Incorporating Gene-Environment Interaction in Testing for Association with Rare Genetic Variants

Han Chen^{a, b} James B. Meigs^{c, d} Josée Dupuis^{a, e}

^aDepartment of Biostatistics, Boston University School of Public Health, ^bDepartment of Biostatistics, Harvard School of Public Health, ^cGeneral Medicine Division, Massachusetts General Hospital, and ^dDepartment of Medicine, Harvard Medical School, Boston, Mass., and ^eThe National Heart, Lung and Blood Institute's Framingham Heart Study, Framingham, Mass., USA

Key Words

Rare variant analysis · Gene-environment interaction · Sequence kernel association test · Joint test · Generalized linear mixed model

Abstract

Objectives: The incorporation of gene-environment interactions could improve the ability to detect genetic associations with complex traits. For common genetic variants, single-marker interaction tests and joint tests of genetic main effects and gene-environment interaction have been well-established and used to identify novel association loci for complex diseases and continuous traits. For rare genetic variants, however, single-marker tests are severely underpowered due to the low minor allele frequency, and only a few gene-environment interaction tests have been developed. We aimed at developing powerful and computationally efficient tests for gene-environment interaction with rare variants. **Methods:** In this paper, we propose interaction and joint tests for testing gene-environment interaction of rare genetic variants. Our approach is a generalization of existing gene-environment interaction tests for multiple genetic variants under certain conditions. **Results:** We show in our simulation studies that our interaction and joint tests have correct type

I errors, and that the joint test is a powerful approach for testing genetic association, allowing for gene-environment interaction. We also illustrate our approach in a real data example from the Framingham Heart Study. **Conclusion:** Our approach can be applied to both binary and continuous traits, it is powerful and computationally efficient.

© 2014 S. Karger AG, Basel

Introduction

Traditional genome-wide association studies (GWAS) have been successfully applied to identify a large number of genetic markers associated with complex diseases and related continuous traits. However, for most complex diseases and continuous traits, all genetic markers identified so far only explain a small proportion of the heritability in these traits, suggesting that a lot of genetic determinants are still undiscovered. Eichler et al. [1] suggested that gene-environment interaction and rare genetic variants may both account for some of the unexplained heritability.

Statistical methods to detect gene-environment interaction have been well-established in the context of single-marker tests [2–4]. To determine if a common genetic

variant interacts with an environmental variable, we can either test the interaction effect only or jointly test both genetic main effect and gene-environment interaction effect. By using the first approach, we are usually interested in detecting gene-environment interaction, regardless of the presence of a significant genetic main effect. However, by using the second approach, we are testing if the genetic marker is associated with the trait of interest, allowing for gene-environment interaction. These methods combined with multivariate meta-analysis have led to the discovery of novel common loci associated with fasting insulin levels by incorporating gene $\times$ body mass index (BMI) interactions [5]. However, the power of single-marker tests greatly depends on the minor allele frequency (MAF) of the genetic marker tested. As a result, single-marker tests have little power in testing gene-environment interaction involving rare genetic variants. On the other hand, rare genetic variants analysis has become a popular research field in genetic association studies, and many statistical methods for rare variants analysis have been proposed [6–12]. Of these methods, the sequence kernel association test (SKAT) [12] has been shown to be a powerful approach in various scenarios. However, all of these methods focus on the main effect association analysis of rare genetic variants.

Compared with common variant analysis, rare variant analysis often requires a larger sample size to attain comparable power. Compared with main effect analysis, interaction analysis also needs a larger sample size. Thus, little attention has been paid to interaction analysis for rare genetic variants, possibly due to the limited sample size in many cohort studies. Tzeng et al. [13] proposed a gene-trait similarity regression approach (SIMreg) to test gene-environment interaction of rare variants. This approach is flexible in the hypothesis testing, and the main effect test, test of interaction and joint test of main effect and interaction are possible. It is available for continuous and binary traits, although the original reference only includes the development for continuous traits. However, in the joint test, it implicitly assumes that the variance component parameters for genetic main effects and gene-environment interactions are of the same magnitude under the alternative hypothesis, which may not be true. In a first attempt to specifically consider gene-environment interactions for rare variants, Kazma et al. [14] proposed 3 joint tests: (1) minimum p value, (2) CAST [6], and (3) SKAT [12]. However, they considered only the scenario where the environmental variable is binary when computing SKAT weights, which limits the applicability of their SKAT approach. They used a single kernel for both

genetic main effects and gene-environment interactions, and pointed out that it would be valuable to model genetic main effects and gene-environment interactions separately. Recently, Lin et al. [15] have proposed the gene-environment set association test (GESAT) as an interaction test in a generalized linear model framework. Although GESAT was originally developed for common variants, it can be easily applied to rare variants by incorporating weights, which we name GESAT with weights (GESAT-W). This test is applicable to both binary and continuous traits, but the joint test was not provided.

In this paper, we propose a general approach for testing gene-environment interaction of rare variants, including two interaction tests and a joint test. Our approach is flexible and works for both binary and continuous traits. Genetic main effects and gene-environment interactions are modeled separately, and no assumption is required on the magnitude of variance component parameters for genetic main effects and gene-environment interactions. We evaluate the performance of our methods in simulation studies, and we also illustrate our approaches in testing gene $\times$ BMI interaction on fasting glucose and fasting insulin levels, adjusting for age, sex, cohort and 10 principal components [16], using an unrelated subset of individuals from the Framingham Heart Study (FHS).

Methods

Interaction Test

Fixed Genetic Main Effects

Our notation is as follows: assuming a sample size of n , let Y_i be the phenotype of individual i ($1 \leq i \leq n$), and follow an exponential family distribution with $E(Y_i) = \mu_i$ and $Var(Y_i) = \phi v(\mu_i)$. Let $X_i = (1, X_{i1}, X_{i2}, \dots, X_{i(p-1)})$ be a row vector of length p , consisting of an intercept and $(p-1)$ covariates, E_i be one of these $(p-1)$ covariates, centered to have a mean of 0, and $G_i = (G_{i1}, G_{i2}, \dots, G_{iq})$ be a row vector of q genetic variants.

Assuming independent observations, we can write the generalized linear mixed model for testing the gene by E_i interaction as

$$g(\mu_i) = X_i\beta + G_iW_1\gamma_1 + E_iG_iW_2\gamma_2,$$

where $g(\cdot)$ is the link function, W_1 and W_2 are $q \times q$ diagonal matrices with elements equal to weights for genetic main effects and gene-environment interaction effects, respectively. β is a vector of fixed-effects parameters for the intercept and $(p-1)$ covariates, γ_1 is a vector of fixed-effects parameters for genetic main effects, γ_2 is a vector of random effects of gene-environment interaction, assumed to have a mean of 0 and covariance matrix $\tau_2 I_q$. The interaction test can be constructed as $H_0: \tau_2 = 0$ versus $H_1: \tau_2 > 0$. Let $Y = (Y_1, Y_2, \dots, Y_n)^T$ be the phenotype vector, $\mu = (\mu_1, \mu_2, \dots, \mu_n)^T$ be the mean vector, X be an $n \times p$ covariates matrix with elements X_{ij} (where $1 \leq i \leq n$ and $0 \leq j \leq p-1$) and $X_{i0} = 1$,

G be an $n \times q$ genotype matrix with elements G_{ij} (where $1 \leq i \leq n$ and $1 \leq j \leq q$), and $E = \text{diag}\{E_i\}$ be an $n \times n$ matrix. We define the working vector under the null hypothesis $y = X\beta + GW_1\gamma_1 + \Delta(Y - \mu)$, where $\Delta = \text{diag}\{g'(\mu_i)\}$ [17]. Let ε be $\Delta(Y - \mu)$, we can write the model as

$$y = X\beta + GW_1\gamma_1 + \varepsilon,$$

where $\text{Var}(\varepsilon) = V = \text{diag}\{\phi v(\mu_i)[g'(\mu_i)]^2\}$. Let $\beta_F, \gamma_{1F}, \phi_F$ be the maximum likelihood (or restricted maximum likelihood) estimates under the null hypothesis $g(\mu_i) = X_i\beta + G_iW_1\gamma_1$, then we can calculate estimates $\mu_{iF} = g^{-1}(X_i\beta_F + G_iW_1\gamma_{1F})$, $\Delta_F = \text{diag}\{g'(\mu_{iF})\}$, $V_F = \text{diag}\{\phi_F v(\mu_{iF})[g'(\mu_{iF})]^2\}$, and the test statistic for $H_0: \tau_2 = 0$ is

$$Q_F = (Y - \mu_F)^T \Delta_F V_F^{-1} E G W_2 W_2^T E V_F^{-1} \Delta_F (Y - \mu_F).$$

Let $Z_i = (X_i, G_iW_1)$ be a row vector of length $(p + q)$, Z be an $n \times (p + q)$ matrix with elements Z_{ij} (where $1 \leq i \leq n$ and $1 \leq j \leq p + q$). Under the null hypothesis,

$$Q_F \sim \sum_{j=1}^q \lambda_j \chi_{1,j}^2,$$

where λ_j s are the eigenvalues of the matrix

$$\Psi_F = W_2 G^T E (V_F^{-1} - V_F^{-1} Z (Z^T V_F^{-1} Z)^{-1} Z^T V_F^{-1}) E G W_2 \quad [12, 18].$$

This interaction test is straightforward; however, when the number of genetic variants q is large, we cannot usually get stable estimates of γ_1 , leaving it impractical. Below, we propose an alternative interaction test which treats genetic main effects γ_1 as random effects.

Random Genetic Main Effects

We use the same notations as before, but now assume that γ_1 is a vector of random effects with a mean of 0 and covariance matrix $\tau_1 I_q$. Again we define the working vector under the null hypothesis $y = X\beta + GW_1\gamma_1 + \Delta(Y - \mu)$, with $\text{Var}(y) = \Sigma = \tau_1 G W_1 W_1^T G^T + V$. We can get the maximum likelihood (or restricted maximum likelihood) estimates $\beta_R, \tau_{1R}, \phi_R$ from the null model, and compute estimates $\mu_{iR} = g^{-1}(X_i\beta_R + G_iW_1\gamma_{1R})$, $\Delta_R = \text{diag}\{g'(\mu_{iR})\}$, $\gamma_R = X\beta_R + G W_1\gamma_{1R} + \Delta_R(Y - \mu_R)$, $V_R = \text{diag}\{\phi_R v(\mu_{iR})[g'(\mu_{iR})]^2\}$, $\Sigma_R = \tau_{1R} G W_1 W_1^T G^T + V_R$, where $\gamma_{1R} = \tau_{1R} W_1 G^T \Sigma_R^{-1} (y_R - X\beta_R)$. The test statistic for $H_0: \tau_2 = 0$ is

$$Q_R = (Y - \mu_R)^T \Delta_R V_R^{-1} E G W_2 W_2^T E V_R^{-1} \Delta_R (Y - \mu_R).$$

Under the null hypothesis,

$$Q_R \sim \sum_{j=1}^q \lambda_j \chi_{1,j}^2,$$

where λ_j s are the eigenvalues of the matrix

$$\Psi_R = W_2 G^T E (\Sigma_R^{-1} - \Sigma_R^{-1} X (X^T \Sigma_R^{-1} X)^{-1} X^T \Sigma_R^{-1}) E G W_2.$$

Joint Test

Kraft et al. [2] suggested simultaneously testing genetic main effect and interaction effect in a joint test. In the context of the generalized linear mixed model (see above), a joint test of main and interaction effects in our context is equivalent to testing the following hypotheses:

$$H_0: \tau_1 = \tau_2 = 0 \text{ versus } H_1: \tau_1 > 0 \text{ or } \tau_2 > 0.$$

Under the null hypothesis, we define the working vector $y = X\beta + \Delta(Y - \mu)$ with $\text{Var}(y) = V$. Let β_J, ϕ_J be the maximum likelihood (or restricted maximum likelihood) estimates from the null model $g(\mu_i) = X_i\beta$, then we can calculate estimates $\mu_{iJ} = g^{-1}(X_i\beta_J)$, $\Delta_J = \text{diag}\{g'(\mu_{iJ})\}$, $\gamma_J = X\beta_J + \Delta_J(Y - \mu_J)$, $V_J = \text{diag}\{\phi_J v(\mu_{iJ})[g'(\mu_{iJ})]^2\}$. If we reparametrize $\tau_1 = \rho\tau$, $\tau_2 = (1 - \rho)\tau$ ($0 \leq \rho \leq 1$), with $\tau = \tau_1 + \tau_2$ and $\rho = \tau_1/(\tau_1 + \tau_2)$, the hypotheses for the joint test may be rewritten as

$$H_0: \tau = 0 \text{ versus } H_1: \tau > 0,$$

where ρ is a nuisance parameter present only under the alternative hypothesis. For any fixed ρ , the test statistic for $H_0: \tau = 0$ is

$$Q_J(\rho) = (Y - \mu_J)^T \Delta_J V_J^{-1} [\rho G W_1 W_1^T G^T + (1 - \rho) E G W_2 W_2^T E] V_J^{-1} \Delta_J (Y - \mu_J).$$

Let P be $V_J^{-1} - V_J^{-1} X (X^T V_J^{-1} X)^{-1} X^T V_J^{-1}$, under the null hypothesis,

$$Q_J(\rho) \sim \sum_{j=1}^n \lambda_j \chi_{1,j}^2,$$

where λ_j s are the eigenvalues of the matrix

$$\Psi_J(\rho) = P^{1/2} [\rho G W_1 W_1^T G^T + (1 - \rho) E G W_2 W_2^T E] P^{1/2}.$$

Note that if $n \geq q$, there are at most q non-zero eigenvalues. When $\rho = 1$, this is the regular SKAT statistic for genetic main effects.

Following Lee et al. [19], we use the minimum p value as the test statistic:

$$Q_J = \inf_{0 \leq \rho \leq 1} p(\rho),$$

where $p(\rho)$ is the p value of $Q_J(\rho)$. For a sequence of S ρ s: $0 \leq \rho_1 < \rho_2 < \dots < \rho_S \leq 1$, the test statistic is $Q_J = \min\{p(\rho_1), p(\rho_2), \dots, p(\rho_S)\}$. To compute the p value of observed $Q_J = q_J$, following Voorman et al. [20], we can write the p value as an integral, which needs to be solved numerically. We use a Monte Carlo method in our implementation (see Appendix).

There are several ways for computing the p value of the sum of χ^2 s [21–24]. In this paper, we generally use Kuonen's saddlepoint method [23] (in interaction tests and evaluating $F(x)$ in the joint test), except that in the joint test, we use a modified moment-matching method (matching the mean, variance, and kurtosis) [19, 24] to compute $p(\rho_S)$ and quantiles $T(\rho_S)$, because it takes much longer to find quantiles using Kuonen's saddlepoint method.

Simulation Studies

We performed simulation studies to evaluate the empirical type I error and power of main effect SKAT [12], the interaction test treating genetic main effects as fixed (INT-FIX), the interaction test treating genetic main effects as random (INT-RAN), the joint test of genetic main effects and gene-environment interactions (JOINT), SIMreg tests by Tzeng et al. [13] (http://www4.stat.ncsu.edu/~jytzeng/software_simreg.php), and GESAT-W by Lin et al. [15]. In a preliminary simulation study for continuous traits, it took up to 26 min to get the p value from 1 SIMreg interaction test on 2,000 individuals and 20 genetic variants. As it would not be feasible to run 1,000,000 simulations to evaluate the type I error, we excluded the SIMreg interaction test from the comparison. We also excluded GESAT-W for binary traits as the current package is only available for continuous traits. We used Wu et al's

weights [12], which are the beta density function with parameters 1 and 25, evaluated at the MAF of the genetic variant, for SKAT, INT-FIX, INT-RAN, GESAT-W and JOINT. In the joint test, we used the Monte Carlo method with $B = 10,000$ to calculate the p value. In the SIMreg main effect test (SR) and joint test (SR-JOINT), we used inverse allele frequency as the weight to calculate the genotype similarity matrix, as recommended by Tzeng et al. [13].

Type I Error

We simulated both continuous and binary phenotypes. For continuous traits, we simulated 5,000 genotype datasets with a sample size of 2,000. In each genotype dataset, we simulated 20 biallelic genetic variants with a MAF randomly sampled from a uniform distribution on the interval (0.005, 0.05), and we fixed the linkage disequilibrium correlation between adjacent markers at $r = 0.5$. For each genotype dataset, we simulated 200 replicates of covariates: $sex \sim \text{Bernoulli}(0.5)$, $age \sim N(50, 5^2)$, $bmi \sim N(25, 4^2)$, then we generated the continuous phenotypes Y from

$$Y = 0.5 \text{ sex} + 0.05 \text{ age} + 0.1 \text{ bmi} + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$ is the random error. The genotypes are not associated with the phenotype. We evaluated the empirical type I errors for SKAT, SR, INT-FIX, INT-RAN, GESAT-W, JOINT, and SR-JOINT. We tested gene $\times$ BMI interaction in INT-FIX, INT-RAN, GESAT-W, JOINT, and SR-JOINT.

For binary traits, we simulated case-control studies with 1,000 cases and 1,000 controls for each replicate. We first simulated 5,000 genotype datasets with a sample size of 20,000. For each genotype dataset, we simulated 200 replicates of covariates. The genetic variants and covariates were simulated following the same procedure as for continuous traits. For each individual, we calculated the probability of disease $P(Y = 1)$ from

$$\log \frac{P(Y = 1)}{1 - P(Y = 1)} = \beta_0 + 0.5 \text{ sex} + 0.05 (age - 50) + 0.1 (bmi - 25),$$

where β_0 was determined such that the baseline prevalence of disease is 10% (individuals with $sex = 0$, $age = 50$, $bmi = 25$). Once we simulated the disease status for all 20,000 individuals in each replicate, we randomly sampled 1,000 cases and 1,000 controls from the cohort. Then we evaluated the empirical type I errors for SKAT, SR, INT-FIX, INT-RAN, JOINT, and SR-JOINT. We tested gene $\times$ BMI interaction in INT-FIX, INT-RAN, JOINT, and SR-JOINT.

Power

For both continuous and binary phenotypes, we simulated 5 scenarios: (1) there are genetic main effects but no gene-BMI interaction effects; (2) there are genetic main effects and weak gene-BMI interaction effects; (3) there are both genetic main effects and gene-BMI interaction effects of the same magnitude; (4) there are gene-BMI interaction effects and weak genetic main effects, and (5) there are gene-BMI interaction effects but no genetic main effects. For continuous traits, we simulated 5 genotype datasets with a sample size of 2,000 in each scenario, and 200 replicates of covariates for each genotype dataset. The genetic variants and covariates

were simulated following the same procedure as in the type I error simulations. For each replicate, we randomly selected 10 causal variants. The continuous phenotypes Y were generated from

$$Y = 0.5 \text{ sex} + 0.05 \text{ age} + 0.1 \text{ bmi} + \sum_{j=1}^q \gamma_{1,j} g_j + \sum_{j=1}^q \gamma_{2,j} g_j (bmi - 25) + \varepsilon,$$

where g_j is the j -th genetic variant, with a MAF of f_j . The genetic main effects $\gamma_{1,j}$ were simulated from $\gamma_{1,j} \sim N(0, \sigma_1^2)$, where constant σ_1 was fixed at 0.2 in Scenarios 1–3, at 0.02 in Scenario 4, and at 0 in Scenario 5. Gene-BMI interaction effects $\gamma_{2,j}$ were simulated from $\gamma_{2,j} \sim N(0, \sigma_2^2)$, where constant σ_2 was fixed at 0 in Scenario 1, at 0.005 in Scenario 2, and at 0.05 in Scenarios 3–5.

For binary traits, we first simulated 5 genotype datasets with a sample size of 20,000, and 200 replicates of covariates for each genotype dataset. The genetic variants and covariates were simulated following the same procedure as in the type I error simulations. For each replicate, we randomly selected 10 causal variants, and we calculated the probability of disease $P(Y = 1)$ from

$$\log \frac{P(Y = 1)}{1 - P(Y = 1)} = \beta_0 + 0.5 \text{ sex} + 0.05 (age - 50) + 0.1 (bmi - 25) + \sum_{j=1}^q \gamma_{1,j} g_j + \sum_{j=1}^q \gamma_{2,j} g_j (bmi - 25),$$

where g_j is the j -th genetic variant, with a MAF of f_j . Parameter β_0 was determined such that the baseline prevalence of disease is 10% (individuals with $sex = 0$, $age = 50$, $bmi = 25$, all $g_j = 0$). After obtaining the disease status for all 20,000 individuals in each replicate, we randomly sampled 1,000 cases and 1,000 controls from the cohort. Genetic main effects $\gamma_{1,j}$ and gene-BMI interaction effects $\gamma_{2,j}$ were calculated in the same way as for continuous traits, except that constant σ_1 was fixed at 0.4 in Scenarios 1–3, at 0.04 in Scenario 4, and at 0 in Scenario 5; and constant σ_2 was fixed at 0 in Scenario 1, at 0.01 in Scenario 2, and at 0.1 in Scenarios 3–5.

Results

Type I Error Simulations

Table 1 includes the empirical type I errors of the 7 tests: (1) genetic main effect test (SKAT); (2) SIMreg main effect test (SR); (3) gene-BMI interaction test adjusting for fixed genetic main effects (INT-FIX); (4) gene-BMI interaction test adjusting for random genetic main effects (INT-RAN); (5) GESAT-W gene-BMI interaction test (GESAT-W); (6) the joint test of genetic main effects and gene-BMI interaction (JOINT), and (7) SIMreg joint test (SR-JOINT) at α levels of 0.05, 0.01, 0.001, and 0.0001 from 1,000,000 simulation replicates. For binary traits, GESAT-W is not available yet.

INT-FIX has slightly inflated type I errors at high α levels, and it is more pronounced for binary traits. SR-

Table 1. Empirical type I errors from simulation studies

α level	SKAT	SR	INT-FIX	INT-RAN	GESAT-W	JOINT	SR-JOINT
<i>Continuous traits</i>							
0.05	0.05093	0.04800	0.05366	0.05033	0.04971	0.04984	0.04839
0.01	0.01008	0.00978	0.01084	0.00992	0.00986	0.00982	0.00993
0.001	0.00102	0.00116	0.00104	0.00096	0.00094	0.00097	0.00108
0.0001	0.00009	0.00013	0.00010	0.00008	0.00008	0.00010	0.00012
<i>Binary traits</i>							
0.05	0.05021	0.04707	0.05799	0.04834	NA	0.04748	0.03647
0.01	0.00972	0.00941	0.01142	0.00915	NA	0.00905	0.00661
0.001	0.00089	0.00102	0.00114	0.00092	NA	0.00086	0.00062
0.0001	0.00007	0.00011	0.00010	0.00008	NA	0.00009	0.00006

Each entry represents the proportion of p values smaller than the corresponding α level from 1,000,000 simulation replicates.

JOINT is conservative for binary traits. All other tests have empirical type I errors close to the corresponding α levels at all 4 levels we studied. For binary traits, SKAT, INT-RAN, and JOINT are slightly conservative at low α levels, but not as conservative as SR-JOINT. The results suggest that INT-RAN and JOINT, which we propose in this paper, are valid tests.

Power Simulations

For continuous traits, the empirical power results at an α level of 0.0001 from 1,000 simulation replicates are shown in figure 1. In Scenario 1, because there are genetic main effects but no gene-BMI interactions, the values for the interaction tests, namely INT-FIX, INT-RAN, and GESAT-W, are very close to 0. The main effect tests (SKAT and SR) are more powerful than the joint tests (JOINT and SR-JOINT) in this case. A similar pattern is observed in Scenario 2, where the association is dominated by genetic main effects and interaction effects are weak. In Scenario 5, there are gene-BMI interactions but no genetic main effects, the interaction tests (INT-FIX, INT-RAN, and GESAT-W) are the most powerful tests. The power of the joint tests (JOINT and SR-JOINT) is slightly lower. The main effect tests (SKAT and SR) show values very close to 0. A similar pattern is observed in Scenario 4. In Scenario 3, where both genetic main effects and gene-BMI interactions are of the same magnitude, SR-JOINT is the most powerful test, followed by JOINT.

For binary traits, the empirical power results at an α level of 0.0001 from 1,000 simulation replicates are shown in figure 2. The results have similar patterns as for continuous traits. Generally, the main effect tests are most powerful when the association is dominated by genetic

main effects, the interaction tests are most powerful when there are no or weak main effects. The joint tests are most powerful when both main effects and interaction effects are of the same magnitude, and they are only slightly less powerful than the most powerful tests in all other scenarios.

Computational Time

Although our simulation studies show that SR-JOINT has comparable power to the JOINT test proposed in this paper, a great advantage of our approach is its computational efficiency. We calculated the average computational time for JOINT and SR-JOINT in our type I error simulation settings with sample sizes from 1,000 to 10,000. Table 2 shows that the computational time for SR-JOINT increases almost cubically with the sample size. In contrast, our JOINT test is much faster than SR-JOINT when the sample size is $\geq 2,000$. The comparison suggests that in real applications, SR-JOINT may only be applicable to small studies.

We also compared the computational time for the 3 interaction tests: INT-FIX, INT-RAN, and GESAT-W. Table 2 shows that INT-RAN is slower than GESAT-W, while INT-FIX is faster. All 3 interaction tests are much faster than the SIMreg interaction test, which takes up to 26 min for 1 gene with a sample size of 2,000. The interaction tests are also faster than the joint tests.

Application to the FHS

In this real data example, we performed a genome-wide sliding window analysis for gene-BMI interaction on fasting glucose and fasting insulin levels. The FHS was initiated in 1948. In 1971, 5,124 individuals who were the

Fig. 1. Power simulation results for continuous traits at an α level of 0.0001. Scenario 1 includes genetic main effects but no gene-BMI interaction effects; Scenario 2 includes genetic main effects and weak gene-BMI interaction effects; Scenario 3 includes both genetic main effects and gene-BMI interaction effects; Scenario 4 includes gene-BMI interaction effects and weak genetic main effects; Scenario 5 includes gene-BMI interaction effects but no genetic main effects.

Fig. 2. Power simulation results for binary traits at an α level of 0.0001. Scenario 1 includes genetic main effects but no gene-BMI interaction effects; Scenario 2 includes genetic main effects and weak gene-BMI interaction effects; Scenario 3 includes both genetic main effects and gene-BMI interaction effects; Scenario 4 includes gene-BMI interaction effects and weak genetic main effects; Scenario 5 includes gene-BMI interaction effects but no genetic main effects.

offspring of the Original Cohort and their spouses were recruited in the Offspring Cohort. By 2005, 4,095 individuals who had at least one parent in the Offspring Cohort had been recruited in the Third Generation Cohort (Gen III). We selected 1,895 unrelated individuals from the FHS Offspring and Gen III Cohorts and analyzed rank-normalized fasting glucose and fasting insulin levels from the most recent physical examination. We adjusted for age, sex, cohort, first 10 principal components, and BMI main effect and tested genetic main effects using SKAT and SR, gene-BMI interaction using INT-FIX, INT-RAN, and GESAT-W, and the joint tests JOINT and SR-JOINT. We used genotype data from the Framingham Single Nucleotide Polymorphism (SNP) Health Association Resource (SHARe) and restricted our analysis to SNPs with a MAF <5%. We defined the gene regions in each test using a sliding window method: each window had a width of 500 kb, with 250 kb each overlapping with the previous and subsequent windows.

We analyzed 11,090 sliding windows from all 22 autosomes. After removing windows with 0 or only 1 rare SNP, we had results from 10,516 windows, with the number of rare SNPs ranging from 2 to 67, with a median of 17. In figure 3, we present quantile-quantile plots from both analyses, using SKAT, SR, INT-FIX, INT-RAN, GESAT-W, JOINT, and SR-JOINT. There was no systematic inflation of type I errors: most p values were very close to the corresponding expected values. We used 4.75×10^{-6} as the genome-wide significance level, based on a Bonferroni procedure to control for a family-wise significance level of 0.05. At this significance level, we found only 1 window associated with fasting glucose levels, allowing for gene-BMI interaction. The region is on chromosome 9, with position (NCBI Build 36) between 114,068,775 and 114,568,775, and there are 9 rare SNPs in this region, which are located in genes *ROD1*, *HSDL2*, *KIAA1958*, *INIP*, *SNX30*, and intergenic regions between them. The JOINT test's p value was 2.06×10^{-6} , but in-

Fig. 3. Quantile-quantile plots from the sliding window analysis on fasting glucose levels (**a, b**) and fasting insulin levels (**c, d**) in an unrelated sample of 1,895 individuals from the FHS. The observed p values from SKAT, SR, JOINT, SR-JOINT, INT-FIX, INT-RAN, and GESAT-W are plotted against the expected p values under the corresponding null hypotheses.

Table 2. Computational time of the joint tests and interaction tests

Sample size	Mean time (SEM), s				
	JOINT	SR-JOINT	INT-FIX	INT-RAN	GESAT-W
1,000	11.53 (0.18)	16.27 (0.31)	0.047 (0.001)	0.779 (0.008)	0.090 (0.001)
2,000	12.74 (0.21)	146.41 (2.49)	0.075 (0.002)	1.179 (0.012)	0.157 (0.001)
5,000	17.13 (0.27)	2,222.42 (35.74)	0.158 (0.003)	3.619 (0.032)	0.406 (0.003)
10,000	33.87 (0.45)	15,779.14 (151.29)	0.292 (0.006)	9.591 (0.095)	0.820 (0.006)

We assumed each gene had 20 genetic variants and calculated the mean computational time for each run and standard error of mean (SEM) from 1,000 replicates.

terestingly, none of the p values from SKAT, SR, INT-FIX, INT-RAN, GESAT-W, or SR-JOINT (3.90×10^{-3} , 2.98×10^{-3} , 3.23×10^{-5} , 1.90×10^{-5} , 3.72×10^{-5} , 8.64×10^{-5} , respectively) reached the genome-wide significance level. We did not find any windows associated with fasting insulin levels, using SKAT, INT-FIX, INT-RAN, and JOINT, at the genome-wide significance level of 4.75×10^{-6} .

Discussion

We propose two interaction tests and one joint test for testing gene-environment interaction of rare genetic variants. These tests are flexible and applicable to both binary and continuous outcomes. In our simulation studies, we showed that INT-FIX has slightly inflated type I errors at high α levels with moderate numbers of rare variants ($q = 20$), while INT-RAN and JOINT have well-controlled type I errors. On the other hand, INT-FIX and INT-RAN have almost the same power in all simulation scenarios, thus we would recommend using INT-RAN instead of INT-FIX for the interaction test in general, and note that INT-FIX results should be interpreted with caution, unless there are only a few variants in the test. An R package rareGE is available from the authors upon request.

The joint tests were slightly less powerful than the main effect tests if there was no interaction, and slightly less powerful than the interaction tests if there was no main effect, which is consistent with our knowledge of the single-marker interaction and joint tests [2]. However, the joint tests were most powerful when both genetic main effect and gene-environment interaction were present, and the power loss compared to the main effect tests or the interaction tests was acceptable when either interaction or main effect was absent or weak, suggesting that the joint tests are an attractive approach for testing genetic association allowing for gene-environment interaction, when we do not have a priori knowledge about the presence of gene-environment interaction.

Tzeng et al. [13, 25] showed that their SIMreg is equivalent to the haplotype random effects model, thus if we replace genotypes by haplotypes in INT-RAN, it is equivalent to the interaction test in gene-trait similarity regression when analyzing continuous traits. Moreover, Tzeng et al.'s joint test would be a special case of our joint test at a fixed $\rho = 0.5$ if we had analyzed haplotypes. It is not surprising that the gene-trait similarity regression joint test has higher power than our joint test when

genetic main effects and interaction effects are of the same magnitude. However, when the sample size is moderate to large, Tzeng et al.'s joint test takes on average 11–465 times as much computational time as our joint test in a single-gene run, which limits its application.

The interaction test proposed by Lin et al. [15] introduces a tuning parameter in the estimation of genetic main effects. They used generalized cross validation [26] to select the tuning parameter. We note that INT-RAN is equivalent to their test if the tuning parameter is unconstrained to their prespecified upper bound and equal to ϕ_R/τ_{1R} , estimated from our null generalized linear mixed model, treating genetic main effects as random. Moreover, our simulation study showed that the performance of their test is very similar to that of INT-RAN.

Although there are several genes (*ROD1*, *HSDL2*, *KIAA1958*, *INIP*, and *SNX30*) overlapping with the region on chromosome 9 found to be associated with fasting glucose levels in the joint test, none of them has previously been reported to influence fasting glucose levels. Little is known about the role of these genes in metabolism regulation, except that *HSDL2* may be involved in the cholesterol-responsive atherosclerosis pathway [27]. However, the genotype data we used in the analysis were from SNP arrays originally designed for GWAS, and most SNPs were common and excluded. For rare variant analysis, it would be ideal if we had sequence data which include more densely distributed rare SNPs. We hope to revisit this example in the future.

Gene-environment interaction has been of great interest in recent years, and gene-environment interaction studies on the whole genome have identified novel association loci for many traits [5, 28–30]. However, most of these studies so far have been performed as single-marker tests and have not been powerful for detecting gene-environment interaction of rare variants. The interaction and joint tests for gene-environment interaction of rare variants proposed in this paper were developed based on the SKAT framework [12], a variance component test for multiple genetic variants, which is powerful regardless of the directions of effects of those variants. Also, the proposed approaches can be viewed as a generalization of existing methods under specific conditions. We believe our general approach, especially the joint test, can be readily used in genetic epidemiological studies to test genetic associations with rare variants, allowing for effect modification by an environmental variable, and also facilitate future discoveries.

Appendix

Calculating the p Value for the Joint Test

To compute the p value of observed $Q_j = q_j$, let $T(\rho_s)$ be the value such that for ρ_s , $1 \leq s \leq S$, $P(Q_j(\rho_s) > T(\rho_s)) = q_j$. Let $u_1 = W_1 G^T V_j^{-1} \Delta_j(Y - \mu_j)$, $u_2 = W_2 G^T E V_j^{-1} \Delta_j(Y - \mu_j)$, then u_1 , u_2 are random vectors with length q , and $Q_j(\rho) = \rho u_1^T u_1 + (1 - \rho) u_2^T u_2$. Since u_1 and u_2 are asymptotically normal, their joint distribution is

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{pmatrix} = \begin{pmatrix} W_1 G^T P G W_1 & W_1 G^T P E G W_2 \\ W_2 G^T E P G W_1 & W_2 G^T E P E G W_2 \end{pmatrix} \right).$$

The conditional distribution of u_2 , given u_1 , is

$$u_2 | u_1 \sim N(\Phi_{21} \Phi_{11}^{-1} u_1, \Phi_{22} - \Phi_{21} \Phi_{11}^{-1} \Phi_{12}).$$

Let λ_j be the eigenvalues of $\Phi_{22} - \Phi_{21} \Phi_{11}^{-1} \Phi_{12}$, $U \Lambda U^T = \Phi_{22} - \Phi_{21} \Phi_{11}^{-1} \Phi_{12}$ be the eigendecomposition, where U is an orthogonal matrix, and $\Lambda = \text{diag}\{\lambda_j\}$ is a $q \times q$ diagonal matrix, then the conditional distribution

$$u_2^T u_2 | u_1 \sim \sum_{j=1}^q \lambda_j \chi_{1,j}^2(\delta_j^2)$$

is the sum of non-central χ^2 s with 1 df, where δ_j is the j -th element of the vector $\Lambda^{-1/2} U^T \Phi_{21} \Phi_{11}^{-1} u_1$. The p value of the test $p_j = P(Q_j < q_j)$ satisfies

$$\begin{aligned} 1 - p_j &= P\left(\min\{p(\rho_1), p(\rho_2), \dots, p(\rho_S)\} > q_j\right) \\ &= P\left(\rho_s u_1^T u_1 + (1 - \rho_s) u_2^T u_2 < T(\rho_s), \forall s = 1, 2, \dots, S\right) \\ &= E\left\{P\left(\rho_s u_1^T u_1 + (1 - \rho_s) u_2^T u_2 < T(\rho_s), \forall s = 1, 2, \dots, S | u_1\right)\right\} \\ &= E\left\{P\left(u_2^T u_2 < \min_{s=1,2,\dots,S} \frac{T(\rho_s) - \rho_s u_1^T u_1}{1 - \rho_s} | u_1\right)\right\}. \end{aligned}$$

References

- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH: Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 2010;11:446–450.
- Kraft P, Yen YC, Stram DO, Morrison J, Gauderman WJ: Exploiting gene-environment interaction to detect genetic associations. *Hum Hered* 2007;63:111–119.
- Manning AK, LaValley M, Liu CT, Rice K, An P, Liu Y, Miljkovic I, Rasmussen-Torvik L, Harris TB, Province MA, Borecki IB, Florez JC, Meigs JB, Cupples LA, Dupuis J: Meta-analysis of gene-environment interaction: joint estimation of SNP and SNP $\times$ environment regression coefficients. *Genet Epidemiol* 2011;35:11–18.
- Aschard H, Hancock DB, London SJ, Kraft P: Genome-wide meta-analysis of joint tests for genetic and gene-environment interaction effects. *Hum Hered* 2010;70:292–300.
- Manning AK, Hivert MF, Scott RA, et al: A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat Genet* 2012;44:659–669.
- Morgenthaler S, Thilly WG: A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 2007;615:28–56.
- Li B, Leal SM: Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 2008;83:311–321.
- Madsen BE, Browning SR: A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 2009;5:e1000384.
- Morris AP, Zeggini E: An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 2010;34:188–193.
- Han F, Pan W: A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered* 2010;70:42–54.
- Hoffmann TJ, Marini NJ, Witte JS: Comprehensive approach to analyzing rare genetic variants. *PLoS One* 2010;5:e13584.
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;89:82–93.
- Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF: Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. *Am J Hum Genet* 2011;89:277–288.

Let

$$F(x) = P\left(u_2^T u_2 > \min_{s=1,2,\dots,S} \frac{T(\rho_s) - \rho_s u_1^T u_1}{1 - \rho_s} | u_1 = x\right)$$

be a function on R^q , which can be easily evaluated using the conditional distribution $u_2^T u_2 | u_1$, and $f(x)$ be the probability density function of the marginal distribution $u_1 \sim N(0, \Phi_{11})$, then the p value is

$$p_j = \int F(x) f(x) dx.$$

This is an integral on R^q . When q is small, we can use numeric methods to calculate the integral, such as an adaptive Gaussian-Hermite quadrature. When q is large, however, quadrature rules become impracticable due to the curse of dimensionality, thus we recommend using Monte Carlo methods to approximate the integral. We randomly sample B u_1 s from $N(0, \Phi_{11})$: $u_{1,1}, u_{1,2}, \dots, u_{1,B}$, then the p value is

$$p_j = \frac{1}{B} \sum_{b=1}^B F(u_{1,b}).$$

Acknowledgements

The authors would like to thank Dr. Xinyi Lin for sharing the R package iSKAT to perform GESAT-W. This research was partially supported by NIH awards R01 DK078616, U01 DK85526, and K24 DK080140. A portion of this research was conducted using the Linux Clusters for Genetic Analysis (LinGA) computing resources at Boston University Medicine Campus. The Framingham Heart Study is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (contract No. N01-HC-25195). This work was partially supported by a contract with Affymetrix, Inc. for genotyping services (contract No. N02-HL-6-4278).

- 14 Kzama R, Cardin NJ, Witte JS: Does accounting for gene-environment interactions help uncover association between rare variants and complex diseases? *Hum Hered* 2012;74: 205–214.
- 15 Lin X, Lee S, Christiani DC, Lin X: Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 2013;14:667–681.
- 16 Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;38:904–909.
- 17 Breslow NE, Clayton DG: Approximate Inference in Generalized Linear Mixed Models. *J Am Stat Assoc* 1993;88:9–25.
- 18 Zhang D, Lin X: Hypothesis testing in semi-parametric additive mixed models. *Biostatistics* 2003;4:57–74.
- 19 Lee S, Wu MC, Lin X: Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;13:762–775.
- 20 Voorman A, Brody J, Chen H, Lumley T: seqMeta: An R package for meta-analyzing region-based tests of rare DNA variants. 2014. <http://cran.r-project.org/web/packages/seqMeta/index.html>.
- 21 Imhof JP: Computing the distribution of quadratic forms in normal variables. *Biometrika* 1961;48:419–426.
- 22 Davies RB: Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *J R Stat Soc Ser C Appl Stat* 1980;29: 323–333.
- 23 Kuonen D: Saddlepoint approximations for distributions of quadratic forms in normal Variables. *Biometrika* 1999;86:929–935.
- 24 Liu H, Tang Y, Zhang HH: A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput Stat Data Anal* 2009;53:853–856.
- 25 Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M: Gene-trait similarity regression for multimarker-based association analysis. *Biometrics* 2009;65:822–832.
- 26 O'Sullivan F, Yandell BS, Raynor WJ Jr: Automatic smoothing of regression functions in generalized linear models. *J Am Stat Assoc* 1986;81:96–103.
- 27 Skogsberg J, Lundstrom J, Kovacs A, Nilsson R, Noori P, Maleki S, Kohler M, Hamsten A, Tegner J, Björkegren J: Transcriptional profiling uncovers a network of cholesterol-responsive atherosclerosis target genes. *PLoS Genet* 2008;4:e1000036.
- 28 Beaty TH, Ruczinski I, Murray JC, et al: Evidence for gene-environment interaction in a genome wide study of nonsyndromic cleft palate. *Genet Epidemiol* 2011;35:469–478.
- 29 Hamza TH, Chen H, Hill-Burns EM, Rhodes SL, Montimurro J, Kay DM, Tenesa A, Kusel VI, Sheehan P, Eaaswarkhanth M, Yearout D, Samii A, Roberts JW, Agarwal P, Bordelon Y, Park Y, Wang L, Gao J, Vance JM, Kendler KS, Bacanu SA, Scott WK, Ritz B, Nutt J, Factor SA, Zabetian CP, Payami H: Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet* 2011;7:e1002237.
- 30 Hancock DB, Artigas MS, Gharib SA, et al: Genome-wide joint meta-analysis of SNP and SNP-by-smoking interaction identifies novel loci for pulmonary function. *PLoS Genet* 2012;8:e1003098.