Powerful Tukey's One Degree-of-Freedom Test for Detecting Gene–Gene and Gene–Environment Interactions

Yaping Wang[1], Donghui Li[2] and Peng Wei[1,3]

[1]Department of Biostatistics, School of Public Health, University of Texas Health Science Center, [2]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, [3]Human Genetics Center, School of Public Health, University of Texas Health Science Center, Houston, TX, USA.

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

ABSTRACT: Genome-wide association studies (GWASs) have identified thousands of single nucleotide polymorphisms (SNPs) robustly associated with hundreds of complex human diseases including cancers. However, the large number of GWAS-identified genetic loci only explains a small proportion of the disease heritability. This "missing heritability" problem has been partly attributed to the yet-to-be-identified gene–gene (G × G) and gene–environment (G × E) interactions. In spite of the important roles of G × G and G × E interactions in understanding disease mechanisms and filling in the missing heritability, straightforward GWAS scanning for such interactions has very limited statistical power, leading to few successes. Here we propose a two-step statistical approach to test G × G/G × E interactions: the first step is to perform principal component analysis (PCA) on the multiple SNPs within a gene region, and the second step is to perform Tukey's one degree-of-freedom (1-df) test on the leading PCs. We derive a score test that is computationally fast and numerically stable for the proposed Tukey's 1-df interaction test. Using extensive simulations we show that the proposed approach, which combines the two parsimonious models, namely, the PCA and Tukey's 1-df form of interaction, outperforms other state-of-the-art methods. We also demonstrate the utility and efficiency gains of the proposed method with applications to testing G × G interactions for Crohn's disease using the Wellcome Trust Case Control Consortium (WTCCC) GWAS data and testing G × E interaction using data from a case–control study of pancreatic cancer.

KEYWORDS: G × E interaction, G × G interaction, GWAS, PCA, SNP, Tukey's 1-df test

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Wang et al. Powerful Tukey's One Degree-of-Freedom Test for Detecting Gene–Gene and Gene–Environment Interactions. *Cancer Informatics* 2015:14(S2) 209–218 doi: 10.4137/CIN.S17305.

RECEIVED: December 04, 2014. **RESUBMITTED:** April 20, 2015. **ACCEPTED FOR PUBLICATION:** April 28, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Original Research

FUNDING: This research was supported by the National Institutes of Health (NIH) grant R01CA169122. PW was also supported by NIH grants R01HL116720 and R21HL126032; DL by a supplemental grant to NIH R01CA98380 and the Sheikh Ahmed Center for Pancreatic Cancer Research Funds. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Peng.Wei@uth.tmc.edu

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Upon submission manuscript was subject to anti-plagiarism scanning. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties. This journal is a member of the Committee on Publication Ethics (COPE).

Published by Libertas Academica. Learn more about this journal.

Introduction

Thanks to the rapidly decreasing cost of high-throughput genotyping technologies, genome-wide association studies (GWASs) have identified thousands of single nucleotide polymorphisms (SNPs) robustly associated with hundreds of complex human diseases and traits,[1] providing novel insights into the disease mechanisms[2] and offering new therapeutic targets.[3] However, the large number of GWAS-identified genetic loci only explains a small proportion of the disease heritability, commonly estimated from twin and family-based studies. For example, a recent large-scale meta-analysis of GWAS has identified 67 SNPs associated with the risk of breast cancer, which, however, only explains 14% of the heritability of breast cancer.[4] This so-called missing heritability problem has been attributed to the yet-to-be-identified susceptibility loci of even smaller effect sizes,[5] rare genetic variants (minor allele frequency (MAF) <1%),[6,7] as well as gene–gene (G × G) and gene–environment (G × E) interactions.[8]

In spite of the important roles of G × G and G × E interactions in understanding disease mechanisms and filling in the missing heritability, there have been very few successes in identifying such interactions.[9,10] Lack of statistical power is one of the main reasons for such limited success. Standard G × G analysis based on GWAS data entails interaction test between each possible pair of SNPs, while standard G × E analysis tests the interaction between the environmental exposure of interest and each of the GWAS SNPs. As G × G and G × E interactions are second-order effects, they are more difficult to detect than genetic main effects. A rule of thumb is that, given the same significance level and comparable magnitude of effect size, detecting a G × E interaction would require a sample size at least 4 times larger than that for detecting a genetic main effect.[11] The lack of power is further exacerbated in the analysis of G × G interactions due to the curse of dimensionality: one million tests of SNP main effects would correspond to 5×10^{11} tests of pairwise SNP interactions. As a result, while a typical GWAS that genotypes around one million SNPs is designed to ensure enough power for detecting genetic main effects, straightforward genome-wide scanning of G × G and G × E interactions can be severely underpowered. In response to this pressing challenge, many new statistical methods have been proposed to improve the power of G × G and G × E interaction tests. Many new methods are aimed at reducing the burden of multiple testing. For example, Kooperberg and Leblanc[12] proposed to perform formal G × G tests only for those SNPs with some marginal/main effects. Other authors proposed some alternative or hybrid strategies to filter SNPs for formal G × E testing.[13–16] Another line of research is to group SNPs into genes or biological pathways to aggregate multiple weak/moderate signals and reduce the total number of interaction tests. He et al.[17] proposed a gene-based G × G test by first performing principal component analysis (PCA) on multiple SNPs in linkage disequilibrium (LD) in a gene region, and then testing interactions between each pair of PCs in the two genes. In addition to several new gene-based G × E tests,[18–20] Tang et al.[21,22] proposed biological pathway-based G × E tests and demonstrated their applications to pancreatic cancer.

In this paper, we propose a parsimonious and powerful gene-based interaction test that can be applicable to both G × G and G × E testing. Our proposed method is motivated by Tukey's one degree-of-freedom (1-df) test for interaction,[23] which was first introduced to statistical genetics by Chatterjee et al.[24] in the context of testing genetic main effects while adjusting for possible G × E effects. The test of Chatterjee et al.[24] is useful in de novo GWAS scanning, aimed at discovering SNPs with any effect, ie, both genetic main and G × E effects or either effect alone. However, in the post-GWAS era, it is often of primary interest to test and discover G × G or G × E interaction effect itself. Here we propose a two-step approach to test G × G/G × E interactions: the first step is to perform PCA on the multiple SNPs within a gene region, along the line of He et al.[17], and the second step is to perform Tukey's 1-df test specifically for the interaction effect. We derive the score test for the latter, which is fast and numerically stable to compute. Using extensive simulations, we show that the proposed approach, which combines the two parsimonious models, namely, the PCA and Tukey's 1-df form of interaction, outperforms other state-of-the-art methods. We also demonstrate the utility and efficiency gains of the proposed method with applications to testing G × G interactions for Crohn's disease using the Wellcome Trust Case Control Consortium (WTCCC) GWAS data and testing G × E (SNP-set by smoking) interaction using data from a case–control study of pancreatic cancer.

Methods

Existing methods: SNP-based and gene-based G × G and G × E tests. We consider a case–control study with a total sample size n including n_0 controls and n_1 cases ($n = n_0 + n_1$). Let Y_i denote the binary disease status of individual i: 0 for controls and 1 for cases ($i = 1,\ldots, n$). Let Z_i denote the covariate vector, including, eg, sex, age, and leading principal components capturing population substructure. Given two SNPs to be tested for G × G interaction, let X_{1i} and X_{2i} denote the genotypes of the two SNPs in subject i, each equal to 0, 1, and 2 for major allele homozygotes, heterozygotes, and minor allele homozygotes, respectively. In this paper, we exclusively focus on multiplicative G × G/G × E interactions in the logistic regression framework; see Ref. 25 for tests of additive interactions. A commonly used SNP-based G × G test is based on the following logistic regression model:

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_Z' Z_i + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_{12} X_{i1} X_{i2} \quad (1)$$

where we assume the additive genetic model for both SNPs. Alternative genetic models, such as the genotypic and dominant models, can also be assumed; see Ref. 26 for details. To test the null hypothesis of no G × G interaction, ie, H_0: $\beta_{12} = 0$, we can perform a 1-df likelihood ratio test (LRT), or its asymptotically equivalent Wald and score tests. As pointed out before, the total number of all pairwise G × G interaction tests is much larger than that of SNP main effect tests in the GWAS setting, leading to prohibitively high computational cost and low statistical power. To conduct SNP-based G × E test between a SNP and an environmental exposure, eg, smoking, we can similarly fit a logistic regression model

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_Z' Z_i + \beta_G X_i + \beta_E E_i + \beta_{GE} X_i E_i, \quad (2)$$

where E_i denotes the exposure of subject i. We test the null hypothesis H_0: $\beta_{GE} = 0$ with a 1-df LRT.

Gene-based G × G and G × E tests have also been proposed to reduce the burden of multiple testing and aggregate weak/moderate signals in genes and biological pathways. Assume two sets of SNPs mapped to two genes of interest, denoted as $\mathbf{X}_{1,i} = \left(X_{1,i1},\ldots,X_{1,iL_1}\right)'$ and $\mathbf{X}_{2,i} = \left(X_{2,i1},\ldots,X_{2,iL_2}\right)'$ for subject i. A SNP is assigned to a gene if it is located within certain range of the gene's transcription start and end sites, eg, ±20,000 base pairs (20 kb), to include SNPs in regulatory regions.[27] Most gene-based multilocus G × G tests are based on the following saturated interaction model:

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_Z' Z_i + \sum_{l_1=1}^{L_1} \beta_{1,l_1} X_{1,il_1} + \sum_{l_2=1}^{L_2} \beta_{2,l_2} X_{2,il_2} + \sum_{l_1=1}^{L_1} \sum_{l_2=1}^{L_2} \beta_{12,l_1 l_2} X_{1,il_1} X_{2,il_2}. \quad (3)$$

The null hypothesis to be tested is $H_0 : \beta_{12} = \left(\beta_{12,11}, \ldots, \beta_{12,L_1 L_2}\right) = 0$. A multivariate score test can be performed based on the test statistic $T_{\text{score}} = U'V^{-1}U$, where U is the efficient score vector of length $L_1 \times L_2$ for β_{12}, and V is its asymptotic covariance matrix under H_0; see Ref. 18 for details. The score test is preferred here over the LRT or Wald test because it requires only fitting the null model without interactions and is more computationally efficient. However, even with the score test, the genetic main effects need to be estimated under the null model and can cause numerical instability due to the multicollinearity among SNPs in high LD within a gene region. In addition, T_{score} approximately follows a χ^2-distribution with a large number of degrees of freedom $(L_1 \times L_2)$ under H_0, suffering from loss of power. Pan et al.[18] proposed to test H_0 with the sum-of-squared score statistic (SSU) $T_{\text{SSU}} = U'U$, assuming an identity covariance matrix in T_{score}. T_{SSU} has an asymptotic distribution of a mixture of $\chi^2(1)$'s, which can be approximated by a scaled and shifted χ^2 distribution.[18] As shown by Pan,[28,29] the SSU test is equivalent to the permutation-based version of a variance component score test for a random-effects logistic regression model for high-dimensional hypothesis testing.[30] In addition, the SSU test is equivalent to the variance component score test in kernel machine regression under a linear kernel, which has been shown to be powerful in rare variant association tests[31] and gene-based G × E tests.[19] Model (3) can be easily modified to perform gene-based G × E test:

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_Z' Z_i + \sum_{l_1=1}^{L_1} \beta_{l_1} X_{il_1} + \beta_E E_i + \sum_{l_1=1}^{L_1} \beta_{El_1} X_{il_1} E_i. \quad (4)$$

Multivariate score test and SSU test for $H_0 : \beta_{E1} = \ldots = \beta_{EL_1} = 0$ can be similarly derived. Of note, the SSU test may still suffer from numerical instability, as the genetic main effects need to be estimated under H_0.

To overcome the multicollinearity problem, researchers have proposed to first perform PCA on the multiple SNPs in a gene region and then test for interactions based on the leading PCs.[17,21,22,27] Specifically, the PCA is used to summarize SNPs in each gene as uncorrelated (orthogonal) linear combinations of the original SNPs accounting for, eg, 90% of the total genetic variation. The number of the resulting PCs, denoted as $PC_1, \ldots, PC_K$, is usually much smaller than the number of the original genotyped SNPs. He et al.[17] proposed the following gene-based G × G model based on the leading K_1 and K_2 PCs of the two genes of interest:

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_Z' Z_i + \sum_{k_1=1}^{K_1} \gamma_{1,k_1} PC_{1,ik_1} + \sum_{k_2=1}^{K_2} \gamma_{2,k_2} PC_{2,ik_2} + \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \gamma_{12,k_1 k_2} PC_{1,ik_1} PC_{2,ik_2}. \quad (5)$$

To test the null hypothesis $H_0 : \gamma_{12} = \left(\gamma_{12,11}, \ldots, \gamma_{12,K_1 K_2}\right)' = 0$, He et al.[17] proposed to employ an LRT whose test statistic approximately follows a χ^2-distribution with $K_1 \times K_2$ degrees of freedom under H_0. Tang et al.[21,22] adapted Model (5) to conduct gene- and pathway-based G × E test:

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_Z' Z_i + \sum_{k_1=1}^{K_1} \gamma_{k_1} PC_{ik_1} + \beta_E E_i + \sum_{k_1=1}^{K_1} \gamma_{Ek_1} PC_{ik_1} E_i, \quad (6)$$

where PC_{ik1}'s are the leading K_1 PCs for the SNPs mapped to a gene or biological pathway. The null hypothesis H_0: $\gamma_{E1} = . = \gamma_{EK1} = 0$ can be tested with an LRT or score test.

New method: Tukey's 1-df interaction test. Although there are typically fewer interaction terms in Model (5) than in Model (3) ($K_1 \times K_2$ versus $L_1 \times L_2$), the former can still be large for interactions involving large genes and may lead to loss of power. Here we propose to employ the parsimonious Tukey's 1-df form of interaction

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_Z' Z_i + \sum_{k_1=1}^{K_1} \gamma_{1,k_1} PC_{1,ik_1} + \sum_{k_2=1}^{K_2} \gamma_{2,k_2} PC_{2,ik_2} + \theta \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} \gamma_{1,k_1} \gamma_{2,k_2} PC_{1,ik_1} PC_{2,ik_2}. \quad (7)$$

Note that the above interaction model assumes that the interaction effect is proportional to the sum of the product of genetic main effects. A single parameter θ is used to capture the interactions between the PCs of the two genes, leading to the parsimonious Tukey's 1-df form of interaction. For numerical stability and computational efficiency, we propose a score test for H_0: $\theta = 0$. We derive the score test statistic as follows:

1. Let $\psi = \left(\beta_0, \beta_Z', \gamma_1', \gamma_2'\right)'$ and $X_i = \left(1, Z_i', PC_{1,i}', PC_{2,i}'\right)'$ be the nuisance parameters and their corresponding covariate vector in Model (7), where $\gamma_1 = \left(\gamma_{1,1}, \ldots, \gamma_{1,K_1}\right)'$ and $\gamma_1 = \left(\gamma_{1,1}, \ldots, \gamma_{1,K_1}\right)'$. We denote the disease probability under Model (7) by

$$P_{\psi,\theta}\left(X_i\right)=\frac{\exp\left\{\beta_0+\beta_Z Z_i+\Sigma_{k_1=1}^{K_1}\gamma_{1,k_1}PC_{1,ik_1}+\Sigma_{k_2=1}^{K_2}\gamma_{2,k_2}PC_{2,ik_2}+\theta\Sigma_{k_1=1}^{K_1}\Sigma_{k_2=1}^{K_2}\gamma_{1,k_1}\gamma_{2,k_2}PC_{1,ik_1}\right\}}{1+\exp\left\{\beta_0+\beta_Z Z_i+\Sigma_{k_1=1}^{K_1}\gamma_{1,k_1}PC_{1,ik_1}+\Sigma_{k_2=1}^{K_2}\gamma_{2,k_2}PC_{2,ik_2}+\theta\Sigma_{k_1=1}^{K_1}\Sigma_{k_2=1}^{K_2}\gamma_{1,k_1}\gamma_{2,k_2}PC_{1,ik_1}\right\}}.$$

2. The log likelihood function for Model (7) is

$$logL=\sum_{i=1}^{n_0+n_1}\left[Y_i logP_{\psi,\theta}\left(X_i\right)+\left(1-Y_i\right)logP_{\psi,\theta}\left(X_i\right)\right].$$

The score function for testing H_0: $\theta=0$ is

$$S\left(\theta=0,\psi=\hat{\psi}_{(0)}\right)=\sum_{i=1}^{n_0+n_1}\left(\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\hat{\gamma}_{1,k_1}\hat{\gamma}_{2,k_2}PC_{1,ik_1}PC_{2,ik_2}\right)\left[Y_i-P_{\hat{\psi}_{(0)},\theta=0}\left(X_i\right)\right],$$

where $\hat{\psi}_{(0)}$ is the maximum likelihood estimate (MLE) of ψ under H_0: $\theta = 0$, which can be easily obtained from the reduced (null) model as a standard logistic regression.

3. Obtain the inverse of the asymptotic variance for $S\left(\theta=0,\psi=\hat{\psi}_{(0)}\right)$ based on the observed information matrix:

$$I^{\theta\theta}|_{\psi=\hat{\psi}(0),\theta=0}=\left[I_{\theta\theta}-I_{\theta\psi}I_{\psi\psi}^{-1}I_{\psi\theta}\right]^{-1}|_{\psi=\hat{\psi}(0),\theta=0},$$

where

$$I_{\theta\theta}|_{\psi=\hat{\psi}_{(0)},\theta=0}=\sum_{i=1}^{n_0+n_1}\left(\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\hat{\gamma}_{1,k_1}\hat{\gamma}_{2,k_2}PC_{1,ik_1}PC_{2,ik_2}\right)^2 P_{\hat{\psi}_{(0)},\theta=0}\left(X_i\right)\left[1-P_{\hat{\psi}_{(0)},\theta=0}\left(X_i\right)\right],$$

$$\begin{aligned}I_{\theta\psi}|_{\psi=\hat{\psi}_{(0)},\theta=0}&=I'_{\psi\theta}|_{\psi=\hat{\psi}_{(0)},\theta=0}\\&=\sum_{i=1}^{n_0+n_1}\left(\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\hat{\gamma}_{1,k_1}\hat{\gamma}_{2,k_2}PC_{1,ik_1}PC_{2,ik_2}\right)P_{\hat{\psi}_{(0)},\theta=0}\left(X_i\right)\left[1-P_{\hat{\psi}_{(0)},\theta=0}\left(X_i\right)\right]X'_i,\end{aligned}$$

and

$$I_{\psi\psi}|_{\psi=\hat{\psi}_{(0)},\theta=0}=\sum_{i=1}^{n_0+n_1}P_{\hat{\psi}_{(0)},\theta=0}\left(X_i\right)\left[1-P_{\hat{\psi}_{(0)},\theta=0}\left(X_i\right)\right]X_iX'_i.$$

4. The score test statistic for H_0: $\theta=0$ is thus

$$T_{Tukey}=S\left(\theta=0,\psi=\hat{\psi}_{(0)}\right)I^{\theta\theta}|_{\psi=\hat{\psi}_{(0)},\theta=0}\,S\left(\theta=0,\psi=\hat{\psi}_{(0)}\right),$$

which approximately follows a $\chi^2(1)$ distribution under H_0.

Model (7) can be easily extended to gene-based testing of G × E:

$$logit\left(P\left(Y_i=1\right)\right)=\beta_0+\beta'_Z Z_i+\sum_{k_1=1}^{K_1}\gamma_{k_1}PC_{ik_1}+\beta_E E_i+\theta\sum_{k_1=1}^{K_1}\gamma_{k_1}\beta_E PC_{ik_1}E_i, \tag{8}$$

where H_0: $\theta = 0$ corresponds to no interaction between the gene and environmental exposure *E*. A score test can be similarly derived, whose test statistic follows a χ^2-distribution with one degree of freedom.

Of note, our proposed Models (7) and (8) are conceptually similar to those proposed by Chatterjee et al.[24]; however, there are two main distinctions. First, while Chatterjee et al directly modeled the genotyped SNPs, we propose to utilize the PCA technique to reduce the dimension and model the uncorrelated PCs, which enjoys numerical stability and improved statistical power (as will be shown in the Results section). Second, and more important, the null hypothesis to be tested in Chatterjee et al.[24] was on the genetic main effects of one gene, ie, $H'_0:\gamma_1=\ldots=\gamma_{K_1}=0$ in Model (8) while accounting for possible G × E interaction effects in Tukey's 1-df form of interaction. In contrast, our focus here is specifically to test the interaction effect, ie, H_0: $\theta=0$. Note that the nuisance parameter θ is not identifiable under H'_0, and some special treatments were proposed by Chatterjee et al.[24]; however, all nuisance parameters are identifiable under H_0: $\theta=0$ in our proposed Models (7) and (8), and standard large-sample likelihood theory can be applied.

Results

A simulation study is often used to evaluate and compare different statistical methods' Type I error rates and powers based on a large number of simulated datasets from some known models, eg, those with or without G × G interactions. We performed extensive simulations to evaluate the proposed gene-based Tukey's 1-df G × G interaction test and compared it with the state-of-the-art methods of He et al.[17] and Pan et al.[18] We also demonstrated the utility of the proposed method with applications to testing G × G interactions for Crohn's disease (CD) using the WTCCC GWAS data and testing G × E (SNP-set by smoking) interaction using data from a case–control study of pancreatic cancer.

Simulation setup. We performed simulations based on the HapMap SNP data of two genes *IL12B* (interleukin 12B; 5q31.1–q33.1) and *IL12RB2* (interleukin 12 receptor, beta 2; 1p31.3–p31.2), which were found to interact with each other influencing the risk of CD in our analysis of the WTCCC GWAS data (results to be shown later). In addition, proteins IL12B and IL12RB2 have been reported to physically interact with each other.[32] We obtained the genotype and phased haplotype data for 60 HapMap CEU individuals (release 22) for *IL12B* and *IL12RB2*. For each gene, we selected common SNPs to be the genotyped/observed SNPs in the simulated datasets if they had MAFs of at least 5%, and were either genotyped in the WTCCC GWAS or picked up as tagSNPs with pairwise tagging $r^2 \geq 0.8$ by the program Tagger.[33] We ended up with 16 and 21 genotyped SNPs for *IL12B* and *IL12RB2*, respectively. We largely followed the simulation setup in He et al.[17] to simulate a case–control study of 2,000 cases and 2,000 controls in each of the 1,000 simulation replications. We designated either one or two causal SNPs with main/interaction effects in each of the two genes. For the scenario of one causal SNP that was randomly selected in each gene, the case–control status was simulated based on the following logistic regression model:

$$logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \beta_1^{(1)} G_{i1}^{(1)} + \beta_1^{(2)} G_{i1}^{(2)} + \beta_{11}^{(12)} G_{i1}^{(1)} G_{i1}^{(2)}, \qquad (9)$$

where $\beta_0 = log(0.01/0.99)$ for a baseline disease prevalence of 1%, $\beta_1^{(1)} = 0$ or 0.5, and $\beta_1^{(2)} = 0$ or 0.5. We let $\beta_{11}^{(12)} = 0$ to evaluate the Type I error rate and let it gradually increase to evaluate the power. For each of the 1,000 simulation replicates, we randomly drew two haplotypes based on their frequencies in the HapMap CEU data to form the genotype data for an individual and generated a large homogeneous study population based on the simulation model (9). We then randomly sampled 2,000 cases and 2,000 controls to form a simulated dataset.

For the scenario of two causal SNPs that were randomly selected in each gene, we simulated the case–control status based on the following logistic regression model:

$$\begin{aligned} logit\left(P\left(Y_i = 1\right)\right) = \beta_0 + \sum_{k=1}^{2} \beta_k^{(1)} G_{ik}^{(1)} \sum_{k=1}^{2} + \beta_k^{(2)} G_{ik}^{(2)} + \beta_{11}^{(12)} \\ \left(G_{i1}^{(1)} G_{i1}^{(2)} + G_{i2}^{(1)} G_{i1}^{(2)} + G_{i1}^{(1)} G_{i2}^{(2)} + G_{i2}^{(1)} G_{i2}^{(2)}\right), \end{aligned} \qquad (10)$$

where $\beta_0 = log\left(0.01/0.99\right), \beta_1^{(1)} = \beta_2^{(1)} = 0$ or 0.5, and $\beta_1^{(2)} = \beta_2^{(2)} = 0$ or 0.5. We let $\beta_{11}^{(12)} = 0$ to evaluate the Type I error rate and gradually increased it to evaluate the power. We have listed the parameter values in Table 1 for the six simulation configurations considered here.

We included the following gene-based G × G tests in the simulation study: (1) the score test for the proposed PC-based Tukey's 1-df interaction Model (7), (2) the generic LRT for PC-based pairwise interaction Model (5) of He et al.[17], and (3) the SSU test for SNP-based pairwise interaction Model (3) of Pan et al.[18] as a representative of variance-component score tests for high-dimensional hypothesis testing. For PC-based tests, we first applied the PCA to the genotype data of each gene region to derive the top PCs that explained at least 90% of the total genetic variation.

Simulation results. As shown in Table 1, all three methods controlled the Type I error rate at the nominal level $\alpha = 0.05$ satisfactorily across the six simulation configurations. We also performed simulations when there was one or two covariates, and the proposed score test for Tukey's 1-df interaction controlled the Type I error rate satisfactorily as well (results not shown).

Figures 1–6 show the power comparison of the three tests across different simulation configurations. Overall, the proposed Tukey's 1-df score test was the most powerful one among the three methods under comparison, and the SNP-based SSU test was the least powerful one, suggesting that the proposed strategy combining the two parsimonious statistical models, namely the PCA and Tukey's 1-df form of interaction, was effective in improving the power of gene-based interaction test. The power difference was larger when there was only one disease locus/causal SNP in each gene (Figs. 1–3) than when there were two disease loci/causal SNPs in each gene (Figs. 4–6) whether or not there was main effect. In particular, there was only a slight power difference between the generic LRT and the SSU in the latter scenario. This suggested that the SSU and its closely related variance-component score tests, such as the kernel machine-based methods, might have more improved power when the interaction signals were less sparse (four true interaction pairs versus one in the simulations here). It is noticeable that the power curve for the SSU test in

Table 1. Empirical Type I error rate based on 1,000 simulation replications (2,000 cases and 2,000 controls in each replication).

CONFIGURATION	$\beta_1^{(1)}$	$\beta_2^{(1)}$	$\beta_1^{(2)}$	$\beta_2^{(2)}$	NO. OF INTERACTIONS*	TUKEY'S 1-DF SCORE TEST	GENERIC LRT	SSU
1	0.5	0	0.5	0	1	0.048	0.051	0.048
2	0.5	0	0	0	1	0.051	0.050	0.050
3	0	0	0	0	1	0.047	0.054	0.054
4	0.5	0.5	0.5	0.5	4	0.054	0.052	0.049
5	0.5	0.5	0	0	4	0.055	0.049	0.049
6	0	0	0	0	4	0.049	0.056	0.056

Notes: *All interaction terms were 0 to evaluate the Type I error rate. Significance level $\alpha = 0.05$.

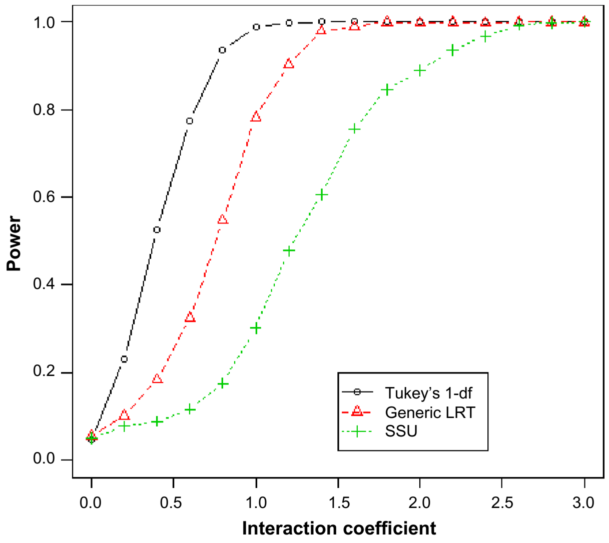

Figure 1. Power curve as a function of the interaction parameter $\beta_{12}^{(12)}$ under simulation Configuration 1. Each gene has only one disease locus, and both loci have main effects.

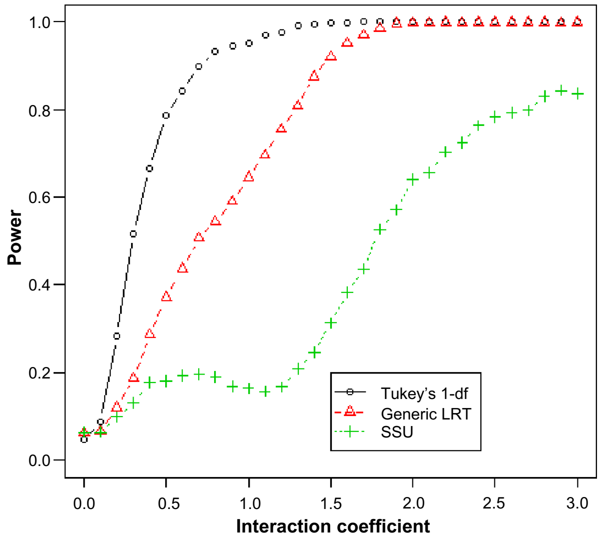

Figure 3. Power curve as a function of the interaction parameter $\beta_{12}^{(12)}$ under simulation Configuration 3. Each gene has only one disease locus, and neither has main effect.

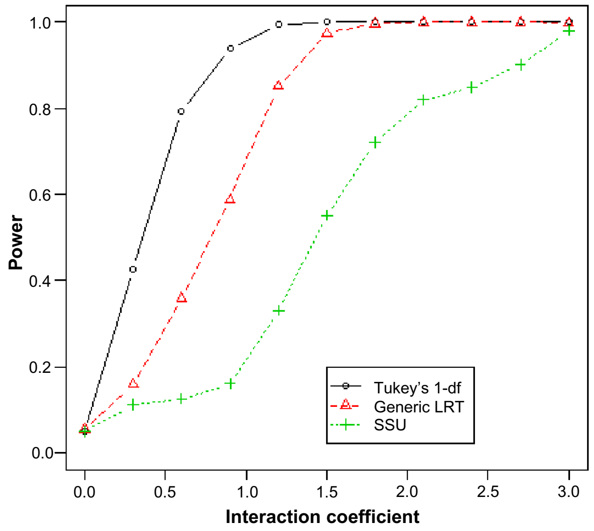

Figure 2. Power curve as a function of the interaction parameter $\beta_{12}^{(12)}$ under simulation Configuration 2. Each gene has only one disease locus, and only one of the loci has main effect.

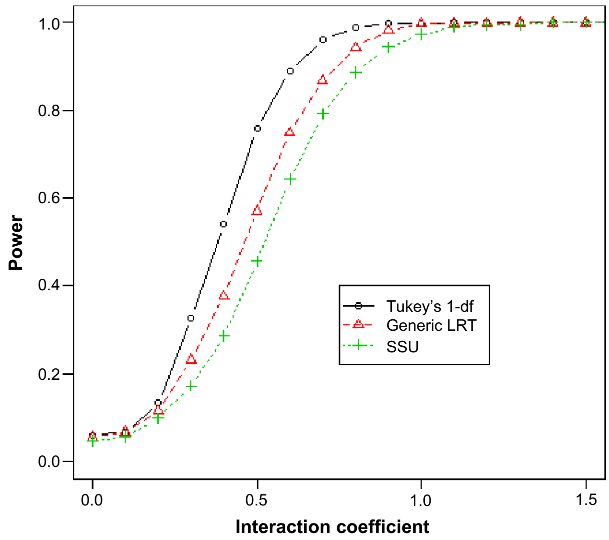

Figure 4. Power curve as a function of the interaction parameter $\beta_{12}^{(12)}$ under simulation Configuration 4. Each gene has two disease loci, and all of the loci have main effects.

Figure 3 was not monotonically increasing as the interaction signal increased. This may be explained by numerical problems in estimating the genetic main effects of multiple SNPs in high LD in Model (3). On the other hand, we did not observe this abnormal phenomenon for the PC-based Tukey's 1-df score test and generic LRT, supporting that using PCs can not only reduce the dimension but also ensure numerical stability. In addition, although the proposed Tukey's 1-df form of interaction is a function of the genetic main effects, it remained more powerful than the other two competing methods even when there was no genetic main effect (Figs. 3 and 6) or only one gene had genetic main effects (Figs. 2 and 5). The reason for this interesting phenomenon is that, even when there was only interaction effect and no main effect in the simulation/true model, the expectation of the marginal genetic effect, ie, the main effect in the fitted null model, was not zero due to the absorption of the true interaction effect into the marginal effect. Therefore, Tukey's test could still retain the statistical power to detect the interaction effect in the absence of main effects. Finally, the ratio of runtime for Tukey's 1-df score test, generic LRT, and SSU test was nearly in the ratio 1:2:3, with the proposed Tukey's test being the fastest.

Application to testing G × G for Crohn's disease. CD is a type of inflammatory bowel disease and is also considered

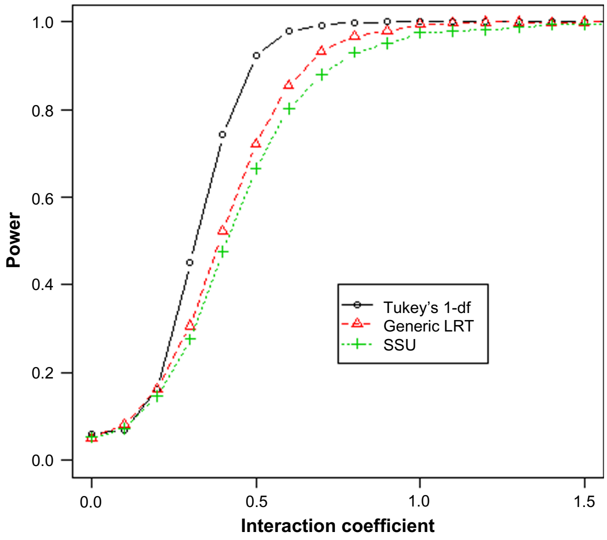

Figure 5. Power curve as a function of the interaction parameter $\beta_{12}^{(12)}$ under simulation Configuration 5. Each gene has two disease loci, and only two of the loci in one gene have main effects.

as an autoimmune disease with a strong genetic component.[34] Although large-scale meta-analyses of GWASs have identified a large number of susceptibility loci for CD, about 78.5% of the estimated heritability is still missing and it has been shown that 80% of the missing heritability could be due to G × G interactions.[8] As a proof of concept, we applied the proposed Tukey's 1-df gene-based G × G test and the generic LRT to the WTCCC case–control GWAS of CD.[35] The GWAS dataset contains 2,000 CD cases and 3,000 controls with a total of 500,568 SNPs. We followed the WTCCC quality control criteria to remove unqualified subjects and SNPs, resulting in 469,612 SNPs in 1,748 cases and 2,938 controls.

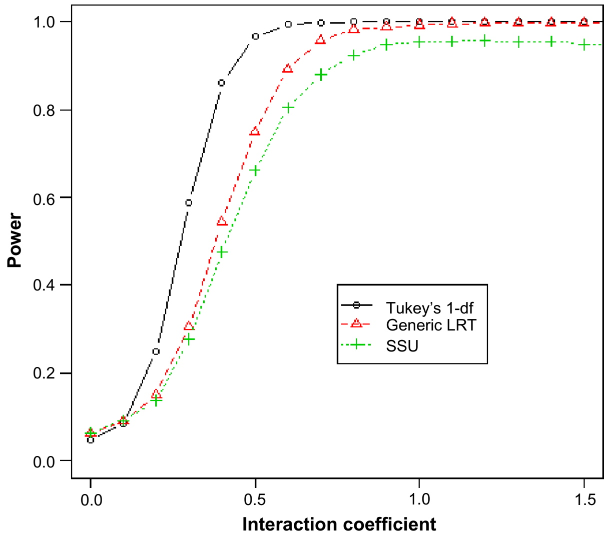

Figure 6. Power curve as a function of the interaction parameter $\beta_{12}^{(12)}$ under simulation Configuration 6. Each gene has two disease loci, and none of the loci has main effect.

To explore the G × G interactions among known CD susceptibility genes, we retrieved the 31 genes reviewed in Ref. 34. We followed Ref. 27 to assign SNPs to genes: an SNP is assigned to a gene if it is located within 20 kb of the gene's transcription start and end sites. A total of 653 SNPs were assigned to the 31 genes. Since these genes were identified by large-scale meta-analysis of GWASs and candidate pathway studies, not all of them may be associated with CD in the WTCCC GWAS. To test for genetic marginal associations, we applied PCA to the SNPs in each of the 31 genes, retrieved the leading PCs that explained at least 90% of the genetic variation, and performed PC-based multilocus association test of the genetic main effects. Twelve nominally significant genes that had *P*-values <0.1 were tested for pairwise G × G interactions, as in Ref. 12. Table 2 lists all pairs of G × G interactions with interaction *P*-values <0.1 by either Tukey's 1-df score test or the generic LRT. As a comparison, we also listed the SSU test *P*-values, which were very similar to those of the generic LRT. As expected with this moderate sample size, none of the tests identified any G × G interaction with a *P*-value less than 0.01. On the other hand, previous research showed that G × G interactions tend to be enriched among neighboring genes on protein–protein interaction (PPI) networks.[36] While none of the five nominally significant G × G interactions by the generic LRT ($P_{\text{LRT}} < 0.1$) was found to coincide with PPIs in the STRING database,[32] three of the nine nominally significant interactions by Tukey's 1-df test ($P_{\text{Tukey}} < 0.1$) appeared to be supported by PPIs: the proteins of three pairs of genes, including *IL12B*–*IL12RB2*, *IL12B*–*IL23A*, and *IL12B*–*IL12RB*1, were found to physically interact with each other. Due to the small counts, this enrichment was not statistically significant. Nevertheless, these biologically plausible G × G interactions would be worth following up in independent and larger samples.

Application to testing G × E in a case–control study of pancreatic cancer. Pancreatic cancer is the fourth leading cause of cancer-related deaths for both men and women in the US with a 5-year survival rate of 6%.[37] Known risk factors for pancreatic cancer include cigarette smoking, long-term Type 2 diabetes, obesity, heavy alcohol consumption, and family history, with smoking conferring the highest risk. It is of interest to investigate whether there exist genes that interact with environmental exposures, such as smoking, influencing the risk of pancreatic cancer. Here we tested the pancreatic cancer susceptibility SNP-set by smoking interaction (G × E) using data from a case–control study conducted at The University of Texas MD Anderson Cancer Center during 2004–2009.[38,39] Cases were patients with pathologically confirmed pancreatic adenocarcinoma, and controls were healthy individuals frequency matched to cases by age, race, and sex. These individuals were genotyped for a total of 19 SNPs in 10 susceptibility genes identified in previous GWAS of pancreatic cancer,[40,41] including *ABO, NR5A2*, and *CLTPM1L-TERT*, as well as candidate genes *FTO, ACDC, PPARG, PRKAA2, PRKAB2,*

Table 2. Top G × G interactions for Crohn's disease (*P*-values <0.1 by either Tukey's 1-df score test or generic LRT) ordered by Tukey's 1-df test *P*-values.

GENE 1	$P_{marginal}$	GENE 2	$P_{marginal}$	P_{Tukey}	P_{LRT}	P_{SSU}	PPI in STRING
IL12RB2	0.0002	*IL12B*	0.0005	**0.012**	0.128	0.128	Yes
STAT3	0.0004	*IL12RB1*	0.058	**0.034**	0.188	0.175	No
IL12B	0.0005	*IL23A*	0.095	**0.050**	0.149	0.157	Yes
IL23R	<0.0001	*IL18R1*	0.017	**0.058**	0.358	0.353	No
STAT3	0.0004	*RORC*	0.0048	**0.061**	0.281	0.294	No
STAT3	0.0004	*IL18R1*	0.017	**0.072**	**0.071**	**0.082**	No
IL23R	<0.0001	*IL18RAP*	0.0111	**0.074**	0.209	0.431	No
IL12B	0.0005	*IL12RB1*	0.058	**0.076**	0.223	0.288	Yes
RORC	0.005	*IL10*	0.093	**0.096**	0.792	0.813	No
STAT3	0.0004	*IL18RAP*	0.011	0.194	**0.016**	**0.023**	No
IL12RB2	0.0002	*TNFα*	0.011	0.295	**0.029**	**0.035**	No
IL18R1	0.017	*IL23A*	0.095	0.370	**0.090**	0.105	No
TNFα	0.011	*IL23A*	0.095	0.929	**0.075**	**0.090**	No

Note: Interaction *P*-values less than 0.1 are in bold.

PRKAB1, and *LOC730242*. Genotyping was performed on genomic DNA from peripheral blood samples using the TaqMan method.[42]

To test the interaction between smoking and the pancreatic cancer susceptibility SNP-set defined by the 19 SNPs in the 10 genes, we applied the proposed Tukey's 1-df score test (Model 8), the generic LRT [Model (6)] and the SSU test [Model (4)] to the above-described UT MD Anderson study of pancreatic cancer with 534 cases and 552 controls. For the former two PC-based tests, we included the leading 11 PCs that explained 87% of the total genetic variation in the interaction model with smoking status (never/ever). Tukey's 1-df test and the generic LRT test gave similarly significant results: the *P*-values were 0.019 and 0.018, respectively. On the other hand, the SSU test result was not significant, with a *P*-value = 0.148. When applied to the 11 PCs, the SSU test had a significant *P*-value of 0.024. Of note, one of the 19 SNPs was the top hit in the GWAS of pancreatic cancer[40] (rs505922 in gene *ABO*), and its interaction with smoking on the risk of pancreatic cancer was highly significant (*P*-value = 0.002 by SNP × smoking test based on Model (2) and *P*-value = 0.038 after the Bonferroni correction). This significant interaction identified by the SNP-set-based G × E tests warrants further replication in independent samples.

Discussion

In this paper we proposed a powerful gene-based test for detecting G × G or G × E interactions by combining two parsimonious statistical models, namely, the PCA and Tukey's 1-df form of interaction. We derived a score test that is computationally fast and numerically stable for the proposed Tukey's 1-df interaction model. Using extensive simulations based on the HapMap phased haplotype data, we showed that the proposed score test for Tukey's 1-df interaction model controlled the Type I error rate at the nominal level satisfactorily and was more powerful than the PC-based generic LRT that tests pairwise interactions and the multiple-SNP-based SSU test as a representative of variance-component score tests. We demonstrated the utility and efficiency gains of the proposed test with applications to detecting G × G interactions for CD using the WTCCC GWAS data and to detecting SNP-set by smoking (G × E) interaction using a case–control study of pancreatic cancer. As demonstrated in the latter application, we recommend first using gene-based interaction tests to identify significant genes, and then performing SNP-based interaction tests within the genes to identify which SNPs significantly interact.

We have focused on gene-based G × G/G × E tests for common SNPs (MAF >5%) based on GWAS data. It would be of interest to extend the Tukey's 1-df test to testing G × G/G × E interactions for rare variants (MAF <5% or 1%) based on the next-generation sequencing (NGS) data. Although PCA is an effective approach to summarize a large number of common SNPs into a few uncorrelated PCs to be used in subsequent testing of genetic main effects or G × G/G × E interactions, it may not work well in capturing the genetic variation dominated by rare variants. A possible alternative is via functional principal component analysis (FPCA), which has been shown to be a powerful method for reducing the dimension of a large number of rare variants.[43,44] Further research is warranted.

The proposed method has some potential limitations. First, the PCA step is solely based on the SNP data and does not take the disease–SNP correlations into account. As a result, it is possible that the leading PCs may not capture the information of the most relevant SNP(s) in association with the disease. Alternative dimension reduction techniques, eg, the partial least squares (PLS) method,[45] could be employed to find linear combinations of the SNPs that are most correlated with the disease status. However, because of the use of the disease status in the first-stage PLS, the test statistic in the second-stage interaction model will no longer follow a χ^2-distribution under the null hypothesis of no interactions, and computationally intensive permutation or parametric bootstrap procedure is needed to obtain the null distribution of the test statistic.[7] Second, the parsimonious Tukey's 1-df form of interaction model assumes that the interaction effect is approximately proportional to the genetic main effects. Although it appeared to be powerful and robust across different simulation scenarios in our study, even in the absence of genetic main effects, it might lose power under some scenarios as investigated in Ref. 45. As the true interaction model is hardly known a priori in real data analysis and likely varies across genes, the proposed method is a competitive and complementary approach to existing gene-based interaction tests.

The current paper focuses on case–control analyses. It is well known that for SNP-based G × E analysis, the case-only test is more powerful than the standard case–control interaction test [Model (2)] if the assumption of gene–environment independence holds in the general population.[46] Specifically, the case-only analysis tests the association between the environmental exposure and the SNP of interest in the cases. It would be interesting to extend the SNP-based case-only test to a gene-based test by testing the association between the environmental exposure and the leading PCs for the multiple SNPs in a gene. Another possible direction of future research is to extend the existing SNP-based tests for nonremovable interactions[47] to gene-based tests. Finally, as the Tukey's 1-df interaction model is in the regression framework, it is not limited to binary disease phenotypes and can be easily extended to G × G/G × E tests for quantitative traits. R programs implementing the proposed Tukey's 1-df score test will be posted on our website at: https://sites.google.com/site/utpengwei/.

Acknowledgments

The authors thank the anonymous reviewers for their constructive comments. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this paper. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

Author Contributions

Conceived and designed the experiment: PW. Analyzed the data: YW, PW. Wrote the first draft of the manuscript: PW YW. Contributed to the writing of the manuscript: PW, YW, DL. Agree with manuscript results and conclusions: PW, YW, DL. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*. 2009;106(23):9362–7.
2. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. *Am J Hum Genet*. 2012;90(1):7–24.
3. Stein EA, Mellis S, Yancopoulos GD, et al. Effect of a monoclonal antibody to PCSK9 on LDL cholesterol. *N Engl J Med*. 2012;366(12):1108–18.
4. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet*. 2013;45(4):353–61.
5. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet*. 2013;45(4):400–5.
6. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A*. 2014; 111(4):E455–64.
7. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. *Genetics*. 2014;197(4):1081–95.
8. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109(4):1193–8.
9. Campa D, Kaaks R, Le Marchand L, et al. Interactions between genetic variants and breast cancer risk factors in the breast and prostate cancer cohort consortium. *J Natl Cancer Inst*. 2011;103(16):1252–63.
10. Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM. Gene-environment interactions in cancer epidemiology: a national cancer institute think tank report. *Genet Epidemiol*. 2013;37(7):643–57.
11. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat Rev Genet*. 2010;11(4):259–72.
12. Kooperberg C, Leblanc M. Increasing the power of identifying gene x gene interactions in genome-wide association studies. *Genet Epidemiol*. 2008;32(3):255–63.
13. Gauderman WJ, Zhang P, Morrison JL, Lewinger JP. Finding novel genes by testing G x E interactions in a genome-wide association study. *Genet Epidemiol*. 2013;37(6):603–13.
14. Hsu L, Jiao S, Dai JY, Hutter C, Peters U, Kooperberg C. Powerful cocktail methods for detecting genome-wide gene-environment interaction. *Genet Epidemiol*. 2012;36(3):183–94.
15. Murcray CE, Lewinger JP, Gauderman WJ. Gene-environment interaction in genome-wide association studies. *Am J Epidemiol*. 2009;169(2):219–26.
16. Murcray CE, Lewinger JP, Conti DV, Thomas DC, Gauderman WJ. Sample size requirements to detect gene-environment interactions in genome-wide association studies. *Genet Epidemiol*. 2011;35(3):201–10.
17. He J, Wang K, Edmondson AC, Rader DJ, Li C, Li M. Gene-based interaction analysis by incorporating external linkage disequilibrium information. *Eur J Hum Genet*. 2011;19(2):164–72.
18. Pan W, Basu S, Shen X. Adaptive tests for detecting gene-gene and gene-environment interactions. *Hum Hered*. 2011;72(2):98–109.
19. 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):667–81.
20. Jiao S, Hsu L, Bzieau S, et al. SBERIA: set-based gene-environment interaction test for rare and common variants in complex diseases. *Genet Epidemiol*. 2013;37(5):452–64.
21. Tang H, Wei P, Duell EJ, et al. Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene and pathway-based interaction analysis of GWAS data. *Carcinogenesis*. 2014;35:1039–45.
22. Tang H, Wei P, Duell EJ, et al. Genes-environment interactions in obesity-and diabetes-associated pancreatic cancer: a GWAS data analysis. *Cancer Epidemiol Biomarkers Prev*. 2014;23(1):98–106.
23. Tukey JW. One degree of freedom for non-additivity. *Biometrics*. 1949;5(3): 232–42.
24. Chatterjee N, Kalaylioglu Z, Moslehi R, Peters U, Wacholder S. Powerful multilocus tests of genetic association in the presence of gene-gene and gene-environment interactions. *Am J Hum Genet*. 2006;79(6):1002–16.
25. Han SS, Rosenberg PS, Garcia-Closas M, et al. Likelihood ratio test for detecting gene (G)-environment (E) interactions under an additive risk model exploiting G-E independence for case-control data. *Am J Epidemiol*. 2012;176(11):1060–7.
26. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet*. 2009;10(6):392–404.

27. Wei P, Tang H, Li D. Insights into pancreatic cancer etiology from pathway analysis of genome-wide association study data. *PLoS One*. 2012;7(10):e46887.
28. Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol*. 2009;33(6):497–507.
29. Pan W. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet Epidemiol*. 2011;35:211–6.
30. Goeman JJ, van de Geer S, van Houwelingen HC. Testing against a high dimensional alternative. *J R Stat Soc B*. 2006;68:477–93.
31. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(l):82–93.
32. Jensen LJ, Kuhn M, Stark M, et al. STRING 8-a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res*. 2009;37(Database issue):D412–6.
33. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. Efficiency and power in genetic association studies. *Nat Genet*. 2005;37(11):1217–23.
34. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet*. 2010;11(12):843–54.
35. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447(7145):661–78.
36. Sun YV. Integration of biological networks and pathways with genetic association studies. *Hum Genet*. 2012;131(10):1677–86.
37. American Cancer Society. *Cancer Facts and Figures* 2013. Atlanta: American Cancer Society; 2013.
38. Li D, Morris JS, Liu J, et al. Body mass index and risk, age of onset, and survival in patients with pancreatic cancer. *JAMA*. 2009;301(24):2553–62.
39. Wei P, Tang H, Li D. Functional logistic regression approach to detecting gene by longitudinal environmental exposure interaction in a case-control study. *Genet Epidemiol*. 2014;38(7):638–51.
40. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet*. 2009;41(9):986–90.
41. Petersen GM, Amundadottir L, Fuchs CS, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, lq32.1 and 5pl5.33. *Nat Genet*. 2010;42(3):224–8.
42. McGuigan FE, Ralston SH. Single nucleotide polymorphism detection: allelic discrimination using TaqMan. *Psychiatr Genet*. 2002;12(3):1336.
43. Luo L, Boerwinkle E, Xiong M. Association studies for next-generation sequencing. *Genome Res*. 2011;21(7):1099–108.
44. Zhang F, Boerwinkle E, Xiong M. Epistasis analysis for quantitative traits by functional regression model. *Genome Res*. 2014;24(6):989–98.
45. Wang T, Ho G, Ye K, Strickler H, Elston RC. A partial least-square approach for modeling gene-gene and gene-environment interactions when multiple markers are genotyped. *Genet Epidemiol*. 2009;33(1):6–15.
46. Mukherjee B, Ahn J, Gruber SB, Chatterjee N. Testing gene-environment interaction in large-scale case-control association studies: possible choices and comparisons. *Am J Epidemiol*. 2012;175(3):177190.
47. Satagopan JM, Elston RC. Evaluation of removable statistical interaction for binary traits. *Stat Med*. 2013;32(7):1164–90.