HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2017 July 10.

Published in final edited form as:

Genet Epidemiol. 2017 April ; 41(3): 233–243. doi:10.1002/gepi.22034.

Detecting association of rare and common variants based on cross-validation prediction error

Xinlan Yang[1], **Shuaichen Wang**[2], **Shuanglin Zhang**[1], and **Qiuying Sha**[1]

[1]Department of Mathematical Sciences, Michigan Technological University, Houghton, MI, USA

[2]BioStat Solutions, Inc., Frederick, MD, USA

Abstract

Despite the extensive discovery of disease-associated common variants, much of the genetic contribution to complex traits remains unexplained. Rare variants may explain additional disease risk or trait variability. Although sequencing technology provides a supreme opportunity to investigate the roles of rare variants in complex diseases, detection of these variants in sequencing-based association studies presents substantial challenges. In this article, we propose novel statistical tests to test the association between rare and common variants in a genomic region and a complex trait of interest based on cross-validation prediction error (PE). We first propose a PE method based on Ridge regression. Based on PE, we also propose another two tests PE-WS and PE-TOW by testing a weighted combination of variants with two different weighting schemes. PE-WS is the PE version of the test based on the weighted sum statistic (WS) and PE-TOW is the PE version of the test based on the optimally weighted combination of variants (TOW). Using extensive simulation studies, we are able to show that (1) PE-TOW and PE-WS are consistently more powerful than TOW and WS, respectively, and (2) PE is the most powerful test when causal variants contain both common and rare variants.

Keywords

rare variants; common variants; association studies; cross-validation prediction error; Ridge regression

1 | INTRODUCTION

The main purpose of genome-wide association studies (GWAS) is to detect common variants by indirect mapping methods. GWAS have identified a large number of common variants that are associated with complex diseases successfully (Bodmer & Bonilla, 2008; Lango Allen et al., 2010; Ng, Turner, & Robertson, 2009; Pritchard, 2001; Pritchard & Cox, 2002; Stratton & Rahman, 2008; Teer & Mullikin, 2010; Walsh & King, 2007). However, the common variants identified by GWAS only account for a small fraction of trait heritability (McCarthy et al., 2008), parts of the missing heritability could be caused by rare variants

Correspondence: Qiuying Sha, Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931. qsha@mtu.edu.

The authors have no conflict of interests to declare.

(Cohen et al., 2006; Ji et al., 2008; Manolio et al., 2009; Marini et al., 2008; Zhu, Feng, Li, Lu, & Elston, 2010). In rare variant association studies, instead of indirect association mapping method, all rare variants need to be tested directly. The new sequencing technology allows sequencing of exome-wide and whole genome of a large amount of individuals (Hodges et al., 2007), which makes directly test for rare variants possible (Andre's et al., 2007). Current exome-wide and whole genome sequencing studies have successfully detected many rare variants responsible for many complex traits, such as low-density lipoprotein (LDL) cholesterol (Lange et al., 2014), bone mineral density (Huang et al., 2015), thyroid function (Zheng et al., 2015), circulating lipid levels (Taylor et al., 2015), and other traits (Walter et al., 2015).

There is an increasing number of researchers who are interested in rare variants association studies (Ahituv et al., 2007; Cohen et al., 2004; Ji et al., 2008; Romeo et al., 2007, 2009). Because the well-developed common variant detecting methods are underpowered for rare variant association tests unless sample sizes or effect sizes are very large, several new methods for rare variant association studies are proposed recently. These methods include burden tests, quadratic tests, and robust tests. Burden tests include the cohort allelic sums test (CAST) (Morgenthaler & Thilly, 2007), the combined multivariate and collapsing (CMC) method (Li & Leal, 2008), the weighted sum statistic (WS) (Madsen & Browning, 2009), and variable threshold (VT) method (Price et al., 2010). Burden tests collapse rare variants in a genomic region into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants in the region (Lee et al., 2012). These tests implicitly assume that all rare variants are causal and that the directions of the effects are all the same. Quadratic tests include tests with statistics of quadratic forms of the score vector such as the sequence kernel association test (SKAT) (Wu et al., 2011), the SKAT for the combined effect of rare and common variants (SKAT-C) (Ionita-Laza, Lee, Makarov, Buxbaum, & Lin, 2013), the test for optimally weighted combination of variants (TOW) (Sha, Wang, Wang, & Zhang, 2012), as well as adaptive weighting (AW) methods such as data-adaptive sum (aSUM) (Han & Pan, 2010), AW methods (Sha, Wang, & Zhang, 2013), and methods proposed by Hoffmann, Marini, and Witte (2010), Lin and Tang (2011), and Yi and Zhi (2011). Quadratic tests are robust to the directions of the effects of causal variants and are less affected by neutral variants than burden tests. Burden tests can only outperform quadratic tests when most of rare variants are causal and the directions of the effects of causal variants are all the same. Robust tests include methods proposed by Derkach, Lawless, and Sun (2012), Greco et al. (2016), Lee et al. (2012), and Sha and Zhang (2014). Robust tests combine information from burden tests, quadratic tests, and possibly other tests aiming to have advantages of burden, quadratic, and possibly other tests.

In this paper, we develop novel statistical methods to test the association between common and rare variants in a genomic region and a complex trait of interest based on cross-validation prediction error (PE). We first propose a PE method based on Ridge regression. Based on PE, we also propose another two tests PE-WS and PE-TOW by testing a weighted combination of variants with two different weighting schemes, the weights suggested by Madsen and Browning (2009) and the optimal weighting scheme developed by Sha, Wang, Wang, and Zhang (2012). By extensive simulation studies, we show that (1) the PE versions of TOW and WS (PE-TOW and PE-WS) are consistently more powerful than TOW and WS,

respectively, and (2) PE is the most powerful test when the causal variants contain both common and rare variants.

2 | METHOD

2.1 | PE model

Consider a sample of n unrelated individuals. Each individual has been genotyped at M variants in a genomic region. Denote y_i as the value of a quantitative trait of the ith individual and denote $g_i = (g_{i1}, \ldots, g_{iM})^T$ as the genotypic scores of the ith individual at M variants, where $g_{im} \in \{0, 1, 2\}$ the number of minor alleles the ith individual has at the mth variant. We assume that there are no covariates. If there are p covariates, $z_{i1}, \ldots, z_{ip}$, for the ith individual, we just genotypes and trait values for the covariates using the method applied by Price et al. (2006) and Sha et al. (2012), that is, adjusting both genotypes and trait values for the covariates through linear models

$$y_i = \alpha_0 + \alpha_1 z_{i1} + \cdots + \alpha_p z_{ip} + \varepsilon_i \text{ and } g_{im} = \alpha_{0m} + \alpha_{1m} z_{i1} + \cdots + \alpha_{pm} z_{ip} + \tau_{im}.$$

Our working model is

$$y_i = \beta_0 + \beta_1 g_{i1} + \cdots + \beta_M g_{iM} + \varepsilon_i \quad (1)$$

To test association under our working model (1), we test the null hypothesis H_0: $\beta_1 = \cdots = \beta_M = 0$.

In the k-fold cross-validation, we divide the data into k equal parts, then use each of the k parts as the testing set and the other $k - 1$ parts as the training set. We use the training set to estimate $\beta = (\beta_0, \ldots, \beta_M)^T$ in equation (1), and use the prediction equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 g_{i1} + \cdots + \hat{\beta}_M g_{iM}$ to predict the trait values in the testing set. Because the genotype data of rare variants are sparse, the smaller the training set is, the more likely the problem of singular of the design matrix will be. Thus, we should try to use the training set as large as possible. In the k-fold cross-validation, the leave-one out- cross-validation (LOOCV) (k equals n) gives the largest training sets. Furthermore, the LOOCV PE has a closed-form formula (James, Witten, Hastie, & Tibshirani, 2013) (also see Appendix A for $\lambda = 0$). Therefore, our proposed tests are based on LOOCV.

In this paper, we construct a novel statistical test to test the association between genotypes of common and rare variants in a genomic region and a complex trait of interest based on the LOOCV PE. We propose to use the LOOCV PE under model (1) as a test statistic. Let $\hat{y}_{ci}$ denote the LOOCV predicted value of y_i under model (1). Then, the statistic can be written as

$$T=\sum_{i=1}^{n}(y_i-\hat{y}_{ci})^2 \quad (2)$$

Note that T is the LOOCV PE. Thus, low values of T would imply significance.

2.2 | Ridge regression

For rare variants, one drawback of the aforementioned LOOCV procedure is that some columns of the design matrix may have all zeros, if we leave one individual out. When the design matrix is not full rank or columns of the design matrix are highly correlated, we can use penalized regressions, such as Ridge regression (Halawa & Bassiouni, 2000; Hoerl, Kannard, & Baldwin, 1975) and Lasso regression (Meier, Van De Geer, & Bühlmann, 2008; Tibshirani, 1996; Yuan & Lin, 2006) among others. Penalized regressions have been applied to the analysis of genetic data (Ayers & Cordell, 2013, 2010; Cule & De Iorio, 2013; Cule, Vineis, & De Iorio, 2011; Malo, Libiger, & Schork, 2008; Warren, Casas, Hingorani, Dudbridge, & Whittaker, 2014). In this paper, we propose to use Ridge regression. Ridge regression penalizes the size of the regression coefficients. Let $x_i = (1, g_{i1}, \ldots, g_{iM})^T$ and $\beta = (\beta_0, \ldots, \beta_M)^T$. In the regression model $y_i=x_i^T\beta+\varepsilon_i$, $i = 1, 2, \ldots n$, the Ridge regression estimator $\hat{\beta}$ is defined as the value of β that minimizes $\sum_i(y_i-x_i^T\beta)^2+\lambda\sum_j\beta_j^2$,where $\lambda \geq 0$ is a tuning parameter. The solution to the Ridge regression is given by $\hat{\beta} = (X^T X + \lambda I)^{-1} X^T y$, where $X = (x_1, \ldots, x_n)^T$. The LOOCV PE for Ridge regression also has a closed-form formula (see Appendix A). For Ridge regression, we denote the test statistic given by equation (2) as T_λ. Let p_λ denote the P-value of T_λ, where p_λ is evaluated using equation (4) in the next section and $p_\lambda=p_\lambda^{(0)}$. We define the LOOCV PE test statistic as

$$T_{PE}=\min_\lambda p_\lambda. \quad (3)$$

In this study, we use a simple method to evaluate the minimization. We divide the interval $[0, \infty)$into subintervals $0 \leq \lambda_1 < \cdots < \lambda_{K-1} < \lambda_K < \infty$. In the simulation studies (see later), we used $K = 10$ and $(\log \lambda_1, \ldots, \log \lambda_{10}) = (1, \ldots, 10)$. Then, $T_{PE} = \min_\lambda p_\lambda = \min_{1 \leq k \leq K} p_{\lambda_K}$.

We use a permutation procedure to evaluate the P-value of T_{PE}. Intuitively, two layers of permutations are needed to estimate p_{λ_k} and the overall P-value for the test statistic T_{PE}. Ge, Dudoit, and Speed (2003) proposed that one layer of permutation can be used to estimate p_{λ_k} and the overall P-value for the test statistic T_{PE}. Here, we use the permutation procedure of Ge et al. to estimate p_{λ_k} and the overall P-value for the test statistic T_{PE}. In each permutation, we randomly shuffle the trait values. Suppose that we perform B replicates of permutations. Let $T_{\lambda_k}^{(b)}$ denote the values of T_{λ_k} based on the bth permuted data for $b = 0, 1,$

…, B and $k = 1, \ldots, K$, where $b = 0$ represents the original data. Then, we transfer $T_{\lambda_k}^{(b)}$ to $p_{\lambda_k}^{(b)}$ by

$$p_{\lambda_k}^{(b)} = \frac{\#\{d : T_{\lambda_k}^{(d)} < T_{\lambda_k}^{(b)} \text{ for } d = 1, \ldots, B\}}{f(b)}, \quad (4)$$

where $f(0) = B$ and $f(b) = B - 1$ for $b = 1, \ldots B$. Let $p^{(b)} = \min_{1 \le k \le K} p_{\lambda_k}^{(b)}$. Then, the P-value of T_{PE} is given by

$$\frac{\#\{b : p^{(b)} < p^{(0)} \text{ for } b = 1, 2, \ldots, B\}}{B}.$$

See Appendix B for a fast algorithm for the permutation procedure.

For testing the effects of common and rare variants, we also propose the following two methods based on the framework of PE. These two methods are to test the effect of a weighted combination of variants with two different weighting schemes:

1. Weighted sum weighting scheme: in this weighting scheme, we replace $g_i = (g_{i1}, \ldots, g_{iM})^T$ with $G_i = \sum_{m=1}^{M} w_m g_{im}$, where $w_m = \frac{1}{\sqrt{p_m(1-p_m)}}$ is the weight suggested by Madsen and Browning (2009) and p_m is the minor allele frequency of the mth variant. The test statistic given by equation (3) based on this weighting scheme is called weighted sum method based on PE (PE-WS).
2. Optimal weighting scheme: in this weighting scheme, we replace, $g_i = (g_{i1}, \ldots, g_{iM})^T$ with $G_i = \sum_{m=1}^{M} w_m g_{im}$, where $w_m = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(g_{im} - \bar{g}_m)}{\sum_{i=1}^{n} (g_{im} - \bar{g}_m)^2}$ is the weight suggested by Sha et al. (2012). The test statistic given by equation (3) based on this weighting scheme is called testing an optimally weighted combination of variants based on PE (PE-TOW).

We use the same permutation procedure as PE to evaluate the P-values of PE-WS and PE-TOW. See Appendix B for fast algorithms for the permutation procedures of PE-WS and PE-TOW.

2.3 | Comparison of tests

We compare the performance of the three proposed tests, PE, PE-WS, and PE-TOW, with that of the WS (Madsen & Browning, 2009), the SKAT (Wu et al., 2011), the SKAT for the combined effect of rare and common variants (SKAT-C) (Ionita-Laza, Lee, Makarov,

Buxbaum, & Lin, 2013), and the test for the optimally weighted combination of variants (TOW) (Sha et al., 2012).

3 | SIMULATION STUDIES

3.1 | Simulation

In simulation studies, we generate genotype data using the Genetic Analysis Workshop 17 (GAW17) data. This dataset contains genotypes of 697 unrelated individuals on 3,205 genes. Similar to Sha et al. (2012), we choose four genes: ELAVL4, MSH4, PDE4B, and ADAMTS4 with 10, 20, 30, and 40 variants, respectively, and then merge the four genes to form a super gene (Sgene) with 100 variants. We generate genotypes based on the genotypes of 697 individuals in the Sgene.

To evaluate type I error, we generate trait values independent of genotypes by using the model:

$$y=0.5Z_1+0.5Z_2+\varepsilon, \quad (5)$$

where Z_1 is a continuous covariate generated from a standard normal distribution, Z_2 is a binary covariate taking values 0 and 1 with a probability of 0.5, and ε follows a standard normal distribution.

To evaluate power, we randomly choose n_c rare variants and one common variant as causal variants and assume that all the n_c rare causal variants have the same heritability. n_r and n_p are the number of risk rare variants and protective rare variants, respectively, then $n_r + n_p = n_c$. Denote x_i^r, x_j^p, and x_c as the genotypes of the ith risk rare variant, the jth protective rare variant, and the common causal variant, respectively. Then, we generate a quantitative trait by the following model:

$$y=0.5Z_1+0.5Z_2+\sum_{i=1}^{n_r}\beta_i^r x_i^r-\sum_{j=1}^{n_p}\beta_j^p x_j^p+\beta_c x_c+\varepsilon, \quad (6)$$

where Z_1, Z_2, ε and are the same as those in equation (5). In equation (6), β_i^r, β_j^p, and β_c are constant coefficients. The values of β_i^r, β_j^p, and β_c depend on the total heritability h_{total} and the ratio of the heritability of rare causal variants to the heritability of the common causal variant R. For given h_{total} and R, based on equation (6), we can calculate the heritability of the rare casual variants and the heritability of the common causal variant. From the heritability of the common causal variant, we can calculate. β_c. From the heritability of the rare casual variants and the assumption that all the rare causal variants have the same heritability, we can calculate the heritability of each rare causal variant. Then, we can calculate β_i^r and β_j^p. The formulae to calculate the values of β_i^r, β_j^p, and β_c are given by $\beta_i^r=\sqrt{\frac{h_{total}R}{\mathrm{var}(x_i^r)n_c(1-h_{total})(1+R)}}$, $\beta_j^p=-\sqrt{\frac{h_{total}R}{\mathrm{var}(x_j^p)n_c(1-h_{total})(1+R)}}$, and $\beta_c=\sqrt{\frac{h_{total}}{\mathrm{var}(x_c)(1-h_{total})(1+R)}}$,

respectively. For power comparisons, we consider two different cases: (1) "Rare" in which all causal variants are rare (minor allele frequency < 0.01) and (2) "Both" in which both rare and common variants contribute to the trait. In each case, we consider two subcases: with covariates and without covariates. In the subcase of without covariates, Z_1, Z_2 are not included in equation (6).

3.2 | Simulation results

For evaluating the type I error of the proposed methods (PE, PE-TOW, and PW-WS), we consider different disease models (with or without covariates), different significance levels, and different sample sizes. The *P*-values are calculated using 10,000 permutations. Type I error rates are evaluated using 10,000 replicated samples. For 10,000 replicated samples, the 95% confidence intervals (CIs) for the estimated type I error rates of nominal levels 0.05, 0.01, and 0.001 are (0.046, 0.054), (0.008, 0.012), and (0.00038, 0.00162), respectively. The estimated type I error rates of the three proposed tests are summarized in Table 1. From this table, we can see that most of the estimated type I error rates are within 95% CIs and those type I error rates not within the 95% CIs are very close to the bound of the corresponding 95% CI, which indicates that the proposed methods are valid.

In power comparisons, the *P*-values of PE, PE-TOW, PEWS, and TOW are calculated using 1,000 permutations, while the *P*-values of WS, SKAT, and SKAT-C are calculated by asymptotic distributions. The powers of all of the seven tests are evaluated using 1,000 replicated samples at a significance level of 0.05 (Figs. 1–3). For Figure 4, the powers of all of the seven tests are evaluated using 1,000 replicated samples at a significance level of 10^{-6} and *P*-values of PE-WS, PE-TOW, PE, and TOW are evaluated by 10^7 permutations.

Power comparisons of the seven tests (PE, PE-TOW, TOW, PE-WS, WS, SKAT, and SKAT-C) for the power as a function of heritability are given in Figure 1. As shown in Figure 1, (1) PE-WS and PE-TOW are consistently more powerful than WS and TOW, respectively; (2) PE is the most powerful test when the causal variants contain both common and rare variants; and PE is the least powerful test when the causal variants are all rare variants; (3) TOW is more powerful than SKAT when the causal variants are all rare variants (MAF < 0.01) and TOW is less powerful than SKAT when the causal variants contain both common and rare variants. The reasons are that (a) TOW and SKAT have different weights, otherwise TOW and SKAT are same and (b) the weights of SKAT are larger than that of TOW only for those variants with MAF in the range (0.01,0.035), and the weights of TOW and SKAT are similar for those variants with MAF > 0.035; and (4) SKAT-C is less powerful than SKAT when the causal variants are all rare variants (MAF < 0.01) and SKAT-C is more powerful than SKAT when the causal variants contain both common and rare variants.

Power comparisons of the seven tests for the power as a function of the percentage of protective variants and for the power as a function of the percentage of causal variants are given in Figures 2 and 3, respectively. These two figures show that the powers of PE, PE-TOW, SKAT, SKAT-C, and TOW are robust to the percentage of protective variants and the percentage of causal variants while powers of PE-WS and WS decrease with the increasing of the percentage of protective variants and increase with the increasing of the percentage of causal variants. Other patterns of power comparisons are similar to that in Figure 1. We also

provide power comparisons of the seven tests using a small significance level of 10^{-6} Fig. 4) and using a large sample size of 5,000 (Fig. 5). Figure 4 shows that the patterns of the power comparisons using significance level 10^{-6} are similar to that using a significance level of 0.05 in Figure 1 (Both; Without covariates). Figure 5 shows that the patterns of the power comparisons using a sample size of 5,000 are similar to that using a sample size of 1,000 in Figure 1 (Both; Without covariates).

In summary, PE-WS and PE-TOW are consistently more powerful than WS and TOW, respectively. When causal variants contain both common and rare variants, PE is the most powerful test, SKAT-C is more powerful than SKAT, and SKAT is more powerful than TOW. When causal variants are all rare variants, TOW is more powerful than SKAT, and SKAT is more powerful than SKAT-C. The powers of PE, PE-TOW, SKAT, SKAT-C, and TOW are robust to the percentage of protective variants and the percentage of causal variants.

4 | ANALYSIS OF THE GAW17 SIMULATED DATASET

The GAW17 simulated dataset consists of a collection of 697 unrelated individuals, their real genotypes, and 200 replicates of the simulated phenotypes. Three quantitative traits Q1, Q2, and Q4 are simulated. Covariates include age, sex, and smoking status. Because quantitative trait Q4 has no genetic components, we do not consider Q4 for the purpose of power comparisons. We perform power comparisons of the seven tests using quantitative traits Q1 and Q2. The *P*-values of TOW, PE-TOW, PE-WS, and PE are evaluated by 10,000 permutations and the *P*-values of WS, SKAT-C, and SKAT are evaluated by asymptotic distributions. The powers of the seven tests are calculated at a significance level of 0.001. We merge every two replicates to one replicate to increase the sample size. In all cases, the minor allele is associated with higher means of the two quantitative traits, which means that all causal variants are risk variants. Q1 has nine causal genes and Q2 has 13 causal genes. We omit causal genes that have one variant, causal genes in which all of the seven tests have 100% power, and causal genes in which all of the seven tests have a power less than 10%. Q1 has five causal genes left and Q2 has seven causal genes left. The powers of TOW, WS, and SKAT to test the association between each of the five causal genes and Q1 are not consistent with that in Table 2 of Sha et al. (2012) because we found that Sha et al. (2012) did not adjust trait values and genotypes for covariates when testing the association for Q1.

The powers of the seven tests to detect association between each of the 12 causal genes and Q1 or Q2 are given in Table 2. As shown in Table 2, WS, TOW, or SKAT-C is the most powerful test in one out of 12 genes, PE-WS, PE-TOW, or SKAT is the most powerful test in two out of 12 genes, and PE is the most powerful test in four out of 12 genes. Three out of four causal genes in which PE or SKAT-C is the most powerful test include common causal variants. Causal variants in the six genes in which TOW, WS, PE-TOW, or PE-WS is the most powerful test are all rare variants (MAF < 1%). Each of the two genes in which SKAT is the most powerful test contains causal variants with MAF in (0.01,0.035). Comparing TOW and WS with PE-TOW and PE-WS, PE-TOW is more powerful than TOW and PE-WS is more powerful than WS in nine out of 12 causal genes. The results from the analysis of the GAW17 simulated dataset are consistent with those from the simulation studies.

5 | DISCUSSION

Based on cross-validation PE under Ridge regression, we developed novel statistical tests to test the association between variants (both common and rare variants) in a genomic region and a complex trait of interest. We proposed PE method based on Ridge regression. Combined with the weighting schemes, we further developed PE-WS and PE-TOW methods. We used extensive simulation studies to compare the performance of PE, PE-WS, and PE-TOW with that of the existing methods: SKAT, SKAT-C, WS, and TOW. Our results showed that (1) the PE versions of TOW and WS (PE-TOW and PE-WS) are consistently more powerful than TOW and WS, respectively; (2) when causal variants contain both common and rare variants, PE is the most powerful test, SKAT-C is more powerful than SKAT, and SKAT is more powerful than TOW. When causal variants are all rare variants, TOW is more powerful than SKAT, and SKAT is more powerful than SKAT-C; and (3) the powers of PE, PE-TOW, SKAT, SKAT-C, and TOW are robust to the percentage of protective variants and the percentage of causal variants.

Each of the three proposed methods PE, PE-TOW, and PEWS has its advantages in some scenarios. PE is more powerful than PE-TOW and PE-WS when causal variants contain both common and rare variants. PE-WS is a burden test and is more powerful than PE-TOW when the percentage of causal variants is large and the directions of the effects of the causal variants are all the same. PE-TOW is more powerful than PE-WS when the percentage of causal variants is small or the directions of the effects of the causal variants are different. We may construct a robust test aiming to have the advantages of all of PE, PE-TOW and PE-WS. Let p_{PE}, $p_{PE\text{–Tow}}$, and $p_{PE\text{–}WS}$ denote the P-values of PE, PE-TOW, and PE-WS, respectively. Then, we define the test statistic of the robust test as $T_{robust} = \min\{p_{PE}, p_{PE\text{–}TOW}, p_{PE\text{–}WS}\}$. However, the performance of the robust test needs further investigation.

PE test statistic does not work well for rare variants. The reason is that many rare variants are singletons. From our simulation results, PE method may be more powerful than existing methods for common variants. The performance of PE for common variants needs further investigation.

Among the three proposed tests (PE, PE-WS, and PE-TOW), PE is most computationally intensive. The computation time required for running PE depends on the sample size, the number of variants in the genomic region, and the number of permutations. The running time of PE with 1,000 permutations on a dataset with 1,000 individuals and 100 variants in a genomic region on a laptop with 4 Intel Cores @ 3.30GHz and 4 GB memory is about 0.1 sec. To perform GWAS, we can first select genomic regions that show evidence of association based on a small number of permutations (e.g. 1,000), and then a large number of permutations are used to test the selected regions.

Acknowledgments

Research reported in this article was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R15HG008209. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences.

Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org).

References

Ahituv N, Kavaslar N, Schackwitz W, Ustaszewska A, Martin J, Hébert S, … Pennacchio LA. Medical sequencing at the extremes of human body mass. American Journal of Human Genetics. 2007; 80:779–791. [PubMed: 17357083]

Andre's A, Clark A, Shimmin L, Boerwinkle E, Sing C, Hixson J. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. Genetics Epidemiology. 2007; 31:659–671.

Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genetic Epidemiology. 2010; 34:879–891. [PubMed: 21104890]

Ayers KL, Cordell HJ. Identification of grouped rare and common variants via penalized logistic regression. Genetic Epidemiology. 2013; 37:592–602. [PubMed: 23836590]

Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. Nature Genetics. 2008; 40(6):695–701. [PubMed: 18509313]

Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science. 2004; 305:869–872. [PubMed: 15297675]

Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, … Affiliations A. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:1810–1815. [PubMed: 16449388]

Cul E, Vineis P, De Iorio M. Significance testing in Ridge regression for genetic data. BMC Bioinformatics. 2011; 12:372. [PubMed: 21929786]

Cule E, De Iorio M. Ridge regression in prediction problems: Automatic choice of the ridge parameter. Genetic Epidemiology. 2013; 37:704–714. [PubMed: 23893343]

Derkach A, Lawless J, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. Genetic Epidemiology. 2012; 37(1):110–121. [PubMed: 23032573]

Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis. Test. 2003; 12:1–77.

Greco B, Hainline A, Arbet J, Grinde K, Benitez A, Tintle N. A general approach for combining diverse rare variant association tests provides improved robustness across a wider range of genetic architectures. European Journal of Human Genetics. 2016; 24:767–773. [PubMed: 26508571]

Halawa AM, El Bassiouni MY. Tests of regression coefficients under ridge regression models. Journal of Statistical Computation and Simulation. 2000; 65:341–356.

Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. Human Heredity. 2010; 70:42–54. [PubMed: 20413981]

Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW, … McCombie WR. Genome-wide in situ exon capture for selective resequencing. Nature Genetics. 2007; 39:1522–1527. [PubMed: 17982454]

Hoerl AE, Kannard RW, Baldwin KF. Ridge regression: Some simulations. Communications in Statistics—Theory and Methods. 1975; 4:105–123.

Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. PLoS One. 2010; 5(11):e13584. [PubMed: 21072163]

Huang J, Howie B, McCarthy S, Memari Y, Walter K, Min JL, … Soranzo N. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. Nature Communications. 2015; 6:8111.doi: 10.1038/ncomms9111

Ionita-Laza I, Lee S, Makarov V, Buxbaum J, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. American Journal of Human Genetics. 2013; 92:841–853. [PubMed: 23684009]

James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning. New York Heidelberg Dordrecht London: Springer; 2013.

Ji W, Foo JN, O'Roak BJ, Zhao H, Larson MG, Simon DB, … Lifton RP. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nature Genetics. 2008; 40:592–599. [PubMed: 18391953]

Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, Tang ZZ. … Project NGOES. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. American Journal of Human Genetics. 2014; 94(2):233–245. [PubMed: 24507775]

Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, Rivadeneira F, … Hirschhorn JN. Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature. 2010; 467:832–838. [PubMed: 20881960]

Lee S, Eond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, … Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. American Journal of Human Genetics. 2012; 91:224–237. [PubMed: 22863193]

Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. American Journal of Human Genetics. 2008; 83:311–321. [PubMed: 18691683]

Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. American Journal of Human Genetics. 2011; 89:354–367. [PubMed: 21885029]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genetics. 2009; 5:e1000384. [PubMed: 19214210]

Malo N, Libiger O, Schork NJ. Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. American Society of Human Genetics. 2008; 82:375–385.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, … Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009; 461:747–753. [PubMed: 19812666]

Marini NJ, Gin J, Ziegle J, Keho KH, Ginzinger D, Gilbert DA, Rine J. The prevalence of folate-remedial MTHFR enzyme variants in humans. Proceedings of the National Academy of Sciences of the United States of America. 2008; 105:8055–8060. [PubMed: 18523009]

McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhornet JN. Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. Nature Reviews Genetics. 2008; 9:356–369.

Meier L, Van De Geer S, Bühlmann P. The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70:53–71.

Morgenthaler S, Thilly WG. A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). Mutation Research. 2007; 615:28–56. [PubMed: 17101154]

Ng SB, Turner EH, Robertson PD. Targeted capture and massively parallel sequencing of 12 human exomes. Nature Letters. 2009; 461:272–276.

Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Lee-JenWei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. American Journal of Human Genetics. 2010; 86:832–838. [PubMed: 20471002]

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics. 2006; 38:904–909. [PubMed: 16862161]

Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? American Journal of Human Genetics. 2001; 69:124–137. [PubMed: 11404818]

Pritchard JK, Cox NJ. The allelic architecture of human disease genes: Common disease-common variant…or not? Human Molecular Genetics. 2002; 11:2417–2423. [PubMed: 12351577]

Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, Hobbs HH, Cohen CJ. Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. Nature Genetics. 2007; 39:513–516. [PubMed: 17322881]

Romeo S, Yin W, Kozlitina J, Pennacchio LA, Boerwinkle E, Hobbs HH, Cohen CJ. Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. Journal of Clinical Investigation. 2009; 119:70–79. [PubMed: 19075393]

Sha Q, Wang S, Zhang S. Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. European Journal of Human Genetics. 2013; 21(3):332–337. [PubMed: 22781093]

Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing optimally weighted combination of variants. Genetic Epidemiology. 2012; 36:561–571. [PubMed: 22714994]

Sha Q, Zhang S. A rare variant association test based on combinations of single-variant tests. Genetic Epidemiology. 2014; 38:494–501. [PubMed: 25065727]

Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. Nature Genetics. 2008; 40:17–22. [PubMed: 18163131]

Taylor PN, Porcu E, Chew S, Campbell PJ, Traglia M, Brown SJ. Consortium UK. Whole-genome sequence-based analysis of thyroid function. Nature Communications. 2015; 6:5681.

Teer JK, Mullikin JC. Exome sequencing: The sweet spot before whole genomes. Human Molecular Genetics. 2010; 19(R2):R145–51. [PubMed: 20705737]

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B (Methodological). 1996; 58:267–288.

Walsh T, King MC. Ten genes for inherited breast cancer. Cancer Cell. 2007; 11:103–105. [PubMed: 17292821]

Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, … Soranzo N. The UK10K project identifies rare variants in health and disease. Nature. 2015; 526(7571):82–90. [PubMed: 26367797]

Warren H, Casas JP, Hingorani A, Dudbridge F, Whittaker J. Genetic prediction of quantitative lipid traits: Comparing shrinkage models to gene scores. Genetic Epidemiology. 2014; 38:72–83. [PubMed: 24272946]

Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. American Journal of Human Genetics. 2011; 89:82–93. [PubMed: 21737059]

Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. Genetic Epidemiology. 2011; 35:57–69. [PubMed: 21181897]

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2006; 68:49–67.

Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello-Diez A, Leo PJ, … Richards JB. Whole-genome sequencing identifies EN1 as a determinant of bone density and fracture. Nature. 2015; 526(7571): 112–117. [PubMed: 26367794]

Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. Genetic Epidemiology. 2010; 34:171–187. [PubMed: 19847924]

APPENDIX A. The closed-form formula of cross-validation prediction error of LOOCV for Ridge regression

Let $x_I = (1, g_{i1}, \ldots g_{iM})^T$ and $X = (x_1, \ldots, x_n)^T$. Let $A_\lambda = (X^TX + \lambda I)^{-1}$, $h_i^\lambda = x_i^T A_\lambda x_i$, and $\hat{\beta}_\lambda = A_\lambda X^T y$. Let $\hat{y}_i^\lambda = x_i^T \hat{\beta}_\lambda$ and $B_\lambda = XA_\lambda X^T$, then $h_\lambda = (h_1^\lambda, \ldots, h_n^\lambda) = diag(B_\lambda)$. Let X_{-i}, $\hat{\beta}_{\lambda,-i}$, and $\hat{y}_{ci}^\lambda$ denote X, $\hat{\beta}_\lambda$, and $\hat{y}_i^\lambda$ when the ith individual leaves out. Noting that $X_{-i}^T X_{-i} = X^T X - x_i x_i^T$, then we have

$$A_{\lambda,-i}=\left(X^TX+\lambda I-x_ix_i^T\right)^{-1}$$
$$=A_\lambda+\frac{1}{1-x_i^TA_\lambda x_i}A_\lambda x_ix_i^TA_\lambda,$$
$$\hat{\beta}_{\lambda,-i}=A_{\lambda,-i}X_{-i}^Ty_{-i}=\left(A_\lambda+\frac{1}{1-a_i^\lambda}A_\lambda x_ix_i^TA_\lambda\right)\left(X^Ty-x_iy_i\right)$$
$$=A_\lambda X^Ty-A_\lambda x_iy_i+\frac{1}{1-h_i^\lambda}A_\lambda x_ix_i^TA_\lambda X^Ty-\frac{h_i^\lambda}{1-h_i^\lambda}A_\lambda x_iy_i,$$
$$\hat{y}_{ci}^\lambda=x_i^T\hat{\beta}_{\lambda,-i}=x_i^T\left(A_\lambda X^Ty-A_\lambda x_iy_i+\frac{1}{1-h_i^\lambda}A_\lambda x_ix_i^TA_\lambda X^Ty-\frac{h_i^\lambda}{1-h_i^\lambda}A_\lambda x_iy_i\right)$$
$$=\hat{y}_i^\lambda-h_i^\lambda y_i+\frac{h_i^\lambda}{1-h_i^\lambda}\hat{y}_i^\lambda-\frac{\left(h_i^\lambda\right)^2}{1-h_i^\lambda}y_i=\frac{1}{1-h_i^\lambda}\hat{y}_i^\lambda-\frac{h_i^\lambda}{1-h_i^\lambda}y_i.$$

Therefore,

$$y_i=\hat{y}_{ci}^\lambda=\frac{1}{1-h_i^\lambda}\left(y_i-\hat{y}_i^\lambda\right).$$

APPENDIX B. The fast algorithms for permutation procedures

PE method

We use the same notations as in Appendix A. Let $\hat{y}_\lambda=(\hat{y}_1^\lambda,\ldots,\hat{y}_n^\lambda)^T$ and then $\hat{y}_\lambda = X(X^TX + \lambda I)^{-1}X^Ty$. For two matrices or vectors A and B, we use $A \times B$ and $\frac{A}{B}$ to denote the element-wise operations. Let m denote the number of columns of matrix X. We assume $n \geq m$. We perform singular value decomposition of X, that is, $X = UDV$, where U is an $n \times m$ matrix with orthonormal columns, D is $m \times m$ diagonal matrix with non-negative real numbers on the diagonal, and V is an $m \times m$ orthogonal matrix. Let $D = diag(d_1, \ldots, d_m)$.

Then $\hat{y}_\lambda = UC_\lambda U^Ty$, where $C_\lambda = diag(c_{\lambda,1}, \ldots, c_{\lambda,m})$ and $c_{\lambda,j}=\frac{d_j^2}{d_j^2+\lambda}$ for $j = 1, \ldots, m$. Let $c_\lambda = (c_{\lambda,1}, \ldots, c_{\lambda,m})^T$ and $y^{(m)} = U^Ty$ be a m-dimensional vector. Then, $\hat{y}_\lambda = UC_\lambda y^{(m)} = U(c_\lambda \times y^{(m)})$ and $h_\lambda = diag(UC_\lambda U^T)$ (in R code, $h_\lambda = rowSums(U \times t(t(U) \times c_\lambda))$). For $0 \leq \lambda_1 < \cdots < \lambda_{K-1} < \lambda_K < \infty$, let $C = c_{\lambda_1}, \ldots, c_{\lambda_K})$ and $H = (h_{\lambda_1}, \ldots, h_{\lambda_K})$. Then, $(\hat{y}_{\lambda_1}, \ldots, \hat{y}_{\lambda_K}) = U(C \times y^{(m)} = U(c_{\lambda_1} \times y^{(m)}, \ldots, c_{\lambda_K} * y^{(m)})$. If we denote $B=\frac{(y-\hat{y}_{\lambda_1},\ldots,y-\hat{y}_{\lambda_K})}{1-H}$, then $(T_{\lambda_1}, \ldots, T_{\lambda_K}) = colSums(B \times B)$. Note that C, U, and H do not change in each permutation.

PE-TOW and PE-WS methods

For PE-TOW, let $X = (x_1, \ldots, x_n)^T$ where $xi = Gi - \bar{G}$. We first centralize the trait values y For simplicity, we still use $y = (y_1, \ldots, y_n)^T$ to denote the trait values after centralization. Let $d=X^TX=\sum_{i=1}^n x_i^2$, $Y=X^Ty=\sum_{i=1}^n x_iy_i$, and $c_\lambda=\frac{1}{d+\lambda}$. Then, $\hat{y}_\lambda = Yc_\lambda X$ and $h_i^\lambda=c_\lambda x_i^2$. Note that we need to recalculate X in each permutation.

For PE-WS, the same formulas for PE-TOW are applied. However, X does not change in each permutation.

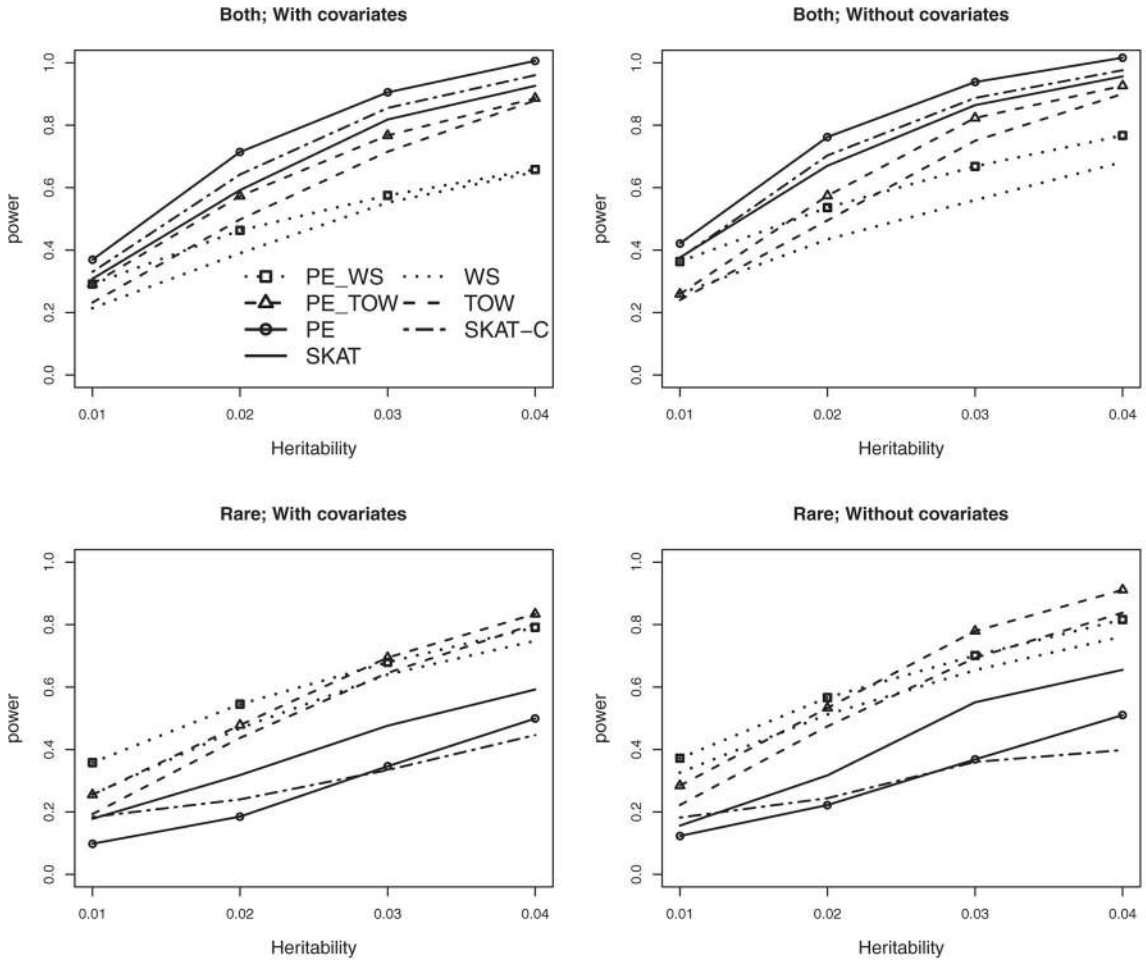

FIGURE 1.

Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT, and SKAT-C) for the power as a function of heritability. "Rare" means that all causal variants are rare. "Both" means that causal variants contain both rare and common (one common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants *x*-axis represents the total heritability of all causal variants. Sample size is 1,000. In this set of simulations, all causal variants are risk variants and 20% of rare variants are causal. The powers are evaluated at a significance level of 0.05

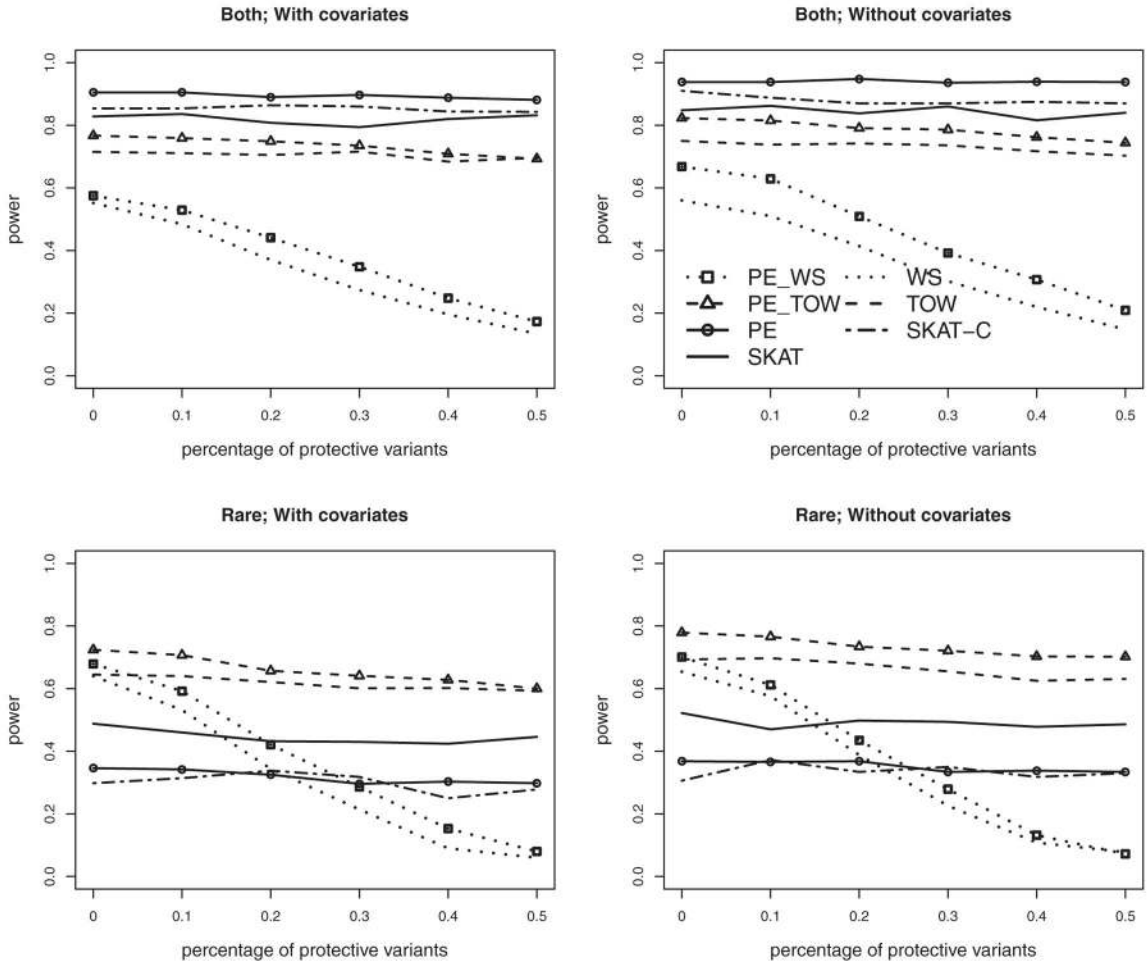

FIGURE 2.

Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT, and SKAT-C) for the power as a function of the percentage of protective variants. "Rare" means that all causal variants are rare. "Both" means that causal variants contain both rare and common (one common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants. *x*-axis represents the percentage of protective variants. Sample size is 1,000. The total heritability is 0.03. Twenty percent of rare variants are causal. The powers are evaluated at a significance level of 0.05

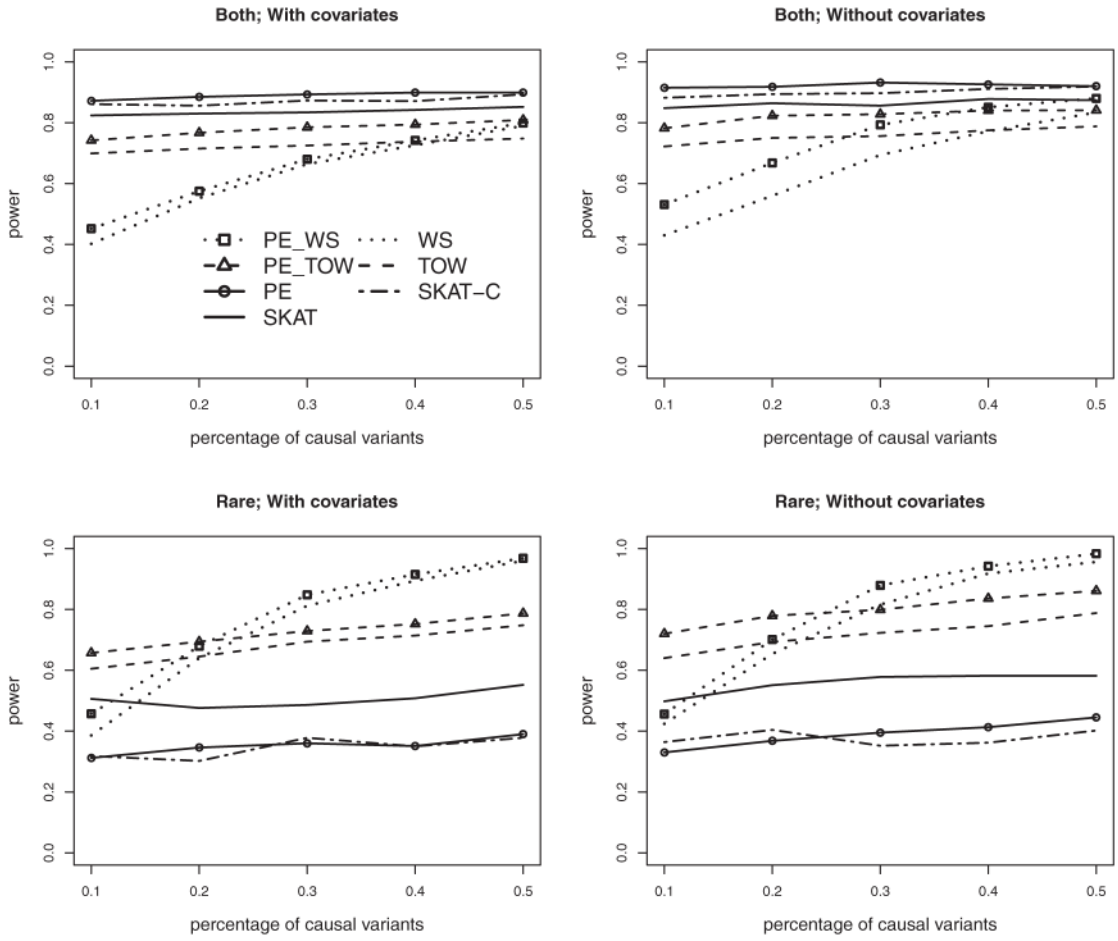

FIGURE 3.

Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT, and SKAT-C) for the power as a function of the percentage of causal variants. “Rare” means that all causal variants are rare. “Both” means that causal variants contain both rare and common (one common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants. *x*-axis represents the total heritability of all causal variants. Sample size is 1,000. The total heritability is 0.03. All causal variants are risk variants. The powers are evaluated at a significance level of 0.05

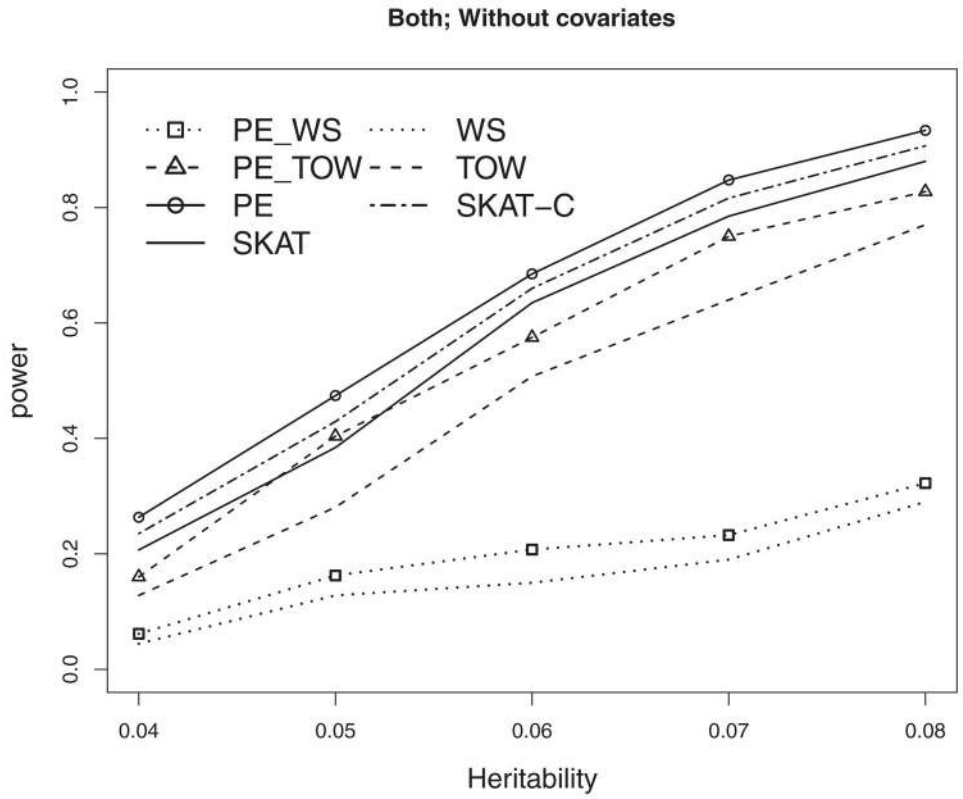

FIGURE 4.

Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT, and SKAT-C) for the power as a function of heritability. "Both" means that causal variants contain both rare and common (one common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants *x*-axis represents the total heritability of all causal variants. Sample size is 1,000. In this set of simulations, all causal variants are risk variants and 20% of rare variants are causal. Powers are evaluated at significance level 10^{-6} and *P*- values of PE-WS, PE-TOW, PE, and TOW are evaluated by 10^{7} permutations

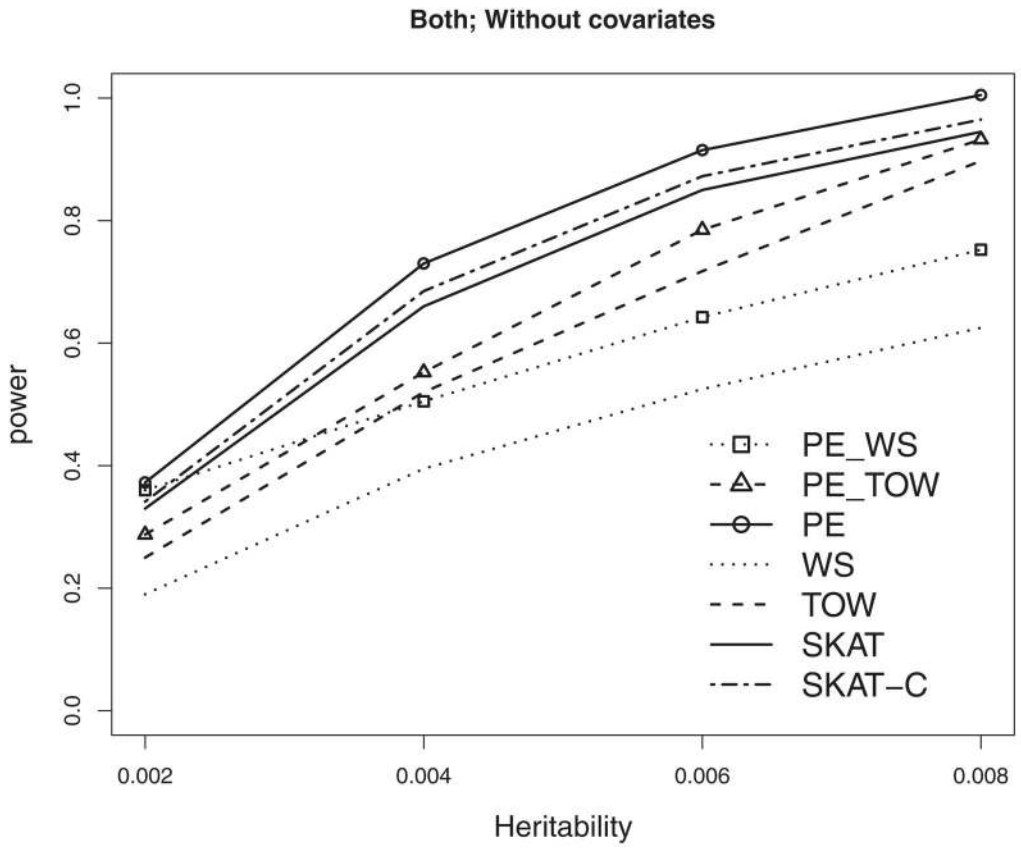

FIGURE 5.

Power comparisons of the seven tests (PE-WS, PE-TOW, PE, WS, TOW, SKAT, and SKAT-C) for the power as a function of heritability. "Both" means that causal variants contain both rare and common (one common variant) and the heritability of the common variant is as twice as the heritability of all the rare causal variants *x*-axis represents the total heritability of all causal variants. Sample size is 5,000. In this set of simulations, all causal variants are risk variants and 20% of rare variants are causal. The powers are evaluated at a significance level of 0.05

TABLE 1

Type I error rates of the three proposed methods with 10,000 replicates

Significance level	Sample size	With covariates			Without covariates		
		PE-WS	PE-TOW	PE	PE-WS	PE-TOW	PE
0.05	500	0.0545	0.0485	0.0525	0.049	0.0506	0.0504
	1,000	0.0503	0.051	0.0519	0.0493	0.0517	0.05
0.01	500	0.0104	0.0091	0.0107	0.0099	0.0088	0.0103
	1,000	0.0112	0.010	0.0102	0.009	0.0097	0.0103
0.001	500	0.0007	0.0009	0.0006	0.0008	0.001	0.0011
	1,000	0.0017	0.0005	0.0016	0.0011	0.0009	0.0008

TABLE 2

Power of the seven tests to detect the association between each of the five causal genes and quantitative trait Q1 and between each of the seven causal genes and quantitative trait Q2

Traits	Gene name	No. of variants, no. of causal variant	Min, max, mean MAF	WS	TOW	PE-WS	PE-TOW	PE	SKAT-C	SKAT
Q1	ARNT	18, 5	0.07, 1.15, 0.33	0.08	0.52	0.05	0.55	0.83	0.84	**0.95**
	ELAVL4	10, 2	0.07, 0.07, 0.07	0.44	0.40	0.48	0.53	**0.72**	0.26	0.00
	FLT4	10, 2	0.07, 0.14, 0.11	**0.85**	0.68	0.70	0.51	0.79	0.14	0.68
	HIF1A	8, 4	0.07, 1.22, 0.39	0.63	0.56	0.46	0.43	**0.91**	0.66	0.88
	VEGFA	6, 1	0.22, 0.22, 0.22	0.33	0.16	**0.45**	0.22	0.22	0.23	0.10
Q2	BCHE	29, 13	0.07, 0.29, 0.10	0.20	0.34	0.23	**0.38**	0.27	0.02	0.14
	LPL	20, 3	0.07, 1.58, 0.60	0.01	0.23	0.05	0.28	0.16	0.33	**0.41**
	PDGFD	11, 4	0.07, 0.86, 0.29	0.07	0.23	0.09	**0.25**	0.15	0.04	0.15
	SIRT1	24, 9	0.07, 0.22, 0.12	0.50	**0.65**	0.52	0.60	0.41	0.61	0.55
	SREBF1	24, 10	0.07, 0.43, 0.22	0.26	0.20	**0.27**	0.25	0.19	0.03	0.06
	VNN1	7, 2	0.57, 17.1, 8.82	0.06	0.68	0.08	0.78	**0.95**	**0.95**	0.02
	VNN3	15, 7	0.07, 9.83, 2.06	0.33	0.55	0.35	0.66	**0.73**	0.68	0.40

Note: Min, max, and mean MAF represent the minimum, maximum, and mean MAF (in percentage) at the causal variants. In each row, the boldfaced number represents the highest power in the row