Optimal tests for rare variant effects in sequencing association studies

Seunggeun Lee
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
Michael C. Wu
Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA
Xihong Lin*
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA
xlin@hsph.harvard.edu

April 5, 2012

Abstract

With development of massively parallel sequencing technologies, there is a substantial need to developing powerful rare variants association tests. Common approaches include burden and non-burden tests. Burden tests assume all rare variants in the target region have effects on the phenotype in the same direction and of similar magnitude. The recently proposed Sequence Kernel association Test (SKAT) [19] an extension of the C-alpha test [12], provides a robust test that is particularly powerful in the presence of protective and deleterious variants and null variants, but is less powerful than burden tests when a large number of variants in a region are causal and in the same direction. As the underlying biological mechanisms are unknown in practice and vary from one gene to another across the genome, it is of substantial practical interest to develop a test that is optimal for both scenarios. In this paper, we propose a class of tests that include burden tests and SKAT as special cases, and derive an optimal test within this class that maximizes power. We show that this optimal test outperforms burden tests and SKAT in a wide range of scenarios. The results are illustrated using simulation studies and triglyceride data from the Dallas Heart Study. In addition, we have derived sample size/power calculation formula for SKAT with new family of kernels to facilitate designing new sequence association studies. Burden tests; Correlated Effects; Kernel Association Test; Rare variants; Score Test

[0]*To whom correspondence should be addressed.

1 Introduction

Advances in high-throughput sequencing technology are reshaping the landscape of medical and human genetics research. In contrast to genome wide association studies (GWAS), which involves genotyping preselected single nucleotide polymorphisms (SNPs) that are relatively common, these new sequencing technologies enable us to also genotype rare genetic variants. Rare genetic variants, here defined as variants with minor allele frequency (MAF) $< 1 - 5\%$, have been shown to play a crucial role in complex trait etiology[2]. Despite their importance, testing for associations between rare variants and traits has proven challenging. Since standard individual variant tests, typically used for analysis of SNPs, are underpowered to detect rare variant effects due to the low allele frequencies and the large numbers of rare variants in the genome, region based analysis has become the standard approach for analyzing rare variants in sequencing studies[6].

The earliest and most commonly used class of region based rare variant tests are the burden tests, which collapse or summarize the rare variants within region as a single genetic variable which can then be tested for association with any trait of interest [6, 10, 11, 14]. For example, the Combined Multivariate and Collapsing (CMC) method [6] collapses information on all rare variants within a region by counting the number of rare alleles. Many extensions and variations on these methods exist. A key limitation of burden tests is that they suffer substantial loss of power in the presence of large number of non-causal variants, or in the presence of both protective and deleterious variants [12].

Recognizing the inherent limitations of burden based methods, Wu *et al.*[19] recently proposed the sequence kernel association test (SKAT), which builds upon the kernel machine regression framework, to test rare variants associations. As a kernel machine based test, SKAT aggregates genetic information across the region using a kernel function and uses a computationally efficient variance component test to test for association. Wu *et al.*[19] also showed that SKAT is a generalization of the classical C-alpha test [12, 13].

Although SKAT offers improved power over burden based tests in many cases, if a large proportion of the rare variants in a region are truly causal and influence the phenotype in the same direction, then burden tests can have higher power than SKAT[1]. This scenario can arise in several current sequencing studies, e.g., whole exome sequencing studies. This is because standard evolutionary principles and population genetics models indicate that majority of rare missense mutations are moderately deleterious [5]. In addition, bioinformatic tools are often used to restrict testing to variants that are likely harmful variants. However, such prior knowledge is often lacking in practice. The underlying biological mechanisms vary from one gene to another across the genome and often unknown. It is hence of substantial practical interest to develop a data-adaptive test that is optimal for both scenarios when scanning the genome in genome-wide sequencing association studies.

In this paper, by exploiting the relationship between burden based tests and SKAT, we propose a data-adaptive optimal test within a class of tests that include both burden tests and SKAT as special cases. Specifically, we consider a class of tests that is an arbitrary linear combination of burden test and SKAT statistics, and identify the optimal test within

this class to maximize power. We show that this new class of tests can be formulated as a generalized family of SKAT tests by incorporating a correlation structure of variants effects through a family of kernels. It reduces to the burden tests when the effects of variants are perfectly correlated. We derive the optimal test (SKAT-O) by estimating the correlation parameter in the kernel matrix to maximize the power, which corresponds to the estimated weight in the linear combination of the burden test and SKAT test statistics that maximizes power. We derive the theoretical distribution of the SKAT-O test statistic, which allows us to calculate the p-value analytically with high accuracy in the tail. This is advantageous for analyzing genome-wide sequencing data by avoiding computationally intensive methods, such as resampling or permutation, to calculate p-values, especially in the tail required to reach genome-wide significance.

In addition we also consider the problem of designing future sequencing association studies. In particular, to design a new sequence association study, it is important to be able to estimate the required sample size to achieve proper statistical power. Although power and sample size calculation can be done via simulation, this computer intensive approach is not desirable. Therefore, we derive the analytical formula for the statistical power of SKAT under the newly proposed family of kernels, and by inverting the power function, we can compute the necessary sample size to adequately power future studies.

2 Optimal Test

For simplicity,we assume that we are interested in testing whether the rare variants in a single region are associated with a complex trait. In a large scale study of multiple regions, the same methods can be applied with the appropriate adjustment of multiple testing.

2.1 Rare variants testing methods

Assume n subjects are sequenced in a region with p genotyped rare variants. For the i^{th} subject, let y_i denote a phenotype variable, $\mathbf{G}_i = (g_{i1}, \ldots, g_{ip})$ the genotypes for the p variants ($g_{ij} = 0, 1, 2$ for 0,1, or 2 copies of the minor allele), $\mathbf{X}_i = (x_{i1}, \ldots, x_{iq})$ the covariates for which we would like to adjust (e.g. demographic or environmental variables). To relate genotypes to continuous/categorical phenotypes, we use the generalized linear model (GLM), such that y_i independently follows an exponential family distribution with first two moments $E(y_i) = \mu_i$ and $Var(y_i) = \phi v(\mu_i)$, and a link function

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{G}_i\boldsymbol{\beta}, \tag{1}$$

where $v(\cdot)$ is a variance function. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the vectors of regression coefficients for the covariates and rare variants, respectively. Under the GLM, the association between the p rare variants and the phenotype y can be tested by evaluating the null hypothesis that H_0: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' = \mathbf{0}$. However, the standard p degree of freedom (d.f.) test may lose power when p is large. To reduce the d.f., additional assumptions need to be made.

Popular burden based tests reduce the d.f. by making the assumption that each β_j is a function of the minor allele frequencies (MAF) such that $\beta_j = w(m_j) = w_j\beta_0$, where m_j is the MAF of the j^{th} variant. Then (1) becomes

$$g(\mu_i) = \mathbf{X}_i\boldsymbol{\alpha} + \beta_0 \sum_{j=1}^{p} w_j g_{ij}, \tag{2}$$

and the association between the genetic variants and the phenotype can be tested by conducting a 1 d.f. test with H_0: $\beta_0 = 0$. We refer to this test as the weighted counting burden test (WBT). The WBT assumes that all variants are causal with the same direction of association and common β_0. Violation of these assumptions can result in significant loss of power.

SKAT takes a different approach to reducing the d.f. It assumes that each β_j independently follows an arbitrary distribution with mean zero and variance $w_j^2\psi$ where w_j is a fixed number that may depend on MAF. Under this assumption, the null hypothesis H_0 : $\boldsymbol{\beta} = 0$ is equivalent to H_0 : $\psi = 0$, i.e., variance component test in generalized linear mixed models by treating β as random effects. Suppose $\mathbf{X}$ is the $n \times q$ covariates matrix, $\mathbf{G} = [\mathbf{G}_1, \ldots, \mathbf{G}_n]'$ is the $n \times p$ genotype matrix, $\mathbf{W} = diag[w_1, \ldots, w_p]$ is a $p \times p$ diagonal matrix of weights, and $\mathbf{K} = \mathbf{GWWG}'$ is an $n \times n$ weighted linear kernel matrix. [19] proposed to use a class of flexible weight functions of the MAF using the beta density function as $w_j = Beta(MAF_j, a_1, a_2)$, where the parameters a_1 and a_2 are pre-specified, and MAF_j are estimated using the sample MAF of the j-th variant.

We define the working vector by $\mathbf{y}^* = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\Delta}(\mathbf{y} - \boldsymbol{\mu})$, where $\boldsymbol{\Delta} = diag\{g'(\mu_i)\}$, and the variance matrix by $\mathbf{V} = diag\{\phi v(\mu_i)[g'(\mu_i)]^2\}$. Their estimates under the null hypothesis are $\widetilde{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\alpha}} + \widehat{\boldsymbol{\Delta}}(\mathbf{y} - \hat{\boldsymbol{\mu}})$, $\widehat{\boldsymbol{\Delta}} = diag\{g'(\widehat{\mu}_i)\}$ and $\widehat{\mathbf{V}} = diag\left\{\hat{\phi} v(\widehat{\mu}_i)[g'(\widehat{\mu}_i)]^2\right\}$, respectively, where $\hat{\boldsymbol{\alpha}}$ is an $q \times 1$ vector of estimates of $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\mu}}$ is an $n \times 1$ vector of estimates of $\boldsymbol{\mu}$, and $\hat{\phi}$ is an estimate of ϕ. All estimates are obtained under the null hypothesis. Following [20], the score test statistic of the variance component ψ is

$$Q = (\widetilde{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\alpha}})'\widehat{\mathbf{V}}^{-1}\mathbf{K}\widehat{\mathbf{V}}^{-1}(\widetilde{\mathbf{y}} - \mathbf{X}\hat{\boldsymbol{\alpha}}) = (\mathbf{y} - \hat{\boldsymbol{\mu}})'\widehat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}^{-1}\mathbf{K}\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}}(\mathbf{y} - \hat{\boldsymbol{\mu}}). \tag{3}$$

When $g(\cdot)$ is a canonical link function, (3) can be simplified to $Q = (\mathbf{y} - \hat{\boldsymbol{\mu}})'\mathbf{K}(\mathbf{y} - \hat{\boldsymbol{\mu}})/\hat{\phi}^2$. For binary and Poisson data, $\phi = 1$.

2.2 New family of kernels

As shown in the previous section, the weighted linear kernel is constructed under the assumption that β_js are independent. If a large percentage of variants in the target region are associated with the phenotype with the same direction of effect, burden tests can outperform SKAT because the current kernels used by SKAT do not account for correlation in $\boldsymbol{\beta}$. Therefore, we propose a new family of kernels that explicitly incorporates correlation among the variant effects.

We propose to allow $\boldsymbol{\beta}$ to follow a multivariate distribution with exchangeable correlation structure. Then the correlation matrix of $\boldsymbol{\beta}$ is $\mathbf{R}_\rho = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}'$. With this correlation structure, the SKAT test statistic is a function of ρ:

$$Q_\rho = (\mathbf{y} - \hat{\boldsymbol{\mu}})'\widehat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}^{-1}\mathbf{K}_\rho\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}}(\mathbf{y} - \hat{\boldsymbol{\mu}}), \tag{4}$$

where $\mathbf{K}_\rho = \mathbf{G}\mathbf{W}\mathbf{R}_\rho\mathbf{W}\mathbf{G}'$. When $\rho = 0$, $\mathbf{K}_\rho$ results in the weighted linear kernel SKAT. When $\rho = 1$, SKAT test statistic is

$$Q_\rho = (\mathbf{y} - \hat{\boldsymbol{\mu}})'\widehat{\boldsymbol{\Delta}}\widehat{\mathbf{V}}^{-1}\mathbf{G}\mathbf{W}\mathbf{1}\mathbf{1}'\mathbf{W}\mathbf{G}'\widehat{\mathbf{V}}^{-1}\widehat{\boldsymbol{\Delta}}(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \left[\sum_{i=1}^{n}\frac{y_i - \hat{\mu}_i}{\hat{\phi}v(\hat{\mu}_i)g'(\hat{\mu}_i)}\sum_{j=1}^{p}w_j g_{ij}\right]^2,$$

which is equivalent to the square of the score test statistic of weighted counting burden test. Thus, both tests can be framed within this new family of kernels. In fact, one can easily show that Q_ρ is a linear combination of SKAT and burden test, i.e, $Q_\rho = (1-\rho)Q_{SKAT} + \rho Q_{burden}$.

For a fixed ρ, Q_ρ follows a mixture of chi-square distributions [17, 8, 7]. Specifically, if $(\lambda_1, \ldots, \lambda_m)$ are the eigenvalues of $\widehat{\mathbf{V}}^{-1/2}\mathbf{K}_\rho\widehat{\mathbf{V}}^{-1/2}$, then the null distribution of Q_ρ can be closely approximated by $\sum\lambda_j\chi^2_{1,j}$, where $\{\chi^2_{1,j}\}$ are independent χ^2_1 random variables. To reduce small sample bias, the restricted maximum likelihood (REML) estimator of the variance component can be used [20]. Define $\mathbf{P} = \widehat{\mathbf{V}}^{-1} - \widehat{\mathbf{V}}^{-1}\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widehat{\mathbf{V}}^{-1}$ where $\widetilde{\mathbf{X}} = [\mathbf{1}\,\mathbf{X}]$, we use the eigenvalues of $\mathbf{P}^{1/2}\mathbf{K}_\rho\mathbf{P}^{1/2}$ to obtain the null distribution of Q_ρ. A p-value can be obtained by matching moments [9] or by inverting the characteristic function [3].

To understand the role of the parameter ρ, we derive in Supplementary Appendix C the analytic relationship between ρ and the given regression coefficients βs as a function of the proportion of causal variants in a region (i.e., the proportion of the β coefficients that are non-zero) and the proportion of causal variants that are protective (i.e., the proportion of the non-zero β coefficients that are negative), assuming the magnitude of the causal variant effects as a function of the MAFs. We illustrate this relationship and its use in the power calculation section.

2.3 Optimal test (SKAT-O)

In practice, we rarely have any information about ρ. Thus, we need a procedure to select ρ to maximize power. The resulting optimal test corresponds to a best linear combination of SKAT and burden tests that maximizes power. This is in general a challenging problem because ρ disappears under the null hypothesis. Davies [4] studied this problem and proposed to use maximum of score test statistic as a test statistic. This approach, however, is not directly applicable here due to the different kurtoses of the Q_ρs. As ρ increase, the null distribution of Q_ρ has a heavier tail, and adjustment for this is difficult.

We employ an different approach and use the minimum of p-values as a test statistic rather than the score statistics. Specifically, the test statistic is

$$T = \inf_{0\le\rho\le1} p_\rho, \tag{5}$$

where p_ρ is the p-value computed based on Q_ρ. T can be obtained by simple grid search across a range of ρ: set a grid $0 = \rho_1 < \rho_2 < \ldots < \rho_b = 1$, then the test statistic $T = min\{p_{\rho_1}, \ldots, p_{\rho_b}\}$.

2.3.1 Null distribution of the test statistic

In order to obtain the null distribution of T and conduct a hypothesis test, let $\mathbf{Z} = \widehat{\mathbf{V}}^{-1/2}\mathbf{GW}$ and $\bar{\mathbf{z}} = (\bar{z}_1, \ldots, \bar{z}_n)'$, where $\bar{\mathbf{z}}_i = \sum_{j=1}^{p} z_{ij}/p$. $\mathbf{M} = \bar{\mathbf{z}}(\bar{\mathbf{z}}'\bar{\mathbf{z}})^{-1}\bar{\mathbf{z}}'$ is a projection matrix onto a space of spanned by $\bar{\mathbf{z}}$. We further let

$$\tau(\rho) = p\rho\bar{\mathbf{z}}'\bar{\mathbf{z}} + \frac{1-\rho}{\bar{\mathbf{z}}'\bar{\mathbf{z}}}\sum_{j=1}^{p}(\bar{\mathbf{z}}'\mathbf{z}_{.j})^2,$$

where $\mathbf{z}_{.j}$ is the j^{th} column of $\mathbf{Z}$, and $Q_{\rho_1}, \ldots, Q_{\rho_b}$ are the score test statistics computed with different $\rho_v, (v = 1, \ldots, b)$. Then we show in Supplementary B that under the null hypothesis Q_{ρ_v} is asymptotically the same as

$$(1-\rho_v)(\sum_{k=1}^{m}\lambda_k\eta_k + \zeta) + \tau(\rho_v)\eta_0 = (1-\rho)\kappa + \tau(\rho_v)\eta_0, \tag{6}$$

where $\{\lambda_1, \ldots \lambda_m\}$ are non-zero eigenvalues of $\mathbf{Z}'(\mathbf{I}-\mathbf{M})\mathbf{Z}$, $\eta_k(k = 0, \ldots, m)$ are i.i.d χ_1^2 random variables, $\kappa = \sum_{k=1}^{m}\lambda_k\eta_k + \zeta$, and ζ satisfies the following conditions:

$$E(\zeta) = 0, \quad Var(\zeta) = 4trace(\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{Z}'(\mathbf{I}-\mathbf{M})\mathbf{Z}),$$
$$Corr(\sum_{k=1}^{m}\lambda_k\eta_k, \zeta) = 0, \quad \text{and} \quad Corr(\eta_0, \zeta) = 0.$$

Since the Pearson correlation between κ and η_0 is zero, we can approximate Q_ρ as the mixture of two independent random variables. We can approximate the distribution of κ by using the characteristic function inversion method [3] after adjusting for the extra variance term of ζ. Letting $q_{min}(\rho_v)$ denote the $(1-T)^{th}$ percentile of the distribution of Q_{ρ_v} for each ρ_v, the p-value based on the test statistic T is

$$\begin{aligned}&1 - P(Q_{\rho_1} < q_{min}(\rho_1), \ldots, Q_{\rho_b} < q_{min}(\rho_b))\\ &= 1 - E\left[P\left(\kappa < min\{(q_{min}(\rho_v) - \rho_v\eta_0)/(1-\rho_v)\}|\eta_0\right)\right],\end{aligned}$$

which can be obtained by one-dimensional numerical integration, which can be easily calculated. When we compute p_{ρ_v} and $q_{min}(\rho_v)$, we approximate each marginal distribution of Q_{ρ_v} by modifying the moment matching method of Liu *et al.*[9]. In particular, instead of matching the first 3 moments, we match the mean, variance and kurtosis to improve the approximation in the tail area. To adjust for small sample bias, one can use REML estimates of the variance components, such that $\mathbf{Z} = \mathbf{P}^{1/2}\mathbf{GW}$. The following algorithm

provides the detailed description of the proposed method.

Step 1: Set a grid $0 = \rho_1 < \rho_2 < \ldots < \rho_b = 1$.
Step 2: Compute $Q_{\rho_1}, \ldots, Q_{\rho_b}$, $\mathbf{Z}$, and $\mathbf{M}$. Here $\mathbf{Z} = \mathbf{P}^{1/2}\mathbf{G}\mathbf{W}$.
Step 3: Compute λ_ks, $\tau(\rho_v)$, and

$$\mu_Q = \sum_{k=1}^{m} \lambda_k^2, \quad \sigma_\zeta = 2\sqrt{trace(\mathbf{Z}'\mathbf{M}\mathbf{Z}\mathbf{Z}'(\mathbf{I}-\mathbf{M})\mathbf{Z})}, \quad \text{and} \quad \sigma_Q = \sqrt{2\sum_{k=1}^{m}\lambda_k^2 + \sigma_\zeta^2}.$$

Step 4: Calculate p_{ρ_v}, T and $q_{min}(\rho_v)$s using the modified moment matching approximation.
Step 5: Numerically integrate $F(\delta(x)|\lambda)f(x|\chi_1^2)$, where

$$\delta(x) = (min\{(q_{min}(\rho_v) - \tau(\rho_v)x)/(1-\rho_v)\} - \mu_Q)\frac{\sqrt{\sigma_Q^2 - \sigma_\zeta^2}}{\sigma_Q} + \mu_Q,$$

$f(x|\chi_1^2)$ is a density function of χ_1^2, and $F(\delta(x)|\lambda)$ is a distribution function of a mixture of chi-square distribution, $\sum \lambda_i \chi_i^2$. The p-value is found as

$$\text{P-value} = 1 - \int F(\delta(x)|\lambda)f(x|\chi_1^2)dx.$$

3 Sample Size and Power Calculations for Designing Sequencing Association Studies

Estimating the necessary sample size to adequately power a study is an important part of designing new sequencing association studies. In this section, we derive the analytical formula for the statistical power of SKAT. We restrict our interest to continuous traits study designs and dichotomous traits study designs, such as case/control study designs, since both are commonly used designs in association studies. We further note that we only consider SKAT with fixed ρ, since at the design stage, researchers generally specify anticipated alternative hypotheses, i.e. a specific ρ based on the scenario they have in mind. The detailed power calculation method can be found in Supplementary Appendix A. The required sample size to achieve a fixed power level can be easily computed by inverting the power function.

The proposed formula can be used to calculate statistical power efficiently for specified sample sizes and α level, given prior information on the genetic architecture of the genomic region of interest. Also required are the proportion of causal variants within the region (i.e, the fraction of non-zero β coefficients), and the proportion of causal variants that are protective (i.e., the fraction of non-zero β coefficients that are negative). Both of these can be easily posited by investigators based on prior belief, and detailed discussions are provided in Wu *et al.*[19]. What is more challenging to specify is the particular

ρ value to be used in power calculations. As discussed in Section 2.2 and Supplementary Appendix C, if the proportion of nonzero βs is p_1, the proportion of positive βs among the non-zero βs is p_2, and the magnitude of the causal variant effects is a function of the MAFs, the optimal ρ can be estimated as

$$\rho = p_1^2(2p_2 - 1)^2. \tag{7}$$

Note that the power computed with this theoretical estimate of ρ will differ slightly from the power of the optimal test (5) which is based on the data driven optimal ρ, since we are fixing ρ to obtain the power. However, our simulation studies (Section 4.3) suggest that this difference tends to be small.

4 Simulations and Real Data Analysis

4.1 Type I error rate

For all simulations, sequence genotypes were generated from 10,000 chromosomes over 1 MB regions by the calibrated coalescent model with mimicking the linkage disequilibrium (LD) structure of European ancestry samples [16]. Continuous phenotypes were generated from the null linear model

$$y_i = 0.5X_1 + 0.5X_2 + \epsilon_i, \quad \epsilon_i \sim N(0,1),$$

and binary phenotypes were generated from the null logistic model

$$logitP(y_i = 1) = \alpha_0 + 0.5X_1 + 0.5X_2,$$

where X_1 was a continuous covariate generated from $N(0,1)$, X_2 was a binary covariate generated from $Bernoulli(0.5)$, and α_0 was chosen to make penetrance 0.01. For binary trait simulations, we generated retrospective case$-$control data with half being cases and half being controls. Since the average length of a gene is around $3kb$, we randomly selected $3kb$ regions across the 1MB chromosome. For each model, we generated a total of $10,000$ data sets.

We applied six different methods to each of the simulated data sets: proposed test with default $w_j = beta(MAF_j; 1, 25)$ weights (SKAT-O); SKAT-O with flat weights (rSKAT-O); original SKAT ($\rho = 0$) with $beta(1, 25)$ weights (SKAT); SKAT ($\rho = 0$) with flat weights (rSKAT); counting based burden test (N); and weighted counting burden test (W). For rSKAT-O, rSKAT and counting based burden test (N), only variants with observed MAF < 0.03 were used. The equal size grid of 11 points (from 0 to 1) were used to obtain test statistics of the SKAT-O and rSKAT-O to search for optimal ρ. Table 1 shows that all six methods well controlled type I error rates with $\alpha = 0.05$ and $\alpha = 0.01$.

To investigate type I error rates at very stringent genome-wide α levels, we conducted extensive simulations under a slightly different setting (Supplementary D). Table 2 shows that SKAT-O can accurately control type I error with moderate α levels, but produces slightly inflated type I error rates at very small α levels.

4.2 Power

As with the type I error simulations, we randomly selected $3kb$ regions from the broader 1Mb region, but we then randomly chose causal variants from among the variants with true MAF < 0.03. The continuous phenotype were simulated from

$$y_i = 0.5X_1 + 0.5X_2 + \beta_1 g_1 + \ldots + \beta_s g_s + \epsilon_i, \quad \epsilon_i \sim N(0,1),$$

and dichotomous phenotypes were simulated from

$$logitP(y_i = 1) = \alpha_0 + 0.5X_1 + 0.5X_2 + \beta_1 g_1 + \ldots + \beta_s g_s,$$

where $(g_1, \ldots, g_s)$ were selected causal variants. Covariates X_1 and X_2 follow the same distribution in the type I error simulation, and α_0 was chosen to make prevalence 0.01.

We considered simulations in which 10%, 20%, or 50% of rare variants were causal. Since it is assumed that rarer variants are more likely to have large effect sizes, we set $\beta_j = c|\log_{10}(m_j)|$, where m_j is the MAF of the j^{th} variant. For continuous trait simulations, we set $c = 0.6$, when 10% of the rare variants were causal, which gives maximum $\beta = 2.4$ for variants with MAF=10^{-4}. We used $c = 0.3$ and $c = 0.2$, when 20% and 50% of the rare variants were causal to compensate for the increased number of causal variants. For dichotomous trait simulations, we set $c = ln13/4 = 0.64$, when 10% of the rare variants were causal, which gives maximum OR=13 for variants with MAF=10^{-4}. We scaled down c with larger percentages of causal variants. We allowed sample size to vary as $n = 1000$, 2000, and 5000. Datasets were generated 1,000 times for each configuration. We applied the same six methods used in the type I error simulations to each data set, and power was estimated as the proportion of p-values less than $\alpha = 2.5 \times 10^{-6}$.

Figure 1 shows the empirical power under all considered configurations when non-zero β coefficients are all positive, i.e., causal variants are all in the same directions. When the percentage of causal SNPs was low, both the original SKATs (SKAT and rSKAT) and the proposed SKAT-O (SKAT-O and rSKAT-O) had higher power than the burden tests. When the proportion of causal SNPs increased, the burden tests performed better, and the original SKAT has lower power than burden tests when 50% of variants are causal. The SKAT-O and rSKAT-O perform better (when 20% of rare variants are causal) or similar to the burden tests (when 50% of rare variants are causal). This suggests that the performance of SKAT-O is closer to the burden tests in the presence of a larger proportion of causal variants. The higher power of the weighted test over the unweighted test also suggests that appropriate weighting can increase the power.

We also conducted simulations in which 20% of causal variants have negative βs (and 80% have positive βs). Results for these simulations are presented in Figure 2 and show that as expected, burden tests lose a significant amount of power since the effects of the causal variants cancel out due to the presence of negative β coefficients. Both the original and the new optimal SKAT and rSKAT outperform the burden tests no matter where the percentage of causal variants is small or large. In this case, the performance of SKAT-O is closer to SKAT. When 10% or 20% of variants were causal variants, SKAT-O and rSKAT-O had slightly lower power than SKAT and rSKAT. With 50% of variants being causal,

SKAT-O and rSKAT-O had slightly higher power than SKAT and rSKAT. Compared to Figure 1, SKAT and rSKAT did not suffer any power loss in continuous trait simulation and had slightly lower power in dichotomous trait simulation which resulted from lower enrichment for causal variants due to fixed prevalence and the presence of protective variants. However, burden tests lose considerable power due to the fact that causal variants β coefficients are in mixed directions. In addition, we present the ρ values selected by SKAT-O in Supplementary Figure 4, which shows that SKAT-O generally selects large ρ values when the percentage of causal variants is high and $\beta + /- = 100/0$ and selects small ρs when either the percentage of causal variants is low or $\beta + /- = 80/20$. Overall, our simulation results confirm that the proposed SKAT-O performed very well under broad circumstances.

4.3 Sample Size and Power Calculation

We investigated the effect of different ρ values on the power under various models (Supplementary E). Supplementary Figures 2 and 3 show that the test with $\rho = 0$ (original SKAT) or $\rho = 1$ (burden tests) can have reduced power, depending on the model assumptions. For example, if only 10% of the variants are causal, then the test with $\rho = 1$ is significantly less powerful than the test with $\rho = 0$. In contrast, when 50% of the variants are causal and all of the non-zero β coefficients are positive, the test with $\rho = 0$ had lower power than the test with $\rho = 1$. In all scenarios, ρ computed from the proposed formula (ρ=estimated) was most powerful. These figures also show that the theoretical power under the theoretical optimal ρ closely approximates the empirical power of SKAT-O verifying the adequacy of using equation (7) to select ρ for power and sample size calculations.

We also conducted simulations to evaluate the accuracy of the power calculation formula given ρ values, and details can be found in Supplementary F.

4.4 Real Data Application

We applied the proposed SKAT-O and other competing methods to the resequencing data from the Dallas Heart Study [18] to test for association between serum triglyceride (TG) levels and rare variants in 3 genes (ANGPTL3, ANGPTL4, and ANGPTL5). The resequencing dataset has sequence information on 93 observed variants in the three genes from each of 3,476 individuals in three ethnic groups (white = 1043, black=1832 and hispanic=601) [15]. 35, 32 and 26 variants reside in the ANGPTL3, ANGPTL4 and ANGPTL5 gene, respectively. All variants except one have MAF < 0.03. The histogram of the estimated allele frequencies of the 92 variants with MAF < 0.03 is presented in Figure 3 A and clearly indicates that the majority of variants are very rare.

We first performed single variant association analysis between each variant and log-transformed TG level to explore whether there were variants with different directions of effect. Specifically, we regressed the trait value on the variant while adjusting for gender

and ethnicity and computed the t-statistic based on the regression coefficient (Figure 3 B-D). There is no clear evidence that ANGPTL3 and ANGPTL5 have variants with opposite effects; however, since some of the t-statistics for the variants in ANGPTL4 have opposing signs and relatively large magnitude, suggesting that there is the potential for variants to have different directions of effect.

We conducted two different analyses. First, we pooled all 93 variants from across the 3 genes and tested for their cumulative effect on log TG level. Second, then we considered each gene separately and tested the association between rare variants in each gene and log TG level. We again applied the 6 methods used in the simulation studies with adjusting gender and ethnicity. The results (Table 3) show that when the variants in all three genes were pooled to form a single region for analysis, SKAT-O was by far the most powerful. Our individual variant analysis results suggested that the variants in ANGPTL3 and ANGPTL5 may affect log TG unidirectionally so it is unsurprising that the burden tests (W,N) had comparable or better performance than SKAT and rSKAT for testing the variants in ANGPTL3 and ANGPTL5. Also as expected due to the apparent presence of variants with opposing effects, SKAT and rSKAT performed better than the burden tests for testing the association between ANGPTL4 and log TG level. However, SKAT-O and rSKAT-O performed very well across all settings and had only slightly larger p-values than the best test for the particular setting.

Although log TG is a continuous variable, purely for illustration we also dichotomized the log TG level by taking the the highest and lowest quartiles of each of the six sex-ethnicity groups and using high/low log TG as a dichotomous outcome. The results were qualitatively similar to the results keeping log TG continuous (Table 3).

5 Discussion and Conclusion

In this paper, we propose a new family of kernels that incorporates correlation among the effects of causal variants. Based on this new family of kernels, we develop the optimal testing procedure that uses the minimum p-values from different kernels as a test statistic. In simulation and real data analysis, we show that the proposed optimal test often outperformed the existing burden test and weighted linear kernel SKAT. In addition, we derived sample size/power formula for SKAT for designing new sequence association studies.

In whole exome or whole genome sequencing studies, we scan the genome by testing a large number of genes. We cannot expect all the genes/regions to follow the same genetic model of association: some are likely to have many causal variants with the effects in the same direction while others may have few causal variants or the causal variants may have the effects in different directions. A good example is the Dallas Heart Study data in which ANGPTL 3 and 4 seem to follow different genetic association models. Thus, the proposed SKAT-O can be an attractive choice for many situations, because it adapts to the underlying biological model by selecting ρ based on the data.

We note that the proposed kernels only consider the compound symmetry covariance

structure for the effect of rare variants. The simple structure of compound symmetry allows us to efficiently compute p-values of the optimal test. For example, it would take only $3 \sim 4$ hours to analyze a whole exome sequencing study with 20,000 genes and 2,000 samples on a laptop. Extension of the method to accommodate other correlation structures is a problem of future interest.

The power/sample size calculation formula presented in this paper are derived with fixed ρ, and thus we acknowledge that this formula may not be applicable for the proposed SKAT-O. To derive power/sample size formula for the optimal test is challenging due to the ζ term, and we suggest a practical approach to pre-select ρ based on assumptions on underlying genetic structure of association. In particular, researchers can use the simple formula we developed in this paper to select a proper ρ parameter based on the expected effects of the variants in power calculations. We note that for simplicity, this formulae was derived under the assumption that the magnitudes of the nonzero β coefficients are equal to the weights used in the test. However, our simulation studies show that the power is robust to this assumption and is not very sensitive to modest changes in ρ, and the proposed approach can closely approximate the empirical power of SKAT-O.

As shown in simulation and real data analysis, proper weighting can improve power to detect rare variants association. In this paper, we use the $beta(1, 25)$ density function as a weight function that has been proposed by Wu *et al.*[19]. Since the weight is fixed prior to conducting the association test, the type I error rate is not inflated. It is possible, however, the power can be lost if the weight is misspecified. Although an attractive solution is to choose the weight adaptively using data and obtain p-values through permutation or resampling, the computationally expense is not desirable, particularly for for genome-wide studies, and it remains of interest to select the weight that maximizes power.

Funding

This work was supported by the National Institutes of Health [R37 CA076404 and P01 CA134294 to SL and XL, and R01 HG006292 to MCW]. We thank the AE and the referees for constructive comments which have improved the paper.

Software

The software can be found at http://www.hsph.harvard.edu/research/skat/

References

[1] S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 2011.

[2] J.C. Cohen, R.S. Kiss, A. Pertsemlidis, Y.L. Marcel, R. McPherson, and H.H. Hobbs. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869, 2004.

[3] R.B. Davies. Algorithm AS 155: The distribution of a linear combination of χ 2 random variables. *Applied Statistics*, 29(3):323–333, 1980.

[4] R.B. Davies. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, 74(1):33, 1987.

[5] G.V. Kryukov, L.A. Pennacchio, and S.R. Sunyaev. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *The American Journal of Human Genetics*, 80(4):727–739, 2007.

[6] B. Li and S.M. Leal. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008.

[7] D. Liu, D. Ghosh, and X. Lin. Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. *BMC Bioinformatics*, 9, 2008.

[8] D. Liu, X. Lin, and D. Ghosh. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, 63(4):1079–1088, 2007.

[9] H. Liu, Y. Tang, and H.H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009.

[10] B.E. Madsen and S.R. Browning. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, 2009.

[11] A.P. Morris and E. Zeggini. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, 34(2):188, 2010.

[12] B.M. Neale, M.A. Rivas, B.F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S.M. Purcell, K. Roeder, M.J. Daly, et al. Testing for an Unusual Distribution of Rare Variants. *PLoS Genetics*, 7(3):161–165, 2011.

[13] J. Neyman and E. Scott. On the use of c() optimal tests of composite hypotheses. *Bulletin of the International Statistical Institute*, 41, 1966.

[14] A.L. Price, G.V. Kryukov, P.I.W. de Bakker, S.M. Purcell, J. Staples, L.J. Wei, and S.R. Sunyaev. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, 2010.

[15] S. Romeo, L.A. Pennacchio, Y. Fu, E. Boerwinkle, A. Tybjaerg-Hansen, H.H. Hobbs, and J.C. Cohen. Population-based resequencing of angptl4 uncovers variations that reduce triglycerides and increase hdl. *Nature genetics*, 39(4):513–516, 2007.

[16] S.F. Schaffner, C. Foo, S. Gabriel, D. Reich, M.J. Daly, and D. Altshuler. Calibrating a coalescent simulation of human genome sequence variation. *Genome Research*, 15(11):1576, 2005.

[17] J.Y. Tzeng and D. Zhang. Haplotype-based association analysis via variance-components score test. *The American Journal of Human Genetics*, 81(5):927–938, 2007.

[18] R.G. Victor, R.W. Haley, D.W.L. Willett, R.M. Peshock, P.C. Vaeth, D. Leonard, M. Basit, R.S. Cooper, V.G. Iannacchione, W.A. Visscher, et al. The dallas heart study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health* 1. *The American journal of cardiology*, 93(12):1473–1480, 2004.

[19] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 2011.

[20] D. Zhang and X. Lin. Hypothesis testing in semiparametric additive mixed models. *Biostatistics*, 4(1):57, 2003.

Table 1: Type I error estimates of six different methods to test an association between randomly selected 3kb regions with continuous and binary traits. Each entry represents type I error rate estimates as the proportion of p-values smaller than α under the null hypothesis based on $10,000$ simulated datasets.

Sample Size	Level α	SKAT-O	SKAT	rSKAT-O	rSKAT	N	W
Continuous Trait							
2000	0.05	0.051	0.051	0.050	0.048	0.048	0.047
	0.01	0.009	0.009	0.009	0.009	0.009	0.009
5000	0.05	0.051	0.049	0.054	0.050	0.049	0.051
	0.01	0.011	0.010	0.011	0.009	0.009	0.009
Dichotomous Trait							
2000	0.05	0.051	0.049	0.049	0.048	0.047	0.047
	0.01	0.011	0.009	0.012	0.010	0.010	0.011
5000	0.05	0.047	0.043	0.046	0.044	0.047	0.048
	0.01	0.010	0.009	0.010	0.008	0.009	0.010

Table 2: Type I error estimates of SKAT-O to test an association between randomly selected 3kb regions with continuous and binary traits at stringent α level. The sample size was $2,000$. Each entry represents type I error rate estimates as the proportion of p-values smaller than α under the null hypothesis based on 10^7 simulated phenotypes.

Level α	Continuous traits	Binary traits
10^{-3}	1.1×10^{-3}	1.1×10^{-3}
10^{-5}	1.7×10^{-5}	1.2×10^{-5}
2.5×10^{-6}	4.1×10^{-6}	3.3×10^{-6}

Table 3: Analysis of Dallas Heart Study sequence data. Each entry represents a p-value from each method after adjusting gender and ethnicity. The selected ρ values by SKAT-O and rSKAT-O are presented in the parentheses. "ALL" indicates analysis results of joint test of all 93 variants in 3 genes.

Gene	SKAT-O	rSKAT-O	SKAT	rSKAT	N	W
Continuous TG Level						
ALL	1.8×10^{-5} ($\rho = 0.1$)	4.6×10^{-5} ($\rho = 0.2$)	9.5×10^{-5}	2.9×10^{-4}	7.2×10^{-5}	2.3×10^{-4}
ANGPTL3	2.6×10^{-3} ($\rho = 0.5$)	1.3×10^{-3} ($\rho = 1$)	8.9×10^{-3}	6.2×10^{-3}	1.1×10^{-3}	1.9×10^{-3}
ANGPTL4	1.7×10^{-4} ($\rho = 0$)	8.5×10^{-4} ($\rho = 0.1$)	9.5×10^{-5}	6.9×10^{-4}	4.9×10^{-3}	3.1×10^{-2}
ANGPTL5	3.4×10^{-1} ($\rho = 1$)	5.0×10^{-1} ($\rho = 1$)	7.6×10^{-1}	8.5×10^{-1}	3.3×10^{-1}	2.0×10^{-1}
Dichotomous TG Level						
ALL	1.1×10^{-4} ($\rho = 0.1$)	2.8×10^{-4} ($\rho = 0.1$)	1.4×10^{-4}	2.4×10^{-4}	2.2×10^{-3}	2.7×10^{-3}
ANGPTL3	4.7×10^{-2} ($\rho = 0.4$)	3.2×10^{-2} ($\rho = 0.8$)	4.7×10^{-2}	4.1×10^{-2}	2.3×10^{-2}	3.6×10^{-2}
ANGPTL4	1.6×10^{-4} ($\rho = 0$)	3.6×10^{-4} ($\rho = 0.1$)	9.4×10^{-5}	3.3×10^{-4}	3.1×10^{-3}	2.2×10^{-2}
ANGPTL5	4.6×10^{-1} ($\rho = 0$)	3.1×10^{-1} ($\rho = 0$)	3.0×10^{-1}	1.9×10^{-1}	8.7×10^{-1}	4.0×10^{-1}

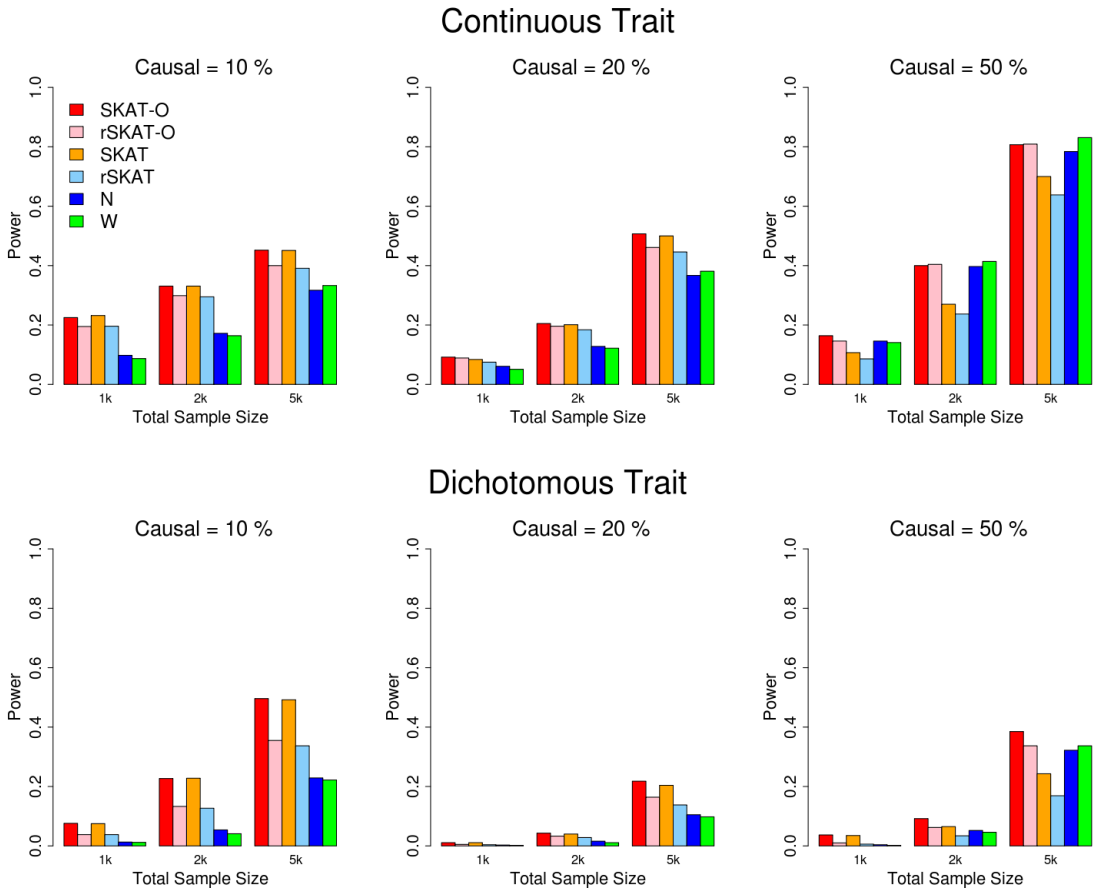

Figure 1: Empirical power of SKAT-O and competing methods at $\alpha = 2.5 \times 10^{-6}$ using simulation studies when region size=3kb and $\beta + /- = 100/0$. Top panel considers continuous phenotypes and bottom panel considers dichotomous phenotypes. From left to right, the plots consider the setting in which 10% of rare variants were causal, 20% of rare variants were causal, and 50% of rare variants were causal. The detailed simulation setups are described in "Simulations and Real Data Analysis".

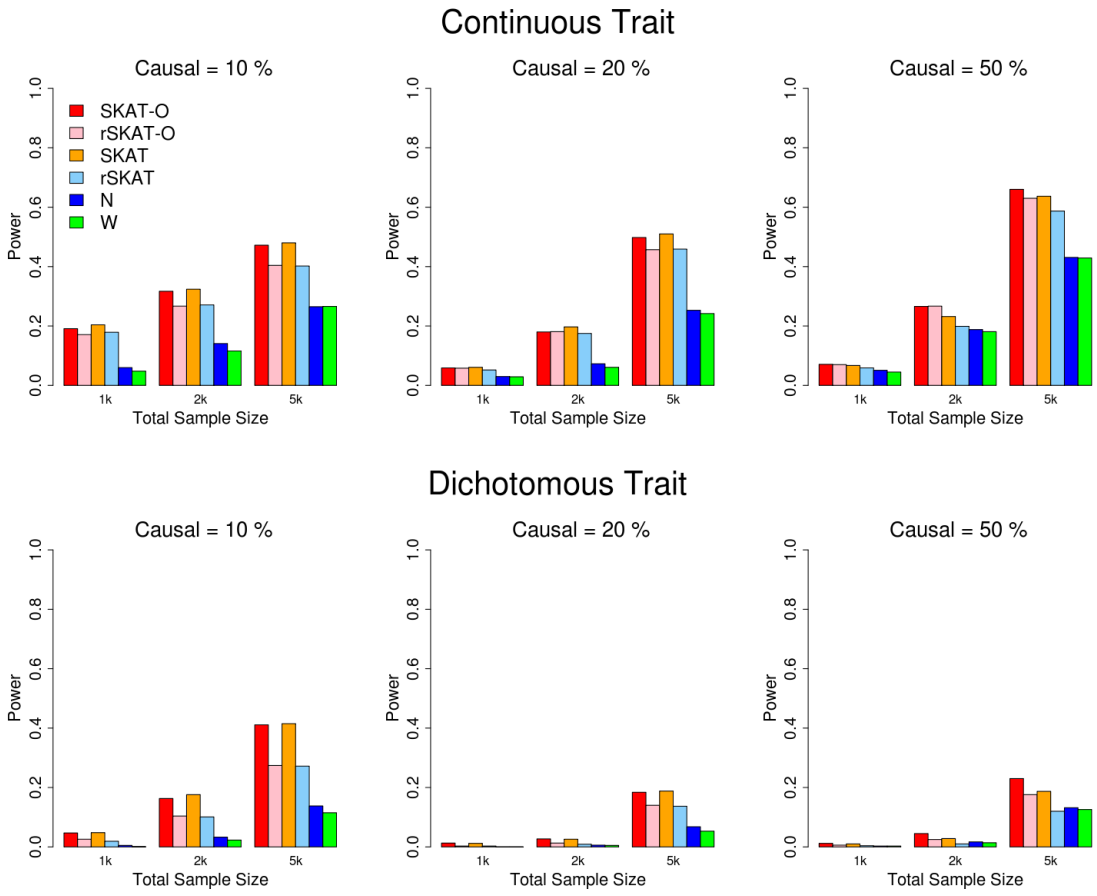

Figure 2: Empirical power of SKAT-O and competing methods at $\alpha = 2.5 \times 10^{-6}$ using simulation studies when region size=3kb and $\beta + /- = 80/20$. Top panel considers continuous phenotypes and bottom panel considers dichotomous phenotypes. From left to right, the plots consider the setting in which 10% of rare variants were causal, 20% of rare variants were causal, and 50% of rare variants were causal. The detailed simulation setups are described in "Simulations and Real Data Analysis".

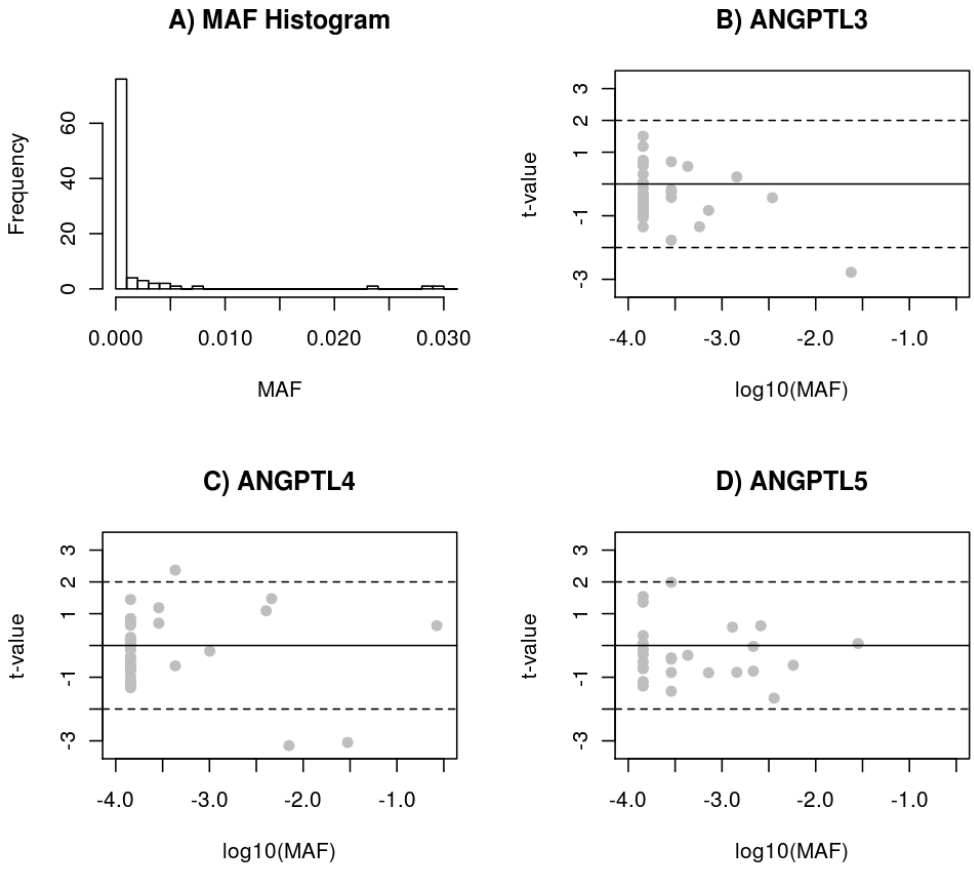

Figure 3: Single variant analysis results of Dallas Heart Study data. A) Histogram of minor allele frequencies of 92 variants with MAF < 0.03. B-D) Plots of log10(MAF) vs. t-statistic values of each variant of ANGPTL 3,4 and 5 genes. The dashed line represents the 95% confidence interval of no-association.