Incorporating Multiple Sets of eQTL Weights into Gene-by-Environment Interaction Analysis Identifies Novel Susceptibility Loci for Pancreatic Cancer

Tianzhong Yang[1,2], Hongwei Tang[3], Harvey A. Risch[4], Sara H. Olson[5], Gloria Petersen[6], Paige M. Bracci[7], Steven Gallinger[8], Rayjean Hung[8], Rachel E. Neale[9], Ghislaine Scelo[10], Eric J. Duell[11], Robert C. Kurtz[12], Kay-Tee Khaw[13], Gianluca Severi[14,15], Malin Sund[16], Nick Wareham[17], Christopher I Amos[18], Donghui Li[3], Peng Wei[†1]

†All correspondence should be addressed to:
Peng Wei, Ph.D.
Department of Biostatistics, The University of Texas MD Anderson Cancer Center,
1400 Pressler St, Unit 1411, Houston, TX 77030.
Email: pwei2@mdanderson.org

[1] Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2] Divison of Biostatistics, University of Minnesota, Minneapolis, MN, USA
[3] Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[4] Yale University School of Public Health, New Haven, CT, USA
[5] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, US
[6] Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA
[7] Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, CA, USA
[8] Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, University of Toronto, Toronto, Canada
[9] Cancer Aetiology and Prevention Group, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia
[10] International Agency for Research on Cancer, Lyon, France
[11] Unit of Nutrition and Cancer, Cancer Epidemiology Research Program Catalan Institute of Oncology - Bellvitge Biomedical Research Institute (ICO-IDIBELL) Avda. Gran Via 199-203 08908 L'Hospitalet de Llobregat, Barcelona, Spain
[12] Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA
[13] Department of Public Health and Primary Care, University of Cambridge, UK
[14] Gustave Roussy, F-94805, Villejuif, France
[15] CESP, Fac. de médecine - Univ. Paris-Sud, Fac. de médecine - UVSQ, INSERM, Université Paris-Saclay, 94805, Villejuif, France
[16] Department of Surgical and Perioperative Sciences, Umeå University, Sweden
[17] MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Cambridge, UK
[18] Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA

Abstract

It is of great scientific interest to identify interactions between genetic variants and environmental exposures that may modify the risk of complex diseases. However, larger sample sizes are usually required to detect gene-by-environment interaction (GxE) than required to detect genetic main association effects. To boost the statistical power and improve the understanding of the underlying molecular mechanisms, we incorporate functional genomics information, specifically, expression quantitative trait loci (eQTLs), into a data-adaptive GxE test, called aGEw. This test adaptively chooses the best eQTL weights from multiple tissues and provides an extra layer of weighting at the genetic variant level. Extensive simulations show that the aGEw test can control the Type 1 error rate, and the power is resilient to the inclusion of neutral variants and non-informative external weights. We applied the proposed aGEw test to the Pancreatic Cancer Case-Control Consortium (discovery cohort of 3585 cases and 3482 controls) and the PanScan II genome-wide association study data (replication cohort of 2021 cases and 2105 controls) with smoking as the exposure of interest. Two novel putative smoking-related pancreatic cancer susceptibility genes, *TRIP10* and *KDM3A*, were identified. The aGEw test is implemented in an R package aGE.

Keywords— Data-adaptive association testing, eQTL, Gene-by-environment interaction, Multiple functional weights, PrediXCan

1 Introduction

Gene-by-Environment interaction (GxE) analysis not only contributes to finding novel genetic loci but also serves as a starting point to understand the underlying biological mechanisms of complex diseases, which could be helpful for prevention and early detection (Hutter et al., 2013; Fleming, 2017; Wolock et al., 2013; Grishkevich and Yanai, 2013). There is accumulating evidence, from animal models (Ayhan et al., 2016) to large genetic epidemiology consortia (Cornelis et al., 2010), that genetic risk factors interact with environmental risk factors to influence complex diseases. However, it is more challenging to detect GxE effects than the genetic main association effects due to the lack of statistical power. As a rule of thumb, the detection of a GxE association effect requires at least four times as many samples as needed for the detection of a genetic main association effect of a comparable magnitude (Smith and Day, 1984). At the same time, large-scale functional genomics studies, such as the Encyclopedia of DNA Elements project (ENCODE Project Consortium, 2004) and the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013), have generated an abundance of functional annotations for a large proportion of the human genome. We are interested in boosting the statistical power of GxE analysis by incorporating such rich external functional information.

We intend to prioritize genetic variants which regulate gene expression levels, i.e., expression quantitative loci (eQTLs), in the GxE analysis. Limited work has been devoted to leveraging transcriptomic information in GxE tests, although there have been many successes in genetic main association effect tests (Li et al., 2013; Nicolae et al., 2010; He et al., 2013; Gamazon et al., 2015; Gusev et al., 2016). Among them, the PrediXCan method has drawn substantial attention (Gamazon et al., 2015), with the underlying assumption that genetic variants around or within a gene influence disease/traits through regulating its expression levels. PrediXCan and the similar methods have shown improved statistical power in detecting genetic main association effects in practice (Gusev et al., 2016; Andaleon et al., 2019). In fact, PrediXcan can be regarded as a weighted burden test (detailed later), putting more weights on the eQTLs strongly associated with their corresponding gene expression levels (Xu et al., 2017). Tissue-specific eQTL weights have been derived for more than 50 tissues and distributed to the public via PredictDB data repository, which could be adapted to GxE testing under the assumption that eQTLs are more likely to interact with an environmental risk factor. However, it remains a challenge as to how to choose the most relevant tissue or combine multiple informative tissues for a given phenotype. We thus propose to develop a GxE test that incorporates eQTLs information from multiple tissues.

There are several challenges to develop a gene-based eQTL-weighted GxE test. First, high linkage disequilibrium (LD) among single nucleotide polymorphisms (SNPs) could be problematic for a gene-based score test. Unlike the likelihood ratio and Wald tests, the score test only requires estimating the genetic main association effects under the null hypothesis of no GxE (Yang et al., 2019). However, the maximum likelihood estimation of the null models was not stable for many genes in our preliminary analysis of real data, which hinders its use in the genome-wide GxE scan. Second, weight misspecification is almost inevitable due to different reasons. Complex diseases are often caused by dysfunction of multiple tissues or cell types (Hao et al., 2018), and very likely, the tissue(s) where a certain gene exerts its function is unknown. Even if we know the most relevant tissue(s) *a priori*, the corresponding weights may not be available or of sub-optimal quality, for example, the weights trained based on a tissue of a small sample size in GTEx. In addition, the eQTLs identified in PrediXcan could be false positives due to systematic variation in the RNA sequencing experiments (Albert and Kruglyak, 2015), or caveats in the weight training processes. Besides the lack of optimal-quality weights, weight misspecification could be caused by the unsatisfied imposed model assumptions. For example, the majority of eQTLs in a gene may not interact with the environmental variable, i.e., neutral variants in terms of GxE. Therefore, we propose a robust and powerful eQTL-weighted data-adaptive GxE (called aGEw) test, as an extension of the adaptive sum of powered test (aSPU) (Pan et al., 2014) and our earlier work on GxE (Yang et al., 2019). Instead of a single tissue, we use weights from all available tissues in a tissue-agnostic manner (Barbeira et al., 2019; Wainberg et al., 2019). With informative eQTL weights, the aGEw test is able to gain substantial statistical power and it controls the Type 1 error rates as shown in the simulation study. In addition, we applied the proposed aGEw test to investigate gene-by-smoking interactions in two of the largest pancreatic cancer genome-wide association study (GWAS) datasets, the Pancreatic Cancer Case-Control Consortium (PanC4) GWAS for discovery and the PanScan II GWAS for replication. Two putative smoking-related pancreatic cancer susceptibility genes, *TRIP10* and *KDM3A*, were identified in the GxE scan based on tissue-specific eQTL weights.

2 Methods

2.1 Notations

The notations are defined as follows. Let y_i be the trait/phenotype of an individual i, where $i = 1, 2, \ldots, n$ and $\mathbf{Y} = [y_1, \ldots, y_n]$. $\mathbf{G}_{\cdot,j}$ is the genotype vector for SNP j: $\mathbf{G}_{\cdot,j} = [G_{1j}, G_{2j}, \ldots, G_{nj}]^T$, where G_{ij} is the number of minor alleles of the j^{th} SNP in a gene region for the i^{th} individual and $j = 1, \ldots, q$. E_i is the environmental variable for individual i: $\mathbf{E} = [E_1, E_2, \ldots, E_n]$. V_i is the gene expression for individual i corresponding to the genotype: $\mathbf{V} = [V_1, \ldots, V_n]$.

$\mathbf{S}_{\cdot,j} = [G_{1j}E_1, G_{2j}E_2, \ldots, G_{nj}E_n]^T$ is a vector of the GxE for the j^{th} SNP. In addition, we denote $\mathbf{G}$ and $\mathbf{S}$ as $n \times q$ matrices containing n subjects and q SNPs: $\mathbf{G} = [\mathbf{G}_{\cdot,1}, \mathbf{G}_{\cdot,2}, \ldots, \mathbf{G}_{\cdot,q}]$ and $\mathbf{S} = [\mathbf{S}_{\cdot,1}, \mathbf{S}_{\cdot,2}, \ldots, \mathbf{S}_{\cdot,q}]$. $\mathbf{X}$ is a $n \times p$ covariate matrix including the intercept term. Suppose we have M sets of weights for each SNP in the region corresponding to M sets of tissues, $\mathbf{W}$ is the $M \times q$ weight matrix with the j^{th} row and m^{th} column as $w_j^{(m)}$, where $w_j^{(m)}$ is the weight for the j^{th} SNP and the m^{th} tissue. Estimated parameter values are denoted with a hat. Weights are scaled to satisfy: $\sum_{j=1}^{q} |w_j^{(m)}| = 1$ for $m = 1, 2, \ldots, M$. Note that vectors and matrices are marked in boldface, whereas scalars are not.

2.2 Review of Existing methods

We review PrediXcan and demonstrate that a direct extension of PrediXcan to GxE is problematic, followed by the review of the aSPU and a data-adaptive GxE (aGE) test as the basis of our proposed aGEw test (Pan et al., 2014; Yang et al., 2019).

2.2.1 PrediXcan

PrediXcan tests the association between genetically regulated gene expression and the phenotype of interest. It hypothesizes that a cis-eQTL with a larger effect on gene expression levels has a stronger genetic main effect. The PrediXcan method is composed of the following steps. First, it estimates weights from an external reference dataset. It uses an additive model for gene expression traits: $V_i = b_0 + \sum_{j=1}^{q} w_j G_{ij} + \varepsilon$, where b_0 is the intercept, w_j is the effect size of SNP j within or near a gene, and ε is the contribution from other factors that influence the expression trait and are independent of the genetic component. $\hat{w}_j$ can be estimated using a penalized regression model and is available via the PredictDB data repository.

In the second step, PrediXcan estimates the genetically regulated gene expression using the genotype information in a GWAS dataset independent of the reference dataset: $\hat{V}_i = \sum_{j=1}^{q} \hat{w}_j G_{ij}$, where $\hat{V}_i$ is the imputed gene expression for individual i. Since the ultimate goal is to identify trait-associated loci, the following model is fitted in the third step:

$$Y_i = \mathbf{X}_i \boldsymbol{b}_{\boldsymbol{X}} + b\hat{V}_i + \nu, \tag{1}$$

where $\boldsymbol{b}_{\boldsymbol{X}}$ is the coefficients for covariates, and b is the coefficient of interest. To test the null hypothesis that there is no association between the genetically regulated gene expression and outcome, that is, $H_0 : b = 0$, a Wald test is usually performed. Equation (1) can be rewritten as follows, from which we can easily recognize its correspondence to a weighted burden test (Xu et al., 2017):

$$Y_i = \mathbf{X}_i \boldsymbol{b}_{\boldsymbol{X}} + b\sum_{j=1}^{q} \hat{w}_j G_{ij} + \nu.$$

Note that $\hat{w}_j$ varies across tissues or reference datasets for each gene and is treated as a vector of given values herein.

2.2.2 Burden test for GxE

A direct extension of PrediXcan to GxE analysis is to test the interaction between an environmental variable and the genetically regulated gene expression. It can be modeled as follows:

$$\begin{aligned} h(\mu_i) &= \mathbf{X}_i \boldsymbol{b}_{\boldsymbol{X}} + \beta_e E_i + b\hat{V}_i + a\hat{V}_i E_i, \\ &= \mathbf{X}_i \boldsymbol{b}_{\boldsymbol{X}} + \beta_e E_i + b\sum_{j}^{q} \hat{w}_j G_{ij} + a\sum_{j}^{q} \hat{w}_j S_{ij}, \end{aligned} \tag{2}$$

where $h(\cdot)$ is the canonical link function in the generalized linear model framework and $\mu_i = E(y_i)$. It was pointed out that one degree-of-freedom burden test with equal weights ($H_0 : a = 0$ vs $H_1 : a \neq 0$) may have inflated Type I error rate for continuous and binary outcomes because the null model is misspecified (Section 3 in Lin et al. (2016)). It could be even worse for the weighted burden test because w_j is prone to misspecification error.

2.2.3 aSPU and aGE tests

The aSPU test was originally developed to test genetic main effect for a set of rare variants (Pan et al., 2014), while the aGE test extended it to test GxE (Yang et al., 2019). The aSPU and aGE tests choose the most powerful test among a family of sum of powered score (SPU) tests, including burden and variance-component tests, such that it can maintain high power under a wide range of association patterns. The SPU tests can be regarded as score tests with a variety of internal or data-driven weighting. Consider the following model for GxE:

$$h(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}_{\boldsymbol{X}} + \beta_e E_i + \sum_{j=1}^{q} \beta_j G_{ij} + \sum_{j=1}^{q} \alpha_j S_{ij}, \tag{3}$$

where $\boldsymbol{\beta}_{\boldsymbol{X}}, \beta_e, \beta_j$, and α_j are the coefficients of $\mathbf{X}$, $\mathbf{E}$, $\mathbf{G}_{\cdot,j}$, and $\mathbf{S}_{\cdot,j}$, respectively. The null hypothesis for the interaction test is $H_0 : \alpha_1 = \ldots = \alpha_q = 0$. The corresponding score function for the j^{th} variant is $U_j = \mathbf{S}_{\cdot,j}^T(\mathbf{y} - \boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is estimated in the generalized linear mixed model (GLMM) framework under the null hypothesis. $\boldsymbol{\beta}_{\boldsymbol{X}}$ and β_e are treated as fixed effects and $\beta_1, \ldots, \beta_q$ are treated as random effects, i.e., $\beta_1, \ldots, \beta_q$ are independent and identically distributed as N(0, τ). Such modeling provides a more stable estimation for main association effects of rare variants or common variants in high LD under the null hypothesis (Yang et al., 2019).

The SPU test with a positive integer power $\gamma \geq 1$ (SPU(γ)) has the test statistic:

$$T_{SPU(\gamma)} = \sum_{j=1}^{q} U_j^{\gamma}.$$

When $\gamma = 1$, the SPU(1) test sums the individual score function component U_j of all the genetic variants in a set of interest, equivalent to the burden test; when $\gamma = 2$, the SPU(2) test uses U_j as the weight and is equivalent to a variance-component test (Chen et al., 2014; Lin et al., 2016). Generally, the test statistics can be written as $T_{SPU(\gamma)} = \sum_{j=1}^{q}(U_j^{\gamma-1})U_j$, where $U_j^{\gamma-1}$ is considered as the internal weights for the j^{th} SNP (Roeder and Wasserman, 2009) and γ is the tuning parameter that controls the amount of weighting. For example, when γ keeps increasing, the SPU(γ) test puts higher weights ($U_j^{\gamma-1}$) on the j^{th} genetic variant with larger U_j, i.e., stronger association strength. When $\gamma = \infty$, the test statistic takes the largest absolute value of the elements of the score vector, and thus, only the variant with the most extreme test statistic in the set is considered (Pan et al., 2014). As a result, aSPU with larger γ could have high power when there are more neutral variants in the SNP set. Since the SPU(γ) with the highest power depends on the unknown underlying association pattern, it is reasonable to choose γ adaptively based on data. To this end, the aGE test selects the most powerful test by taking the minimum of p-values P_{SPU} with different γ's: $T_{aGE} = min_{\gamma \in \Gamma} P_{SPU(\gamma)}$, where $P_{SPU(\gamma)}$ is the corresponding p-value of $T_{SPU(\gamma)}$. Since the minimum p-value is no longer a genuine p-value, a computationally efficient Monte Carlo simulation method is used to calculate p-values in the aGE test (Yang et al., 2019).

2.3 New Method: adaptive weighted GxE test (aGEw)

Directly applying aGE test by replacing the unweighted GxE term S_{ij} with the weighted term $w_j S_{ij}$ in model (3) may be viable for incorporating a single set of weights. However, the new aGEw method can incorporate external eQTL weights from multiple tissues and provide flexible weights on each tissue. The full model corresponding to the m^{th} tissue ($m = 1, 2, \ldots, M$) is as follows:

$$h(\mu_i) = \mathbf{X}_i \boldsymbol{\beta}_{\boldsymbol{X}} + \beta_e E_i + \sum_{j=1}^{q} \beta_j G_{ij} + \sum_{j=1}^{q} \alpha_j^{(m)} w_j^{(m)} S_{ij}. \tag{4}$$

The null hypothesis is $a_j^{(m)} = 0$ for $j = 1, \ldots, q$ and $m = 1, \ldots, M$. The score function for the j^{th} SNP and m^{th} tissue is $U_{Sj}^{(m)} = w_j^{(m)} \mathbf{S}_{\cdot,j}^T (\mathbf{y} - \boldsymbol{\mu})$. Similar to aGE, $\boldsymbol{\mu}$ is estimated under H_0 using the GLMM due to the SNPs in high LD. We then define the following test statistic for the GEw(γ_1, γ_2) test:

$$T_S(\gamma_1; \gamma_2) = \sum_{m=1}^{M} \big[\big(\sum_{j=1}^{q} (U_{Sj}^{(m)})^{\gamma_1} \big)^{1/\gamma_1} \big]^{\gamma_2}, \tag{5}$$

where both $\gamma_1 \in \Gamma_1$ and $\gamma_2 \in \Gamma_2$ are positive integers for controlling each SNP and each set of tissue-specific weights' contribution to the overall test statistic, respectively. The inner sum of the weighted score function with power γ_1 can be regarded as the aGE test statistic for weighted genotype in each tissue, and they are normalized to the power of $1/\gamma_1$ before taking a second power γ_2. Based on our previous investigations (Pan et al., 2014; Kwak and Pan, 2016; Ma and Wei, 2019) and simulation study to be detailed later, we would recommend using $\Gamma_1 = \{1, 2, 3, 4, 5, 6\}$ and $\Gamma_2 = \{1, 2, 4\}$, which suffice to strike a balance between statistical power and computational complexity. When $\gamma_1 = 1$, the GEw$(\gamma_1 = 1, \cdot)$ test is equivalent to the adaptive weighted burden test, assuming that the interaction effects of SNPs are proportional to the external weights, i.e., the full model for the m^{th} tissue is reduced to

$$h(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}_{\boldsymbol{X}} + \beta_e E_i + \sum_j^q \beta_j G_{ij} + \alpha^{(m)} \sum_j^q w_j^{(m)} S_{ij}. \tag{6}$$

Note that formulation (6) avoids the potential problem of an inflated Type 1 error rate in model (2). Even values of γ_1 provide aGEw the robustness to the varying signs of the functional weights and the score function components.

As for γ_2, it can provide adaptive weighting on functional weights from different tissues. When $\gamma_2 = 1$, the GEw$(\cdot, \gamma_2 = 1)$ test statistic gives equal preference to the weights from different tissues. The aGE test is a special case of aGEw test with $\gamma_2 = 1, M = 1$ and equal weights. A larger γ_2 favors scenarios when fewer tissues are relevant to the phenotype of interest. The two-layer test statistic T_{aGEw} combines T_S of different γ_1 and γ_2 using the minimum p-value as follows: $T_{aGEw} = min_{\gamma_1 \in \Gamma_1, \gamma_2 \in \Gamma_2} P_{S(\gamma_1;\gamma_2)}$, where $P_{S(\gamma_1;\gamma_2)}$ is the p-value of the GEw(γ_1, γ_2) test. Since the underlying association pattern and the best set of relevant tissues for each gene are usually unknown, we conduct a grid search over a set of possible values of γ_1 and γ_2. In the evaluation of single sets of weights, no tuning on the tissue level is needed. Therefore, we used $\Gamma_1 = \{1, 2, 3, 4, 5, 6\}$ and $\Gamma_2 = \{1\}$ for such cases in the real-data application and simulation study, and $\Gamma_1 = \{1, 2, 3, 4, 5, 6\}$ and $\Gamma_2 = \{1, 2, 4\}$ otherwise.

For p-value calculation, we utilize the asymptotic distribution of the score functions with a closed-form formula for the variance-covariance matrix for binary traits (Supplementary Materials Section 1) to perform a single-layer Monte-Carlo simulation. The detailed steps are presented in Supplementary Materials Section 2. Note that our method can be used for continuous traits with a straightforward modification of the variance-covariance matrix of the score functions (Supplementary Materials Section 1) and is implemented in the R package “aGE”.

2.4 Simulation setup

To evaluate the Type I error rates and power for our proposed tests, we simulated case-control studies with varying sample size scenarios: 500 cases and 500 controls, 1000 cases and 1000 controls, 1000 cases and 2000 controls, and 3500 cases and 3500 controls with eQTL weights of different qualities. Genotypes were simulated as in Wang and Elston with six independent LD blocks (Wang and Elston, 2007). We denote SNPs with genetic main effects as causal SNPs (csSNPs). Each of the first five blocks consisted of one csSNP and the rest SNPs were in LD with the csSNP, while the last block consisted of neutral SNPs only. There were a total of $2l$ non-causal SNPs (non-csSNPs), including l neutral SNPs and l SNPs in LD with the csSNPs, distributed evenly across the 5 blocks. l varied in different simulation scenarios. In each block, the SNPs had an autoregressive (AR)-1 correlation structure with $\rho = 0.8$. The minor allele frequency of SNPs ranged from 5% to 45%. The environmental variable was assumed to be independent of the genotypes: $E \sim N(0, 1)$. We also simulated two covariates: age $\sim N(55, 5.7^2)$ and gender $\sim$ Bernoulli(0.5).

We investigate a single set of weights ($M = 1$) and multiple sets of correlated weights ($M > 1$) for Type I error rates and power assessment. We denote the weights for csSNPs, SNPs in LD with csSNPs, and neutral SNPs as $\mathbf{W}_{cs}$, $\mathbf{W}_{tag}$, and $\mathbf{W}_{noise}$, respectively, and $\mathbf{W}_1 = [\mathbf{W}_{cs}, \mathbf{W}_{tag}, \mathbf{W}_{noise}]$. To take account into the correlation among the weights from multiple tissues, we simulated $\mathbf{W}^{(m)}$ by setting $\mathbf{W}^{(m)} = \mathbf{W}_1 + u_1 + \epsilon_m$, where $u_1 \sim N(0, 0.1^2)$, $\epsilon_m \sim N(0, 0.05^2)$, and $u_1 \perp \epsilon_m$ for $m = 1, \ldots, M$. We then standardized $\mathbf{W}^{(m)}$ to $\mathbf{w}^{(m)}$ such that $\sum_{j=1}^{q} |w_j^{(m)}| = 1$ for tissue m.

The Type I error rates were evaluated based on 10,000 replications when there was no GxE.

The phenotype was simulated from the following model:

$$\text{logit}(\mu_i) = t_0 + 0.5 \times \text{sex}_i + 0.05 \times \text{age}_i + 0.25 \times E + \sum_{j=1}^{5} \beta_j \text{csSNP}_{ij},$$

where $t_0 = \log(\frac{p_0}{1-p_0})$, baseline disease prevalence $p_0 = 0.05$, and $\boldsymbol{\beta}^{\boldsymbol{T}} = [0.43, -0.41, 0.51, 0.25, -1.76]$. $\exp(\boldsymbol{\beta})$ was a random draw from a uniform distribution on the odds ratio interval $[0.5, 2]$ and fixed across simulations. Cases and controls were sampled until the number of each reached the desired sample size. We considered the Type 1 error rates of the aGEw tests with a single set of weights (scenarios T.1 - T.2) and multiple sets of weights (scenarios T.3 - T.4), and we varied the number of non-csSNPs, as specified in Table 1.

Power for the aGEw test was assessed based on 1000 replications. The simulation model for power evaluation was:

$$\text{logit}(\mu_i) = t_0 + 0.5 \times \text{sex}_i + 0.05 \times \text{age}_i + 0.25 \times E_i + \sum_{j=1}^{5} \beta_j \text{csSNP}_{ij} + \sum_{j=1}^{5} \alpha_j \text{csSNP}_{ij} \times E_i,$$

where $\boldsymbol{\beta}^{\boldsymbol{T}} = [0.43, -0.41, 0.51, 0.25, -1.76]$ and $\boldsymbol{\alpha}^{\boldsymbol{T}} = [-0.20, 1.18, -0.43, -1.16, 0.34]$. $\exp(\boldsymbol{\alpha})$ was another random draw from the uniform distribution on the odds ratio interval $[0.5, 2]$ and fixed across simulations. We further added l SNPs in LD with the csSNPs and l neutral SNPs, where l was set at 0, 50, 100, and 200, i.e., the total number of non-csSNPs was 0, 100, 200, and 400, respectively. We first evaluated the power with a single set of weights across five scenarios: S.1) accurate weights ($\mathbf{W}_{cs} = \boldsymbol{\alpha}, \mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{0}^T$), S.2) informative weights with attenuation bias ($\mathbf{W}_{cs} = \boldsymbol{\alpha} + \epsilon_\alpha, \mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{0}^T$), S.3) weights with correct signs but wrong magnitudes ($\mathbf{W}_1 = \text{sign}(\boldsymbol{\alpha})$), S.4) 0-1 weights (weights of all csSNPs were 1 and 0 otherwise), and S.5) incorrect weights (neutral SNPs had larger weights than csSNPs and SNPs in LD with csSNPs).

We then evaluated the power with three types of correlated weights M.1)-M.3), corresponding to informative weights, incorrect weights, and a combination with informative and incorrect weights, respectively. The specific weights in S.1) - S.5) and M.1) - M.3) are summarized in Table 1. The power performance was compared to the benchmark S.0), i.e., the unweighted aGE test.

2.5 Application to Gene-by-Smoking Interaction Analysis of Pancreatic Cancer

To understand the biological mechanism of pancreatic cancer and identify new genetic loci, we applied our proposed aGEw tests to the PanC4 GWAS with a sample size of 7076 (Childs et al., 2015), and replicated the findings in PanScan II GWAS with a sample size of 4126 (Tang et al., 2014; Amundadottir et al., 2009; Petersen et al., 2010). Descriptive summary statistics are presented in Table S1, and a detailed description of the studies is provided elsewhere (Klein et al., 2018). The disease outcome of interest was having primary adenocarcinoma of the pancreas or not, and the environmental exposure of interest was ever vs never smoking. Covariates included sex, age, study sites, top five principal components capturing the population substructure, diabetes status, and categorized body mass index (≤ 25, $25 - 29.9$, and ≥ 30). Imputation of the genotypes was performed for PanScan II based on a 1000 Genomes reference panel (The 1000 Genomes Project Consortium, 2010), and imputation for PanC4 was performed based on the Haplotype Reference Consortium panel (McCarthy et al., 2016). Weights of all available tissues derived from the GTEx project were downloaded from PredictDB (GTEx V6): a total of 7789 genes whose expression level could be predicted by SNPs using the elastic net method with a cross-validated $R^2 > 0.01$ were included. Notably, only 1058 out of these genes had weights available in the pancreas. To control family-wise error rates, the genome-wide significance threshold in the discovery cohort (PanC4) was set as 0.05/7789=6.42E-06. The top five genes were carried out to be replicated in PanScan II with the replication significance threshold 0.05/5=0.01. We performed the aGEw tests with continuous weights, weights of the pancreas, and 0-1 weights; for the 0-1 weights, the weights for a SNP was 0 if the SNP had zero weights across all tissues, and one otherwise. For comparison, the unweighted aGE test was conducted on all SNPs included in building the gene expression prediction models in PredictDB. As most of the genes were not expected to interact with smoking, we used a step-up Monte Carlo simulation procedure: we started with 1000 Monte Carlo simulations for all genes and re-ran the tests with 1E+06 times for the genes with the initial p-values <5E-03. With the recent release of GTEx V8, a new set of elastic net models is available during the revision of this manuscript, thus were used for secondary comparison purpose. A total of 20,351 genes were included based on the recommended criteria: nested cross-validated $R^2 > 0.01$ and correlation test p-value < 0.05 (Gamazon et al., 2015; Barbeira et al., 2018, 2019). Among them, 3591 genes had weights available in the pancreas.

2.6 Software and Data Availability

The aGEw test has been implemented in an R package called "aGE" (version $>= 0.2$), available at https://github.com/ytzhong/projects. The PanC4 and PanScan II GWAS data are accessible from NCBI dbGaP with accession numbers phs000648.v1.p1 and phs000206.v5.p3. The eQTL weights used in the data application are based on GTEx V6 prediction models at the time of the development of the aGEw test and the recently available GTEx V8 models, both accessible at http://predictdb.org.

3 Results

3.1 Simulation studies

Table 2 shows the empirical Type I error rates with balanced and unbalanced case-control samples of different sample sizes. GEw($\gamma_1 = 1, \gamma_2 = 1$) is equivalent to the burden test extended to GxE. In each scenario, the GEw($\gamma_1 = 1, \gamma_2 = 1$) and aGEw tests showed satisfactory control of the Type 1 error rates at the nominal level.

Fig.1A shows the power curves of the aGEw tests with different single sets of weights based on 1000 cases and 1000 controls. As expected, the power of the aGE test decreased as the number of non-csSNPs increased. Informative weights increased the statistical power and made the test more robust to the inclusion of non-csSNPs. For example, the power of using informative weights in S.1) was 0.81 with 200 SNPs in LD with csSNPs and 200 neutral SNPs, while the power of aGE test was 0.22. The overall power was clearly higher when the weights were more informative, i.e., the power in S.1) $>$ S.2) $>$ S.3) $>$ S.4) $>$ benchmark. On the other hand, incorrect weights (weights of neutral SNPs higher than those of csSNPs and SNPs in LD with csSNPs) in S.5) led to a large amount of power loss, suggesting that selection on weights is important to maintain high power of the weighted tests. We were particularly interested in the 0-1 weights (S.4) because it did not assume that SNPs with larger effects on gene expression levels had larger GxE effects. We found that there was still a noteworthy power gain as compared with the benchmark aGE test. In Fig.1B, the aGEw test with multiple sets of correlated informative weights (M.1) had the highest power and little power loss in the presence of increasing number of neutral SNPs, while the test with incorrect magnitudes (M.2) had the worst power, slightly worse than the unweighted aGEw test. Remarkably, the test with informative weights contaminated by incorrect weights as in M.3) maintained high power, demonstrating the effectiveness and robustness of our proposed test. The empirical power evaluation for other sample sizes is available in Table S2 and Table S3, where we observed a clear increase of power as the sample size increased. When the sample size approached 7000, the empirical power was close to 1 with informative or partially informative weights.

3.2 Gene-by-smoking Interaction Analysis for Pancreatic Cancer

Fig S1 presents the quantile-quantile plots for the tests based on the PanC4 GWAS data with no indication of p-value inflation. None of the genes reached genome-wide significance, suggesting that identifying gene-by-smoking interaction is a difficult task even with our uniquely large pancreatic cancer GWAS datasets. As an alternative, we investigated the top five significant genes in the discovery cohort. The top five genes ordered by the p-values of aGEw tests with continuous weights and aGEw tests with 0-1 weights are presented in Table 3 and 4, respectively. Among the top five genes identified from the interaction tests (Table 4), Lysine-specific demethylase 3A (*KDM3A*) had a p-value of less than 0.01 in the replication dataset. The only set of weights available for *KDM3A* was derived from the thyroid, a tissue found to be relevant with pancreatic cancer (Sarosiek et al., 2016; Gullo et al., 1991). This gene would be missed by using the aGE test (p-value $>$ 0.05). Additionally, the two sets of weights used for Thyroid hormone receptor interactor 10 (*TRIP10*) were obtained from tibial nerve and skin on the lower leg that was exposed to the sun. Tibial nerve seemed to be a more relevant tissue because analysis based on the tissue showed more significant p-values in both the discovery and replication cohorts (p-values $<$ 0.01). In fact, the GEw test with the smallest p-value for this gene had $\gamma_2 = 4$, suggesting that one tissue was much more informative than the other. This gene would also be missed by the unweighted test whose p-values in both datasets were less than 0.05, but not small enough to reach the significance thresholds. None of these top genes had weights from the pancreas, which was expected to be one of the most relevant tissues for pancreatic cancer, likely due to the small sample size available in the GTEx study (V6). In addition, none of the top five genes identified from the unweighted aGE test were successfully replicated.

It is interesting that the aGEw test with continuous weights for *KDM3A* was less significant than the aGEw test with 0-1 weights and that the aGEw test with continuous weights for *TRIP10* was more significant than the aGEw test with 0-1 weights in the discovery dataset. The same pattern was also observed in the replication dataset. It suggests that the magnitude of the weights could be informative for some genes, but not for others. This could be because some of the underlying model assumptions were violated or the available weights were not very informative. A detailed examination of the p-value distribution of the single-variant GxE analyses in the most relevant tissue showed that eQTLs had much smaller overall p-values than non-eQTLs for *KDM3A*, but not for *TRIP10* (Fig S2). The continuous weights of *TRIP10* showed a more clear increasing trend on the -log10 scale of the p-values than those of *KDM3A* (Fig S3).

Regarding the computation speed, it took less than 36 hours to perform the four tests for all genes on a single-core processor.

3.3 Comparison of GTEx V6 and V8 models in the aGEw tests

The results of the aGEw tests based on weights trained using GTEx V6 and V8 data were quite different not only in the top signals (Table S5 and Table S6), but also systematically (Figure S4). The discrepancy between the p-values was due to the fact that the V6 and V8 models identified different and non-overlapping sets of eQTLs. With a larger training size in GTEx V8 data, there were more tissues available and genes satisfying the inclusion criteria: the average (median) number of tissues in V8 was 9 (4), while the average (median) number of tissues for V6 was 7 (3). Additionally, V8 models were sparser than the V6 models: the average (median) number of SNPs per gene for V8 model was 28 (25), and the number was 78 (54) in V6. Among the 7346 genes available in both V6 and V8, 83% of the genes had more SNPs selected in V6, although both were trained by the elastic net regression method. More importantly, on average, only 5% of the eQTLs in V8 were present in V6 (1st quantile = 2% and 3rd quantile = 9%; see Table S4). For the two pancreatic cancer susceptibility genes identified based on the V6 models, out of 68 SNPs in *TRIP10*, four SNPs were included in the V8 models and out of 14 SNPs in *KDM3A*, only one SNP was included in V8. Therefore, the corresponding p-values either did not reach or only reached the marginal significance level: in PanC4 study, the p-value of aGEw (0-1) for *KDM3A* was 3.80E-02, and the p-value of aGEw was 5.09E-02 for *TRIP10*; in PanScan II study, the p-value of aGEw (0-1) for *KDM3A* was 2.70E-02, and the p-value of aGEw was 0.33 for *TRIP10*. Furthermore, many of the top genes in V8 were not available in V6 and none of the genes was successfully replicated (Table S5 and Table S6). Although more genes that were identified to have genetic main effects in previous GWAS (Low et al., 2010; Petersen et al., 2010; Wu et al., 2012; Wolpin et al., 2014; Childs et al., 2015; Klein et al., 2018) had more weights available in V8 than V6, none of them was statistically significant in the interaction analysis (Table S7).

4 Discussion

A subset of a population who carry certain genetic variants may be in a higher risk of developing a disease if they are exposed to certain modifiable environmental risk factors, such as smoking, drinking, and physical activities. However, identifying the subpopulation is challenging due to the lack of statistical power, and many existing studies do not have adequate sample sizes to address GxE (Hunter, 2005; Ritz et al., 2017). Therefore, it is critical to improve power for GxE testing with a given sample size. In this paper, we hypothesize that eQTLs are more likely to be involved in GxE, and thus incorporating transcriptomic information could be helpful. To use transcriptomic data from multiple tissues, we have proposed a powerful and doubly-adaptive gene-based aGEw test. We used the mixed model to handle the high LD among SNPs, achieving estimation stability for the genetic main effects under the null model of no GxE. Our proposed aGEw test has two layers of adaptive weighting: one adaptive weighting on SNPs and the other on external weights from different tissues. We showed via extensive simulations that our test could control the Type I error rates. As expected, we found that informative weights could increase power, but incorrect weights could diminish power. Remarkably, if both correct and incorrect weights were included, the power loss of our proposed test was negligible. On the other hand, we showed that the aGEw test was sensitive to the identify of eQTLs selected for a gene in the real data application. The eQTLs identified for each gene often had a small intersection between the two versions of weights: many eQTLs included in V6 were no longer under consideration in V8, leading to the discrepancy in p-values. If the true eQTLs were not included, aGEw was not able to capture the signal.

Limited work has been devoted to incorporating functional annotation to GxE analysis, especially for pancreatic cancer. Although cigarette smoking is a well-established risk factor for pancreatic cancer, there have been very few reported findings in gene-by-smoking interactions (Tang et al., 2014). In the analysis of pancreatic cancer, we borrowed transcriptomic information from all available tissues and used two types of weights, i.e., the continuous weights directly obtained from PredictDB and 0-1 weights. While none of the genes reached genome-wide significance, with an alternative approach examining the top significant genes, the aGEw tests identified two putative genes *KDM3A* and *TRIP10* that may interact with smoking; in contract, the unweighted aGE test failed to identify any significant gene. *KDM3A*, which belongs to the Lysine demethylases family, has been shown to play a role in carcinogenesis in multiple cancers and to be an optimal pharmacological target for pancreatic cancer (Lomberk et al., 2016). A recent mice study showed knockdown of *KDM3A* in pancreatic cancer cell line reduced the invasive and sphere-forming activities in culture and formation of orthotopic tumors (Dandawate et al., 2019). *TRIP10*, whose overexpression may enhance pancreatic cancer cell migration, could be involved in the carcinogenesis of pancreatic cancer (Roignot et al., 2010). Smoking could increase the methylation levels in the upstream of *TRIP10* and thus upregulate its expression (Walters et al., 2014; Wu et al., 2019). To our knowledge, these two genes have not been reported to interact with cigarette smoking. For further investigation, a possible way is to increase statistical power by combining studies through meta-analysis, in which the study heterogeneity needs to be carefully accommodated (Yang et al., 2020).

Another insight gained from the real-data application is the usefulness of 0-1 weights, when quantitative weights were not available or sometimes not as informative. It enjoys a more relaxed and probably more realistic assumption for GxE analysis and helps to identify a novel smoking-related pancreatic cancer susceptibility gene. Our simulation study confirmed that correctly using 0-1 weights improved power as compared to the unweighted test. Such a finding has also been reported in other functional studies.(Li et al., 2013; Su et al., 2017; Wu and Pan, 2018; Yang et al., 2020).

As demonstrated here, informative weights can greatly increase the power of GxE analysis; however, it still remains a question as to what would be the best possible functional weight for GxE. In this paper, we mainly focused on using weights derived from eQTL datasets, but our method is not limited to a particular source or training models. For example, the aGEw test can use weights derived from different kinds of QTLs (Idaghdour and Awadalla, 2013), such as protein QTLs or methylation QTLs. In addition, some long-range enhancer regions may also be a potential target for GxE. Different sources of weights can be incorporated to utilize various external biological knowledge, and we expect further improvement in power with more relevant weights.

Acknowledgments

We thank the two anonymous reviewers for their constructive comments. This research was supported by the National Institutes of Health (NIH) grant R01CA169122; P.W. was supported by NIH grants R01HL116720 and R21HL126032. S.H.O. was supported by NIH grant P30CA008748. R.E.N. and the Queensland Pancreatic Cancer Study were funded by the Australian National Health and Medical Research Council. The authors thank Ms. Jessica Swann and the National Institute of Statistical Sciences writing workshop for editorial assistance and suggestions. The authors acknowledge the Texas Advanced Computing Center at The University of Texas at Austin for providing computing resources. The authors alone are responsible for the views expressed in this article and they do not necessarily represent the views, decisions or policies of the institutions with which they are affiliated. The authors declare that there is no conflict of interest.

References

Albert, F. W. and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics* **16,** 197.

Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R. Z., Fuchs, C. S., Petersen, G. M., Arslan, A. A., Bueno-de Mesquita, H. B., Gross, M., Helzlsouer, K., Jacobs, E. J., et al. (2009). Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nature Genetics* **41,** 986–990.

Andaleon, A., Mogil, L. S., and Wheeler, H. E. (2019). Genetically regulated gene expression underlies lipid traits in hispanic cohorts. *PloS One* **14,** e0220827.

Ayhan, Y., McFarland, R., and Pletnikov, M. V. (2016). Animal models of gene–environment interaction in schizophrenia: A dimensional perspective. *Progress in Neurobiology* **136,** 1–27.

Barbeira, A. N., Bonazzola, R., Gamazon, E. R., Liang, Y., Park, Y., Kim-Hellmuth, S., Wang, G., Jiang, Z., Zhou, D., Hormozdiari, F., et al. (2019). Widespread dose-dependent effects of RNA expression and splicing on complex diseases and traits. *BioRxiv* page 814350.

Barbeira, A. N., Dickinson, S. P., Bonazzola, R., Zheng, J., Wheeler, H. E., Torres, J. M., Torstenson, E. S., Shah, K. P., Garcia, T., Edwards, T. L., et al. (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9,** 1–20.

Barbeira, A. N., Pividori, M. D., Zheng, J., Wheeler, H. E., Nicolae, D. L., and Im, H. K. (2019). Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genetics* **15,** e1007889.

Chen, H., Meigs, J. B., and Dupuis, J. (2014). Incorporating gene-environment interaction in testing for association with rare genetic variants. *Human Heredity* **78,** 81–90.

Childs, E. J., Mocci, E., Campa, D., Bracci, P. M., Gallinger, S., Goggins, M., Li, D., Neale, R. E., Olson, S. H., Scelo, G., et al. (2015). Common variation at 2p13. 3, 3q29, 7p13 and 17q25. 1 associated with susceptibility to pancreatic cancer. *Nature Genetics* **47,** 911.

Cornelis, M. C., Agrawal, A., Cole, J. W., Hansel, N. N., Barnes, K. C., Beaty, T. H., Bennett, S. N., Bierut, L. J., Boerwinkle, E., et al. (2010). The gene, environment association studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genetic Epidemiology* **34,** 364–372.

Dandawate, P., Ghosh, C., Palaniyandi, K., Paul, S., Rawal, S., Pradhan, R., Sayed, A. A. A., Choudhury, S., Standing, D., Subramaniam, D., et al. (2019). The histone demethylase KDM3A, increased in human pancreatic tumors, regulates expression of DCLK1 and promotes tumorigenesis in mice. *Gastroenterology* **157,** 1646–1659.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia of DNA elements) project. *Science* **306,** 636–640.

Fleming, S. M. (2017). Mechanisms of gene-environment interactions in parkinsons disease. *Current Environmental Health Reports* **4,** 192–199.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Nicolae, D. L., Cox, N. J., et al. (2015). A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics* **47,** 1091.

Grishkevich, V. and Yanai, I. (2013). The genomic determinants of genotype× environment interactions in gene expression. *Trends in Genetics* **29,** 479–487.

Gullo, L., Pezzilli, R., Bellanova, B., D'Ambrosi, A., Alvisi, V., and Barbara, L. (1991). Influence of the thyroid on exocrine pancreatic function. *Gastroenterology* **100,** 1392–1396.

Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W., Jansen, R., De Geus, E. J., Boomsma, D. I., Wright, F. A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics* **48,** 245–252.

Hao, X., Zeng, P., Zhang, S., and Zhou, X. (2018). Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genetics* **14,** e1007186.

He, X., Fuller, C. K., Song, Y., Meng, Q., Zhang, B., Yang, X., and Li, H. (2013). Sherlock: detecting gene-disease associations by matching patterns of expression QTL and GWAS. *The American Journal of Human Genetics* **92,** 667–680.

Hunter, D. J. (2005). Gene–environment interactions in human diseases. *Nature Reviews Genetics* **6,** 287.

Hutter, C. M., Mechanic, L. E., Chatterjee, N., Kraft, P., and Gillanders, E. M. (2013). Gene-environment interactions in cancer epidemiology: A national cancer institute think tank report. *Genetic Epidemiology* **37,** 643–657.

Idaghdour, Y. and Awadalla, P. (2013). Exploiting gene expression variation to capture gene-environment interactions for disease. *Frontiers in Genetics* **3,** 228.

Klein, A. P., Wolpin, B. M., Risch, H. A., Stolzenberg-Solomon, R. Z., Mocci, E., Zhang, M., Canzian, F., Childs, E. J., Hoskins, J. W., Jermusyk, A., et al. (2018). Genome-wide meta-analysis identifies five new susceptibility loci for pancreatic cancer. *Nature Communications* **9,** 556.

Kwak, I.-Y. and Pan, W. (2016). Gene-and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics* **33,** 64–71.

Li, L., Kabesch, M., Bouzigon, E., Demenais, F., Farrall, M., Moffatt, M. F., Lin, X., and Liang, L. (2013). Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Frontiers in Genetics* **4,** 103.

Lin, X., Lee, S., Wu, M. C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. *Biometrics* **72,** 156–164.

Lomberk, G. A., Iovanna, J., and Urrutia, R. (2016). The promise of epigenomic therapeutics in pancreatic cancer. *Epigenomics* **8,** 831–842.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The genotype-tissue expression (GTEx) project. *Nature Genetics* **45,** 580.

Low, S.-K., Kuchiba, A., Zembutsu, H., Saito, A., Takahashi, A., Kubo, M., Daigo, Y., Kamatani, N., Chiku, S., Totsuka, H., et al. (2010). Genome-wide association study of pancreatic cancer in Japanese population. *PloS One* **5,** e11824.

Ma, Y. and Wei, P. (2019). FunSPU: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. *PLoS Genetics* **15,** e1008081.

McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A. R., Teumer, A., Kang, H. M., Fuchsberger, C., Danecek, P., Sharp, K., et al. (2016). A reference panel of 64,976 haplotypes for genotype imputation. *Nature Genetics* **48,** 1279.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated snps are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genetics* **6,** e1000888.

Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics* **197,** 1081–1095.

Petersen, G. M., Amundadottir, L., Fuchs, C. S., Kraft, P., Stolzenberg-Solomon, R. Z., Jacobs, K. B., Arslan, A. A., Bueno-de Mesquita, H. B., Gallinger, S., Gross, M., et al. (2010). A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22. 1, 1q32. 1 and 5p15. 33. *Nature Genetics* **42,** 224.

Ritz, B. R., Chatterjee, N., Garcia-Closas, M., Gauderman, W. J., Pierce, B. L., Kraft, P., Tanner, C. M., Mechanic, L. E., and McAllister, K. (2017). Lessons learned from past gene-environment interaction successes. *American Journal of Epidemiology* **186,** 778–786.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics* **24,** 398.

Roignot, J., Taieb, D., Suliman, M., Dusetti, N., Iovanna, J., and Soubeyran, P. (2010). CIP4 is a new ArgBP2 interacting protein that modulates the ArgBP2 mediated control of WAVE1 phosphorylation and cancer cell migration. *Cancer Letters* **288,** 116–123.

Sarosiek, K., Gandhi, A. V., Saxena, S., Kang, C. Y., Chipitsyna, G. I., Yeo, C. J., and Arafat, H. A. (2016). Hypothyroidism in pancreatic cancer: Role of exogenous thyroid hormone in tumor invasion preliminary observations. *Journal of Thyroid Research* **2016,** 2454989.

Smith, P. and Day, N. (1984). The design of case-control studies: the influence of confounding and interaction effects. *International Journal of Epidemiology* **13,** 356–365.

Su, Y.-R., Di, C.-Z., and Hsu, L. (2017). A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics* **18,** 119–131.

Tang, H., Wei, P., Duell, E. J., Risch, H. A., Olson, S. H., Bueno-de Mesquita, H. B., Gallinger, S., Holly, E. A., Petersen, G., Bracci, P. M., et al. (2014). Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene- and pathway-based interaction analysis of GWAS data. *Carcinogenesis* **35,** 1039–1045.

The 1000 Genomes Project Consortium (2010). A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061.

Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A. N., Knowles, D. A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. *Nature Genetics* **51,** 592.

Walters, M. S., De Bishnu, P., Salit, J., Buro-Auriemma, L. J., Wilson, T., Rogalski, A. M., Lief, L., Hackett, N. R., Staudt, M. R., Tilley, A. E., et al. (2014). Smoking accelerates aging of the small airway epithelium. *Respiratory Research* **15,** 94.

Wang, T. and Elston, R. C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. *The American Journal of Human Genetics* **80,** 353–360.

Wolock, S. L., Yates, A., Petrill, S. A., Bohland, J. W., Blair, C., Li, N., Machiraju, R., Huang, K., and Bartlett, C. W. (2013). Gene× smoking interactions on human brain gene expression: finding common mechanisms in adolescents and adults. *Journal of Child Psychology and Psychiatry* **54,** 1109–1119.

Wolpin, B. M., Rizzato, C., Kraft, P., Kooperberg, C., Petersen, G. M., Wang, Z., Arslan, A. A., Beane-Freeman, L., Bracci, P. M., Buring, J., et al. (2014). Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nature Genetics* **46,** 994.

Wu, C., Miao, X., Huang, L., Che, X., Jiang, G., Yu, D., Yang, X., Cao, G., Hu, Z., Zhou, Y., et al. (2012). Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nature Genetics* **44,** 62.

Wu, C. and Pan, W. (2018). Integration of enhancer-promoter interactions with GWAS summary results identifies novel schizophrenia-associated genes and pathways. *Genetics* **209,** 699–709.

Wu, X., Huang, Q., Javed, R., Zhong, J., Gao, H., and Liang, H. (2019). Effect of tobacco smoking on the epigenetic age of human respiratory organs. *Clinical Epigenetics* **11,** 1–9.

Xu, Z., Wu, C., Wei, P., and Pan, W. (2017). A powerful framework for integrating eQTL and GWAS summary data. *Genetics* **207,** 893–902.

Yang, T., Chen, H., Tang, H., Li, D., and Wei, P. (2019). A powerful and data-adaptive test for rare-variant–based gene-environment interaction analysis. *Statistics in Medicine* **38,** 1230–1244.

Yang, T., Kim, J., Wu, C., Ma, Y., Wei, P., and Pan, W. (2020). An adaptive test for meta-analysis of rare variant association studies. *Genetic Epidemiology* **44,** 104–116.

Yang, T., Wu, C., Wei, P., and Pan, W. (2020). Integrating DNA sequencing and transcriptomic data for association analyses of low-frequency variants and lipid traits. *Human Molecular Genetics* **29,** 515–526.

Tables

Table 1: Weights specified in the simulation study. M is the number of sets of weights, l is the number of SNPs in LD with causal SNPs and neutral SNPs.

Scenario	Key features	M	l	Weights
	Type 1 error rate			
T.1)	benchmark	1	0	$\mathbf{W}_{cs} = \mathbf{1}^T, \mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{0}^T$
T.2)	benchmark + neutral variants	1	50	$\mathbf{W}_{cs} = \mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{1}^T$
T.3)	multiple weights	8	0	$\mathbf{W}_1 \sim N(0, 0.1^2)$ for $m = 1, 2, 3, 4$, and $\mathbf{W}_1 = [-0.1, 1, -0.2, -1, 0.3]$ for $m = 5, 6, 7, 8$
T.4)	multiple weights + neutral variants	8	50	A combination of M.1) and M.2)
	Power: a single set of weights			
S.0)	benchmark/equal weights	1	0,50, 100,200	$\mathbf{W}_{cs} = \mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{1}^T$
S.1)	accurate weights	1	0,50,100,200	$\mathbf{W}_{cs} = \boldsymbol{\alpha} = [-0.20, 1.18, -0.43, -1.16, 0.34]$, $\mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{0}^T$
S.2)	informative weights	1	0,50,100,200	$\mathbf{W}_{cs} = \boldsymbol{\alpha} + \epsilon_\alpha = [-0.1, 1, -0.2, -1, 0.3]$, $\mathbf{W}_{tag} \sim N(0, 0.1^2)$, $\mathbf{W}_{noise} \sim N(0, 0.05^2)$.
S.3)	correct sign but wrong magnitude	1	0,50,100,200	$\mathbf{W}_{cs} = \text{sign}(\boldsymbol{\alpha}) = [-1, 1, -1, -1, 1]$, $\mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{0}^T$
S.4)	0-1 weights	1	0,50,100,200	$\mathbf{W}_1 = I(\boldsymbol{\alpha} \neq 0)$, I is the identity function
S.5)	incorrect weights	1	0,50,100,200	$\mathbf{W}_{cs} \sim N(0, 0.1^2)$, $\mathbf{W}_{tag} \sim N(0, 0.1^2)$, and $\mathbf{W}_{noise} \sim N(0, 0.5^2)$
	Power: multiple sets of weights			
M.1)	informative weights	4	0,50,100,200	$\mathbf{W}_{cs} = [-0.1, 1, -0.2, -1, 0.3]$,$\mathbf{W}_{tag} = \mathbf{W}_{noise} = \mathbf{0}^T$
M.2)	incorrect weights	4	0,50,100,200	$\mathbf{W}_{cs} \sim N(0, 0.1^2)$, $\mathbf{W}_{tag} \sim N(0, 0.1^2)$, $\mathbf{W}_{noise} \sim N(0, 0.5^2)$.
M.3)	informative + incorrect weights	8	0,50,100,200	A combination of M.1) and M.2)

Table 2: Empirical Type I error rates at the significance level of 0.05 based on 10,000 replications.

			Single weights		Multiple weights	
Scenarios	case	control	T.1) benchmark	T.2) benchmark + neutral variants	T.3) multiple weights	T.4) multiple weights + neutral variants
GEw($\gamma_1 = 1, \gamma_2 = 1$)	500	500	0.047	0.047	0.053	0.045
aGEw	500	500	0.046	0.033	0.050	0.038
GEw($\gamma_1 = 1, \gamma_2 = 1$)	1000	1000	0.049	0.045	0.053	0.044
aGEw	1000	1000	0.047	0.035	0.046	0.043
GEw($\gamma_1 = 1, \gamma_2 = 1$)	1000	2000	0.047	0.047	0.055	0.044
aGEw	1000	2000	0.049	0.039	0.049	0.046
GEw($\gamma_1 = 1, \gamma_2 = 1$)	3500	3500	0.051	0.045	0.055	0.050
aGEw	3500	3500	0.048	0.039	0.053	0.046

Table 3: The top five genes for the aGEw test identified in PanC4. nTissue is the number of tissues with heritable gene expression; NSNP is the number of SNPs per gene included in the training of GTEx V6 models; nSNP is the number of SNPs with non-zero weights among any available tissues; the aGEw test used all of the available qualified continuous weights and the pair of γ_1, γ_2 with the smallest p-value is reported in parenthesis; aGEw (0-1) is the aGEw test that used the 0-1 binary weight; aGE is the unweighted test based on the NSNP available variants. The gene achieving statistical significance in the replication dataset is marked in boldface. Weights are from GTEx V6 models.

					PanC4			PanScan II		
Gene	Chr	nTissue	NSNP	nSNP	aGEw (γ_1, γ_2)	aGEw (0-1)	aGE	aGEw	aGEw (0-1)	aGE
DYNC1LI1	3	1	445	3	2.00E-04 (1,1)	3.40E-04	2.40E-02	2.30E-02	4.00E-02	1.40E-01
BACH1	21	2	631	25	6.60E-04 (1,4)	4.30E-03	3.90E-02	6.70E-01	8.42E-01	9.05E-01
PAPSS1	4	13	386	142	8.00E-04 (1,1)	9.99E-03	7.99E-03	3.49E-01	5.46E-01	4.14E-01
SMURF1	7	1	355	10	1.30E-03 (2,1)	2.60E-03	1.90E-02	6.41E-01	4.28E-01	3.34E-01
TRIP10	19	2	461	68	1.30E-03 (1,4)	3.00E-02	2.50E-02	**5.99E-03**	1.62E-01	1.90E-02

Table 4: The top five genes for the aGEw (0-1) test identified in PanC4. nTissue is the number of tissues with heritable gene expression; NSNP is the number of SNPs per gene included in the training of GTEx V6 models; nSNP is the number of SNPs with non-zero weights among any available tissues; the aGEw test used all of the available qualified continuous weights; aGEw (0-1) is the aGEw test that used the 0-1 weight; aGE is the unweighted test based on all available SNPs. The pair of γ_1, γ_2 with the smallest p-value is reported in parenthesis. The gene achieving statistical significance in the replication dataset is marked in boldface. Weights are from GTEx V6 models.

					PanC4			PanScan II		
Gene	Chr	nTissue	NSNP	nSNP	aGEw	aGEw (0-1) (γ_1, γ_2)	aGE	aGEw	aGEw (0-1)	aGE
KDM3A	2	1	361	14	3.10E-02	2.90E-04 (2,1)	8.29E-02	4.70E-02	**9.99E-03**	5.69E-02
DYNC1LI1	3	1	445	3	2.00E-04	3.40E-04 (4,1)	2.40E-02	2.30E-02	4.00E-02	1.40E-01
TMCO3	13	9	361	124	3.90E-02	3.70E-04 (5,1)	2.10E-03	1.80E-01	3.54E-01	4.21E-01
DCUN1D2	13	4	372	110	1.60E-01	4.40E-04 (6,1)	5.00E-03	9.91E-01	5.26E-01	3.87E-01
RAPGEF5	7	5	856	108	7.69E-02	5.90E-04 (5,1)	1.10E-02	4.35E-01	5.65E-01	8.34E-01

Figures

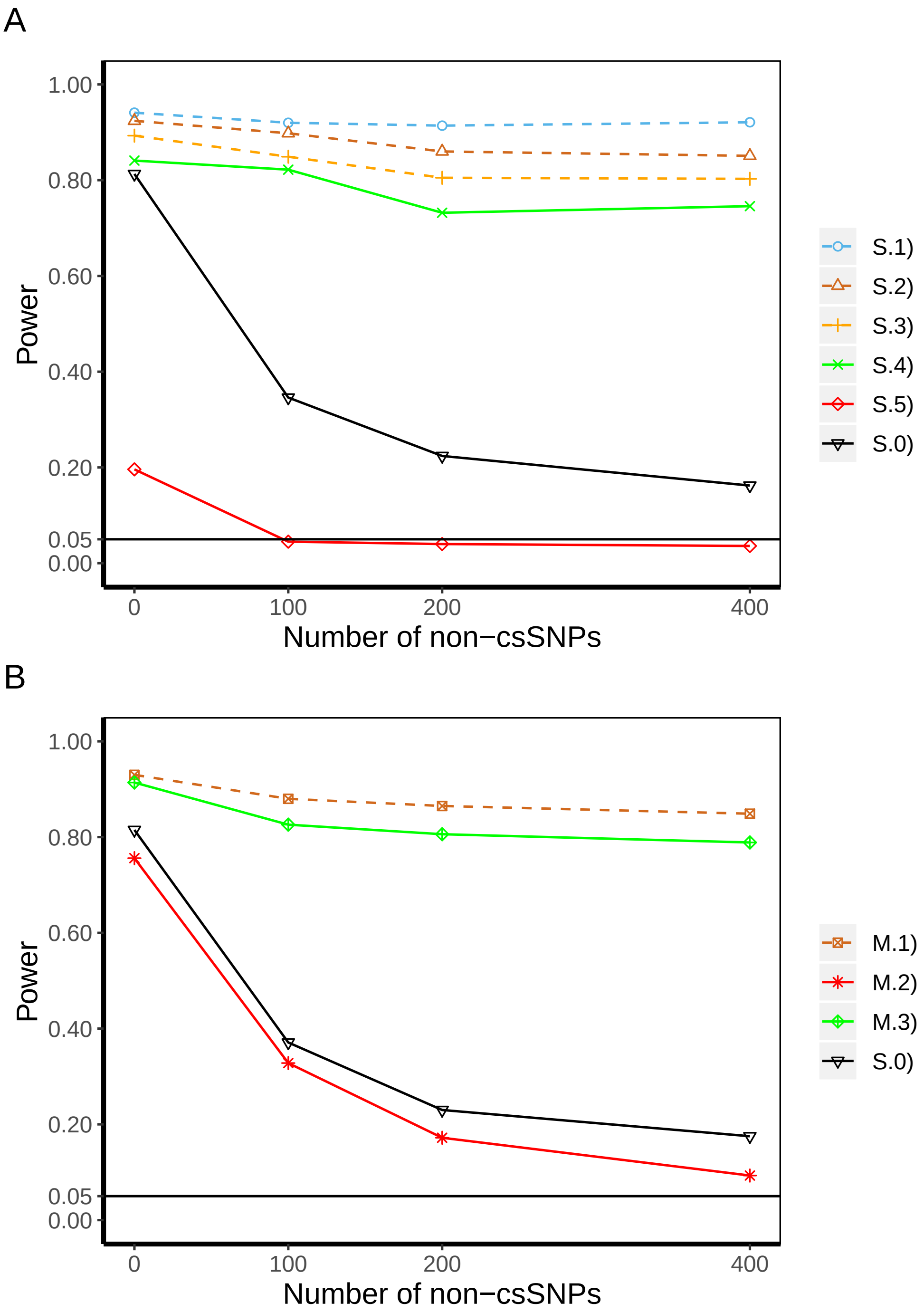

Figure 1: Empirical power curves for aGEw tests based on 1000 cases and 1000 controls. A: aGEw with single sets of weights; B: aGEw with multiple sets of correlated weights. Weights S.1)-S.5) and M.1)-M.3) are specified in Table 1 : S.1) accurate weights, S.2) informative weights, S.3) correct sign but wrong magnitude, S.4) 0-1 weights, and S.5) incorrect weights; M.1) informative weights, M.2) incorrect weights, and M.3) informative + incorrect weights.