601620, 601621). Genome-wide association studies have identified variants in the *TBX*3-*TBX*5 gene region that influence heart rate and cardiac electrical activity (Pfeufer et al. (2010); Sotoodehnia et al. (2010)). *CACNB*2 encodes the beta-2 subunit of a voltage-dependent calcium channel protein, and is expressed in the heart. Mutations in *CACNB*2 have been shown to cause Brugada syndrome, characterized by cardiac electrical abnormalities and sudden cardiac death (OMIM 600003, 611876). The detailed analysis of the top SNPs in these two identified regions can be found in section 4.7 of the Supplementary Materials (Supplementary Figure 3).

Genome-wide GEI analysis

We applied the proposed test to 24743 genes (Section 2) for a set-based analysis and compared it with a single SNP analysis of 1011876 SNPs in these sets via GEE. Supplementary Figures 4 – 5 presents the QQ-plots summarizing the results of the set-based analysis using our proposed method. The set based analysis identified a highly suggestive interaction between region *LOC*100129138 and perceived walkability on systolic blood pressure (p-value = 2.04×10^{-6}, Bonferroni threshold = 2.02×10^{-6}). However, the single SNP analysis did not identify any interaction between any SNP and perceived walkability. The smallest p-value equals 8.12×10^{-6} which is much higher than the Bonferroni threshold = 4.94×10^{-8}. This illustrates the potential advantage of a genome-wide set-based GEI analysis compared to a genome-wide single SNP-based GEI analysis. In addition, we observed that iSKAT QQ plots are substantially inflated for a genome-wide analysis in MESA (Supplementary Figures 6 – 7), which is consistent with our simulation studies. This is because the CHN ethnic group only has 775 subjects, but there are many large regions in the genome-wide analysis. The performance of iSKAT with $m < q$ is less than optimal.

7 Conclusion and Discussion

We have developed a statistical framework for set-based inference for testing gene-environment interaction with quantitative traits in both cross-sectional and longitudinal studies. We showed that a generalized score test similar to the tests derived from more sophisticated approaches (e.g., kernel machine regression, similarity regression and genetic random field model) could be postulated using the most commonly used fixed effect model for multivariate regression. Instead of a hybrid model like iSKAT where the main effects are considered as fixed effects but the interactions are considered as random effects, the proposed fixed effect model presents a direct unified framework. We also demonstrated

improved properties of a set-based test compared to a single SNP analysis when multiple causal SNPs exist.

Although many set-based tests have been proposed for evaluating genetic association, our test is the first set-based test for GEI that is able to utilize the rich time varying outcome and exposure data. Our numerical studies show that substantial power gain can be achieved by using the proposed test, compared to methods only using a single outcome/exposure measure (e.g., average/baseline value). The test is also robust to misspecification of within-subject correlation, which is a desirable property in studies with longitudinal measures.

We studied the role of gene-environment independence, and developed methods for main effect adjustment of E and G that permits more robust and powerful inference. Under the independence condition, we showed that the proposed test is robust to misspecification of main effect of E by simply using a sandwich estimator and weighted centered E and G. When the independence condition does not hold, we proposed the method of sieves to model the main effect of E correctly. An interesting finding is that flexibly modeling the main effect does not hurt power for tests of GEI significantly. To remedy the curse of dimensionality in the potentially high dimensional G space, we developed the weighted PCA approach for dimension reduction that allows us to apply the test to large regions where the number of SNPs is close to or larger than the sample size.

We illustrated the method by a targeted GEI analysis and a genome-wide GEI analysis of MESA neighborhood study, where both time varying outcome and exposure data are available. The application illustrates that the longitudinal approach utilizing the full trajectory of longitudinal outcome and exposure measures is substantially more powerful than the approach using a single measurement. It also shows the advantage of a genome-wide set-based GEI analysis compared to a genome-wide single SNP-based GEI analysis. The application is novel in its rich longitudinal neighborhood data and the findings may aid in prevention of chronic diseases by modifying the built environment around us and creating new healthy food resources and recreational facilities and provide public health recommendations for susceptible genetic sub-groups in terms of their neighborhood choice. More importantly, neighborhood interventions or changes in the built environment can impact many people at the same time instead of recommending changes towards lifestyle factors of an individual.

The proposed generalized score test is computationally efficient, because it only fits the model under the null hypothesis. We developed an R package LGEWIS for its scalable implementation in future studies which is freely available at the Comprehensive R Archive Network (CRAN): https://cran.r-project.org/web/packages/LGEWIS/.

There are several limitations of the proposed method. First, the weighted PCA method is an ad-hoc method proposed for dimension reduction of G, only studied through simulation and data analyses. The optimality of this method has not been established in this paper. It will be desirable to develop an optimal method for the main effect adjustment of G in the future and establish its theoretical properties more rigorously. Second, we only considered linear GEI terms in this paper. Directly adding more flexible non-linear GEI terms will certainly lead to

loss of power, which is different from flexibly modeling the main effects. It will be interesting to investigate efficient strategies for modeling non-linear interaction terms. Third, the method was proposed for quantitative traits. Future extension to generalized linear models will be important to develop. Moreover, we note that our result is closely related to the work by Vansteelandt et al. (2008), where they proposed multiply robust inference for statistical interactions by not only modeling the main effect of E, but also the conditional distribution of E given X and G. Future work that develops a multiply robust set-based inference for GEI, boosted with dimension reduction in the G space will be of great interest.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by *NSF DMS 1406712 and NIH/NIEHS grant ES020811, NIH/NHLBI grant R00HL113164, NIH/NHLBI HL101161, and NIMHHD Grant 2P60MD002249 Center for Integrative Approaches to health Disparities (CIAHD). MESA and the MESA SHARe project are conducted and supported by the National Heart, Lung, and BloodInstitute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166,N01-HC-95167, N01-HC-95168, N01-HC-95169 and CTSAUL1-RR-024156. Funding for SHARe genotyping was provided by NHLBI Contract N02-HL-64278. Genotyping was performed at Affymetrix (Santa Clara, California, USA) and the Broad Institute of Harvard and MIT (Boston, Massachusetts, USA) using the Affymetrix Genome-Wide HumanSNP Array 6.0.*

References

Akaike, H. Selected Papers of Hirotugu Akaike. Springer; 1998. Information theory and an extension of the maximum likelihood principle; p. 199-213.

Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. Genetic epidemiology. 2011; 35(7):606–619. [PubMed: 21769936]

Bild DE, Bluemke DA, Burke GL, Detrano R, Roux AVD, Folsom AR, Greenland P, JacobsJr DR, Kronmal R, Liu K, et al. Multi-ethnic study of atherosclerosis: objectives and design. American Journal of Epidemiology. 2002; 156(9):871–881. [PubMed: 12397006]

Boonstra PS, Mukherjee B, Gruber SB, Ahn J, Schmit SL, Chatterjee N. Tests for gene-environment interactions and joint effects with exposure misclassification. American journal of epidemiology. 2016:kwv198.

Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Briefings in Bioinformatics. 2007; 8(1):32–44. [PubMed: 16772269]

Chen H, Meigs JB, Dupuis J. Incorporating gene-environment interaction in testing for association with rare genetic variants. Human heredity. 2014; 78(2):81–90. [PubMed: 25060534]

Christine PJ, Auchincloss AH, Bertoni AG, Carnethon MR, Sánchez BN, Moore K, Adar SD, Horwich TB, Watson KE, Roux AVD. Longitudinal associations between neighborhood physical and social environments and incident type 2 diabetes mellitus: The multi-ethnic study of atherosclerosis (mesa). JAMA Internal Medicine. 2015; 175(8):1311–1320. [PubMed: 26121402]

Cornelis MC, Tchetgen EJT, Liang L, Qi L, Chatterjee N, Hu FB, Kraft P. Gene-environment interactions in genome-wide association studies: a comparative study of tests applied to empirical studies of type 2 diabetes. American Journal of Epidemiology. 2012; 175(3):191–202. [PubMed: 22199026]

Cui JS, Hopper JL, Harrap SB. Antihypertensive treatments obscure familial contributions to blood pressure variation. Hypertension. 2003; 41(2):207–210. [PubMed: 12574083]

Derkach A, Lawless JF, Sun L, et al. Pooled association tests for rare genetic variants: a review and some new results. Statistical Science. 2014; 29(2):302–321.

Fisher, RA. Statistical Methods for Research Workers. Genesis Publishing Pvt Ltd; 1925.

Goeman JJ, Van De Geer SA, Van Houwelingen HC. Testing against a high dimensional alternative. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68(3):477–493.

Grenander, U. Abstract inference. Wiley; New York: 1981.

He Z, Zhang M, Lee S, Smith JA, Guo X, Palmas W, Kardia SL, Roux AVD, Mukherjee B. Set-based tests for genetic association in longitudinal studies. Biometrics. 2015; 71(3):606–615. [PubMed: 25854837]

He Z, Zhang M, Zhan X, Lu Q. Modeling and testing for joint association using a genetic random field model. Biometrics. 2014; 70(3):471–479. [PubMed: 24628067]

ICBP. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011; 478(7367):103–109. [PubMed: 21909115]

Kaiser P, Mujahid M, Carnethon M, Bertoni A, Adar S, Shea S, McClelland R, Lisbeth L, Diez Roux A. Neighborhood environments and incident hypertension in the multi-ethnic study of atherosclerosis. American Journal of Epidemiology. 2015 forthcoming.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu Y, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ, et al. The ucsc genome browser database. Nucleic Acids Research. 2003; 31(1): 51–54. [PubMed: 12519945]

Lee S, Wright FA, Zou F. Control of population stratification by correlation-selected principal components. Biometrics. 2011; 67(3):967–974. [PubMed: 21133882]

Li S, Cui Y, et al. Gene-centric gene–gene interaction: A model-based kernel machine method. The Annals of Applied Statistics. 2012; 6(3):1134–1161.

Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73(1):13–22.

Lin X, Lee S, Christiani DC, Lin X. Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics. 2013; 14(4):667–681. [PubMed: 23462021]

Lin X, Lee S, Wu MC, Wang C, Chen H, Li Z, Lin X. Test for rare variants by environment interactions in sequencing association studies. Biometrics. 2016; 72(1):156–164. [PubMed: 26229047]

Mallows CL. Some comments on cp. Technometrics. 1973; 15(4):661–675.

Marceau R, Lu W, Holloway S, Sale MM, Worrall BB, Williams SR, Hsu FC, Tzeng JY. A fast multiple-kernel method with applications to detect gene-environment interaction. Genetic Epidemiology. 2015; 39(6):456–468. [PubMed: 26139508]

Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics. 2007; 39(7):906–913. [PubMed: 17572673]

McKusick, VA. Mendelian inheritance in man: a catalog of human genes and genetic disorders. Vol. 1. JHU Press; 1998.

Moore LV, Roux AVD, Nettleton JA, Jacobs DR. Associations of the local food environment with diet qualitya comparison of assessments based on surveys and geographic information systems the multi-ethnic study of atherosclerosis. American Journal of Epidemiology. 2008; 167(8):917–924. [PubMed: 18304960]

Mujahid MS, Roux AVD, Morenoff JD, Raghunathan TE, Cooper RS, Ni H, Shea S. Neighborhood characteristics and hypertension. Epidemiology. 2008; 19(4):590–598. [PubMed: 18480733]

Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. American Journal of Epidemiology. 2012; 175(3):177–190. [PubMed: 22199027]

Newey WK. Convergence rates and asymptotic normality for series estimators. Journal of Econometrics. 1997; 79(1):147–168.

Pan W. Asymptotic tests of association with multiple snps in linkage disequilibrium. Genetic Epidemiology. 2009; 33(6):497–507. [PubMed: 19170135]

Papas MA, Alberg AJ, Ewing R, Helzlsouer KJ, Gary TL, Klassen AC. The built environment and obesity. Epidemiologic Reviews. 2007; 29(1):129–143. [PubMed: 17533172]

Pfeufer A, van Noord C, Marciante KD, Arking DE, Larson MG, Smith AV, Tarasov KV, Müller M, Sotoodehnia N, Sinner MF, et al. Genome-wide association study of pr interval. Nature Genetics. 2010; 42(2):153–159. [PubMed: 20062060]

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR, Willer CJ. Locuszoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26(18):2336–2337. [PubMed: 20634204]

Sallis JF, Floyd MF, Rodríguez DA, Saelens BE. Role of built environments in physical activity, obesity, and cardiovascular disease. Circulation. 2012; 125(5):729–737. [PubMed: 22311885]

Schwarz G. Estimating the dimension of a model. The Annals of Statistics. 1978; 6(2):461–464.

Sotoodehnia N, Isaacs A, de Bakker PI, Dörr M, Newton-Cheh C, Nolte IM, Van der Harst P, Müller M, Eijgelsheim M, Alonso A, et al. Common variants in 22 loci are associated with qrs duration and cardiac ventricular conduction. Nature Genetics. 2010; 42(12):1068–1076. [PubMed: 21076409]

Tchetgen Tchetgen EJ, Kraft P. On the robustness of tests of genetic associations incorporating gene-environment interaction when the environmental exposure is misspecified. Epidemiology. 2011; 22(2):257–261. [PubMed: 21228699]

Thomas D. Gene–environment-wide association studies: emerging approaches. Nature Reviews Genetics. 2010; 11(4):259–272.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. The American Journal of Human Genetics. 2011; 89(2):277–288. [PubMed: 21835306]

VanderWeele TJ, Ko YA, Mukherjee B. Environmental confounding in gene-environment interaction studies. American Journal of Epidemiology. 2013; 178(1):144–152. [PubMed: 23821317]

VanderWeele TJ, Shpitser I. On the definition of a confounder. The Annals of Statistics. 2013; 41(1): 196–220. [PubMed: 25544784]

Vansteelandt S, VanderWeele TJ, Tchetgen EJ, Robins JM. Multiply robust inference for statistical interactions. Journal of the American Statistical Association. 2008; 103(484):1693–1704. [PubMed: 21603124]

Voorman A, Lumley T, McKnight B, Rice K. Behavior of qq-plots and genomic control in studies of gene-environment interaction. PloS One. 2011; 6(5):e19416. [PubMed: 21589913]

Wang L. Gee analysis of clustered binary data with diverging number of covariates. The Annals of Statistics. 2011; 39(1):389–417.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful snp-set analysis for case-control genome-wide association studies. The American Journal of Human Genetics. 2010; 86(6):929–942. [PubMed: 20560208]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics. 2011; 89(1): 82–93. [PubMed: 21737059]

Table 1

Simulation study evaluating the main effect adjustment of E (2526 subjects, 190 SNPs). iSKAT: set based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-linear: the proposed test with a linear main effect of E. GE-spline: the proposed test using natural cubic-spline smoothing for E with $\sqrt{m}$ basis functions. GE-true: the proposed test with the correct model, which correctly specifies the main effect of E. The GE methods were implemented using the weighted PCA approach for the main effect of G. Each cell presents type I error rate or power based on 1000 replicates evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07. The calibrated powers with value zero correspond to very high type I errors.

Cross-sectional data

Type I error rate

	C1 holds				**C1 does not hold**			
$h_M(E)$	E	E^2	$E + E^2$	exp(E)	E	E^2	$E + E^2$	exp(E)
iSKAT	0.055	0.170	0.109	0.181	0.054	0.899	0.751	0.947
rareGE	0.057	0.174	0.108	0.195	0.050	0.905	0.762	0.951
GE-linear	0.055	0.046	0.046	0.032	0.050	0.780	0.663	0.618
GE-spline	0.053	0.053	0.054	0.057	0.050	0.051	0.048	0.051
GE-true	0.053	0.053	0.052	0.055	0.051	0.050	0.048	0.053

Power

	C1 holds				C1 does not hold			
$h_M(E)$	E	E^2	$E + E^2$	exp(E)	E	E^2	$E + E^2$	exp(E)
iSKAT	0.751	0.549*	0.666*	0.553*	0.761	0*	0*	0*
rareGE	0.760	0.565*	0.681*	0.553*	0.771	0.329*	0.445*	0.185*
GE-linear	0.746	0.556	0.669	0.572	0.754	0.949	0.932	0.829
GE-spline	0.745	0.745	0.750	0.741	0.733	0.726	0.727	0.721
GE-true	0.750	0.747	0.754	0.743	0.756	0.740	0.739	0.744

Longitudinal data ($d = 4$)

Type I error rate

$h_M(E)$	C1 holds				C1 does not hold			
	E	E^2	$E + E^2$	$\exp(E)$	E	E^2	$E + E^2$	$\exp(E)$
GE-linear	0.045	0.042	0.040	0.039	0.049	0.906	0.729	0.869
GE-spline	0.043	0.042	0.043	0.042	0.048	0.048	0.048	0.048

Power

$h_M(E)$	C1 holds				C1 does not hold			
	E	E^2	$E + E^2$	$\exp(E)$	E	E^2	$E + E^2$	$\exp(E)$
GE-linear	0.793	0.606	0.731	0.681	0.796	0*	0*	0*
GE-spline	0.786	0.786	0.786	0.784	0.774	0.774	0.775	0.769
GE-true	0.789	0.788	0.788	0.789	0.796	0.780	0.780	0.782

Table 2

Simulation study evaluating the main effect adjustment of G (775 subjects). A linear main effect of E was fitted. Each cell presents the type I error rate/power based on 1000 replicates. MinP: single SNP analysis using GEE adjusted by the effective number of independent tests (Gao et al., 2008). iSKAT: region based test proposed by Lin et al. (2013). rareGE: rareGE test proposed by Chen et al. (2014) assuming a random main effect of G. GE-none: the proposed test adjusting for none of the SNPs. GE-wPCA/wPCAM-$\sqrt{m}$: the proposed test adjusting for the leading $\sqrt{m}$ components using the weighted PCA and robust(wPCA)/model-based(wPCAM) inference. GE-true: the proposed test with the correct model, which correctly correctly includes all SNPs with non-zero main effects. Type I error rate and power were both evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate > 0.07.

Cross-sectional data

	Type I error rate				Power			
	C1 holds		C1 does not hold		C1 holds		C1 does not hold	
q	400	700	400	700	400	700	400	700
MinP	0.052	0.047	0.116	0.129	0.426	0.355	0.676*	0.636*
iSKAT	0.050	0.066	0.055	0.060	0.557	0.505	0.780	0.747
rareGE	0.054	0.055	0.072	0.059	0.634	0.561	0.821*	0.810
GE-none	0.025	0.038	0.189	0.181	0.446	0.402	0.769	0.778
GE-wPCA-$\sqrt{m}$	0.030	0.038	0.040	0.030	0.540	0.514	0.771	0.741
GE-wPCAM-$\sqrt{m}$	0.041	0.038	0.048	0.048	0.591	0.561	0.805	0.785
GE-true	0.036	0.038	0.045	0.041	0.584	0.509	0.797	0.759

Longitudinal data

	Type I error rate				Power			
	C1 holds		C1 does not hold		C1 holds		C1 does not hold	
q	400	700	400	700	400	700	400	700
MinP	0.028	0.033	0.088	0.108	0.472	0.462	0.568*	0.575*
GE-none	0.045	0.039	0.187	0.173	0.560	0.564	0.477*	0.435*
GE-wPCA-$\sqrt{m}$	0.033	0.034	0.031	0.034	0.588	0.569	0.687	0.682
GE-wPCAM-$\sqrt{m}$	0.034	0.037	0.032	0.040	0.626	0.589	0.736	0.708

Longitudinal data								
	Type I error rate				Power			
	C1 holds		C1 does not hold		C1 holds		C1 does not hold	
q	400	700	400	700	400	700	400	700
GE-true	0.040	0.034	0.038	0.039	0.615	0.589	0.724	0.707

Table 3

Simulation study evaluating the use of longitudinal data. GE-CS/AR1: the proposed test with different working correlation (compound symmetric/first order autoregressive). GE-avg.: the proposed test using the average/baseline value of repeated measurements. iSKAT-avg.: cross-sectional iSKAT using the average value of repeated measurements. rareGE-avg.: rareGE test proposed by Chen et al. (2014) using the average value of repeated measurements, and assuming a random main effect of G. The GE methods were implemented using the weighted PCA approach for the main effect of G and natural cubic-spline for the main effect of E adjusting for $\sqrt{m}$ terms. Each cell presents type I error rate or power based on 1000 replicates evaluated at $\alpha = 0.05$. Power is empirically calibrated to $\alpha = 0.05$ and marked as "*" when a method has type I error rate >0.07.

Type I error rate

	Methods using full trajectory		Methods using average value		
d	**GE-CS**	**GE-AR1**	**GE-avg.**	**iSKAT-avg.**	**rareGE-avg.**
1	0.047	0.047	0.047	0.047	0.048
2	0.046	0.046	0.041	0.050	0.050
3	0.048	0.048	0.058	0.072	0.063
4	0.039	0.041	0.052	0.092	0.078

Power

	Methods using full trajectory		Methods using average value		
d	**GE-CS**	**GE-AR1**	**GE-avg.**	**iSKAT-avg.**	**rareGE-avg.**
1	0.379	0.379	0.379	0.387	0.400
2	0.581	0.581	0.390	0.415	0.430
3	0.722	0.719	0.431	0.422*	0.466
4	0.805	0.803	0.414	0.409*	0.456*

Table 4

Analysis of Multi-Ethnic Study of Atherosclerosis (MESA) data: interactions between neighborhood variables and the region indexed by rs10850411 on systolic blood pressure. Each cell shows the p-value. EUR: European Americans; AFA: African Americans; HIS: Hispanics; CHN: Asians of Chinese descent. Meta: Meta-analysis combining the results of four ethnic groups using Fisher's combined probability test. GE-linear: the proposed test with a linear main effect of *E*. GE-spline: the proposed test using $\sqrt{n}$ natural cubic-spline basis functions for the main effect of *E*. MinP: minimum p-value test based on GEE. The assumed working correlation is compound symmetric. iSKAT-avg./base.: cross-sectional iSKAT using the average/baseline value of repeated measurements as the outcome. rareGE-avg./base.: cross-sectional rareGE using the average/baseline value of repeated measurements as the outcome, where a random main effect of G is assumed. Bonferroni correction threshold is 0.00043.

Systolic Blood Pressure - Region Indexed by rs10850411 (190 -214 SNPs)										
	Density of favorable food stores					Density of recreational facilities				
	EUR	CHN	AFA	HIS	Meta	EUR	CHN	AFA	HIS	Meta
GE-linear	0.0427	0.0890	0.8651	0.6256	0.1353	0.7857	0.9631	0.4937	0.8130	0.9670
GE-spline	0.0570	0.0544	0.9320	0.7231	0.1366	0.8480	0.9640	0.5405	0.8891	0.9848
MinP	0.0602	0.5753	1.0000	1.0000	0.5664	1.0000	1.0000	1.0000	1.0000	1.0000
iSKAT-avg.	0.2416	0.3695	0.7435	0.9257	0.6942	0.5134	0.3400	0.3869	0.5226	0.5707
iSKAT-base.	0.3953	0.2459	0.7200	0.9298	0.7070	0.7215	0.3239	0.5886	0.7561	0.8068
rareGE-avg.	0.2421	0.3144	0.9448	0.7851	0.6754	0.5281	0.5386	0.5169	0.3989	0.6839
rareGE-base.	0.4524	0.1045	0.9670	0.8334	0.5875	0.8591	0.3717	0.5458	0.7175	0.8426
	Perceived Healthy Food Availability					Perceived walkability				
	EUR	CHN	AFA	HIS	Meta	EUR	CHN	AFA	HIS	Meta
GE-linear	0.0005	0.9736	0.7591	0.9270	0.0446	0.2812	0.3235	0.3384	0.1678	0.2297
GE-spline	0.0009	0.9067	0.8241	0.9034	0.0608	0.2127	0.3746	0.3166	0.2058	0.2303
MinP	0.0047	1.0000	1.0000	1.0000	0.2177	1.0000	1.0000	1.0000	1.0000	1.0000
iSKAT-avg.	0.8205	0.4296	0.4285	0.6091	0.7817	0.9028	0.6702	0.6318	0.7490	0.9617
iSKAT-base.	0.4331	0.5049	0.4503	0.5283	0.6571	0.8422	0.8531	0.6441	0.6475	0.9658
rareGE-avg.	0.7542	0.3500	0.5124	0.2821	0.5878	0.8363	0.4632	0.6999	0.6264	0.8956
rareGE-base.	0.5336	0.2642	0.3120	0.2036	0.3073	0.9141	0.7050	0.6569	0.3873	0.8900

J Am Stat Assoc. Author manuscript; available in PMC 2018 May 16.