G3, 2021, 11(10), jkab220
DOI: 10.1093/g3journal/jkab220
Advance Access Publication Date: 1 July 2021
Investigation

A two-stage testing strategy for detecting genes×environment interactions in association studies

Jiabin Zhou,[1] Shitao Li,[2] Ying Zhou ,[1,*] and Xiaona Sheng[3,*]

[1]Department of Statistics, School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China,
[2]Department of Basic Course, Shenyang University of Technology, Liaoyang 111000, China, and
[3]School of Information Engineering, Harbin University, Harbin 150086, China

*Corresponding author: Department of Statistics, School of Mathematical Sciences, Heilongjiang University, Harbin 150080, China. Email: yzhou@aliyun.com (Y.Z.); School of Information Engineering, Harbin University, Harbin 150086, China. Email: shengxiaona2006@163.com (X.S.)

Abstract

Identifying gene×environment (G×E) interactions, especially when rare variants are included in genome-wide association studies, is a major challenge in statistical genetics. However, the detection of G×E interactions is very important for understanding the etiology of complex diseases. Although currently some statistical methods have been developed to detect the interactions between genes and environment, the detection of the interactions for the case of rare variants is still limited. Therefore, it is particularly important to develop a new method to detect the interactions between genes and environment for rare variants. In this study, we extend an existing method of adaptive combination of *P*-values (ADA) and design a novel strategy (called iSADA) for testing the effects of G×E interactions for rare variants. We propose a new two-stage test to detect the interactions between genes and environment in a certain region of a chromosome or even for the whole genome. First, the score statistic is used to test the associations between trait value and the interaction terms of genes and environment and obtain the original *P*-values. Then, based on the idea of the ADA method, we further construct a full test statistic via the *P*-values of the preliminary tests in the first stage, so that we can comprehensively test the interactions between genes and environment in the considered genome region. Simulation studies are conducted to compare our proposed method with other existing methods. The results show that the iSADA has higher power than other methods in each case. A GAW17 data set is also applied to illustrate the applicability of the new method.

Keywords: genes×environment interactions; *P*-value; rare variant; score test; two-stage approach

Introduction

Due to the availability of genetic variants, genome-wide association studies (GWAS) have successfully detected a large number of common variants associated with many human traits and diseases (Visscher *et al.* 2012; Welter *et al.* 2014). However, common variants can only explain a small proportion of disease heritability (Maher 2008; McCarthy *et al.* 2008; Bansal *et al.* 2010) and additional disease heritability can be explained by rare variants (Pritchard 2001; Pritchard and Cox 2002; Manolio *et al.* 2009). In recent years, many statistical methods have been developed for rare-variant association testing, for example, the cohort allelic sums test (Morgenthaler and Thilly 2007) which belongs to burden tests. When all rare variants in a given region are causal ones and impact on the trait in the same direction, the burden tests are more powerful (Bansal *et al.* 2010). The C-alpha method (Neale *et al.* 2011) and the sequence kernel association test (SKAT) (Wu *et al.* 2011) belong to nonburden tests. These two methods are robust to the different directions of variants effects. Unfortunately, however, when the given region includes neutral variants, both the burden tests and the nonburden tests may suffer from loss of power. To deal with this problem, Sha *et al.* (2012) proposed a novel test for detecting the collective effect of an optimally weighted combination of variants (TOW). Lin *et al.* (2014) proposed an association test by adaptive combination of *P*-values, called "ADA." Based on the ADA method, for different types of traits and variants, some extended methods have been proposed (Zhou and Wang 2015; Wang *et al.* 2018).

In biology and genetics, many complex phenotypes/diseases are usually affected by both genetic factors and environmental factors, for example, the ADH7 variants and alcohol consumption in upper aerodigestive cancers (Hashibe *et al.* 2008), or the GRIN2A variants and coffee consumption in Parkinson's disease (Hamza *et al.* 2011). Therefore, the main effects of genes and the interaction effects of G×E are major and necessary parts in statistical modeling when performing association studies. To date, increasing attention has been paid to detecting the interactions between genes and genes, and the interactions between genes and the environment aiming at common and rare variants. Jiao *et al.* (2013) proposed a two-stage method "SBERIA" for case-control studies, which can be applied to cases of common variants and rare variants. Wang *et al.* (2017) considered a set-based mixed effect model for gene-environment interactions based on the method of Jiao *et al.* (2013), and the authors discussed the

Received: May 19, 2021. **Accepted:** June 22, 2021

construction of score statistic for the terms associated with fixed and random effects of G×E to avoid direct parameter estimation in the MixGE model. Lin *et al.* (2013) proposed a gene-environment interaction association test (GESAT), which allowed easy adjustment of covariates, but was less powerful when naively applied to case of rare variants. To test rare variants by environment interactions, Lin *et al.* (2016) developed the interaction sequence kernel association test (iSKAT) to assess the effects of rare variants by environment interactions. Because the weight of iSKAT is specially selected in the simulation process, when the weight of the interaction term between gene and environment is too small, the power of this method will decrease. Based on the method TOW (Sha *et al.* 2012), Zhao *et al.* (2020) presented a novel method of test the interactions between genes and environment. However, when the weights of the interaction terms between genes and environment are in the same direction, this method also has the limitation of power loss.

In this study, we extend the ADA method of adaptive combination of *P*-values (Lin *et al.* 2014) and design a novel strategy (iSADA) to test the effects of G×E interactions for case of rare variant association studies. A new two-stage test is proposed to detect the interactions between genes and environment in a certain region of a chromosome or even for the whole genome. We first apply the score statistic to test the associations between trait value and the interaction terms of genes and environment and obtain the original *P*-values. Then, based on the idea of the ADA method, we further construct a full test statistic via the *P*-values of the preliminary test in the first step, so that we can comprehensively test the interactions between genes and environment in the target genome region. Simulation studies are conducted to compare our proposed method with other existing methods. The results show that the iSADA method has higher power than other methods in each case. Finally, we used the GAW17 data set to evaluate the performance of the method that we proposed in this study.

Materials and methods

Consider n unrelated individuals sequenced in a testing region with M rare variants. We are interested in testing the interaction effects between causal rare variants and a certain environment variable on a trait. The relationship between quantitative trait value y_i and genotype vector $\boldsymbol{G_i} = (G_{i1}, G_{i2}, \ldots, G_{iM})^T$, environmental variable E_i and the G×E interaction terms $S_{ik}(i = 1, 2, \ldots, M)$ can be constructed by the following full statistical model:

$$y_i = \sum_{k=1}^{M} \beta_k^G G_{ik} + \beta^E E_i + \sum_{k=1}^{M} \beta_k^{GE} S_{ik} + \varepsilon_{0i}, i = 1, 2, \ldots, n, \qquad (1)$$

where subscript i refers to the ith individual and k refers to the kth variant of the genetic set. β_k^G denotes the kth genotype main effect, β^E denotes environment effect, β_k^{GE} denotes the kth G×E interaction effect, $G_{ik} \in \{0, 1, 2\}$ denotes the genotype score (number of minor alleles) at the kth variant, $S_{ik} = E_i G_{ik}$ denotes the interaction term between the kth variant and the environment for individual i, and $\varepsilon_{0i} \sim N(0, \sigma_0^2)$, where $i = 1, 2, \ldots, n, k = 1, 2, \ldots, M$. We next present a two-stage testing strategy to identify interaction effects of G×E on the quantitative trait.

Remark: The true generation mechanism of any trait value is unknown, so it is difficult to seek an optimal statistical model for describing the relationship between a considered trait and the genotypes, as well as environmental variables. We first assume that model (1) holds in our method, in fact the true statistical model between the trait values y_i and genotypes G_{ik}, environmental variable E_i, and G×E interactions S_{ik} can be a more general model. Without loss of generality, the generalized linear model (GLM)

$$g(\mu_i) = \sum_{k=1}^{M} \beta_k^G G_{ik} + \beta^E E_i + \sum_{k=1}^{M} \beta_k^{GE} S_{ik} \qquad (2)$$

can be considered, where $\mu_i = E(y_i | G_{ik}, E_i, S_{ik})$ is the conditional expectation of phenotype y_i given G_{ik}, E_i, and S_{ik}; and $g(\cdot)$ is a canonical link function. Through our extensive investigation, we validate that our new testing strategy can maintain higher power relative to other methods in the case of model misspecification, although it will decrease some power in a certain degree, which guarantees the robustness of the new method.

We next present a two-stage testing strategy on the basis of linear model (1) to identify interaction effects of G×E on the quantitative trait. GLM (2) will be used to verify the robustness of the proposed method in the part of Simulation studies.

Stage one: Obtain preliminary test *P*-value for single site

In the first stage, we test associations between the trait value and each interaction term in the given genome region. To simplify the test problem, we consider the following local statistical model:

$$y_i = \beta_1^G G_{i1} + \cdots + \beta_M^G G_{iM} + \beta^E E_i + \beta_k^{GE} S_{ik} + \varepsilon_i, \qquad (3)$$

where $i = 1, 2, \cdots, n$, the explanations of variables G_{ik}, E_i, S_{ik}, and parameters β_k^G, β^E, β_k^{GE} are the same with those given below model (1); $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent and identically distributed, and $\varepsilon_i \sim N(0, \sigma^2)$.

Let $\boldsymbol{Y} = (y_1, y_2, \ldots, y_n)^T$, $\boldsymbol{X}_i = (G_{i1}, G_{i2}, \ldots, G_{iM}, E_i)^T$, where $i = 1, 2, \ldots, n$, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$.

For the kth rare variant, we consider the test problem $H_0 : \beta_k^{GE} = 0 (k = 1, 2, \ldots, M)$. The score statistic is used to test the multiple hypothesis, and the basic form of the score test is given by the following theorem.

Theorem. The score statistic to test the null hypothesis $H_0 : \beta_k^{GE} = 0$ under model (3) is given by

$$Q_k = \frac{1}{\hat{\sigma}^2} \frac{W^2}{U},$$

where $\hat{\sigma}^2 = \frac{1}{n} \boldsymbol{Y}^T (\boldsymbol{I_n} - \boldsymbol{V}) \boldsymbol{Y}$, $W = \boldsymbol{S}_k^T (\boldsymbol{I_n} - \boldsymbol{V}) \boldsymbol{Y}$, $U = \boldsymbol{S}_k^T (\boldsymbol{I_n} - \boldsymbol{V}) \boldsymbol{S}_k$, $\boldsymbol{V} = \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T$, $\boldsymbol{S}_k = (S_{1k}, S_{2k}, \ldots, S_{nk})^T$, and $\boldsymbol{I_n}$ is an $n \times n$ identify matrix.

Proof Under model (3), $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{S}_k \beta_k^{GE} + \boldsymbol{\varepsilon}$, and $\boldsymbol{Y} \sim N(\boldsymbol{X\beta} + \boldsymbol{S}_k \beta_k^{GE}, \boldsymbol{I_n} \sigma^2)$, where $\boldsymbol{\beta}^T = ((\boldsymbol{\beta^G})^T, \beta^E) = (\beta_1^G, \cdots, \beta_M^G, \beta^E)$. The log-likelihood is given by

$$lnL(\boldsymbol{\theta}) = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \beta_k^{GE}, \sigma^2)^T$. Then

$$\frac{\partial lnL(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^T\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}\right),$$

$$\frac{\partial lnL(\boldsymbol{\theta})}{\partial \beta_k^{GE}} = \frac{1}{\sigma^2}\mathbf{S_k}^T\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}\right),$$

$$\frac{\partial lnL(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}),$$

$$\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}, \frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\partial \beta_k^{GE}} = -\frac{1}{\sigma^2}\mathbf{X}^T\mathbf{S_k},$$

$$\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}\partial \sigma^2} = -\frac{1}{\sigma^4}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}),$$

$$\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial (\beta_k^{GE})^2} = -\frac{1}{\sigma^2}\mathbf{S_k}^T\mathbf{S_k}, \frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \beta_k^{GE}\partial \boldsymbol{\beta}^T} = -\frac{1}{\sigma^2}\mathbf{S_k}^T\mathbf{X},$$

$$\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \beta_k^{GE}\partial \sigma^2} = -\frac{1}{\sigma^4}\mathbf{S_k}^T\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}\right),$$

$$\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}),$$

$$\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \sigma^2\partial \boldsymbol{\beta}^T} = -\frac{1}{\sigma^4}\mathbf{X}^T\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}\right),$$

$$\frac{\partial^2 lnL(\boldsymbol{\theta})}{\partial \sigma^2\partial \beta_k^{GE}} = -\frac{1}{\sigma^4}\mathbf{S_k}^T\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{S_k}\beta_k^{GE}\right).$$

Let $\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}^2$ denote the maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}$ and σ^2 under null hypothesis $H_0 : \beta_k^{GE} = 0$. The log-likelihood under H_0 is given by

$$lnL(\boldsymbol{\theta}_0) = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}),$$

where $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}, \sigma^2)^T$. Let

$$\frac{\partial lnL(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0,$$

$$\frac{\partial lnL(\boldsymbol{\theta}_0)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = 0,$$

then we can obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \ \hat{\sigma}^2 = \frac{1}{n}\mathbf{Y}^T(\mathbf{I_n} - \mathbf{V})\mathbf{Y},$$

where $\mathbf{V} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, and $\mathbf{I_n}$ is an $n \times n$ identify matrix. Based on the above computation, then the score matrix and information matrix can be obtained as follows,

$$\boldsymbol{H} = \frac{\partial lnL(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \beta_k^{GE}=0, \sigma^2=\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}(\mathbf{0}^T, \mathrm{W}, 0)^T,$$

and

$$\boldsymbol{I} = -\mathrm{E}\frac{\partial^2 ln(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}\beta_k^{GE}=0, \sigma^2=\hat{\sigma}^2} = \frac{1}{\hat{\sigma}^2}\begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{S_k} & \mathbf{0} \\ \mathbf{S_k}^T\mathbf{X} & \mathbf{S_k}^T\mathbf{S_k} & \frac{\mathrm{W}}{\hat{\sigma}^2} \\ \mathbf{0}^T & \frac{\mathrm{W}}{\hat{\sigma}^2} & \frac{n}{2\hat{\sigma}^2} \end{bmatrix},$$

respectively. Therefore, the score test statistic is given by

$$Q_k = \boldsymbol{H}^T\boldsymbol{I}^{-1}\boldsymbol{H} = \frac{1}{\hat{\sigma}^2}\frac{\mathrm{W}^2}{\mathrm{U}},$$

where $\mathrm{W} = \mathbf{S_k}^T(\boldsymbol{I_n} - \mathbf{V})\mathbf{Y}$, and $\mathrm{U} = \mathbf{S_k}^T(\boldsymbol{I_n} - \mathbf{V})\mathbf{S_k}$.

After performing the above score test for each interaction effect β_k^{GE} $(k = 1, 2, \ldots, M)$, we can obtain the corresponding *P*-value of each test. Let $P_1, P_2, \ldots, P_M$ respectively represent the *P*-values of testing for the associations of G×E interaction terms with the trait value. Next, we give stage two of the proposed strategy, so that the G×E interactions in the given region can be tested comprehensively.

Stage two: Integrate the *P*-values of each site and construct a test statistic for a certain region

In the second stage, based on the ADA method, we give the statistic of the combination of *P*-values to test the G×E interactions of a certain region. In detail, we first assign an attribute to each variant site. For each site, we calculate the sample covariance of the genotypes and the trait values of all subjects. If the covariance is greater than 0, we name the site as "deleterious-inclined variant site." To more effectively guard against the noise caused by neutral variants (Zaykin *et al.* 2002; Yang and Chen 2011), following the idea in Lin *et al.* (2014), we then consider J candidate truncation thresholds $\rho_1, \rho_2, \ldots, \rho_J$. For the jth candidate truncation threshold, the significance score of the deleterious-inclined sites is calculated by

$$S_j^+ = -\sum_{k=1}^{M}\xi_k \cdot I_{\{p_k < \rho_j\}} \cdot \omega_k \cdot \log p_k, \ j = 1, \ldots, J, \tag{4}$$

where ξ_k is an indicator variable coded as 1 if the kth site is deleterious-inclined and 0 otherwise, $I_{\{p_k < \rho_j\}}$ is an indicator variable coded as 1 if $p_k < \rho_j$ and 0 otherwise, and ω_k is a weight given to the kth site. In this study, we use the weight $\omega_k = \mathrm{Beta}(\mathrm{MAF}_k; 1, 25)$ that proposed by Ionita-Laza *et al.* (2013). Similarly, when the sample covariance between the genotypes of a variant site and the trait values of all individuals is less than 0, we name the site as "protective-inclined variant site." For the jth candidate truncation threshold, the significance score of the protective-inclined sites is calculated by

$$S_j^- = -\sum_{k=1}^{M}\varphi_k \cdot I_{\{p_k < \rho_j\}} \cdot \omega_k \cdot \log p_k, \ j = 1, 2, \ldots, J, \tag{5}$$

where φ_k is an indicator variable coded as 1 if the kth site is protective-inclined and 0 otherwise, $I_{\{p_k < \rho_j\}}$ is an indicator variable coded as 1 if $p_k < \rho_j$ and 0 otherwise. By Equations (4) and (5), we obtain the significance score of deleterious-inclined

variants S_j^+ and protective-inclined variants S_j^-, respectively, for each candidate truncation threshold. Following the ADA method (Lin *et al.* 2014), we specify $J = 11$ equally spaced candidate truncation thresholds, that is, $(\rho_1, \ldots, \rho_J) = (0.10, 0.11, \ldots, 0.20)$. Then the test statistic regardless of the effect directions (deleterious or protective) is $S_j = \max(S_j^+, S_j^-)$, $j = 1, 2, \ldots, J$.

In order to obtain the P-value of the observed statistic $S_j = \max(S_j^+, S_j^-)$ for each j $(j = 1, \ldots, J)$, following the method in Lin *et al.* (2014), we first conduct B permutations since the distribution of the statistic S_j is cannot be obtained directly. For the rth permutation $(1 \le r \le B)$, we randomly shuffle the trait values without changing the genotypes of each individual. In this way, we can get a new set of permutation samples, and then we can obtain the value of the permuted statistic $S_j^{(r)} = \max(S_j^+(r), S_j^-(r))$, $r = 1, 2, \ldots, B$.

By comparing each $S_j^{(r)}$ $(r = 1, 2, \ldots, B)$ with S_j among the B permutations, we can further estimate the P-value p_j^* of S_j $(j = 1, 2, \ldots, J)$ with the corresponding frequency. The final statistic is the minimum P-value among p_j^* across the J candidate truncation thresholds for the observed samples, *i.e.*, $T = \min\{p_1^*, p_2^*, \ldots, p_J^*\}$. Second, in a similarly way of permutation by comparing $T^{(r)}$ $(r = 1, \ldots, B)$ with T, we can obtain the "adjusted P-value" $p_T = \frac{\sum_{r=1}^{B} I(T^{(r)} \le T) + 1}{B+1}$ for statistic T (Lin *et al.* 2014; Wang *et al.* 2018).

The proposed strategy in this study is referred to as "iSADA," since we first use the score statistic to test the associations between trait value and the interaction terms of genes and environment and obtain the original P-values, and then, based on the idea of the ADA method, we further construct a comprehensive test statistic via the P-values of the preliminary test in the first stage. By the newly proposed strategy, we can effectively exclude the impact of neutral variants in a given genome region and comprehensively test the interactions between genes (or variants) and environment in the given region.

Simulation studies using linear model

In this section, we conduct numerical studies to evaluate the performance of the proposed iSADA for detecting G×E interactions of rare variants. The R code for implementing our method and the user's manual of it can be found in Supplementary materials. In our simulation design, we simulate $M = (20, 100)$ rare variants of 500 subjects and generate quantitative trait values. The minor allele frequency (MAF) of the kth rare variant site is randomly generated by $\text{MAF}_k \sim U(0.005, 0.05)$, $k = 1, 2, \ldots, M$, and then the genotype can be generated under the assumption of Hardy-Weinberg equilibrium (HWE). The environmental factor is assigned to be binary variable that take values -1 and 1 with probability 0.3 and 0.7, respectively. The quantitative trait value Y is generated by the full model (1), with a residual variable of standard normal distribution. We compared the performance of our proposed method with five other methods: iSKAT (Lin *et al.* 2016), iSKAT0 (β_k^{GE}s are assumed to be independent in the SKAT method with $\rho = 0$), iSKAT1 (β_k^{GE}s are perfectly correlated in the SKAT method with $\rho = 1$), MixGE-f and MixGE-t (Wang *et al.* 2017).

In our simulation, the P-value of the iSADA is obtained with 300 permutations, and the type I error rates and powers are evaluated by 1000 replications at the nominal significance level α of 0.01 and 0.05, respectively.

For $M = 20$ loci, for parameters β^E and $\boldsymbol{\beta}^G$ we give three settings when calculating the estimated type I error rates or powers.

Table 1 Simulation results of powers for quantitative trait with G×E effects of opposite directions (20 loci)

Sig. level	Method[a]	Setting I	Setting II	Setting III
0.01	iSKAT	0.007	0.007	0.007
	iSKAT0	0.008	0.008	0.008
	iSKAT1	0.006	0.006	0.006
	MixGE-f	0.171	0.251	0.177
	MixGE-t	0.141	0.197	0.129
	iSADA	0.238	0.305	0.291
0.05	iSKAT	0.046	0.047	0.047
	iSKAT0	0.049	0.049	0.048
	iSKAT1	0.056	0.057	0.055
	MixGE-f	0.379	0.508	0.393
	MixGE-t	0.330	0.442	0.321
	iSADA	0.449	0.513	0.508

[a]Note: $\boldsymbol{\beta}^{GE} = 0.2 \times (1, 0, 1, 0, -1, 1, 0, 1, 0, -1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0)^T$.

Simulation settings for the case of 20 loci

Setting I: $\beta^E = 0$, $\boldsymbol{\beta}^G = \mathbf{0}$;
Setting II: $\beta^E = 0$, $\boldsymbol{\beta}^G = 0.03 \times (1, 1, 1, 0, -1, 1, 0, 1, 0, -1, 0, 1, 1, 0, -1, 1, 1, 1, 1, 1)^T$;
Setting III: $\beta^E = 0.01$, $\boldsymbol{\beta}^G = 0.01 \times (1, 1, 1, 0, -1, 1, 0, 1, 0, -1, 0, 1, 1, 0, -1, 1, 1, 1, 1, 1)^T$.

For $M = 100$ loci, for parameters β^E and $\boldsymbol{\beta}^G$ we also give three settings when calculating the estimated type I error rates or powers.

Simulation settings for the case of 100 loci

Setting I: $\beta^E = 0$, $\boldsymbol{\beta}^G = \mathbf{0}$;
Setting II: $\beta^E = 0$, $\boldsymbol{\beta}^G = 0.03 \times \boldsymbol{\beta}_0^G$;
Setting III: $\beta^E = 0.01$, $\boldsymbol{\beta}^G = 0.01 \times \boldsymbol{\beta}_0^G$;

where elements of $\boldsymbol{\beta}_0^G$ include 25 zeros, 60 ones, and 15 (-one)s. The true value of parameter vector $\boldsymbol{\beta}^{GE} = (\beta_1^{GE}, \beta_2^{GE}, \ldots, \beta_M^{GE})^T$ is zero vector under H_0. Under H_1, we design different cases of G×E interactions and the values of $\boldsymbol{\beta}^{GE}$ can be found below Tables 1–4, where the proportions of negative, neutral and positive interactions are 10, 55, and 35% in Tables 1 and 3; and the proportions of neutral and positive interactions are 55, and 45% in Tables 2 and 4.

Simulation studies using generalized linear model

In this section, in order to further evaluate the performance robustness of the proposed method, we conduct simulation studies under situation of model misspecification. The quantitative trait value Y is randomly generated by the following GLM model

Table 2 Simulation results of powers for quantitative trait with G×E effects of same directions (20 loci)

Sig. level	Method[a]	Setting I	Setting II	Setting III
0.01	iSKAT	0.009	0.010	0.009
	iSKAT0	0.007	0.007	0.007
	iSKAT1	0.004	0.005	0.006
	MixGE-f	0.322	0.237	0.248
	MixGE-t	0.246	0.188	0.189
	iSADA	0.485	0.571	0.552
0.05	iSKAT	0.048	0.047	0.051
	iSKAT0	0.052	0.053	0.053
	iSKAT1	0.051	0.005	0.051
	MixGE-f	0.551	0.455	0.503
	MixGE-t	0.495	0.387	0.429
	iSADA	0.709	0.764	0.762

[a]Note: $\boldsymbol{\beta}^{GE} = 0.2 \times (1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0)^T$.

Table 3 Simulation results of powers for quantitative trait with G×E effects of opposite directions (100 loci)

Sig. level	Method[a]	Setting I	Setting II	Setting III
0.01	iSKAT	0.011	0.011	0.011
	iSKAT0	0.007	0.008	0.008
	iSKAT1	0.010	0.011	0.010
	MixGE-f	0.839	0.850	0.805
	MixGE-t	0.685	0.754	0.638
	iSADA	1.000	1.000	1.000
0.05	iSKAT	0.056	0.054	0.054
	iSKAT0	0.055	0.052	0.055
	iSKAT1	0.051	0.052	0.050
	MixGE-f	0.943	0.948	0.929
	MixGE-t	0.900	0.917	0.863
	iSADA	1.000	1.000	1.000

[a]Note: $\boldsymbol{\beta}^{GE}$=0.2× $\boldsymbol{\beta}_0^{GE}$, and $\boldsymbol{\beta}_0^{GE}$ includes 55 zeros, 35 ones, and 10 (−1)s.

Table 4 Simulation results of powers for quantitative trait with G×E effects of same directions (100 loci)

Sig. level	Method[a]	Setting I	Setting II	Setting III
0.01	iSKAT	0.014	0.011	0.014
	iSKAT0	0.015	0.010	0.015
	iSKAT1	0.013	0.014	0.012
	MixGE-f	0.954	0.940	0.939
	MixGE-t	0.914	0.876	0.869
	iSADA	1.000	1.000	1.000
0.05	iSKAT	0.071	0.067	0.071
	iSKAT0	0.075	0.074	0.075
	iSKAT1	0.049	0.056	0.048
	MixGE-f	0.990	0.983	0.980
	MixGE-t	0.981	0.967	0.967
	iSADA	1.000	1.000	1.000

[a]Note: $\boldsymbol{\beta}^{GE}$=0.2× $\boldsymbol{\beta}_0^{GE}$, and $\boldsymbol{\beta}_0^{GE}$ includes 55 zeros and 45 ones.

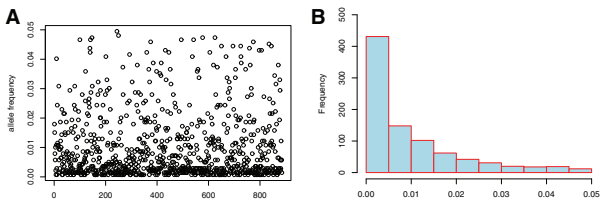

Figure 1 The MAFs of 839 rare SNP variants on chromosome 4. (A) presents individual site's MAF and (B) shows the histogram of all these variants.

$$g(\mu_i) = \exp^{-\{G_i^T\boldsymbol{\beta}^G + E_i\beta^E + S_i^T\boldsymbol{\beta}^{GE}\}},$$

where the settings of parameters $\boldsymbol{\beta}^G$, β^E, and $\boldsymbol{\beta}^{GE}$ are same as the previous simulation; The environmental factor is still the binary variable and the residual variable follows a standard normal distribution.

We still use the six methods iSADA, iSKAT, iSKAT0, iSKAT1, MixGE-f, and MixGE-t to analyze the simulation data, and the whole process is replicated for 1000 times to estimate type I error rates and powers. The part of simulation results were summarized in Supplementary Tables S1–S6 of the Supplementary Tables of this study, because of the contextual restriction.

Statistical analysis of GAW17 data set

In this section, we apply the proposed iSADA method to analyze the data set of 697 independent subjects of the Genetic Analysis Workshop 17 (GAW17), which was in fact obtained from the sequence alignment files supplied by the 1000 Genomes Project (http://www.1000genomes.org). The GAW17 data set includes genotypes of 24,487 autosomal markers (SNPs) assigned to 3205 genes, simulated affection status, quantitative traits or risk factors (Q1, Q2, and Q4), age, sex, and smoking status. The GAW17 has reported that trait Q1 was influenced by 9 genes and there exist 1 G×E (smoking) interaction effect on trait Q1, where the KDR gene on chromosome 4 has a significant interaction with smoking (Almasy *et al.* 2011).

Table 5 The corresponding relationship between the original SNP codes with the new ones in GAW17 data set analysis

Region	SNP[a]				
161–165	C4S1818	C4S1824	C4S183	C4S1830	C4S1831
166–170	C4S1834	C4S1837	C4S1838	C4S1839	C4S1842
171–175	C4S1843	C4S1847	C4S1859	**C4S1861**	C4S1868
176–180	C4S1872	**C4S1873**	**C4S1874**	**C4S1877**	**C4S1879**
181–185	C4S1881	**C4S1884**	**C4S1887**	**C4S1889**	**C4S1890**
186–190	C4S1892	C4S1893	C4S19	C4S1916	C4S1917
191–195	C4S1927	C4S1928	C4S1929	C4S193	C4S1936
196–200	C4S1938	C4S194	C4S197	C4S1996	C4S1997

[a]Note: SNPs in bold type are contained in the KDR gene.

Table 6 Summary results of the detected genome regions via 6 methods based on the GAW17 data set

P-value[a]						
iSKAT	**iSKAT0**	**iSKAT1**	**MixGE-f**	**MixGE-t**	**iSADA**	**Region**
0.042	0.360	**0.025**	**0.018**	**0.037**	**0.006****	121–140
1.000	0.816	1.000	0.367	0.545	**0.029***	161–180
0.323	0.772	0.200	0.204	0.192	**0.009****	181–200
0.076	**0.039**	1.000	**0.015**	**0.011**	**0.016***	221–240
0.044	0.109	**0.027**	0.306	0.407	0.405	421–440
0.097	0.054	0.196	**0.010**	**0.011**	0.395	541–560
0.064	**0.035**	0.182	**0.045**	**0.034**	0.312	581–600
0.044	0.290	**0.027**	0.388	0.545	**0.049***	821–839

[a]Note: **: *P*-value <0.01; *: *P*-value <0.05.

Here, we validated the feasibility of using the proposed iSADA method to detect the gene×smoking interaction effect on Q1. We selected the genetic data of chromosome 4 of 697 unrelated individuals in the GAW17 data, and mainly studied the 839 rare SNP variants given in the data set, the MAFs of which are ≤ 0.05. Figure 1 shows the allele frequency distribution of the 839 rare variant sites (see Figure 1). Figure 1A presents the scatter plot of the allele frequencies in site order, and Figure 1B shows the histogram of all these allele frequencies. For the 839 rare variant sites, we chose every continuous 20 rare SNP sites to compose a testing region and there are a total of 42 such regions (*i.e.*, regions 1–20, 21–40,⋯, and 821–839). All sites of the reported KDR gene are contained within the assigned regions of 161–180 and 181–200 in this study (see Table 5 for the corresponding relationship between the original SNP codes with the new ones in this analysis)(Almasy *et al.* 2011). Then the iSADA method, as well as the other five methods (iSKAT, iSKAT0, iSKAT1, MixGE-f, and MixGE-t) was used to detect the interaction effects between each region of rare variants and smoking status on Q1. We listed all significant interaction regions that were detected by the above methods and the corresponding *P*-values in Table 6.

Results

Simulated data result for case of linear model

a. Evaluating type I error rates:

The results for the type I error rates are shown in Tables 7 and 8. For 1000 replicated samples, the confidence interval (CI) of

Table 7 Simulation results of Type I error rates for quantitative trait with binary environment (20 loci)

Sig. level	Method[a]	Setting I	Setting II	Setting III
0.01	iSKAT	0.006	0.007	0.006
	iSKAT0	0.008	0.008	0.008
	iSKAT1	0.005	0.005	0.005
	MixGE-f	0.008	0.013	0.016
	MixGE-t	0.008	0.016	0.008
	iSADA	0.011	0.015	0.011
0.05	iSKAT	0.045	0.045	0.044
	iSKAT0	0.049	0.049	0.049
	iSKAT1	0.048	0.048	0.048
	MixGE-f	0.044	0.059	0.056
	MixGE-t	0.038	0.053	0.060
	iSADA	0.047	0.050	0.049

[a]Note: $\boldsymbol{\beta}^{GE}=0$.

Table 8 Simulation results of Type I error rates for quantitative trait with binary environment (100 loci)

Sig. level	Method[a]	Setting I	Setting II	Setting III
0.01	iSKAT	0.006	0.006	0.007
	iSKAT0	0.004	0.004	0.004
	iSKAT1	0.009	0.009	0.010
	MixGE-f	0.008	0.013	0.016
	MixGE-t	0.009	0.009	0.008
	iSADA	0.006	0.010	0.008
0.05	iSKAT	0.044	0.045	0.043
	iSKAT0	0.032	0.034	0.031
	iSKAT1	0.049	0.049	0.050
	MixGE-f	0.036	0.038	0.043
	MixGE-t	0.046	0.037	0.037
	iSADA	0.052	0.061	0.053

[a]Note: $\boldsymbol{\beta}^{GE}=0$.

type I error rate is $(\alpha - 2\sqrt{\frac{\alpha(1-\alpha)}{1000}}, \alpha + 2\sqrt{\frac{\alpha(1-\alpha)}{1000}})$, where α denotes nominal significance level. So the CIs of type I error rates for significance levels 0.01 and 0.05 are (0.0037, 0.0163) and (0.0363, 0.0637), respectively. When there are no main effects and environmental effect (see Setting I of Tables 7 and 8), all methods can control type I error rates well. We also evaluated the performance of these G×E tests in the presence of genetic main effects. When $\boldsymbol{\beta}^{G} \neq \boldsymbol{0}$ (see Setting II, Setting III of Tables 7 and 8), all the simulation results of the estimated type I error rates are close to the nominal significance levels. It can be found from the two tables that all the estimated type I error rates are all located in their own CIs. In addition, even if the number of variant sites increases from 20 to 100, the type I error rates are also stable for all tests, which shows that the type I error rates do not depend on the number of variant sites. Therefore, these results indicate that all the tests are valid ones when conducting test of G×E interactions.

By comparing the type I error rates among Settings I–III, we find for most of the methods that there is a small difference in the estimated values, but the values in Setting II are a little bigger than the corresponding ones in Settings I and III, which means that the main effect of genes (rare variants) have a certain impact on the G×E interaction testing.

b. Power comparison:

The simulated power results of testing G×E interactions are shown in Tables 1–4. First, in each panel of all the tables, it can be clearly found that methods MixGE-f, MixGE-t, and iSADA are more powerful than methods iSKAT, iSKAT0, and iSKAT1, since the power values of the former are greater than those of the latter. The powers of the iSKAT and its related methods may further improve, when the true G×E interaction effects increase with the fixed sample size, or increasing the sample size with the fixed G×E interaction effects. As expected, all the powers increase as the significance level varies from 0.01 to 0.05. Among all these methods, the proposed iSADA method is the most powerful in each scenario.

Different settings of β^{E} and $\boldsymbol{\beta}^{G}$ have certain impacts on the G×E interaction test of the iSADA method. In Tables 1 and 2, the test powers in Setting II ($\beta^{E}=0$ and $|\beta_k^{G}|=0.03$) are higher than the corresponding ones in Setting I ($\beta^{E}=0$ and $|\beta_k^{G}|=0$) and Setting III ($\beta^{E}=0.01$ and $|\beta_k^{G}|=0.01$), which further validates that the main effects of genes (rare variants) have more positive impact than the main effect of environment variable when performing the G×E interaction test by the iSADA method. The powers of the other methods are affected by more factors besides parameters β^{E} and $\boldsymbol{\beta}^{G}$, however, their pattern is not apparent.

In the simulation, different proportions of positive and negative G×E interaction effects were designed, and the proportion of neural variants was set as 55% for all scenarios. When the number of variant loci is 20, by comparing the corresponding power values of Table 1 (opposite directions of interactions) and Table 2 (same direction of interactions), we find that the iSKAT and its related methods are less impacted by the proportions of positive and negative G×E interaction effects; The powers of methods MixGE-f and the MixGE-t increase with the proportion of positive G×E interaction effect increasing in Settings I and III, but it shows opposite trend in Setting II; For the iSADA method, the powers in Table 2 are higher than the corresponding ones in Table 1, that is to say, the iSADA method performs better when the directions of G×E interaction effects are same. When the number of variant loci is 100, by comparing the power results in Tables 3 and 4, we find that the iSKAT and its related methods are still less impacted by the proportion of G×E interaction effects. But methods MixGE-f and the MixGE-t perform better when the G×E interaction effects have the same direction. For both cases, the powers of the iSADA method reach the maximum 1.

In addition, the number of variant loci also affects the test powers of each method. Under the condition of same proportion 55% of neural G×E interaction effects, through comparing the powers in Table 1 with those in Table 3, it is not difficult to find that the testing regions with more variants correspond to higher test powers. The same results hold when comparing the powers in Table 2 with those in Table 4.

Simulated data result for case of generalized linear model

The results are shown in Supplementary Tables. From Supplementary Tables S1 and S2, it is not difficult to find that all methods can control type I error rates, even if we are choosing a new model to generate simulation data, which means all methods are still credible. From Supplementary Tables S3 and S6, we find that the power results of other methods are close to those obtained under simulation situations of linear model, but the results of the MixGE-f and MixGE-t drop too much. The powers of the iSADA decrease a little and the powers of the iSKAT and related methods get somewhat increase. Among all the methods, the proposed iSADA method still performs best. These results sufficiently show that the new method is less impacted by model misspecification, and it is a robust and powerful method for testing the G×E interaction effects.

Table 9 Computing time of the iSADA method with typical parameters

Parameter[a]	Number of loci	Type	Setting I (hours)	Setting II (hours)	Setting III (hours)
	20	Type I error rate	1.199	1.192	1.211
	20	Power (G×E effects of opposite directions)	1.835	1.841	1.835
	20	Power (G×E effects of same directions)	1.833	1.845	1.831
	100	Type I error rate	8.366	8.368	8.372
	100	Power (G×E effects of opposite directions)	8.412	8.391	8.412
	100	Power (G×E effects of same directions)	8.424	8.426	8.423

[a]Note: sample size: 500; replication: 1000 times; permutation: 300 times.

GAW17 data result

It can be found from the testing results in Table 6 that 8 significant interaction regions are detected by these methods. The "*" behind *P*-value means that the corresponding region was detected significantly at nominal level 0.05, and the "**" means more significantly at nominal level 0.01. The proposed iSADA method found 5 significant interaction regions, and the other method found less than our method. The iSADA method successfully detected the target interaction regions of 161–180 and 181–200 that contains rare variants of the KDR gene (C4S1861, C4S1873, C4S1874, C4S1877, C4S1879, C4S1884, C4S1887, C4S1889, and C4S1890); However, the other methods did not give a significant result. From this view, it shows that the iSADA method performs better than the other methods when detecting G×E interaction on a quantitative trait in a certain genome region.

Besides the above two regions, the iSADA method also found G×E interactions in the regions of 121–140, 221–240 and 821–839. In fact, all the other methods except the iSKAT0 also found G×E interaction in this region of 121–140. Methods iSKAT0, MixGE-f, and MixGE-t also detected G×E interaction region of 221-240. And methods iSKAT, iSKAT1 found G×E interaction region of 821–839. Unfortunately, the regions of 121–140, 221–240, and 821–839 are false positive ones, *i.e.*, there is no true sites with G×E interactions in these regions (Almasy *et al.* 2011).

Discussion

Increasing evidence shows that G×E interactions of rare variants may play an important role in explaining the etiology of complex disease (Shields and Harris 2000; Ramos and Olden 2008; Aschard *et al.* 2012). In this study, we extend an existing method of adaptive combination of *P*-values (Lin *et al.* 2014), and design a novel strategy (iSADA) to test the G×E interaction effects for rare variants. Simulation studies show that the newly proposed method in this study is powerful and robust in each scenario, and the results of GAW17 data set analysis validate the application of the iSADA method.

In the simulation part, we employ a linear model to model the relationship between the trait value and the G×E interactions as well as the other two terms. Meanwhile, we also use the GLM to model the relationship between the trait value and those variables in order to evaluate the impact of model misspecification to the proposed iSADA method, which further verifies that the proposed method has good robustness under different models. Besides, G×G interaction terms can also be considered in the statistical model, which will not affect the simulation results. We also analyze the data set of GAW17, compared with the other 5 methods, the iSADA method effectively detected the interaction effects of rare variants and smoking status on quantitative trait Q1, which shows that the iSADA method has better detecting ability.

Some recent studies have shown that most complex disorders are potentially caused by both rare and common variants (Walsh and King 2007; Stratton and Rahman 2008). Although we mainly focus on detecting of G×E interactions for rare variants in this study, in fact, our proposed method can also be extended and applied to test G×E interactions for common variants. When simultaneously considering rare variants and common variants in the proposed method, one can properly adjust the weights in the statistics S_j^+ and S_j^-, and we suggest that the weights of common variants are set as $\omega_k = \text{Beta}(\text{MAF}_k; 0.5, 0.5)$, which was proposed by Ionita-Laza *et al.* (2013). For the trait type, although we focus on continuous traits in this study, our method can also be applied to other traits, where the GLM can be used to fit the relationship between the trait and G×E interaction terms. Besides, for stage one of our method, it is feasible to use the Levene test statistic to test the equality of variances under different genotypes for each variant site, which equivalently tests the G×E interaction for that site. But it should be noted that different local statistical model needs to be considered in the part of theoretical analysis.

The proposed method also has some limitations. For example, the running time of the iSADA method is a little longer than that of the other methods. The computation of one simulated data containing 500 subjects (considering 100 rare variant sites) takes about 30 seconds on average in the Windows System with Intel Core i5-3470 3.20 GHz processor and 4 GB memory, since more permutations are used in the iSADA method. The computing times of the iSADA method are shown in Table 9. For the same data the other methods only take no more than 10 seconds. In addition, using Levene test in stage one may suffer a similar computing time problem. Therefore, in the future we will conduct studies to improve the running speed of the iSADA method.

Data availability

The simulated data can be generated by the R code which are provided in the Supplementary material (Name: R code for iSADA and User's manual.zip). GAW17 data can be found at http://www.gaworkshop.org.

Supplementary material is available at *G3* online.

Acknowledgments

The authors would like to thank the joint Editor and referees for their helpful comments that greatly improved the presentation of the paper. We thank the Genetic Analysis Workshop 17 for providing the data set analyzed in this paper.

Funding

The National Natural Science Foundation of China (12071114) and the Natural Science Foundation of Heilongjiang Province of China (LH2019A020).

Conflicts of interest

None declared.

Literature cited

Almasy L, Dyer TD, Peralta JM, Kent JW, Charlesworth JC, *et al.* 2011. Genetic analysis workshop 17 mini-exome simulation. BMC Proc. 5:S2.

Aschard H, Chen J, Cornelis M, Chibnik L, Karlson E, *et al.* 2012. Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. Am J Hum Genet. 90:962–972.

Bansal V, Libiger O, Torkamani A, Schork N. 2010. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 11:773–785.

Hamza T, Chen H, Hill-Burns E, Rhodes S, Montimurro J, *et al.* 2011. Genome-wide gene-environment study identifies glutamate receptor gene grin2a as a Parkinson's disease modifier gene via interaction with coffee. PLoS Genet. 7:e1002237.

Hashibe M, McKay J, Curado M, Oliveira J, Koifman S, *et al.* 2008. Multiple ADH genes are associated with upper aerodigestive cancers. Nat Genet. 40:707–709.

Ionita-Laza I, Lee S, Makarov V, Buxbaum J, Lin X. 2013. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 92:841–853.

Jiao S, Hsu L, Bézieau S, Brenner H, Chan A, *et al.* 2013. Sberia: set-based gene-environment interaction test for rare and common variants in complex diseases. Genet Epidemiol. 37:452–464.

Lin W, Lou X, Gao G, Liu N. 2014. Rare variant association testing by adaptive combination of p-values. PLoS One. 9:e85728.

Lin X, Lee S, Christiani D, Lin X. 2013. Test for interactions between a gene/snpset and environment/treatment in generalized linear models. Biostatistics. 14:667–681.

Lin X, Lee S, Wu M, Wang C, Chen H, *et al.* 2016. Test for rare variants by environment interactions in sequencing association studies. Biometrics. 72:156–164.

Maher B. 2008. Personal genomes: the case of the missing heritability. Nature. 456:18–21.

Manolio T, Collins F, Cox N, Goldstein D, Hindorff L, *et al.* 2009. Finding the missing heritability of complex diseases. Nature. 461: 747–753.

McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, *et al.* 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 9:356–369.

Morgenthaler S, Thilly W. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (cast). Mutat Res. 615:28–56.

Neale B, Rivas M, Voight B, Altshuler D, Devlin B, *et al.* 2011. Testing for an unusual distribution of rare variants. PLoS Genet. 7: e1001322.

Pritchard J. 2001. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 69:124–137.

Pritchard J, Cox N. 2002. The allelic architecture of human disease genes: common disease-common variant. . .or not? Hum Mol Genet. 11:2417–2423.

Ramos R, Olden K. 2008. Gene-environment interactions in the development of complex disease phenotypes. Int J Environ Res Public Health. 5:4–11.

Sha Q, Wang X, Wang X, Zhang S. 2012. Detecting association of rare and common variants by testing an optimally weighted combination of variants. Genet Epidemiol. 36:561–571.

Shields P, Harris C. 2000. Cancer risk and low-penetrance susceptibility genes in gene-environment interactions. J Clin Oncol. 18: 2309–2315.

Stratton M, Rahman N. 2008. The emerging landscape of breast cancer susceptibility. Nat Genet. 40:17–22.

Visscher P, Brown M, McCarthy M, Yang J. 2012. Five years of GWAS discovery. Am J Hum Genet. 90:7–24.

Walsh T, King M. 2007. Ten genes for inherited breast cancer. Cancer Cell. 11:103–105.

Wang C, Sun J, Guillaume B, Ge T, Hibar D, *et al.* 2017. A set-based mixed effect model for gene-environment interaction and its application to neuroimaging phenotypes. Front Neurosci. 11:191.

Wang M, Ma W, Zhou Y. 2018. Association detection between ordinal trait and rare variants based on adaptive combination of p values. J Hum Genet. 63:37–45.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, *et al.* 2014. The nhgri gwas catalog, a curated resource of snp-trait associations. Nucl Acids Res. 42:D1001–D1006.

Wu M, Lee, S Cai, T Li, Y Boehnke, M, *et al.* 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 89:82–93.

Yang H, Chen C. 2011. Region-based and pathway-based qtl mapping using a p-value combination method. BMC Proc. 5:S43.

Zaykin D, Zhivotovsky L, Westfall P, Weir B. 2002. Truncated product method for combining p-values. Genet Epidemiol. 22:170–185.

Zhao Z, Zhang J, Sha Q, Hao H. 2020. Testing gene-environment interactions for rare and/or common variants in sequencing association studies. PLoS One. 15:e0229217.

Zhou Y, Wang Y. 2015. Detecting association of rare and common variants by adaptive combination of p-values. Genet Res. 6:e20.

Communicating editor: R. Cantor