Published in final edited form as:

Stat Methods Med Res. 2019 September ; 28(9): 2868–2875. doi:10.1177/0962280218791205.

A novel gene-set association test based on variance-gamma distribution

Zhongxue Chen^{1,*}, Qingzhong Liu², Kai Wang³

¹Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington. 1025 E. 7th street, Bloomington, IN, 47405, USA.

²Department of Computer Science, Sam Houston State University, 1803 Avenue I. Huntsville, Texas 77341, USA

³Department of Biostatistics, College of Public Health, University of Iowa. 145 N. Riverside Drive, Iowa City, IA, 52242, USA.

Abstract

Several gene- or set-based association tests have been proposed recently in the literature. Powerful statistical approaches are still highly desirable in this area. In this paper we propose a novel statistical association test, which uses information of the burden component and its complement from the genotypes. This new test statistic has a simple null distribution, which is a special and simplified variance-gamma distribution, and its p-value can be easily calculated. Through a comprehensive simulation study, we show that the new test can control type I error rate and has superior detecting power compared with some popular existing methods. We also apply the new approach to a real data set; the results demonstrate that this test is promising.

Keywords

burden test; genetic association; SKAT; variance-gamma distribution

1. Introduction

Detecting associated genetic variants is still a challenging but important step when we study some common diseases. Nowadays hundreds of genome-wide association studies (GWASs) have been conducted and a huge amount of genetic data are available for researchers to identify significant genetic factors which play an important role in understanding the diseases' etiology and finding possible treatments. To eliminate false positives, very stringent significance levels are usually chosen when we perform statistical tests to detect significant single nucleotide polymorphisms (SNPs) from a GWAS. As a consequence, we may also lose power in detecting true positives. This is especially true for SNPs with small effect sizes and/or low allele frequencies. One solution to overcome the abovementioned

*Corresponding author: Zhongxue Chen, PhD Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington. 1025 E. 7th street, PH C104, Bloomington, IN, 47405, USA Phone:(812) 855-1163 Fax: (812) 855-7109 zc3@indiana.edu.

difficulties is to perform statistical tests based on a set of genetic variants (e.g., gene). Gene- or set-based statistical analyses can use less stringent significance levels as fewer tests are conducted. In addition, the statistical detecting power can potentially increase when accumulative effects from a set of variants are investigated together.

Recently, several gene- or SNP-set-based association tests have been proposed in the literature^{1–11}. In general, those methods are based on the overall association between the adjusted phenotype and the transformed genotype. The burden test⁷, the sequencing kernel association test (SKAT)⁵, and the optimal SKAT (SKAT-O)⁴ are among the most popular ones in this area. The SKAT can be viewed as a method based on principal component analysis¹⁰. The association between each principal component and the adjusted phenotype is measured and then an overall association is calculated from the weighted individual associations with the weights determined by the genotype data. The burden test can be viewed as a special case of SKAT where the transformed genotype is just the sum of genotypes among the SNPs in a set or a gene⁵. It has been shown that under some conditions, the burden test may be more powerful than the SKAT⁴. Because of this, the SKAT-O uses a linear combination of the statistics from the burden test and the SKAT to make it more robust. Unfortunately, under the null hypothesis the burden test statistic and the SKAT statistic are correlated⁴. As a consequence, the null distribution of the SKAT-O statistic is intractable and numeric method is required to estimate the p-value⁴. More seriously, our study shows that the SKAT-O may have inflated type I error rate (see the simulation study section).

In this report, we propose a novel statistical test which combines the information from the burden test and its complement in an efficient way so that the new test statistic has a simple null distribution from which the p-value can be easily calculated. In addition, through a comprehensive simulation study, we show that the proposed test is more powerful than others under many situations. We also apply our new test to a real data to demonstrate its application.

2. Methods

In this report, we use the following notations: m , the number of SNPs in a set, n , the number of subjects in the study, $y = (y_1, y_2, \dots, y_n)^T$, the adjusted phenotypes (standardized residuals from a GLM which adjusts for non-genetic co-variables), where T denotes the matrix transpose. We use $n \times m$ matrix $G_{n \times m}$ to denote the genotype data, where the element of G , $(g_{ij}) = 0, 1$, or 2 ($i = 1, 2, \dots, n; j = 1, 2, \dots, m$), the number of minor alleles of the j^{th} SNP from the i^{th} subject. Note that asymptotically, y can be treated as n independent samples from a standard normal distribution $N(0,1)$. In addition, we define $K = GG^T$.

2.1 Some existing tests

The SKAT statistic is defined as⁵

$$SKAT = y^T K y \quad (1)$$

It can be shown that under the null hypothesis that none of the m SNPs is associated with the adjusted phenotype y , $SKAT$ asymptotically follows a linear combination of chi-square distributions. More specifically, $SKAT \sim \sum_{i=1}^k \lambda_i \chi_{i,1}^2$ where λ_i are the none-zero eigenvalue of K and $\chi_{i,1}^2$ are independently identically distributed (iid) chi-square distribution with degree of freedom (df) equals to 1. The set of none-zero eigenvalues λ_i can also be calculated from the $m \times m$ matrix $G^T G$ as $G^T G$ and GG^T have the same set of none-zero eigenvalues. The latter calculation is more efficient when $n > m$. The p-value for $SKAT$ can be estimated using the Davis method^{5,12}.

The burden test statistic is defined as⁷

$$B = y^T G \mathbf{1} \mathbf{1}^T G^T y \quad (2)$$

Denote $\bar{z} = (\bar{z}_1, \bar{z}_2, \dots, \bar{z}_n)^T$ where $\bar{z}_i = \sum_{j=1}^m g_{ij}/m$ and g_{ij} is the (i, j) element of matrix G , then $\bar{z} = G \mathbf{1}/m$ where $\mathbf{1}$ is the $m \times 1$ column vector with all elements equal to 1. Let $\bar{z} \bar{z}^T / (\bar{z}^T \bar{z}) = G \mathbf{1} \mathbf{1}^T G^T / \mathbf{1}^T G^T G \mathbf{1}$, then $M = M^T$ and $M \bar{z} = \bar{z}$. The burden test statistic is proportional to the following statistic

$$B' = y^T M y \quad (3)$$

It is easy to show that under the null hypothesis, the test statistic B' asymptotically follows a chi-square distribution with df equals to 1.

For any real number ρ satisfying $0 \leq \rho \leq 1$, define $S_\rho = \rho B + (1 - \rho) SKAT$. The SKAT-O test statistic is defined as⁴

$$SKATO = \min_{0 \leq \rho \leq 1} p_\rho \quad (4)$$

where p_ρ is the p-value from the test S_ρ . The two statistics B and $SKAT$ are correlated and the null distribution of $SKATO$ is intractable⁴. A numeric method is usually used to estimate the p-value for $SKATO$ ⁴.

2.2 The proposed test

We define the following statistic:

$$T = y^T B y \quad (5)$$

where $B = (A + A^T)/2$ and $A = (I - M)GG^TM$. Note that M determinates the contribution from the burden test, and $I - M$ is complementary to M as it can be shown that M is an

idempotent matrix and $(I - M)M = 0$. To study the properties of T , we need the following lemmas.

Lemma 1. If $X \sim N(\mu, \Sigma)$, then for symmetric matrices A and B , we have

- (i) $E(X^T A X) = \text{trace}[A \Sigma] + \mu^T A \mu$, $V(X^T A X) = 2\text{trace}[A \Sigma A \Sigma] + 4\mu^T A \Sigma A \mu$
- (ii) $\text{Cov}(X^T A X, X^T B X) = 2\text{trace}[A \Sigma B \Sigma] + 4\mu^T A \Sigma B \mu$.

Proof: See 20.18 (a) (p434) and 20.25 (c) (ii) and (iii) (p438) of Seber¹³.

For the variance of the statistic T in (5), we have the following property.

Lemma 2. $V(T) = \text{trace}[KMK(I - M)]$.

Proof: Note that here $y \sim N(0, I)$, from lemma 1, $V(T) = 2\text{trace}[B^2]$, where $B = (A + A^T)/2$, $A = (I - M)KM$. But $B^2 = [A^2 + AA^T + A^T A + (A^T)^2]/4$, $A^2 = (I - M)KM(I - M)KM = 0$ (as $M(I - M) = 0$). Similarly, we have $(A^T)^2 = (A^2)^T = 0$. But, $AA^T = (I - M)KMMK(I - M) = (I - M)KMK(I - M)$ and $A^T A = MK(I - M)(I - M)KM = MK(I - M)KM$. Hence,

$$V(T) = 2\text{trace}[B^2] = \frac{\text{tr}[AA^T + A^T A]}{2} = \frac{\text{tr}[AA^T] + \text{tr}[A^T A]}{2} = \text{trace}[AA^T]. \text{ But } \text{trace}[AA^T] = \text{trace}[(I - M)KMK(I - M)] = \text{trace}[KMK(I - M)(I - M)] = \text{trace}[KMK(I - M)]. \text{ Therefore, } V(T) = \text{trace}[KMK(I - M)].$$

From the proof of lemma 2, we have the following result.

Corollary 1. $V(T) = 0$ if and only if $A = 0$.

For the statistic T in (5), we have the following result.

Theorem 1. Asymptotically, under the null hypothesis, if $A \rightarrow 0$, where $A = (I - M)KM$, $T \sim \lambda Y_1^2 - \lambda Y_2^2$, where Y_1 and Y_2 are iid $N(0, 1)$, $\lambda = \sqrt{V(T)}/2$, and $V(T) = \text{trace}[KMK(I - M)]$.

Proof: Since $T = y^T B y$ and $B = (A + A^T)/2 \rightarrow 0$ is symmetric, the quadratic form of T can be written as $T = \sum_i \lambda_i (y^T \mu_i)^2$, where λ_i and μ_i are the eigenvalues and eigenvectors of B . Under

the null hypothesis, asymptotically, $y^T \mu_i \sim N(0, 1)$ so $(y^T \mu_i)^2 \sim \chi_{1,1}^2$ and $\chi_{1,1}^2$ are iid χ_1^2 . But $\text{rank}(A) = \text{rank}(A^T) = 1$, and $\text{rank}(B) = \text{rank}(A) + \text{rank}(A^T) = 2$. Hence, since $B \rightarrow 0$, the number of non-zero eigenvalues of B is $\text{rank}(B) = 2$ (See p95, 6.16 (a) of Seber¹³). Let $T = \lambda_1 \chi_{1,1}^2 + \lambda_2 \chi_{2,1}^2$, where $\lambda_1 \neq 0$, we have, $E(T) = \lambda_1 + \lambda_2$, and $V(T) = 2\lambda_1^2 + 2\lambda_2^2$. But

from lemma 1, $E(T) = E(y^T B y) = \text{trace}[B] = \frac{\text{tr}[A] + \text{tr}[A^T]}{2} = 0$. Therefore,

$$V(T) = 2\lambda_1^2 + 2\lambda_2^2 = 4\lambda_1^2. \text{ Finally, } \lambda_1 = -\lambda_2 = \lambda = \sqrt{V(T)}/2.$$

The proposed test is called VGAT (Variance-Gamma Association Test) and defined as

$$S = \begin{cases} T/\lambda, & \text{if } \lambda \neq 0 \\ SKAT, & \text{if } \lambda = 0 \end{cases} \quad (6)$$

For S in (6) we have the following results.

Theorem 2. Asymptotically, under the null hypothesis, if $\lambda = 0$, the new test statistic in (6) S -variance-gamma distribution with the location parameter $c = 0$, the spread parameter $\sigma = 1$, the asymmetry parameter $\theta = 0$, and the shape parameter $\nu = 2$ for the probability density

function $f(x) = c(c, \sigma, \theta, \nu) \times e^{-\frac{\theta(x-c)}{\sigma^2}} |x-c|^{\frac{1}{\nu}-\frac{1}{2}} K_{\frac{1}{\nu}-\frac{1}{2}}\left(\frac{|x-c|\sqrt{\frac{2\sigma^2}{\nu} + \theta^2}}{\sigma^2}\right)$, where $K_\nu(\cdot)$ is the

modified Bessel function of the third kind of order ν , and

$$c(c, \sigma, \theta, \nu) = \frac{2}{\sigma\sqrt{2\pi\nu}^\nu \Gamma(1/\nu)} \left(\frac{1}{\sqrt{\frac{2\sigma^2}{\nu} + \theta^2}}\right)^{\frac{1}{\nu}-\frac{1}{2}}.$$

Proof: from theorem 1, when $\lambda = 0$, $S = \chi_{1,1}^2 - \chi_{2,1}^2$, the difference of two iid random variables each follows a χ_1^2 . Therefore, it follows the claimed variance-gamma distribution 14.

The two-sided p-value from the new test statistic S with the observed statistic s is defined as $P = \text{Prob}(S > |s|)$, which can be approximated using the Davis method when $\lambda = 0$ ¹². For instance, we can use function `davis(·)` from the R package “CompQuadForm” (<https://cran.r-project.org/web/packages/CompQuadForm/index.html>).

Since the type I error rates using the Davis approximation for both S with (i.e., SKAT) and without $\lambda = 0$ under the null hypothesis asymptotically equal to the preset significance level, we have the following result.

Theorem 3. Using the Davis approximation, the test statistic S can asymptotically control type I error rate.

3. Simulations study

In this section, we investigate the property of the new test through simulation study based on genotypes of real data. In this report, for the SKAT, SKAT-O and the burden tests, we use the R package “SKAT” (<https://cran.r-project.org/web/packages/SKAT/index.html>) to get their p-values.

3.1 The GWA17 data

First, we apply the new method and others to the Genetic Analysis Workshop 17 (GAW17) data. The GAW17 uses the information of a subset of genes with sequencing data available in the 1000 Genomes Project. In GAW17, SNPs from the gene ELAVL4 influence the simulated quantitative phenotype Q_1 . Except for the genetic risk factors, Q_1 was also

assumed to be associated with some covariates, such as age, gender, and smoking status. For this gene, a total of 200 simulated phenotypes (Q_1 values) were conducted in the GAW17 data set. First we used a linear regression model to account for the effects of those non-genetic factors, then we applied our proposed test, along with SKAT, SKAT-O and the burden test, to the standardized residuals obtained from the regression step.

Figure 1 plots the $-\log_{10}(\text{p-values})$ obtained by the four methods. This plot clearly show that the proposed test produces smaller p-values compared to SKAT, SKAT-O, and the burden test, for most of the cases. This indicates that the proposed test VGAT is more powerful than the other methods. For some cases, the improvements of the VGAT were substantial.

3.2 Simulation study

To conduct sensible simulation study, we use the genotypes of gene ELAVL4 from the GAW17 study as they are from real genetic data. For the gene ELAVL4, there are 10 SNPs and 697 subjects without missing genotypes. To keep the true genotype structure, we will remain the genotype data unchanged in the simulation study.

To mimic the real situation, we simulate two none-genetic covariates: one is binary and the other is from a normal distribution. For each subject, we simulate the phenotype using the following formula:

$$y_i = x_{1,i} + 2x_{2,i} + \sum_{j=1}^{10} \beta_j g_{ij} + e_i (i = 1, 2, \dots, 697) \quad (7)$$

where $x_{1,i} \sim \mathcal{N}(1,2)$, $x_{2,i} \sim \text{Binomial}(n=1, p=0.5)$, g_{ij} is the genotype of the j^{th} SNP from the i^{th} subject, and $e_i \sim \mathcal{N}(0,1)$. For the coefficients β_j ($j=1,2, \dots, 10$), we consider several different settings. To obtain the empirical type I error rates, we set all $\beta_j = 0$. To compare the detecting powers, we randomly choose a portion (θ) of SNPs as causal SNPs and the rest as non-causal variants whose coefficients are set to be 0. Among those causal variants, we randomly assign a portion (r) to have positive effects and let the remaining with negative effects. For those causal SNPs, we first follow the suggestion from the SKAT paper⁵ and set the effect size as a function of the MAF, $|\beta_j| = -d \times \log_{10}(\text{MAF}_j)$, where MAF_j is the MAF of the j^{th} SNP. We then consider the situation where the effect sizes are independent of MAF, $|\beta_j| = 3d$. For the constant d , we use different values, 0.5 and 1. In this simulation study we use different significances levels so that the power values are appreciable.

Table 1 reports the empirical type I error rates from the four methods using different significance levels (α). The SKAT test, the burden test and the proposed test can control type I error rate for most of the situations considered. However, surprisingly, we notice that the SKAT-O test had inflated type I error rates for all of the conditions. To further investigate their performance under the null hypothesis, we plotted the histograms of the null p-values in Figure 2. The null p-values from the VGAT, the SKAT test and the burden test evenly distribute between 0 and 1; while SKAT-O produced disproportionally small p-values. We used Anderson-Darling method to test whether the p-values from each method under the null

hypothesis follow a uniform distribution between 0 and 1. The p-values from the Anderson-Darling test were 0.071, 6.0×10^{-9} , 0.49, and 0.30 for the SKAT, the SKAT-O, the burden test and the proposed test, respectively. These results indicate that the SKAT-O produced null p-values that were significantly different from a uniform distribution.

Tables 2 and 3 list the empirical power values of the four methods under different situations (various values for θ and r) when $d=1$, significance level $\alpha = 10^{-4}$ were used. From those tables, we see that for most of the conditions, the proposed test had the highest empirical power values. The SKAT-O in general had slightly better power values than the SKAT which in turn performed better than the burden test. However, caution should be taken when we compare SKAT-O with others as we have already noticed that the SKAT-O may have inflated type I error rate. With different values of d and α , we had similar observations (see Supplementary Tables S1 and S2 where $d=0.5$ and $\alpha = 10^{-2}$ were used).

4. Real data application

In this section, we apply the proposed test along with others to a real data, the ocular hypertension treatment study (OHTS) data¹⁵. OHTS is a National Eye Institute-sponsored multi-center, randomized clinical trial. Its goal was to investigate the efficacy of medical treatment in delaying or preventing the onset of primary open angle glaucoma (POAG) in individuals with elevated intraocular pressure. Data set for this genetic study is available at Database of Genotypes and Phenotypes (dbGaP, Study Accession phs000240.v1.p1). There were 1,051,295 genotyped SNPs. The HGNC gene symbols were obtained using the R/Bioconductor package biomaRt (version 2.26.1) and 30,562 autosomal genes were included.

In this application, we want to detect the association between each gene and the outcome central corneal thickness (CCT), which is used to assess POAG in this study. Two hundred and forty-nine non-Hispanic Black samples are used in this application. After adjusting for covariates age and gender using a linear regression, the standardized residues from the regression analysis are used for the association tests. To detect possible associated genes, we list all genes whose smallest p-values from the four methods are less than 2.5×10^{-5} . Using this criterion three genes were found and listed in Table 4 which also reports the p-values from the four methods. For all of the three listed genes, the proposed test had the smallest p-values. For two genes (IKBKE and GHDC), all of the three other methods obtained very large p-values (all greater than 0.05). However, it warrants further investigations to determine whether the three listed genes are true positives.

5. Discussion and conclusion

In this report, we propose a novel gene- or set-based association test, VGAT. This new test has relatively simple null distribution and its p-value can be easily calculated using currently available statistical tool. Through simulation study and real data application, we have shown that the new method could control type I error rate and had superior detecting power compared with some popular existing methods. We expect that more significant associated genes or sets of variants can be identified when apply our proposed test to existing genetic data. In addition, the proposed method can be applied to common variants, rare variants, or a

set with both common and rare variants. However, like any population based association tests, when apply the proposed test to real data, we need adjust for the possible population stratification effect, which may produce spurious association. For example, the principal component analysis based method can be used when we have subjects from multiple races/ethnics¹⁶.

From the definition of T in (5), we can rewrite it as $T = [y^T M] K [y^T (I - M)]^T$, where $y^T M$ is the contribution from the burden test while $y^T (I - M)$ is from the complement of the burden test. Therefore, the proposed test combines information for both the burden test and its complementary part. On the other hand, the SKAT-O tries to combine information from the burden test and another term which is not independent of the burden test. As a consequence, the SKAT-O has intractable null distribution and the p-value from the numeric method implemented in the R package “SKAT” may cause inflated type I error rate as observed from our simulation study. Because of this, we should use SKAT-O with caution.

Our proposed method is also computational efficient. We used the real data set to compare the computational burden for those methods. All of the tests were implemented in R. The running times were obtained when they were run in a computer with 3.60 GHz Intel Core processors and 8 GB of RAM memory. The running times are 0.47, 3.62, 0.45, and 0.42 hours for the SKAT, SKAT-O, burden, and the proposed method, respectively.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Wang K Boosting the Power of the Sequence Kernel Association Test by Properly Estimating Its Null Distribution. *The American Journal of Human Genetics* 2016; 99: 104–14. [PubMed: 27292111]
2. Fan R, Wang Y, Boehnke M, et al. Gene level meta-analysis of quantitative traits by functional linear models. *Genetics*. 2015; 200: 1089–104. [PubMed: 26058849]
3. Pan W, Kim J, Zhang Y, Shen X and Wei P. A powerful and adaptive association test for rare variants. *Genetics*. 2014; 197: 1081–95. [PubMed: 24831820]
4. Lee S, Wu MC and Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012; 13: 762–75. [PubMed: 22699862]
5. Wu MC, Lee S, Cai T, Li Y, Boehnke M and Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*. 2011; 89: 82–93. [PubMed: 21737059]
6. Neale BM, Rivas MA, Voight BF, et al. Testing for an unusual distribution of rare variants. *PLoS Genet* 2011; 7: e1001322. [PubMed: 21408211]
7. Li B and Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* 2008; 83: 311–21. [PubMed: 18691683]
8. Chen Z and Wang K. A gene-based test of association through an orthogonal decomposition of genotype scores. *Hum Genet* 2017; 136: 1385–94. [PubMed: 28864915]
9. Chen Z, Lin T and Wang K. A Powerful Variant-Set Association Test Based on Chi-Square Distribution. *Genetics* 2017; 207: 903–10. [PubMed: 28912342]

10. Chen Z, Han S and Wang K. Genetic association test based on principal component analysis. *Stat Appl Genet Mol Biol* 2017; 16: 189–98. [PubMed: 28672760]
11. Chen Z, Lu Y, Lin T, Liu Q and Wang K. Gene-based genetic association test with adaptive optimal weights. *Genet Epidemiol* 2018; 42: 95–103. [PubMed: 29178441]
12. Davies RB. Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)* 1980; 29: 323–33.
13. Seber GA. *A matrix handbook for statisticians* John Wiley & Sons, 2008.
14. Kotz S, Kozubowski T and Podgorski K. *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance* Springer Science & Business Media, 2012.
15. Gordon MO and Kass MA. The Ocular Hypertension Treatment Study: design and baseline description of the participants. *Arch Ophthalmol* 1999; 117: 573–83. [PubMed: 10326953]
16. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA and Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; 38: 904. [PubMed: 16862161]

Figure 1.
-log₁₀(p-values) obtained by methods from gene ELAVL4.

Figure 2.
Histogram of p-values by methods from simulation study under the null

Table 1.

Empirical type I error rate (p-value) using 100,000 replicates where p-value is obtained from the binomial test when testing for the true type I error equals the preset significance level α .

α	SKAT	SKAT-O	Burden	VGAT
0.05	0.04986 (0.58)	0.05226 (5.7×10^{-4})	0.05079 (0.13)	0.05062 (0.19)
0.01	0.01043 (0.089)	0.01270 (9.6×10^{-17})	0.01057 (0.037)	0.01017 (0.30)
0.005	0.00509 (0.35)	0.00686 (1.7×10^{-15})	0.00543 (0.030)	0.00509 (0.35)
0.001	0.00112 (0.13)	0.00156 (1.3×10^{-7})	0.00111 (0.15)	0.00106 (0.29)
0.0005	0.00052 (0.41)	0.00088 (7.5×10^{-7})	0.00053 (0.35)	0.00050 (0.52)
0.0001	0.00011 (0.42)	0.00019 (7.2×10^{-3})	0.00011 (0.42)	0.00013 (0.21)

Table 2.

Empirical powers for all methods when $d=1$, $\alpha = 10^{-4}$, and effect size is a function of MAF.

θ	r	SKAT	SKAT-O	Burden	VGAT
0.1	0	0.10	0.10	0.10	0.15
0.1	1	0.11	0.11	0.11	0.16
0.2	0.5	0.22	0.22	0.22	0.31
0.2	1	0.20	0.21	0.20	0.30
0.4	0.5	0.44	0.45	0.40	0.59
0.4	1	0.40	0.45	0.45	0.61
0.6	0.5	0.62	0.63	0.53	0.80
0.6	1	0.62	0.71	0.71	0.81
0.8	0.5	0.82	0.83	0.63	0.92
0.8	1	0.79	0.88	0.88	0.90
1	0.5	1.00	1.00	0.71	0.97
1	1	1.00	1.00	1.00	1.00

Table 3.

Empirical powers for all methods when $d=1$, $\alpha = 10^{-4}$, and effect size is independent of MAF.

θ	r	SKAT	SKAT-O	Burden	VGAT
0.1	0	0.11	0.11	0.11	0.29
0.1	1	0.11	0.11	0.11	0.28
0.2	0.5	0.18	0.19	0.18	0.48
0.2	1	0.19	0.20	0.19	0.47
0.4	0.5	0.34	0.35	0.31	0.70
0.4	1	0.31	0.37	0.34	0.74
0.6	0.5	0.50	0.51	0.37	0.76
0.6	1	0.37	0.46	0.46	0.84
0.8	0.5	0.64	0.65	0.41	0.67
0.8	1	0.27	0.34	0.34	0.76
1	0.5	0.73	0.74	0.39	0.53
1	1	0.00	0.13	0.17	0.26

Table 4.

Genes from the OHTS black samples with the smallest p-value among the four methods less than 2.5×10^{-5} .

Chr	Gene	SKAT	SKAT-O	Burden	VGAT
1	IKBKE	4.23e-1	3.42e-1	2.49e-1	1.25e-5
17	GHDC	7.79e-2	7.79e-2	7.79e-2	1.98e-5
19	EIF5AP3	5.09e-4	5.18e-4	7.02e-4	2.41e-5