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

Complete Effect-Profile Assessment in Association Studies With Multiple Genetic and Multiple Environmental Factors

Zhi Wang,[1] Arnab Maity,[2] Yiwen Luo,[1] Megan L. Neely,[3] and Jung-Ying Tzeng[1,2]*

[1] Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina, United States of America; [2] Department of Statistics, North Carolina State University, Raleigh, North Carolina, United States of America; [3] Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, United States of America

Received 7 May 2014; Revised 20 October 2014; accepted revised manuscript 5 November 2014.
Published online 23 December 2014 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21877

ABSTRACT: Studying complex diseases in the post genome-wide association studies (GWAS) era has led to developing methods that consider factor-sets rather than individual genetic/environmental factors (i.e., Multi-G-Multi-E studies), and mining for potential gene-environment (G×E) interactions has proven to be an invaluable aid in both discovery and deciphering underlying biological mechanisms. Current approaches for examining effect profiles in Multi-G-Multi-E analyses are either underpowered due to large degrees of freedom, ill-suited for detecting G×E interactions due to imprecise modeling of the G and E effects, or lack of capacity for modeling interactions between two factor-sets (e.g., existing methods focus primarily on a single E factor). In this work, we illustrate the issues encountered in constructing kernels for investigating interactions between two factor-sets, and propose a simple yet intuitive solution to construct the G×E kernel that retains the ease-of-interpretation of classic regression. We also construct a series of kernel machine (KM) score tests to evaluate the complete effect profile (i.e., the G, E, and G×E effects individually or in combination). We show, via simulations and a data application, that the proposed KM methods outperform the classic and PC regressions across a range of scenarios, including varying effect size, effect structure, and interaction complexity. The largest power gain was observed when the underlying effect structure involved complex G×E interactions; however, the proposed methods have consistent, powerful performance when the effect profile is simple or complex, suggesting that the proposed method could be a useful tool for exploratory or confirmatory G×E analysis.

Genet Epidemiol 39:122–133, 2015. © 2014 Wiley Periodicals, Inc.

KEY WORDS: factor-set association analysis; kernel machine regression; genetic-environmental interactions; joint and conditional tests

Introduction

Complex human diseases are influenced not only by genetic and environmental factors, but also by the interplay between the two. Investigating gene-environment (G×E) interactions can facilitate understanding the etiology of these phenotypes by providing insight into biological mechanisms associated with diseases [Mechanic et al., 2012; Murcray et al., 2009], by explaining heterogeneity across populations, by classifying risk subgroups based on differential environmental exposures [Kraft et al., 2007; Murcray et al., 2009; Thomas, 2010a,b], and by identifying novel genes acting through interactions but exhibiting minimal marginal effects [Thomas, 2010a].

Many statistical methods for evaluating single nucleotide polymorphism – environment interactions (SNP×E) effects have been developed (see Hutter et al. [2013] for a comprehensive review). Studying complex diseases in the post genome wide association studies (GWAS) era has led to the development of methods that consider factor-sets rather than individual genetic and environmental factors (referred to in this text as multi-G and multi-E methods, respectively). For example, assessing G×E with multi-G has drawn great attention in recent array- and sequence-based association studies [e.g., Jiao et al., 2013; Lin et al., 2013; Tzeng et al., 2011]. By pooling information across a set of genetic markers, these methods have improved power, either through aggregating genetic signals or by reducing degrees of freedom, to detect G×E signals missed by individual single-nucleotide polymorphism (SNP) analyses. However, most existing multi-G methods consider only a single E factor, and the need for Multi-G-Multi-E methods is becoming apparent as investigators have turned their attention to mining existing genomic data for potential G×E interactions. This shift in research focus is evidenced by the emergence of Multi-G-Multi-E studies in the recent literature [Dai et al. 2013; Edwards et al., 2013; Naj et al., 2013; Patel et al., 2013; Wu et al., 2012].

Focusing on regression-based approaches, the commonly adopted strategies for examining effect profiles in

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

*Correspondence to: Jung-Ying Tzeng, Department of Statistics and Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh, NC 27695, USA. E-mail: jytzeng@stat.ncsu.edu

Multi-G-Multi-E analyses are classic (naïve) regression, principal component (PC) regression, and kernel machine (KM) regression. Naïve regression uses each element of the factor-sets as an individual covariate in a regression model. PC regression first performs PC analysis on the genetic and environmental factor-sets and then uses a subset of the resulting PCs (e.g., the first PC or the top PCs that explain 80% of the variation in the set) in a regression analysis [Gauderman et al., 2007; Wang et al., 2009]. Investigating G×E interactions can easily be incorporated into these frameworks, but with potential drawbacks. Although being straight forward and easy to apply, naïve regression is usually underpowered due to the large number of predictors and the small effect sizes of individual factors. PC regression typically has better power than naïve regression; however, its application can be subjective when deciding how many PCs to include in the downstream analyses and valuable information can be lost. Moreover, naïve and PC regressions do not incorporate any prior information about the elements of the factor-sets.

In contrast, KM regression does not suffer from the burden of dimensionality-like naïve regression, and less information is lost to data reduction compared to PC regression. KM regression first computes pairwise similarities between subjects based on their covariate values using a prespecified kernel function and then performs least squares regression of the response on the similarity measures. In the KM framework, testing for the factor-set effects is reduced to testing for the nullity of the corresponding variance components rather than the actual parameters in the mean model. Typically, the corresponding degrees of freedom used by the KM test are much smaller than the number of model parameters, resulting in increased power [Kwee et al., 2008; Liu et al., 2007, 2008]. Kernels can be constructed to incorporate important biological information [Kwee et al., 2008; Wu et al., 2010]. In addition, KM regression can handle correlated factors, such as PC regression, and allows for the investigation of nonlinear effects, which is not practical in naïve regression and not feasible in PC regression.

KM methods have been developed to study interaction effects. However, most approaches are developed for detecting gene-gene interactions [Larson and Schaid, 2013; Maity and Lin, 2011; Wang et al., 2014]. For detecting G×E effects, methods similar to KM regressions have been developed under the framework of generalized linear-mixed model [e.g., Lin et al., 2013] or similarity regression [Tzeng et al., 2011] to examine the interaction between a single environmental factor and a set of genetic markers. Care needs to be taken when extending these approaches to appropriately investigate interactions under the multi-G and multi-E setting. Depending on the kernel used to model the multi-E factors, the G×E kernel may not be directly constructed by crossing the G and E kernel like when constructing the G×G kernels due to the potential risk of modeling duplicate G and E terms in the G×E kernel. (As illustrated in the Methods section, such risk often does not exist for gene-gene interaction if the linear or the identity-by-state (IBS) kernel is used to model the gene effect.) Others have proposed KM regression approaches that circumvent the need for directly constructing a kernel for interaction effects. For example, Clark et al. (2014) are in the process of developing a KM regression approach that attempts to separate the individual G and E effects from the combined joint effect through careful modeling. Then, using the remaining signal to capture the G×E interaction effect, they attempt to build a test based on this component, avoiding the need to specify any explicit kernel for the G×E interaction.

In this work, we illustrate the issues encountered in constructing kernels for investigating interactions between two factor-sets, and propose a simple yet intuitive solution to construct the G×E kernel that retains the ease-of-interpretation of classic regression. We also construct a series of score tests to evaluate the complete effect profile. That is, each component of the effect profile (i.e., G, E, and G×E interaction) can be examined individually or in conjunction with other components to understand the effect patterns. As such, the approach can be used as an exploratory tool to investigate the effect profile in G×E studies. Unlike naïve regression or PC regression, developing tests under a KM regression may be nontrivial depending on the null hypothesis, which dictates the number of nuisance variance components that need to be estimated when deriving the test statistic. We propose an expectation–maximization (EM) algorithm to efficiently estimate the variance components. Finally, we demonstrate through simulation studies and a data application that our proposed method can have markedly improved power over the commonly used approaches for investigating G×E interactions in a Multi-G-Multi-E setting.

Methods

For individual i with $i = 1, \ldots, n$, let Y_i be the continuous trait value, $G_i = (g_{i1}, g_{i2}, \ldots, g_{iL})$ be a $L \times 1$ genotype vector for the L SNPs of interests, $E_i = (e_{i1}, e_{i2}, \ldots, e_{iM})$ be a $M \times 1$ design vector for the M environmental factors, and $X_i = (x_{i1}, x_{i2}, \ldots, x_{iQ})$ be a $Q \times 1$ design vector for covariates that are not included in either G_i or E_i. We consider the following regression to model the relationship between the trait value Y_i and the genetic factors G_i, the environmental factors E_i, and their interactions GE_i after adjusting for the additional covariates X_i:

$$Y_i = X_i^T\beta + h_G(G_i) + h_E(E_i) + h_{GE}(GE_i) + \varepsilon_i, \qquad (1)$$

where β is a $Q \times 1$ vector of regression coefficients describing the effects of the covariates X_i, ε_i's are independent random errors that follow a $\mathcal{N}(0, \sigma)$ distribution, and $h_G(\cdot)$, $h_E(\cdot)$, and $h_{GE}(\cdot)$ are smooth, vector-valued functions that capture the genetic, environmental, and G×E effects, respectively. There are many possible choices for the functions $h_*(\cdot)$; e.g., specifying a linear function corresponds to traditional linear regression. Under the KM framework, $h_*(\cdot)$ can be specified through a linear combination of a positive definite kernel function $K_*(\cdot, \cdot)$ and is assumed to lie in the functional space generated by that kernel [Kimeldorf and Wahba, 1970;

Liu et al., 2008]. Following the representation theorem, the functions $h_G(\cdot)$, $h_E(\cdot)$, and $h_{GE}(\cdot)$ can be written as $h_G(G_i) = \sum_{(i'=1)}^{n} \alpha_{Gi'} K_G(G_i, G_{i'})$, $h_E(E_i) = \sum_{(i'=1)}^{n} \alpha_{Ei'} K_E(E_i, E_{i'})$, and $h_{GE}(G_i, E_i) = \sum_{i'=1}^{n} \alpha_{GEi'}\ K_{GE}((G_i, E_i), (G_{i'}, E_{i'}))$, respectively, where $\alpha_{*i'}$ is the unknown parameter and the kernel function $K_*(\cdot, \cdot)$ is a distance metric that quantifies the similarity between subject i and subject i'. Some commonly used kernel functions include the linear kernel function, given by $K_z(z_i, z_{i'}) = z_i^T z_{i'}$, the IBS kernel for genetic data, given below in (2), and the polynomial kernel function, given by $K_z(z_i, z_{i'}) = (1 + z_i^T z_{i'})^d$ with a constant d.

For the genetic kernel, we consider the commonly adopted IBS kernel [Kwee et al., 2008]:

$$\begin{aligned} K_G(G_i, G_{i'}) &= \frac{1}{2\sum w_\ell} \sum_{\ell=1}^{L} w_\ell IBS(g_{i\ell}, g_{i'\ell}) \\ &= \frac{1}{2\sum w_\ell} \sum_{\ell=1}^{L} w_\ell \{2 \cdot I(g_{i\ell} = g_{i'\ell}) \\ &\quad + I(|g_{i\ell} - g_{i'\ell}| = 1)\}, \end{aligned} \tag{2}$$

where $IBS(g_{i\ell}, g_{i'\ell})$ denotes the number of alleles shared by subject i and subject i' at SNP ℓ and w_ℓ is the SNP-specific weight. The weight can be used to incorporate prior information about the genetic variants such as based on allele frequencies, functionality, or degree of evolutionary conservation. Because similarity in rare alleles is more informative than similarity in common alleles, we use $w_\ell = f_\ell^{-3/4}$ as recommended by Pongpanich et al. [2012], where f_ℓ is the minor allele frequency (MAF) of SNP ℓ. With this weight, the contribution of rare variants is up-weighted, but not too strongly so that the contribution of common variants can still be retained. For the environmental kernel, we use the interactive kernel [Maity and Lin, 2011]:

$$K_E(E_i, E_{i'}) = 1 + \sum_{m=1}^{M} e_{im} e_{i'm} + \sum_{m<k} e_{im} e_{ik} e_{i'm} e_{i'k}, \tag{3}$$

which explicitly includes two-way interactions along with the main effects of the environmental variables. If one wishes to include quadratic effects in the model, then the second-order polynomial kernel can be specified instead.

G × E Interaction Kernel

We aim to construct the G×E kernel, $K_{GE}(\cdot, \cdot)$, directly based on $K_G(\cdot, \cdot)$ and $K_E(\cdot, \cdot)$ so that a G×E analysis under the KM framework can enjoy the same ease-of-interpretation as the naïve or PC regression. However, the construction is not a straightforward task due to the concern of introducing duplicate G and E terms in the G×E kernel. That is, when constructing $K_{GE}(\cdot, \cdot)$ as a function of $K_G(\cdot, \cdot)$ and $K_E(\cdot, \cdot)$, such as taking the product, marginal genetic and environmental effects often appear in $K_{GE}(\cdot, \cdot)$, while these terms are already captured in $K_G(\cdot, \cdot)$ or $K_E(\cdot, \cdot)$. For example, in our setting, if we constructed $K_{GE}(\cdot, \cdot)$ as the product of $K_G(\cdot, \cdot)$ and $K_E(\cdot, \cdot)$, the G×E kernel would be defined as:

$$\begin{aligned} &K_{GE}((G_i, E_i), (G_{i'}, E_{i'})) \\ &\quad = K_G(G_i, G_{i'}) K_E(E_i, E_{i'}) \\ &\quad = K_G(G_i, G_{i'}) \left(1 + \sum_{m=1}^{M} e_{im} e_{i'm} + \sum_{m<k} e_{im} e_{ik} e_{i'm} e_{i'k} \right) \\ &\quad = K_G(G_i, G_{i'}) + K_G(G_i, G_{i'}) \\ &\qquad \times \left(\sum_{m=1}^{M} e_{im} e_{i'm} + \sum_{m<k} e_{im} e_{ik} e_{i'm} e_{i'k} \right), \end{aligned} \tag{4}$$

where the first term duplicates the genetic main effect and is introduced by the constant in $K_E(\cdot, \cdot)$. The presence of duplicate terms in the G×E kernel causes colinearity in Model (1) and leads to invalid conclusion about the interaction effects.

The overlap in (4) suggests a simple yet effective solution for directly constructing $K_{GE}(\cdot, \cdot)$ using $K_G(\cdot, \cdot)$ and $K_E(\cdot, \cdot)$: redefine the environmental kernel as $K_E^*(E_i, E_{i'}) = K_E(E_i, E_{i'}) - 1$ and then calculate the G×E kernel as:

$$\begin{aligned} &K_{GE}((G_i, E_i), (G_{i'}, E_{i'})) \\ &\quad = K_G(G_i, G_{i'}) K_E^*(E_i, E_{i'}) \\ &\quad = K_G(G_i, G_{i'}) \times \left(\sum_{m=1}^{M} e_{im} e_{i'm} + \sum_{m<k} e_{im} e_{ik} e_{i'm} e_{i'k} \right). \end{aligned} \tag{5}$$

The approach in (5) can also be used for general kernel specification. For example, if one chooses to use a polynomial kernel for $K_G(\cdot, \cdot)$ or $K_E(\cdot, \cdot)$ instead of the IBS or interaction kernel, respectively, one can define $K_z^*(z_i, z_{i'}) = K_{z,polynomial}(z_i, z_{i'}) - 1$ for $Z = G$ or E first, and then construct $K_{GE}(\cdot, \cdot)$ by taking the element-wise product of $K_G(\cdot, \cdot)$ and $K_E(\cdot, \cdot)$.

Score Tests for Assessing Multi-G-Multi-E Effects

In a Multi-G-Multi-E setting, there may be several null hypotheses of interest depending on the goal of the analysis. For example, in a confirmatory analysis, the goal may be to replicate a G×E interaction signal and the G×E effect would be tested individually. However, in an exploratory analysis, the goal maybe to look for any evidence of a relationship between the genetic and environmental factors and the response and the G, E, and G×E effects would be tested jointly. To address these needs, we develop a series of tests based on the Model (1).

First, when little is known a priori, a joint test, defined as $H_0^{Joint} : h_G(\cdot) = h_E(\cdot) = h_{GE}(\cdot) = 0$, serves as a good tool to detect the overall association induced by genetic and environmental main effects or by G×E interaction effects. Instead of beginning with a scan of genetic main effects, investigators can begin with a full scan using the joint test for the overall association, involving both genetic and environmental effects. A scan by joint tests may lead to increased flexibility and power to detect a signal because some genes can exhibit negligible marginal effects, but strong effects among particular exposure groups [Kraft et al., 2007; Thomas, 2010a].

If the joint test is rejected, a G×E test, defined as H_0^{GE} : $h_{GE}(\cdot) = 0$, can then be used to identify whether the effect of the genetic variables are modified by the environmental variables. The ability to model the G×E effects separately from the main effects can be extremely useful in the Multi-G-Multi-E setting for several reasons. First, the interaction test can aid in understanding biological mechanisms and pathways [Mechanic et al., 2012]. For example, Vineis et al. 2001 found interactions between gene *NAT*2 and tobacco smoking and other occupational exposures when studying bladder cancer. These results revealed a potential role of arylamines (found in tobacco smoke and other occupational materials such as hair dyes) in the pathogenesis of bladder cancer—*NAT2* is involved in the detoxification of arylamine, and only subjects with the "slow-detox" genotype were at increased risk of cancer when exposed to arlymines. Second, interaction tests can be used to identify novel genes functioning through interactions and to explain "missing heritability." Many studies of complex diseases, including childhood asthma, breast cancer, and colorectal cancer, are currently under way to search for genes interfering with different environmental factors [Thomas, 2010a]. Third, although statistical interaction is not entirely consistent with biological interaction [Thompson, 1991], interaction tests can still help to improve the performance of risk prediction models for disease therapies by identifying genotypes that respond differently for given treatments—a key task in pharmacogenomics studies [Murcary et al., 2009].

Furthermore, if there is no evidence of a G×E interaction, conditional genetic and environmental effects can be further evaluated by testing $H_0^{G|E}$: $h_G(\cdot) = 0$ without constraining $h_E(\cdot)$ but under the constraint of $h_{GE}(\cdot) = 0$, and testing $H_0^{E|G}$: $h_E(\cdot) = 0$ without constraining $h_G(\cdot)$ but under the constraint of $h_{GE}(\cdot) = 0$, respectively. We develop conditional tests, $G|E$ and $E|G$, rather than marginal tests of G or E, because they are usually more meaningful and interpretable in the context of Multi-G-Multi-E studies. That is, researchers are often interested in investigating the incremental information about the response provided by genetic effects, e.g., over and above other covariates in the mean model. Additionally, the conditional tests can be used to understand the inconsistent association findings because marginal associations, often called the crude associations [Robins and Morgenstern, 1987], may disappear after taking differences in other genetic and environmental factors into consideration.

Using an argument similar to Liu et al. [2007], we show in the Supplementary Note that Model (1) has an equivalent linear mixed model representation and can be expressed as:

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{h_G} + \boldsymbol{h_E} + \boldsymbol{h}_{\boldsymbol{GE}} + \boldsymbol{\varepsilon} \qquad (6)$$

where $h_G \sim \mathcal{N}(0, \tau_G K_G)$, $h_E \sim \mathcal{N}(0, \tau_E K_E)$, $h_{GE} \sim \mathcal{N}(0, \tau_{GE} K_{GE})$, and $\varepsilon \sim \mathcal{N}(0, \sigma I_n)$. In addition, testing $H_0{:}h_z(z) = 0$ is equivalent to testing $H_0{:}\tau_z = 0$ for $z \in \{G, E, GE\}$. Based on these results, we derive the score tests for the joint test, G×E test, the conditional G test (G|E test), and the conditional E test (E|G test) based on the restricted maximum likelihood (REML) of Model (6) in the Supplementary Note. Specifically, the test statistic for the joint test, G×E test, G|E test, and E|G test are given as follows:

$$\begin{aligned} T_{joint} &= \frac{1}{2} Y^T P_0 (K_G + K_E + K_{GE}) P_0 Y|_{\tau_{GE}=\tau_G=\tau_E=0, \widehat{\sigma_{joint}}}, \\ T_{GE} &= \frac{1}{2} Y^T P_1 K_{GE} P_1 Y|_{\tau_{GE}=0, \tau_G=\widehat{\tau_G}, \tau_E=\widehat{\tau_E}, \sigma=\widehat{\sigma_{GE}}}, \\ T_{G|E} &= \frac{1}{2} Y^T P_2 K_G P_2 Y|_{\tau_{GE}=0, \tau_G=0, \tau_E=\widetilde{\tau_E}, \sigma=\widetilde{\sigma_{G|E}}}, \quad \text{and} \\ T_{E|G} &= \frac{1}{2} Y^T P_3 K_E P_3 Y|_{\tau_{GE}=0, \tau_G=\widetilde{\tau_G}, \tau_E=0, \sigma=\widetilde{\sigma_{E|G}}}, \end{aligned} \qquad (7)$$

where $Y^T = (Y_1, \ldots, Y_n)$, $P_t = V_t^{-1} - V_t^{-1} X (X^T V_t^{-1} X)^{-1} X^T V_t^{-1}$ for $t = \{0, 1, 2, 3\}$, $K_z = K_z(\cdot, \cdot)$ for $z \in \{G, E, GE\}$, $V_0 = \sigma I_n$, $V_1 = \tau_G K_G + \tau_E K_E + \sigma I_n$, $V_2 = \tau_E K_E + \sigma I_n$, and $V_3 = \tau_G K_G + \sigma I_n$. In the Supplementary Note, we also derive the EM algorithms, following similar steps as described in Tzeng et al. [2011], to obtain the estimates of the nuisance variance components (i.e., $\widehat{\sigma_{joint}}$, $\widehat{\tau_G}$, $\widehat{\tau_E}$, $\widehat{\sigma_{GE}}$, $\widetilde{\tau_E}$, $\widetilde{\sigma_{G|E}}$, $\widetilde{\tau_G}$, and $\widetilde{\sigma_{E|G}}$) under certain null hypotheses. As shown in the Supplementary Note, these test statistics can be rewritten as a quadratic form of $Z^T A Z$, where Z follows a standard multivariate normal distribution. By eigenvalue decomposition of A (and obtaining nonzero eigenvalues $\eta_1, \ldots, \eta_c$ and the corresponding eigenvectors $e_1, \ldots, e_c$), we have $Z^T A Z = \sum_{i=1}^{c} \eta_i (e_i^T Z)^2 \equiv \sum_{i=1}^{L} \eta_i \tilde{Z}_i^2$ with $\tilde{Z}_i^2$ following a 1 degrees of freedom (df) chi-square distribution. Therefore, the distribution of the score-test statistics can be approximated by a weighted chi-squared distribution, and *P* values can be obtained by moment matching approaches [Duchesne and Lafaye De Micheaux, 2010].

Simulation Studies

We performed simulation studies to compare the performance of the proposed KM regression method to the performance of methods currently available for performing analyses in a Multi-G-Multi-E setting. Our simulation studies were based on a haplotype distribution of 12 SNPs from the AGRT1 gene [French et al., 2006]. The haplotype distribution as well as the MAFs and linkage disequilibrium (LD) coefficients for each SNP are given in Supplementary Table S1. The LD was quantified by the average pairwise R^2 between each SNP and the remaining 11 SNPs.

We generated the 12-SNP genotypes of an individual by randomly drawing two haplotypes based on the haplotype distribution in Supplementary Table S1 and then forming the unphased genotypes. We generated five environmental factors (E factors in short) from the multivariate normal distribution $MVN_5(0, V)$ where $V_{\ell m} = \sigma_E^2 \cdot \rho_E^{I\{\ell \neq m\}}$ with $\sigma_E^2 = 1$ and $\rho_E = 0.3$ (low correlation among E factors) and 0.7 (high correlation among E factors). We considered two causal SNPs out of the 12 SNPs under four scenarios (Supplementary Table S2), i.e., the two causal variants had low vs. high MAFs and were in low vs. high LD. We randomly selected two of the five E factors and set them as causal for $\rho_E = 0.3$ and

Table 1. Type I error rates of the joint test and the G×E test averaged over 10,000 replicates

$(\gamma_G, \gamma_E, \gamma_{GE})$	(a_1, a_2, a_{12})	Nominal level	KM_G^{IBS} [a]	KM_G^{INT}	PC1	PC80	LM
Joint test		0.05	0.045	0.046	0.050	0.053	0.053
(0,0,0)	(0,0,0)	0.005	0.0046	0.0042	0.0051	0.0052	0.0048
		0.001	0.0010	0.0009	0.0010	0.0009	0.0009
G×E *test*							
(1,1,0)	(1,1,1)	0.05	0.044	0.038	0.095	0.218	0.282
		0.005	0.0045	0.0034	0.0192	0.0628	0.0929
		0.001	0.0008	0.0009	0.0073	0.0273	0.0410
	(1,1,0.5)	0.05	0.041	0.042	0.055	0.107	0.121
		0.005	0.0046	0.0042	0.0084	0.0191	0.0228
		0.001	0.0012	0.0012	0.0020	0.0065	0.0065
	(0.5,0.5,1)	0.05	0.040	0.041	0.128	0.248	0.311
		0.005	0.0041	0.0033	0.0299	0.0727	0.0932
		0.001	0.0011	0.0007	0.0111	0.0293	0.0402
	(0,0,1)	0.05	0.038	0.046	0.143	0.261	0.288
		0.005	0.0046	0.0056	0.0385	0.0742	0.0951
		0.001	0.0013	0.0011	0.0164	0.0327	0.0424

[a] KM_G^{IBS}, GE kernel machine regression with the IBS kernel for the genetic factors; KM_G^{INT}, GE kernel machine regression with the interactive kernel for the genetic factor; PC1, principal component regression using only the first G and E PC and its interaction; PC80, principal component regression using the top G and E PCs that explain 80% of the variation in the set and their interactions. LM, linear regression with covariates, including the 12 genetic and five environment factors as well as the 60 G×E interaction terms.

for $\rho_E = 0.7$. Then the phenotype values were generated by $Y_i = \mu(G_i, E_i) + \varepsilon_i$, where ε_i were generated from a $\mathcal{N}(0, \sigma)$ distribution and $\mu(G, E) = \gamma_G(a_1G_1 + a_2G_2 + a_{12}G_1G_2) + \gamma_E(a_1E_1 + a_2E_2 + a_{12}E_1E_2) + \gamma_{GE}(G_1E_1 + G_2E_2 + G_1G_2E_1E_2)$. We considered four sets of (a_1, a_2, a_{12}) : (1,1,1), i.e., equal effect size of main and interaction effects; (1,1,0.5), i.e., weaker interaction effects; (0.5,0.5,1), i.e., weaker main effects; and (0,0,1), i.e., interaction effect only. To explore the scale dependence of interaction tests, we also considered

$$\begin{aligned}\mu(G, E) = {} & \gamma_G(a_1G_1 + a_2G_2 + a_{12}G_1G_2)\\ & + \gamma_E\{a_1(E_1 - \min(E1)) + a_2(E_2 - \min(E2))\\ & + a_{12}(E_1 - \min(E1))(E_2 - \min(E2))\}\\ & + \gamma_{GE}\left\{\begin{array}{c}\log(G_1 + 0.001)\log(E_1 - \min(E_1))\\ + \log(G_2 + 0.001)\log(E_2 - \min(E2))\\ + \log(G_1 + 0.001)\log(G_2 + 0.001)\\ \log(E_1 - \min(E1))\log(E_2 - \min(E2))\end{array}\right\}.\end{aligned} \tag{8}$$

We set $(\gamma_G, \gamma_E, \gamma_{GE}) = (1,1,0.5)$, $(a_1, a_2, a_{12}) = (1,1,1)$, $\sigma_E^2 = 1$, and $\rho_E = 0.3$ so that the results for the "log-scale" model would be comparable to the results from the "nonlog-scale" model.

When examining the Type I error rates of the proposed tests, we set $(\gamma_G, \gamma_E, \gamma_{GE}) = (0,0,0)$ for the joint test, (1,1,0) for the G×E test, (0,1,0) for the G|E test, and (1,0,0) for the E|G test. When evaluating power, we set $(\gamma_G, \gamma_E, \gamma_{GE}) = (0.15,0.15,0.15)$ for the joint test, (1,1,0.5) for the G×E test, (0.2,1,0) for the G|E test, and (1,0.2,0) for the E|G test. These γ values were chosen so that the power of detecting a signal fell within a reasonable range for comparisons.

In each simulation setting, samples of size 200 were generated. We generated 100 replicates for power simulations and 10,000 replicates for evaluating Type I error rates at the nominal rates of 0.05, 0.005, and 0.001. Each replicate was analyzed using the following methods: (1) GE-KM: the proposed KM regression that constructed K_E based on the five E factors, using the interactive kernel [Maity and Lin, 2011]; constructed K_G based on the 12 SNPs, using the IBS kernel (with weights $w_\ell = f_\ell^{-3/4}$ [Pongpanich et al., 2012]; referred to as KM_G^{IBS}) or the interactive kernel (referred to as KM_G^{INT}); and constructed K_{GE} directly using K_G and K_E as described in the method section; (2) PC1: PC regression, including the first PCs of the G set and the E set, respectively, and their interaction; (3) PC80: PC regression, including the top PCs that explain 80% of the variation in the G set and in the E set, respectively, and their two-way interactions; and (4) LM: linear regression, including all 12 SNPs, five environmental factors, and the 5 × 12 pairwise G×E terms. All tests were performed at an α-level of 0.05.

Results

Type I error rates are presented in Table 1. Power results are presented in Figures 1–4 (for $\rho_E = 0.3$) and Supplementary Figures S1–S4 (for $\rho_E = 0.7$). Each figure contains four plots for the four simulation scenarios—low vs. high MAF (across rows) and low vs. high LD (across columns). Effect sizes in terms of (a_1, a_2, a_{12}) for the causal genetic and environmental factors are shown on the x-axis.

The Joint Test

The Type I error analysis suggested that all methods had desirable and similar performances under a null model (Table 1). For power (Fig. 1 for $\rho_E = 0.3$ and Supplementary Fig. S1 for $\rho_E = 0.7$), GE-KM methods had the best power in all scenarios, and the power was similar between the IBS kernel (KM_G^{IBS}) and interactive kernel (KM_G^{INT}). The large power gain of the GE-KM methods tended to occur when there existed strong $G_1 \times G_2$ effects and $E_1 \times E_2$ effects, i.e., relatively large a_{12} compared to a_1 and a_2. This is likely

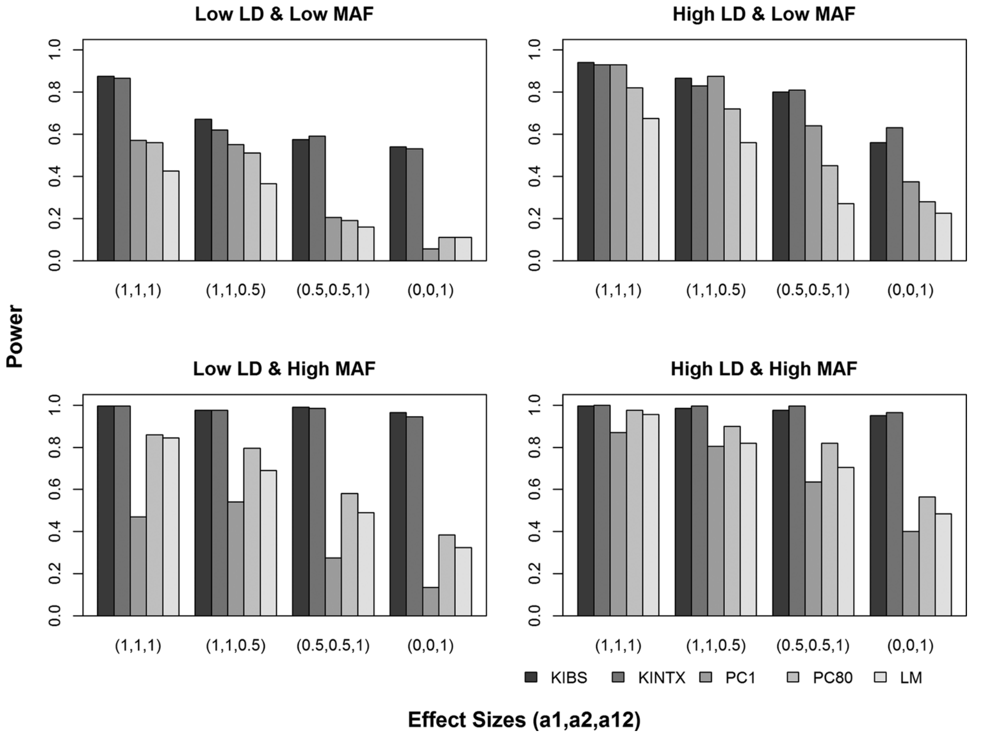

Figure 1. Power results for the Joint Test with low correlation among the environmental factors ($\rho_E = 0.3$). The results were based on 100 runs of the joint test H_0^{Joint} : $h_G(\cdot) = h_E(\cdot) = h_{GE}(\cdot) = 0$ at $\alpha = 0.05$. *KIBS* and *KINTX* are the proposed GE kernel machine method with IBS kernel and interactive kernel for the genetic factors, respectively, and the interactive kernel was used for the environmental factors; PC1 is the PC regression using only the first PC of the G and E effects; PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set; LM is the linear regression including all 12 SNPs, five environmental factors, and the 5 × 12 pairwise G×E terms.

because PC regressions captured the additive effects and had limited power when the underlying effect structure involved only complex interactions. Unlike GE-KM that showed consistent and good power performance across all simulation scenarios, the relative performance of benchmark methods depended on the MAF and LD of causal SNPs and the complexity of effect structure. Specifically, with low MAF, the relative power was PC1 > PC80 > LM, which seemed to reflect the impact of the df used in the joint test (i.e., the smaller df, the higher power). With high MAF, the relative power became PC80 > LM > PC1, which seemed to reflect a trade-off between df and information captured. PC80 appeared to be the least sensitive among PC80, PC1, and LM. The results suggested that GE-KM could be a powerful and robust analytic tool in the Multi-G-Multi-E setting.

The G×E Test

The Type I error rates (Table 1) for PC1, PC80, and LM were much higher than the nominal level. This is possibly because the G×E interaction term fitted in these models did not capture the G×G and E×E effects that then led to spurious G×E association. Similar results have been reported under single-G-single-E analysis [Voorman et al., 2011]. The Type I error rates for the GE-KM methods were around or slightly less than the nominal level. The power results are shown in Figure 2 (for $\rho_E = 0.3$) and Supplementary Figure S2 (for $\rho_E = 0.7$). The GE-KM approaches had higher power than benchmark methods, which had inflated Type I error rates. The power loss of the PC and LM methods suggests that the benchmark methods did not correctly model the G×E effects and resulted in the inflated false-positive and false-negative rates. Between the two GE-KM methods, KM_G^{INT} tended to have similar or slightly higher power than KM_G^{IBS}. This pattern is somewhat expected because KM_G^{INT} explicitly captured the interaction terms in the G set, so that the G×E effects can be more precisely modeled and efficiently detected by KM_G^{INT}. In summary, GE-KM is able to maintain the Type I error rate while achieving higher power to detect a G×E signal comparing to the benchmarks.

The power of the G×E test from the log-scale model is summarized in Table 2. Comparing the power of the G×E tests from the "log-scale" model (Table 2) and "nonlog-scale" model (i.e., the first set of five bars in Fig. 2), we found the KM_G^{IBS} seems to be more robust to interaction scales than KM_G^{INT}. Specifically, when (1) the true interaction effects and the modeled interaction effects are both on the original scale (i.e., $G \times E$), KM_G^{INT} has similar/slightly better power than KM_G^{IBS} and higher power than the inflated benchmark methods, PC1, PC80, and LM. However, when (2) the true

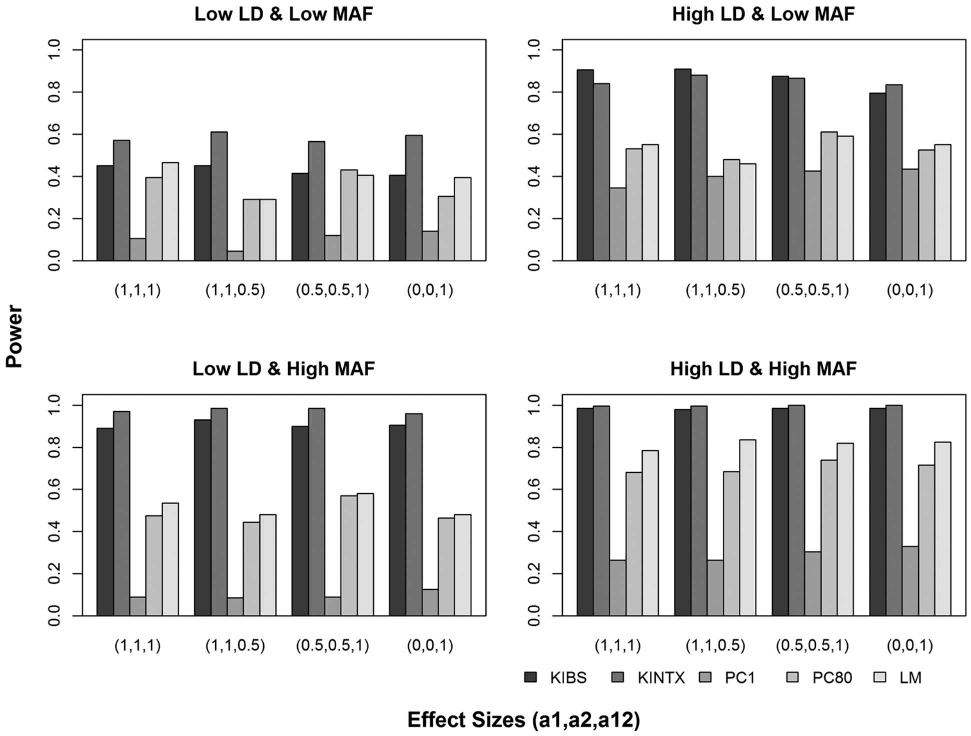

Figure 2. Power results for the G×E test with low correlation among the environmental factors ($\rho_E = 0.3$). The results were based on 100 runs of the G×E test $H_0^{G\times E} : h_{GE}(\cdot) = 0$ at $\alpha = 0.05$. *KIBS* and *KINTX* are the proposed GE kernel machine method with IBS kernel and interactive kernel for the genetic factors, respectively, and the interactive kernel was used for the environmental factors; PC1 is the PC regression using only the first PC of the G and E effects; PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set; LM is the linear regression, including all 12 SNPs, five environmental factors, and the 5 × 12 pairwise G×E terms.

interaction effects are on log scale (i.e., $\log G \times \log E$) and the modeled interaction effects are on the original scale (i.e., $G \times E$), KM_G^{INT} has lower power than KM_G^{IBS}; while KM_G^{IBS} still has power higher than the inflated benchmark tests just as in scenario (1), KM_G^{INT} has lower power than PC80 and LM.

The G|E Test (When no G×E Effect)

All methods had desirable and similar Type I error rates around the nominal level under the null models (Table 3). The power results for detecting the effect of the G set (i.e., $a_1G_1 + a_2G_2 + a_{12}G_1G_2$), after conditioning on the effect of the E set, are shown in Figure 3 (for $\rho_E = 0.3$) and Supplementary Figure S3 (for $\rho_E = 0.7$). First, we observed that KM_G^{INT} consistently had the best or near best power across all scenarios to detect the G effects, while KM_G^{IBS} had similar or considerably less power than KM_G^{INT}. Compared to KM_G^{INT}, the power loss of KM_G^{IBS} tended to occur with high MAF, for which scenarios the G×G interactions were stronger. The relative performance of PC1, PC80, and LM varied from scenario to scenario. Generally speaking, PC80 tended to have the best power among PC1, PC80, and LM, while PC1 tended to have the worst power among the three benchmark methods.

The E|G Test (When no G×E Effect)

All methods had desirable and similar Type I error rates around the nominal level under the null models (Table 3). The power results for detecting the E-set effect (i.e., $a_1E_1 + a_2E_2 + a_{12}E_1E_2$) after adjusting for the G-set effect are shown in Figure 4 (for $\rho_E = 0.3$) and Supplementary Figure S4 (for $\rho_E = 0.7$). Across all scenarios, GE-KM performed better than or comparable to the benchmark methods. The two GE-KM methods had similar power, suggesting the ability of evaluating the E effects is not very sensitive to the choices of kernels for the G set. Among the benchmark methods, the relative performance varied from scenario to scenario. Generally speaking, PC1 tended to have the worst power; the power of LM tended to be similar or slightly better (e.g., when strong E main effects) than that of PC80. Finally, the large power gain of the KM methods over PC1, PC80, and LM tended to occur in the presence of a relatively strong $E_1 \times E_2$ effect (e.g., $a_{12} > a_1$ and a_2), especially when $(a_1, a_2, a_{12}) = (0, 0, 1)$ (i.e., no main effects).

We also considered a grid of $(\gamma_G, \gamma_E, \gamma_{GE})$ for $(a_1, a_2, a_{12}) =$ (1,1,1) and (0,0,1) with $\rho_E = 0.3$ for a subset of the methods (i.e., KM_G^{IBS}, PC1, and PC80). The results (shown in Supplementary Figs. S5–S8) suggested a very similar pattern of relative performance as presented above. In summary, GE-KM was able to yield powerful and robust performance when evaluating the effect profile of a group of G factors and E

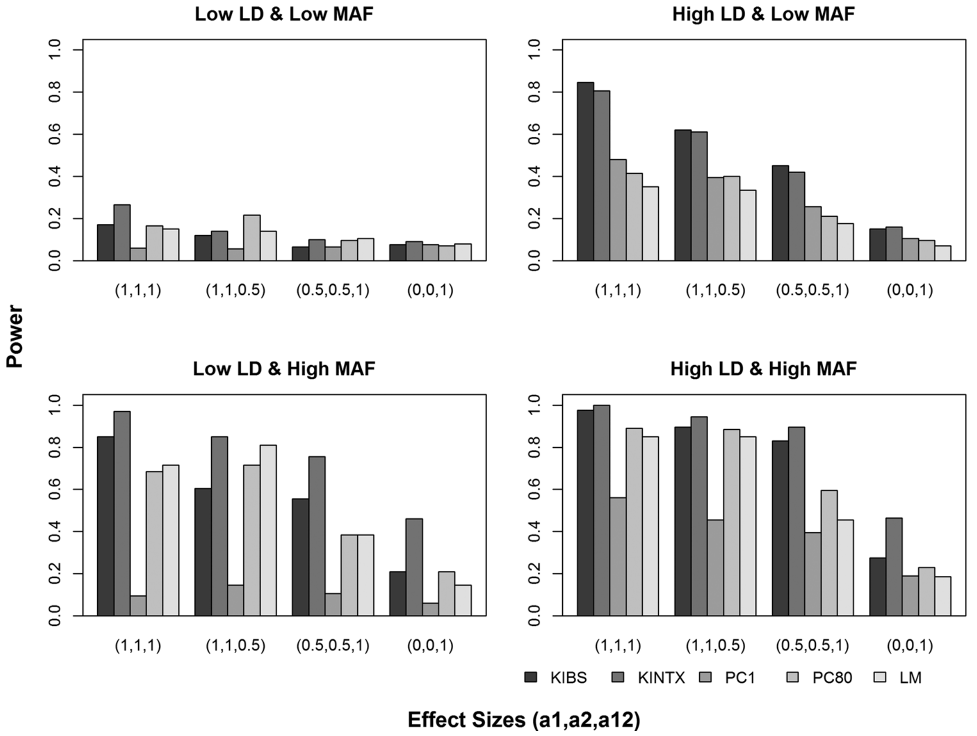

Figure 3. Power results for the G|E test with low correlation among the environmental factors (ρ_E = 0.3). The results were based on 100 runs of the G|E test $H_0^{G|E}$: $h_G(\cdot) = 0$ at α = 0.05. *KIBS* and *KINTX* are the proposed GE kernel machine method with IBS kernel and interactive kernel for the genetic factors, respectively, and the interactive kernel was used for the environmental factors; PC1 is the PC regression using only the first PC of the G and E effects; PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set; LM is the linear regression, including all 12 SNPs, five environmental factors, and the 5 × 12 pairwise G×E terms.

factors. Comparing to the benchmark methods, GE-KM consistently had the best or near best power in detecting the joint effect, G×E effect, the conditional G effect, and the conditional E effect with different causal SNP MAFs, with different correlation pattern among the G factors and among the E factors, and when the effect profile was simple or involved complex interactions between the G and E effects.

Real Data Example: Application to the CoLaus Study Data

We used the proposed GE-KM regression method to analyze data from the CoLaus study [Song et al., 2011]. The dataset contains the measured plasma level of the lipoprotein-associated phospholipase A2 (Lp-PLA2) enzyme in 87 subjects, which has been shown to be associated with risk of coronary heart disease. Song et al. 2011 studied the effect of the nonsynonymous rare variants (MAF < 0.05) in *PLA2G7* on Lp-PLA2 activity based on 29 carriers of the rare nonsynonymous variants and 58 matched noncarriers who were matched based on age, sex, and low-density lipoprotein cholesterol level. The study found significantly lower Lp-PLA2 activity in carriers compared to the noncarriers.

In our analysis, we aimed to examine the association between Lp-PLA2 enzyme activity and 11 common variants typed in the *PLA2G7* gene from the CoLaus study, a group of clinical risk factors of cardiovascular disease (CVD), and potential interactions between these genetic variants and CVD risk factors (i.e., the G, E, and G×E effects, respectively, estimated in the GE-KM method). The CVD risk factors considered in this analysis include homocystein, insulin, glucose, aspartate amino-transferase, triglycerides, apolipoprotein B, glomerula filtration rate, and body mass index. We also adjust for set of potential confounders: ethnic background, carrier status, age, gender, storage duration of serum, physical activity, smoking status, and alcohol assumption. We assess the association between the 11 common variants, the eight risk factors, and their interactions using the GE-KM methods (KM_G^{IBS} and KM_G^{INT}) as well as the benchmark methods (PC1 and PC80). These methods were as described in the Simulation section and the results of the series of tests are given in Table 4. We did not perform the LM analysis because the number of parameters (11 SNPs + 8 risk factors + 88 interaction terms) exceeds the sample size (87).

We first applied the joint test and found a significant signal (P-value = 0.009 for KM_G^{IBS} and 0.006 for KM_G^{INT}). We next performed the G×E test to assess the interactions between the *PLA2G7* variants and the CVD risk factors. The results suggest that there is no evidence of an interaction (P-value = 0.252 for KM_G^{IBS} and 0.174 for KM_G^{INT}) and that the major source of the joint association signal may arise from main genetic or environment effects. To further explore the joint

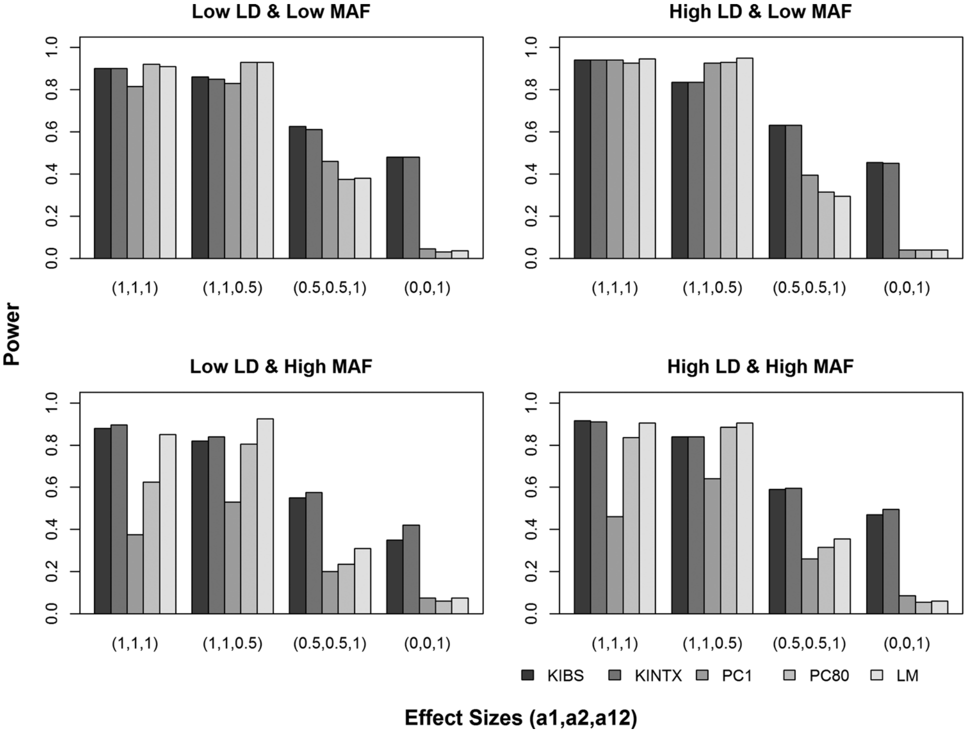

Figure 4. Power results for the E|G test with low correlation among the environmental factors (ρ_E = 0.3). The results were based on 100 runs of the E|G test $H_0^{E|G}$: $h_E(\cdot)$ = 0 at α = 0.05. *KIBS* and *KINTX* are the proposed GE kernel machine method with IBS kernel and interactive kernel for the genetic factors, respectively, and the interactive kernel was used for the environmental factors; PC1 is the PC regression using only the first PC of the G and E effects; PC80 denotes the PC regression using the top PCs that explain 80% of the variation in the set; LM is the linear regression, including all 12 SNPs, five environmental factors, and the 5 × 12 pairwise G×E terms.

Table 2. Power of G×E tests when the true G×E effect was on the log scale (i.e., log *G* × log *E*) while analyzed on the original scale (i.e., *G* × *E*)

$(\gamma_G, \gamma_E, \gamma_{GE})$	(a_1, a_2, a_{12})	(LD, MAF)	KM_G^{IBS} [a]	KM_G^{INT}	PC1	PC80	LM
(1,1,0.5)	(1,1,1)	Low, low	0.99	0.9	0.19	0.96	0.98
		Low, high	0.98	0.87	0.77	0.91	0.89
		High, low	0.7	0.39	0.37	0.55	0.87
		High, high	1	0.79	0.4	0.87	0.91

[a] KM_G^{IBS}, GE kernel machine regression with the IBS kernel for the genetic factors; KM_G^{INT}, GE kernel machine regression with the interactive kernel for the genetic factor; PC1, principal component regression using only the first G and E PC and its interaction; PC80, principal component regression, using the top G and E PCs that explain 80% of the variation in the set and their interactions. LM, linear regression with covariates, including the 12 genetic and five environment factors as well as the 60 G×E interaction terms.

signal, we then performed the G|E test and the E|G test to examine the effect of the *PLA2G7* variants and the CVD risk factors on Lp-PLA2 activity levels, respectively. The results of the conditional E test revealed no association between CVD risk factors and Lp-PLA2 levels (*P*-value = 0.326 for KM_G^{IBS} and 0.318 for KM_G^{INT}), while the conditional G test found evidence of an association between the *PLA2G7* variants and LP-PLA2 activity levels (*P*-value = 0.006 for both KM_G^{IBS} and KM_G^{INT}). These results suggest that the common variants may provide additional diagnostic information about enzyme activity after accounting for differences in CVD risk factors and rare *PLA2G7* variants as well as other potential confounders. We also performed SNP pairwise analyses, where for the 11 SNPs (denoted by as SNP1 to SNP11), we analyzed one SNP pair at a time (labeled as M1 and M2) by regressing the enzyme activities on M1, M2 and M1 × M2 and adjusting for the CVD risk factors and confounders. The significance of M1, M2, and their interactions was assessed at the significant level of $0.05/\binom{11}{2} = 3 \times 10^{-4}$. We found that SNP10 and SNP11 have significant main effects and there is no evidence of an interactive effect between any SNP pairs. The significant main effects of SNP10 and SNP11 showed up consistently when conditioning on SNP1, SNP2, SNP3, SNP4, SNP6, and SNP7, and had relatively small *P* values when conditioning on SNP5 and SNP8–SNP11. The conditional effect sizes of SNP10 and SNP11 are –34.80 and –28.78, respectively, on average (averaging across the 10 SNP pair

Table 3. Type I error rates of the G|E and E|G tests averaged over 10,000 replicates

$(\gamma_G, \gamma_E, \gamma_{GE})$	(a_1, a_2, a_{12})	Nominal level	KM_G^{IBS} [a]	KM_G^{INT}	PC1	PC80	LM
G|E test							
(0,1,0)	(1,1,1)	0.05	0.053	0.047	0.048	0.046	0.046
		0.005	0.0049	0.0052	0.0052	0.0053	0.0054
		0.001	0.0011	0.0011	0.0012	0.0011	0.0016
	(1,1,0.5)	0.05	0.043	0.047	0.055	0.053	0.057
		0.005	0.0046	0.0041	0.0051	0.0055	0.0049
		0.001	0.0008	0.0008	0.0011	0.0011	0.0008
	(0.5,0.5,1)	0.05	0.049	0.048	0.053	0.048	0.044
		0.005	0.0050	0.0049	0.0050	0.0045	0.0051
		0.001	0.0009	0.0008	0.0011	0.0009	0.0013
	(0,0,1)	0.05	0.047	0.050	0.044	0.043	0.056
		0.005	0.0050	0.0045	0.0058	0.0058	0.0054
		0.001	0.0009	0.0008	0.0012	0.0011	0.0014
E|G test							
(1,0,0)	(1,1,1)	0.05	0.040	0.041	0.040	0.048	0.050
		0.005	0.0041	0.0042	0.0051	0.0051	0.0057
		0.001	0.0010	0.0011	0.0013	0.0011	0.0012
	(1,1,0.5)	0.05	0.041	0.045	0.046	0.049	0.048
		0.005	0.0046	0.0048	0.0047	0.0045	0.0048
		0.001	0.0012	0.0011	0.0007	0.0006	0.0010
	(0.5,0.5,1)	0.05	0.037	0.043	0.042	0.056	0.056
		0.005	0.0043	0.0043	0.0051	0.0052	0.0050
		0.001	0.0009	0.0007	0.0010	0.0007	0.0011
	(0,0,1)	0.05	0.046	0.046	0.050	0.057	0.053
		0.005	0.0044	0.0047	0.0050	0.0060	0.0056
		0.001	0.0012	0.0013	0.0012	0.0011	0.0010

[a] KM_G^{IBS}, GE kernel machine regression with the IBS kernel for the genetic factors; KM_G^{INT}, GE kernel machine regression with the interactive kernel for the genetic factor; PC1, principal component regression using only the first G and E PC and its interaction; PC80, principal component regression, using the top G and E PCs that explain 80% of the variation in the set and their interactions. LM, linear regression with covariates, including the 12 genetic and five environment factors as well as the 60 G×E interaction terms.

Table 4. The *P* values of various tests in the analysis of the CoLaus study data

	Null hypothesis being tested			
	Joint	G×E	G|E	E|G
KM_G^{IBS} [a]	0.009	0.252	0.006	0.326
KM_G^{INT}	0.006	0.174	0.006	0.318
PC1	0.321	Not applicable	0.899	0.280
PC80	0.055	Not applicable	0.002	0.904

[a] KM_G^{IBS}, GE kernel machine regression with the IBS kernel for the genetic factors; KM_G^{INT}, GE kernel machine regression with the interactive kernel for the genetic factor; PC1, principal component regression using only the first G and E PC and its interaction; PC80, principal component regression using the top G and E PCs that explain 80% of the variation in the set and their interactions. The LM (linear model) analysis was not performed because the number of parameters (11 Gs + 8Es + 88 G×Es) exceeded the sample size (87 subjects).

models), and the marginal effect sizes of SNP10 and SNP11 are −36.94 (SD 8.65) and −28.83 (SD 7.75), respectively.

When applying the benchmark methods to the same dataset with the same analysis procedure, we did not conduct the G×E test due to the inflated Type I error rates. PC1 resulted in no significant findings for the joint test, the conditional E test, or the conditional G test. These results indicate that the multi-G and multi-E information may not be sufficiently captured by the first PC. In contrast, the PC80 was able to detect a near significant result in the joint test (*P*-value = 0.055). Although the approach detected a significant association in the conditional G test (*P*-value = 0.002), the signal would have been missed if a hierarchical analysis structure was undertaken where effect-specific tests (i.e., the G×E, G|E, and E|G tests) were performed only if the joint test detected a signal. The *P*-value of the G|E test based on PC80 is on the same order, but smaller than the *P*-value based on GE-KM, which may imply that an additive linear effect exists among the common variants within *PLA2G7*.

For verification purposes, we also fitted a multiple linear regression model, including the main effects of the potential confounders, the 11 common *PLA2G7* variants, and the CVD risk factors. An *F*-test was performed to access the effect of the 11 variants in PLA2G7 and the resulting *P*-value was 0.035, agreeing with the analyses from GE-KM and PC80 methods.

Discussion

Studying complex diseases in the post-GWAS era has led to the development of methods that consider factor-sets rather than individual genetic and environmental factors. Much of the work in the recent literature has focused on Multi-G approaches, and KM regression has emerged as a powerful and flexible approach for performing analyses in a Multi-G setting. However, the need for Multi-G-Multi-E methods is becoming apparent as investigators have turned their attention to mining genomic data for potential G×E interactions. In this work, we propose a KM regression approach that directly constructs a kernel for G×E interactions based on the genetic and environmental kernels and incorporates a series of score tests to evaluate the complete effect profile (i.e., the G, E, and G×E effect individually or in combination). We find that our method can have markedly better power than the currently available approaches when performing analyses in a Multi-G-Multi-E setting. The largest gains were observed when the underlying effect structure involved complex interactions, a scenario believed to be plausible for complex human diseases. As such, the proposed KM approach is a powerful and flexible tool for performing exploratory or confirmatory analyses for investigating G×E interactions. The R code that implemented the proposed KM methods is available on the author's Web site: http://www4.stat.ncsu.edu/~jytzeng/software.php.

Approaches currently available for examining effect profiles in Multi-G-Multi-E analyses include naïve regression, PC regression, and KM regression. We found that the Naïve and PC regressions, although simple and straightforward to apply, have inflated Type I error rates when performing test involving the G×E effects. This could possibly be due to the fact that the PCs cannot adequately capture the nonadditive genetic and environmental effects (i.e., G×G and E×E effects) in the underlying model. As a result, the G×E interaction term in the PC model absorbed the G×G and E×E effects and led to false G×E findings when there were no G×E interactions. This observation agrees with the findings of Voorman et al. 2011 who studied the implications of using a misspecifed mean model when investigation G×E interactions. Specifically, Voorman et al. 2011 pointed out severe Type I error rate inflation can occur in G×E-GWAS studies when the fitted model used to screen for G×E interactions does not correctly reflect the true underlying G and E effects and suggested a model-robust estimate of the variance to address the issue. The proposed KM models can offer an alternative solution to the inflation problem. To minimize the impact of model misspecification, one can use polynomial kernels, interactive kernels, or IBS kernels to capture the nonadditive and nonlinear effects of multi-G (or multi-E) factors; one can also use the nonparametric kernels, such as Gaussian kernels to model the effects in a nonparametric fashion. Our simulation results indicated that the proposed GE-KM maintained control on the Type I error rate across all simulation settings and had comparable or better power than the PC approaches.

Choosing the "correct" or "optimal" kernel in general kernel machine regression is still an important open problem. In our experience, when using kernel-based approach, exactly "correct" kernels do not need to be specified for the effects in the null model (e.g., the main effects for a G×E test). As long as the kernel used for the null model effects is correctly specified, the test will still be valid (i.e., produce nominal Type I error) even if there is a misspecification in the kernel matrix for the effects to be tested. Specifically, for G×E tests, as long as the main-effect kernel used contains the true model, the G×E test will still be valid. For example, if the correct main effect is linear (i.e., the linear kernel is optimal for the main effect), one can still use a quadratic kernel for the main effect and still be able to form a valid test. On the other hand, even if the G×E kernel is not correctly specified, the G×E test will still be valid, although the test based on the "wrong G×E kernel" would be less powerful than the test based on "correct G×E kernel." To assure the "near correctness" of the kernels used to model the effects in the null model, one might apply the methodology proposed in Wu et al. [2013]. To be specific, we can start with several candidate kernels (e.g., for SNP data, we might start with the linear, quadratic, and IBS kernels) and then create a "composite kernel" based on the candidate kernels. The composite kernel based tests have performance close to the optimal kernel while showing substantial improvement over the "wrong" kernels.

The proposed GE-KM approach can be applied in a variety of research settings where the definition of the "environmental" factors can vary based on the application area. For example, in the analysis of the CoLaus Study data, the environmental factors were taken to be clinical risk factors that are known to be related the trait being studied. The environmental factors could be taken to be factors that are truly external to the individual, such as air pollution measures when studying the interaction between genetics and pollution and the risk of lung cancer, or could be taken to be other biomarkers, such as expression or metabolite levels when studying the interaction between genetics and transcriptomics or metabolomics on the risk of adverse events in a clinical trial. Finally, the proposed framework can be directly applied to study the G×G interactions among a few G sets, or the E×E interactions among a few E sets. However, it is incapable of detecting SNP × SNP interactions within a G set or the E×E interactions within an E set because the proposed KM method is an approach for factor-set analysis that can only be used to detect set level of signals. Obtaining factor level of signals is a logical next step once set-level signals are detected; however, unraveling a set-level signal is a complex topic and requires future research.

Acknowledgments

The authors thank Drs. Peter Vollenweider and Gerard Waeber, PIs of the CoLaus study, and Drs. Meg Ehm and Matthew Nelson, collaborators at GlaxoSmithKline, for providing the CoLaus phenotype and sequence data. This work was supported by National Institutes of Health grants R00 ES017744 (to A.M.), R01 MH084022 (to J.Y.T.), and P01 CA142538 (to J.Y.T.).

References

Clark JJ, Maity A, Harmon QE, Engel SM, Epstein MP, Wu MC. 2014. Gene and region based testing of gene-gene interactions for quantitative traits with the SNP-set kernel interaction test (SKIT). Submitted.

Dai X, Wu C, He Y, Gui L, Zhou L, Guo H, Yuan J, Yang B, Li J, Deng Q and others. 2013. A genome-wide association study for serum bilirubin levels and gene-environment interaction in a Chinese population. *Genet Epidemiol* 37:293–300.

Duchesne P, Lafaye De Micheaux P. 2010. Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Comput Stat Data Anal* 54:858–862.

Edwards DRV, Naj AC, Monda K, North KE, Neuhouser M, Magvanjav O, Kusimo I, Vitolins MZ, Manson JE, O'Sullivan MJ and others. 2013. Gene-environment interactions and obesity traits among postmenopausal African-American and Hispanic women in the Women's Health Initiative SHARe Study. *Hum Genet* 13:323–336.

French B, Lumley T, Monks SA, Rice KM, Hindorff LA, Reiner AP, Psaty BM. 2006. Simple estimates of haplotype relative risks in case-control data. *Genet Epidemiol* 30:485–494.

Gauderman WJ, Murcray C, Gilliland F, Conti DV. 2007. Testing association between disease and multiple SNPs in a candidate gene. *Genet Epidemiol* 32:108–118.

Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM, NCI Gene-Environment Think Tank. 2013. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. *Genet Epidemiol* 37:643–657.

Jiao S, Hsu L, Bézieau 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* 37:452–64.

Kimeldorf G, Wahba G. 1970. A correspondence between Bayesian estimation on stochastic processes and smoothing splines. *Ann Math Stat* 41:495–502.

Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ. 2007. Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* 63:111–119.

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. *Am J Hum Genet* 82:386–397.

Larson NB, Schaid DJ. 2013. A kernel regression approach to gene-gene interaction detection for case-control studies. *Genet Epidemiol* 37:695–703.

Lin X, Lee S, Christiani DC, Lin X. 2013. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 14:667–681

Liu D, Liu X, Ghosh D. 2007. Semiparametric regression of multi-dimensional genomic pathway data: least square kernel machines and linear mixed models. *Biometrics* 63:1079–1088.

Lui D, Ghosh D, Lin X. 2008. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinform* 9:292.

Maity A, Lin X. 2011. Powerful tests for detecting a gene effect in the presences of possible gene-gene interactions using garrote kernel machines. *Biometrics* 67:1271–1284.

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.

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

Naj AC, Scott WK, Courtenay MD, Cade WH, Schwartz SG, Kovach JL, Agarwal A, Wang G, Haines JL, Pericak-Vance MA. 2013. Genetic factors in nonsmokers with age-related macular degeneration revealed through genome-wide gene-environment interaction analysis. *Ann Hum Genet* 77:215–231.

Patel CJ, Chen R, Kodama K, Ioannidis JP, Butte AJ. 2013. Systematic identification of interaction effects between genome-and environment-wide associations in type 2 diabetes mellitus. *Hum Genet* 132:1–14.

Pongpanich M, Neely ML, Tzeng JY. 2012. On the aggregation of multimarker information for marker-set and sequencing data analysis: genotype collapsing vs. similarity collapsing. *Front Genet* 2:1–14.

Robins JM, Morgenstern H. 1987. The foundations of confounding in epidemiology. *Comp Math Appl*, 14:869–916.

Song K, Nelson MR, Aponte J, Manas ES, Bacanu SA, Yuan X, Kong X, Cardon L, Mooser VE, Whittaker JC and others. 2011. Sequencing of Lp-PLA2-encoding PLA2G7 gene in 2000 Europeans reveals several rare loss-of-function mutations. *Pharmacogenomics J* 12:425–431.

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

Thomas D. 2010b. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu Rev Public Health* 31:21–36.

Thompson WD. 1991. Effect modification and the limits of biological inference from epidemiologic data. *J Clin Epidemiol* 44:221–232.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 89:277–288.

Vineis P, Marinelli D, Autrup H, Brockmoller J, Cascorbi I, Daly AK, Golka K, Okkels H, Risch A, Rothman N and others. 2001. Current smoking, occupation, N-acetyltransferase-2 and bladder cancer: a pooled analysis of genotype-based studies. *Cancer Epidemiol Biomarkers Prev* 10:1249–1252.

Voorman A, Lumley T, McKnight B, Rice K. 2011. Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS ONE* 6(5): e19416.

Wang T, Ho G, Ye K, Strickler H, Elston RC. 2009. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genet Epidemiol* 33:6–15.

Wang X, Epstein MP, Tzeng JY. 2014. Analysis of gene-gene interactions using gene-trait similarity regression. *Hum Hered* 78:17–26.

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* 44:1090–1097.

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* 86:929–942.

Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Molldrem JJ, Armistead PM. 2013. Kernel machine SNP-set testing under multiple candidate kernels. *Genet Epidemiol* 37:267–275.