Boosting the Power of the Sequence Kernel Association Test by Properly Estimating Its Null Distribution

Kai Wang[1,*]

The sequence kernel association test (SKAT) is probably the most popular statistical test used in rare-variant association studies. Its null distribution involves unknown parameters that need to be estimated. The current estimation method has a valid type I error rate, but the power is compromised given that all subjects are used for estimation. I have developed an estimation method that uses only control subjects. Named SKAT+, this method uses the same test statistic as SKAT but differs in the way the null distribution is estimated. Extensive simulation studies and applications to data from the Genetic Analysis Workshop 17 and the Ocular Hypertension Treatment Study demonstrated that SKAT+ has superior power over SKAT while maintaining control over the type I error rate. This method is applicable to extensions of SKAT in the literature.

Introduction

Since the proliferation of DNA microarray technology about a decade ago, genome-wide association studies (GWASs) have successfully discovered many genetic variants associated with numerous human diseases and traits.[1] However, these identified variants explain only a small fraction of the overall heritability for most complex traits.[2–5] Because DNA microarray technology targets only common SNPs, it is possible that the "missing heritability" is due to variants that are rare. The rapid development of next-generation sequencing technology provides a great opportunity for studying rare variants.

Because a single SNP can provide only limited power, it is important to simultaneously leverage multiple SNPs in a gene or a region in rare-variant association studies. One approach is to test for the cumulative effect of rare variants.[6–10] Such burden tests are most powerful when the effects of all variants are in the same direction. When some variants are protective but others are deleterious, this approach is no longer optimal.

An alternative approach is to test for the cumulative quadratic effect of rare variants. Examples include the C-alpha test[11] and the sequence kernel association test (SKAT).[12] SKAT is a score-based variance-component test and is computationally efficient. It includes the C-alpha test as a special case. It has a solid theoretical foundation.[13,14] It can handle both dichotomous traits and continuous traits and is able to control for covariates. SKAT has been generalized in many ways for the purposes of incorporating burden tests,[15] conducting meta-analysis,[16] analyzing extreme continuous traits[17] and survival outcomes,[18–20] performing family-based association tests,[21] and studying gene-gene and gene-environmental interactions.[22,23]

One advantage of SKAT is that the form of its limiting distribution is known. It is a linear combination of a certain number of independent and identically distributed random variables, each of which follows a chi-square distribution with 1 degree of freedom (df). Although these combination coefficients are unknown, once they are estimated, one is able to compute the theoretical p value right away[24,25] without using computation-intensive techniques such as permutation. The focus of this report concerns ways to estimate these coefficients.

The current practice used by SKAT is to use all subjects in estimating these coefficients. However, as I will show, although the type I error rate is maintained, the power of SKAT is compromised in such a practice. Recognizing this fact, I propose a general approach to estimating these coefficients by using only control subjects. I demonstrate how this approach can be applied to dichotomous traits with and without covariates, as well as continuous traits. Its performance is assessed by extensive simulation studies. I further illustrate the utility of this method by applying it to data from the Genetic Analysis Workshop 17 (GAW17)[26] and the Ocular Hypertension Treatment Study (OHTS).[27]

Material and Methods

Let m be the number of SNPs in a candidate region or a pathway. The genotype score of subject i at SNP j is denoted by g_{ij}, which takes value 0, 1, or 2, the number of copies of the minor allele. The genotypes of all subjects can be summarized by an $n \times m$ matrix $\mathbf{G} = (g_{ij})$, where n is the number of subjects.

Case-Control Data with No Covariates

Let n_1 be the number of case subjects and n_0 be the number of control subjects. The total number of subjects is $n = n_1 + n_0$. Phenotypes are represented by an $n \times 1$ vector $\mathbf{y}$, whose elements are equal to 1 for case subjects and 0 for control subjects. Define

[1] Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, IA 52242, USA
*Correspondence: kai-wang@uiowa.edu

http://dx.doi.org/10.1016/j.ajhg.2016.05.011.

$\widehat{\pi} = n_1/n$ as the proportion of case subjects. The most popular version of the SKAT statistic is the one with linear kernel[12] and has the following form:

$$\mathrm{SKAT} = (\mathbf{y} - \widehat{\pi}1)^t\mathbf{G}\mathbf{G}^t(\mathbf{y} - \widehat{\pi}1).$$

When none of the m SNPs is associated with the case-control status (i.e., the null hypothesis), the distribution of SKAT can be approximated by that of $\sum_{k=1}^{n}\lambda_k\chi^2_{k,1}$, where $\{\lambda_k\}_{k=1,2,\ldots,n}$ are eigenvalues of the $n \times n$ matrix $\mathbf{P}^{1/2}\mathbf{G}\mathbf{G}^t\mathbf{P}^{1/2}$, which has $\mathbf{P} = \widehat{\pi}(1-\widehat{\pi})(\mathbf{I} - n^{-1}11^t)$, an $n \times n$ matrix, and $\{\chi^2_{k,1}\}_{k=1,2,\ldots,n}$ as independent 1-df chi-square-distributed random variables.[12] Here, $\mathbf{I}$ is an $n \times n$ identity matrix, and $\mathbf{1}$ is an $n \times 1$ vector of values of 1. The rank of matrix $\mathbf{P}^{1/2}\mathbf{G}\mathbf{G}^t\mathbf{P}^{1/2}$ is min $\{m, n\}$. Because the non-zero eigenvalues $\{\lambda_k\}_{k=1,2,\ldots,n}$ are the same as the non-zero eigenvalues from $\mathbf{G}^t\mathbf{P}\mathbf{G}$, the distribution of $\sum_{k=1}^{n}\lambda_k\chi^2_{k,1}$ is completely determined by the eigenvalues of $\mathbf{G}^t\mathbf{P}\mathbf{G}$. Note that the dimension of $\mathbf{G}^t\mathbf{P}\mathbf{G}$, which is $m \times m$, is smaller than that of $\mathbf{P}^{1/2}\mathbf{G}\mathbf{G}^t\mathbf{P}^{1/2}$ in the usual case of $m < n$. It is computationally more convenient to use $\mathbf{G}^t\mathbf{P}\mathbf{G}$.

Define the sample variance matrix of $\mathbf{G}$ as $\mathbf{S}$. We have

$$\mathbf{S} = \frac{1}{n-1}\mathbf{G}^t\left(\mathbf{I} - n^{-1}11^t\right)\mathbf{G}$$

$$= \frac{1}{(n-1)\widehat{\pi}(1-\widehat{\pi})}\mathbf{G}^t\mathbf{P}\mathbf{G}.$$

That is, $\mathbf{G}^t\mathbf{P}\mathbf{G} = (n-1)\widehat{\pi}(1-\widehat{\pi})\mathbf{S}$. The matrix $\mathbf{S}$ is an estimate of the underlying variance matrix of m SNPs when the null hypothesis of no association is true. When none of the SNPs is associated with the phenotype (the null hypothesis), it makes sense to use all subjects to compute $\mathbf{S}$ for the null distribution because none of the subjects carries a variant, and all of them are expected to share a common variance of genotype scores. However, when some SNPs are associated with the phenotype, using all subjects is no longer appropriate because case and control subjects are expected to have different mean SNP scores at these SNPs. Their variance matrices of the genotype scores are expected to be different as well. In this situation, a reasonable estimation of the null variance matrix of the m SNPs is the sample variance matrix of the genotypes scores for the control subjects.

I propose the following method for approximating the null distribution of SKAT. Let $\mathbf{S}_0$ be the sample variance matrix of the genotype scores on the m SNPs among the n_0 control subjects:

$$\mathbf{S}_0 = \frac{1}{n_0-1}\mathbf{G}_0^t\left(\mathbf{I}_0 - n_0^{-1}1_0 1_0^t\right)\mathbf{G}_0,$$

where $\mathbf{G}_0$ is an $n_0 \times m$ matrix of genotype scores of control subjects, $\mathbf{I}_0$ is an $n_0 \times n_0$ identity matrix, and $\mathbf{1}_0$ is an $n_0 \times 1$ vector of values of 1. Let $\{\tilde{\lambda}_k\}_{k=1,2,\ldots,n_0}$ be the eigenvalues of matrix $(n-1)\widehat{\pi}(1-\widehat{\pi})\mathbf{S}_0$. The null distribution of the SKAT statistic is approximated by $\sum_{k=1}^{m}\tilde{\lambda}_k\chi^2_{k,1}$ provided that $n_0 > m$. I call this method SKAT+.

Apparently, when none of the SNPs is associated with the phenotype, $\mathbf{S}_0$ and $\mathbf{S}$ converge to the same variance matrix as n_0 and n get large. That is, using either $\mathbf{S}_0$ or $\mathbf{S}$ will give a valid type I error rate. However, when the null hypothesis is not true, $\mathbf{S}_0$ and $\mathbf{S}$ no longer converge to the same variance matrix. The trace of the limiting matrix of $\mathbf{S}$ is larger than that of $\mathbf{S}_0$. Using $\mathbf{S}$ instead of $\mathbf{S}_0$ tends to get a distribution with larger variance and a larger p value and hence a less powerful testing procedure.

Case-Control Data with Covariates

Let $\mathbf{x}_i$ be a $p \times 1$ vector of the values of p covariates for subject i. To remove their confounding effect, the following logistic regression is fit first:

$$\mathrm{logit}\quad \Pr(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}^t\mathbf{x}_i.$$

Let $\widehat{\pi}_i$ be the estimated probability of $y_i = 1$:

$$\widehat{\pi}_i = \frac{\exp\{\widehat{\alpha}_0 + \widehat{\boldsymbol{\alpha}}^t\mathbf{x}_i\}}{1 + \exp\{\widehat{\alpha}_0 + \widehat{\boldsymbol{\alpha}}^t\mathbf{x}_i\}},$$

where $\widehat{\alpha}_0$ and $\widehat{\boldsymbol{\alpha}}$ are estimates of α_0 and $\boldsymbol{\alpha}$, respectively. Let $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \widehat{\pi}_2, \ldots, \widehat{\pi}_n)^t$. The SKAT[12] statistic is now defined by

$$\mathrm{SKAT} = (\mathbf{y} - \widehat{\boldsymbol{\pi}})^t\mathbf{G}\mathbf{G}^t(\mathbf{y} - \widehat{\boldsymbol{\pi}}).$$

When the sample size is large enough, the distribution of SKAT can be approximated by that of $\sum_{k=1}^{n}\lambda_k\chi^2_{k,1}$, where $\{\lambda_k\}_{k=1,2,\ldots,n}$ are eigenvalues of $\mathbf{P}^{1/2}\mathbf{G}\mathbf{G}^t\mathbf{P}^{1/2}$ or, equivalently, those of $\mathbf{G}^t\mathbf{P}\mathbf{G}$. Here,

$$\mathbf{P} = \mathbf{V} - \mathbf{V}\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^t\mathbf{V}\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}^t\mathbf{V},$$

which has $\tilde{\mathbf{X}} = [1 \quad \mathbf{X}]$, an $n \times (p+1)$ matrix, and

$$\mathbf{V} = \mathrm{diag}(\widehat{\pi}_1(1-\widehat{\pi}_1), \widehat{\pi}_2(1-\widehat{\pi}_2), \ldots, \widehat{\pi}_n(1-\widehat{\pi}_n)).$$

Define $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \frac{1}{n-p-1}\mathbf{G}^t\mathbf{P}\mathbf{G}.$$

$\boldsymbol{\Sigma}$ is the variance matrix of the residuals of the projection of $\mathbf{V}^{1/2}\mathbf{G}$ on the linear space spanned by $\mathbf{V}^{1/2}\tilde{\mathbf{X}}$, where $\mathbf{V}^{1/2} = \mathrm{diag}(\sqrt{\widehat{\pi}_1(1-\widehat{\pi}_1)}, \ldots, \sqrt{\widehat{\pi}_n(1-\widehat{\pi}_n)})$. Following the same idea as in the previous subsection, we substitute $\boldsymbol{\Sigma}$ with its counterpart computed with subjects whose value of $y_i - \widehat{\pi}_i$ is low. One decision we have to make is the threshold for selection. For instance, we can use those subjects whose $y_i - \widehat{\pi}_i$ value is below the median or third quartile of $\{(y_i - \widehat{\pi}_i)\}_{i=1,\ldots,n}$. Another choice is to use control subjects only. This is equivalent to using 0 as the threshold given that all control subjects satisfy $y_i - \widehat{\pi}_i \leq 0$.

Let n_0 be the number of such chosen subjects. Let $\mathbf{G}_0$, $\mathbf{V}_0$, and $\tilde{\mathbf{X}}_0$ denote the sub-matrices of $\mathbf{G}$, $\mathbf{V}$, and $\tilde{\mathbf{X}}$, respectively, corresponding to the selected subjects. Let $\mathbf{P}_0$ and $\boldsymbol{\Sigma}_0$ be defined as

$$\mathbf{P}_0 = \mathbf{V}_0 - \mathbf{V}_0\tilde{\mathbf{X}}_0\left(\tilde{\mathbf{X}}_0^t\mathbf{V}_0\tilde{\mathbf{X}}_0\right)^{-1}\tilde{\mathbf{X}}_0^t\mathbf{V}_0.$$

and

$$\boldsymbol{\Sigma}_0 = \frac{1}{n_0-p-1}\mathbf{G}_0^t\mathbf{P}_0\mathbf{G}_0,$$

respectively. Let $\{\tilde{\lambda}_k\}_{k=1,2,\ldots,m}$ be the eigenvalues of matrix $(n-p-1)$ $\boldsymbol{\Sigma}_0$. Because $\mathbf{G}^t\mathbf{P}\mathbf{G}$ can be expressed as $\mathbf{G}^t\mathbf{P}\mathbf{G} = (n-p-1)\boldsymbol{\Sigma}$, the null distribution of the SKAT statistic is approximated by $\sum_{k=1}^{m}\tilde{\lambda}_k\chi^2_{k,1}$ when $n_0 > m$. This is because when the null is true, the residuals $\{y_i - \widehat{\pi}_i\}_{i=1,\ldots,n}$ are independent of the covariate-corrected genotypes. Hence, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}$ converge to the same matrix as n_0 and n get large.

When there are no covariates, we have $\tilde{\mathbf{X}} = 1$ and $\mathbf{P} = \widehat{\pi}(1-\widehat{\pi})(\mathbf{I} - n^{-1}11^t)$, where $\widehat{\pi} = n_1/n$. This method reduces to the method proposed in the previous subsection.

Continuous Traits

First, we fit the following null linear regression that has covariates only (that is, no genotype scores are used):

$$y_i = \alpha_0 + \alpha^t \mathbf{X}_i + \epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma^2)$. Let $\widehat{y}_i$ denote the predicted value of y_i and $\widehat{\mathbf{y}} = (\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n)^t$. The estimates of α_0, $\boldsymbol{\alpha}$, and σ^2 are denoted by $\widehat{\alpha}_0$, $\widehat{\boldsymbol{\alpha}}$, and $\widehat{\sigma}^2$, respectively. The SKAT statistic with linear kernel is defined as[12]

$$\text{SKAT} = (\mathbf{y} - \widehat{\mathbf{y}})^t \mathbf{G}\mathbf{G}^t (\mathbf{y} - \widehat{\mathbf{y}}).$$

Its asymptotic distribution is $\sum_{k=1}^{n} \lambda_k \chi^2_{k,1}$, where $\{\lambda_k\}_{k=1,...,n}$ are the eigenvalues of $\mathbf{G}^t\mathbf{P}\mathbf{G}$, in which $\mathbf{P} = \widehat{\sigma}^2(\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^t\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^t)$ and $\tilde{\mathbf{X}} = [1 \quad \mathbf{X}]$.

We write the matrix $\mathbf{G}^t\mathbf{P}\mathbf{G}$ as $\mathbf{G}^t\mathbf{P}\mathbf{G} = (n - p - 1)\widehat{\sigma}^2\mathbf{S}$, where

$$\mathbf{S} = \frac{1}{n - p - 1}\mathbf{G}^t\left(\mathbf{I} - \tilde{\mathbf{X}}(\tilde{\mathbf{X}}^t\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^t\right)\mathbf{G}$$

denotes the variance matrix of the genotype scores after the effects of the covariates are removed. Assuming that higher phenotypic value is positively associated with disease severity, we can choose subjects whose residual $y_i - \widehat{y}_i$ is low. Denote the matrices of the genotypes and the covariates of these selected subjects as $\mathbf{G}_0$ and $\mathbf{X}_0$, respectively. Define $\tilde{\mathbf{X}}_0 = [1 \quad \mathbf{X}_0]$. Furthermore, denote their covariate-corrected genotype-score variance matrix as $\mathbf{S}_0$. The asymptotic distribution of SKAT is then estimated by $\sum_{k=1}^{n} \tilde{\lambda}_k \chi^2_{k,1}$, where $\{\tilde{\lambda}_k\}$ are the eigenvalues of $(n - p - 1)\widehat{\sigma}^2\mathbf{S}_0$. $\mathbf{S}_0$ and $\mathbf{S}$ converge to the same matrix under the null given that the residuals $\{y_i - \widehat{y}_i\}$ are independent of the genotypes, and selection based on $\{y_i - \widehat{y}_i\}$ doesn't bias $\mathbf{S}_0$ as an estimate of the underlying genotype covariance matrix.

A Resampling Procedure

To obtain a resampling p value, we can use the following procedure. Given the n_0 selected subjects and their genotype-score matrix $\mathbf{G}_0$, we randomly select n_0 phenotype residuals $\{y_i - \widehat{\pi}_i\}$ (or $\{y_i - \widehat{\mathbf{y}}_i\}$ for continuous traits). Let $\tilde{\mathbf{y}}_0 = \mathbf{y}_0 - \widehat{\boldsymbol{\pi}}_0$ denote the vector of selected phenotype residuals. A resampling version of the SKAT statistic is computed as

$$\frac{n - p - 1}{n_0 - p - 1}\tilde{\mathbf{y}}_0^t(\mathbf{I}_0 - n_0^{-1}1_0 1_0^t)\mathbf{G}_0\mathbf{G}_0^t(\mathbf{I}_0 - n_0^{-1}1_0 1_0^t)\tilde{\mathbf{y}}_0.$$

Repeat this process the desired number of times. The resampling p value is computed as the proportion of the resampled SKAT statistic that is equal to or larger than the observed SKAT statistic.

Results

Simulation Studies

I simulated dichotomous traits and continuous traits under various configurations in order to evaluate the performance of the proposed SKAT+ method. I calculated both the p value based on the asymptotic distribution and the resampling p value based on 1,000 resamples. I compared SKAT+ with SKAT, SKAT-O,[28] and aSPU,[29] a recent competitor to SKAT. The aSPU method depends on the best p value of a number of sums of powered score statistics at each SNP. For the SKAT method, I used the R package SKAT (version 1.1.2). Small-sample adjustment was turned off, and no weights were applied to SNPs, although such tricks can be used in SKAT+ as well. For the aSPU method, I downloaded (on November 30, 2015) the permutation-based aSPU R code aSPUperm.R from Dr. Wei Pan's website. This aSPU R code works only for dichotomous traits, although in theory it works for continuous traits as well. The number of permutations for aSPU is fixed at 1,000. Under each configuration, the number of simulation

Table 1. Simulated Type I Error Rate for Dichotomous Traits over 1,000 Replicates

	$\alpha = $ **0.05**					$\alpha = $ **0.01**				
	No. of Neutral Rare Variants					**No. of Neutral Rare Variants**				
Method	**0**	**4**	**8**	**16**	**32**	**0**	**4**	**8**	**16**	**32**
Without Covariates										
SKAT+	0.046	0.040	0.047	0.043	0.040	0.007	0.005	0.013	0.005	0.006
SKAT+ (rs)	0.063	0.046	0.059	0.062	0.049	0.017	0.012	0.017	0.014	0.013
SKAT	0.050	0.036	0.049	0.049	0.040	0.007	0.006	0.011	0.004	0.006
SKAT-O	0.041	0.041	0.052	0.052	0.045	0.009	0.008	0.010	0.004	0.005
aSPU	0.037	0.037	0.047	0.048	0.041	0.005	0.007	0.009	0.005	0.005
With Covariates										
SKAT+	0.050	0.042	0.038	0.051	0.039	0.004	0.011	0.002	0.012	0.004
SKAT+ (rs)	0.062	0.047	0.049	0.066	0.050	0.012	0.016	0.010	0.017	0.008
SKAT	0.042	0.039	0.044	0.049	0.045	0.006	0.008	0.003	0.013	0.008
SKAT-O	0.055	0.046	0.054	0.057	0.051	0.006	0.013	0.005	0.012	0.008
aSPU	0.045	0.046	0.049	0.053	0.051	0.005	0.011	0.007	0.016	0.011

SKAT+ (rs) represents the SKAT+ method based on resampling p values.

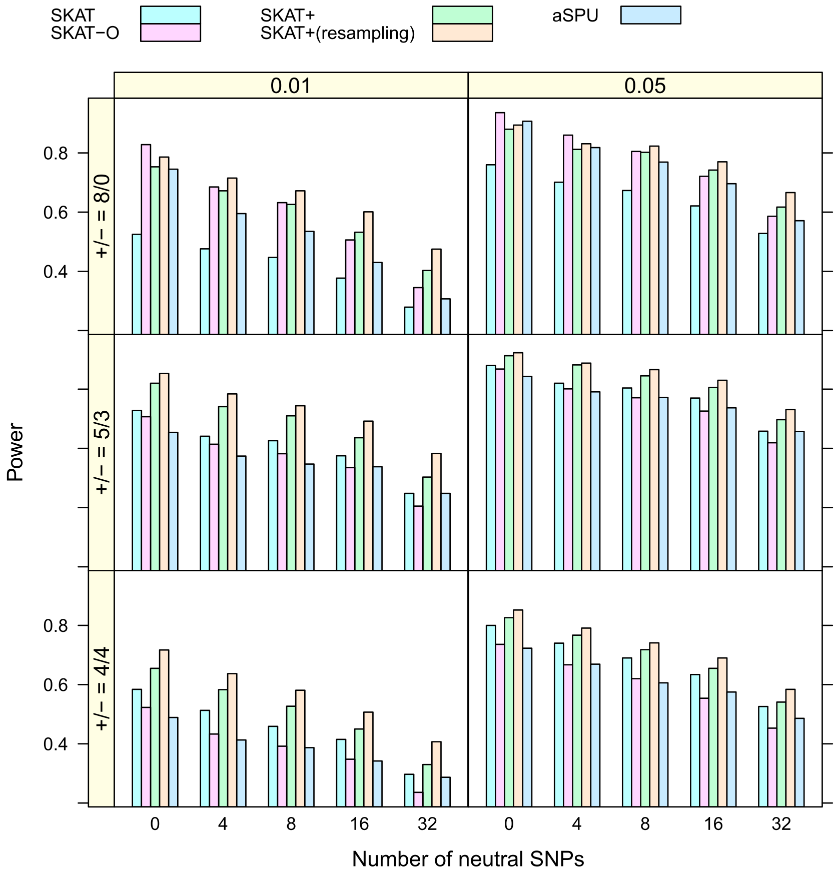

Figure 1. Simulated Power of Dichotomous Traits over 1,000 Replicates
There are no covariates. Only control subjects were selected for SKAT+.

replications for all methods is fixed at 1,000. Rejection rates at significance levels 0.05 and 0.01 are reported unless otherwise noted.

For the assessment of type I error rate and power for dichotomous traits, I used an R code used previously[30] and many other publications coauthored by Dr. Wei Pan. This R code (simRareSNP.R) was downloaded (on November 30, 2015) from Dr. Wei Pan's website. Specifically, there are eight disease-susceptibility variants whose odds ratios for each additional copy of the minor allele are pre-specified. For instance, an odds ratio equal to 1 for each of these variants implies that none of them is associated with the simulated trait. So, this is a setup used for the study of type I error rate. The disease probability can be expressed via the following logistic-regression model:

$$\text{logit}[\Pr(\text{disease})] = \beta_0 + \beta_1 g_1 + \ldots + \beta_8 g_8,$$

where $\beta_0 = -\log(0.05/0.95)$ corresponds to a background disease probability of 0.05, and $\exp(\beta_1), \ldots, \exp(\beta_8)$ are the odds ratios for the eight disease-susceptibility variants, respectively. In addition, a varying number of neutral rare variants whose minor allele frequencies (MAFs) were randomly chosen from a uniform distribution on interval (0.001, 0.01) were simulated. All variants were in linkage equilibrium. A total of 500 case and 500 control subjects were simulated. A similar simulation setup has also been used elsewhere.[12,31] Only control subjects were used for the SKAT+ method. Simulated type I error rates with odds ratios equal to 1 for each of the eight disease-susceptibility variants are reported in the first half of Table 1. They are under control for all four statistics except that the resampling p values for SKAT+ are slightly liberal in some situations.

I also considered the case where there are two covariates, X_1 and X_2. The disease probability is determined by

$$\begin{aligned}\text{logit}[\Pr(\text{disease})] = \beta_0 + 0.5\,X_1 \\ + 0.5\,X_2 + \beta_1 g_1 \\ + \ldots + \beta_8 g_8,\end{aligned}$$

where X_1 follows a standard normal distribution, and X_2 follows a binary distribution taking value 0 or 1 with probability 0.5. This way of adding covariates has been used elsewhere.[12] To accommodate covariates X_1 and X_2, I modified Dr. Wei Pan's R code accordingly. Subjects whose $y_i - \widehat{\pi}$ values were not higher than the third quartile of $\{(y_i - \widehat{\pi})\}_{i=1,\ldots,1{,}000}$ were selected for the SKAT+ method. Simulated type I error rates are reported in the second half of Table 1. Again, they are under control except that the resampling p values for SKAT+ are slightly liberal in some situations.

For the power analysis, I considered three scenarios in terms of the odds ratios of the eight disease-causing SNPs. They were (1) ORs = (2, 2, 2, 2, 2, 2, 2, 2), (2) ORs = (3, 3, 2, 2, 2, 1/2, 1/2, 1/2), and (3) ORs = (3, 1/3, 2, 2, 2, 1/2, 1/2, 1/2). These are typical scenarios used in the literature.[29,30] Scenario 1 mimics the situation where all causal variants work in the same direction, whereas the other two allow their directions to be different with different deleterious/protective ratios. The power is presented in Figure 1 and Figure 2 for situations without and with covariates, respectively. The SKAT+ method performed apparently better—in some cases much better—than SKAT across the board. It also performed better than SKAT-O and aSPU in almost all cases even when ORs = (2, 2, 2, 2, 2, 2, 2, 2), a case in favor of SKAT-O and aSPU because this is a situation for which they are optimized.[28,30]

Continuous traits were simulated according to the following model:[12]

$$y = 0.5\,X_1 + 0.5\,X_2 + \beta_1 g_1 + \ldots + \beta_8 g_8, \qquad \text{(Equation 1)}$$

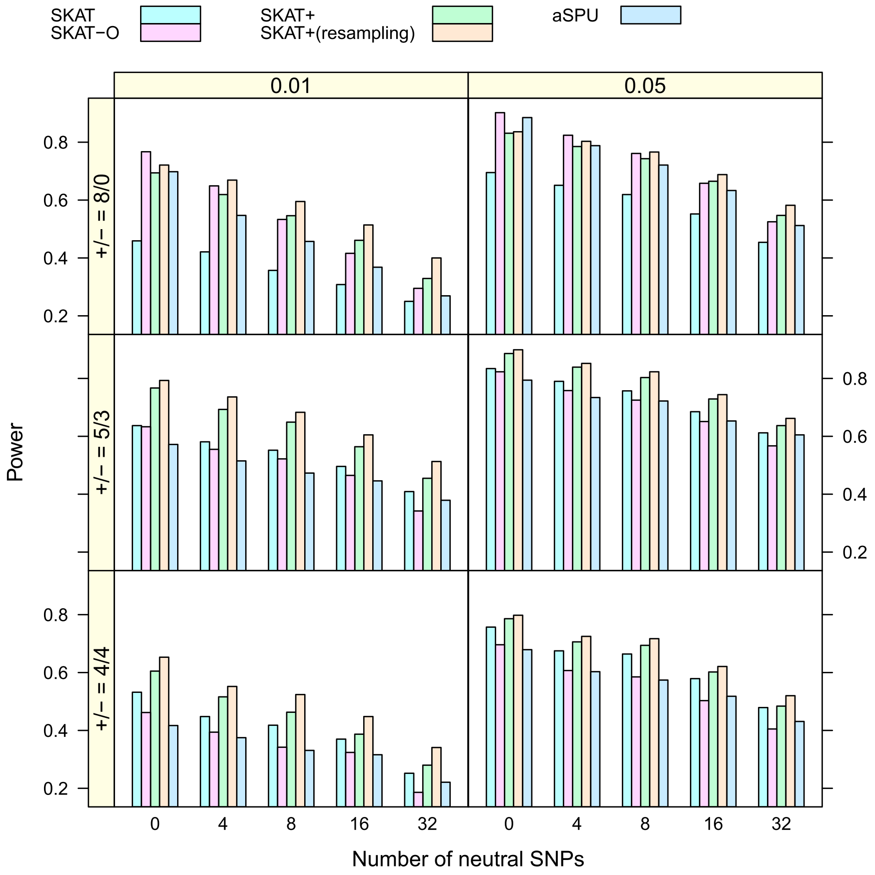

Figure 2. Simulated Power of Dichotomous Traits over 1,000 Replicates
There are two covariates. The third quartile of $\{y_i - \widehat{\pi}_i\}_{i=1,\ldots,1{,}000}$ was used for selecting subjects for SKAT+.

where X_1 and X_2 were generated in the same way as in the case of dichotomous traits, and g_1, …, g_8 were genotype scores at eight causal rare variants. The MAFs of these causal variants were selected uniformly from the interval (0.001, 0.01). For the study of type I error rate, $\beta_1, \ldots, \beta_8$ were set at 0. For the study of power, the magnitude of β_j was equal to $|0.2\log_{10}(\text{MAF})|$, and the signs of these β values were in one of the following three configurations: (1) signs = (+, +, +, +, +, +, +, +), (2) signs = (+, +, +, +, +, +, −, −), or (3) signs = (+, +, +, +, −, −, −, −). Additional neutral SNPs were simulated with MAFs uniformly chosen from the interval (0.001, 0.01). Sample size was 1,000, and subjects whose $y_i - \widehat{y}_i$ values were below the third quartile of the residuals were selected for SKAT+. That is, 75% were selected. The type I error rates are presented in Table 2. All are in line with the nominal levels. The simulated power is presented in Figure 3. SKAT+ performed better than SKAT when the ratio of positive β values to negative β values was 8:0 (configuration 1) or 6:2 (configuration 2). When this ratio was 4:4 (configuration 3), the performance of SKAT+ was almost identical to that of SKAT. This is because at this ratio, the simulated trait according to model 1 is symmetric with respect to 0. After the effect of covariates is removed, the sample covariance matrix of the selected subjects is expected to be the same as that of the whole sample. Compared with SKAT-O, SKAT+ performed worse only in configuration 1, a situation optimized for SKAT-O. In other situations, SKAT-O performed even worse than SKAT.

Application to GAW17 Data

GAW17 provided 200 phenotype datasets[26] simulated with mini-exome genotype data on 697 subjects selected from the 1000 Genomes Project. There were 24,487 SNPs in 3,205 genes. Four phenotypes were simulated on the basis of the genotypes of these subjects. Three of them (denoted by Q1, Q2, and Q4) were quantitative, and one was dichotomous such that the affection status was determined by Q1, Q2, and Q4 and a latent liability through a liability threshold model. Q1 was influenced by 39 SNPs in 9 genes. Q2 was influenced by 72 SNPs in 13 genes.

Table 2. Simulated Type I Error Rate for Continuous Traits over 1,000 Replicates

	α = **0.05**					α = **0.01**				
	No. of Neutral Rare Variants					**No. of Neutral Rare Variants**				
Method	**0**	**4**	**8**	**16**	**32**	**0**	**4**	**8**	**16**	**32**
SKAT+	0.054	0.055	0.052	0.042	0.048	0.009	0.010	0.007	0.011	0.006
SKAT+ (rs)	0.051	0.058	0.051	0.040	0.056	0.011	0.011	0.008	0.013	0.010
SKAT	0.048	0.048	0.045	0.036	0.043	0.008	0.006	0.005	0.009	0.006
SKAT-O	0.039	0.049	0.037	0.038	0.043	0.008	0.008	0.005	0.009	0.006

Subjects whose $y_i - \widehat{y}_i$ values were among the lowest 75% were selected for the SKAT+ method. SKAT+ (rs) represents the SKAT+ method based on resampling p values. The aSPU method was not computed because the R code from its authors doesn't handle continuous traits.

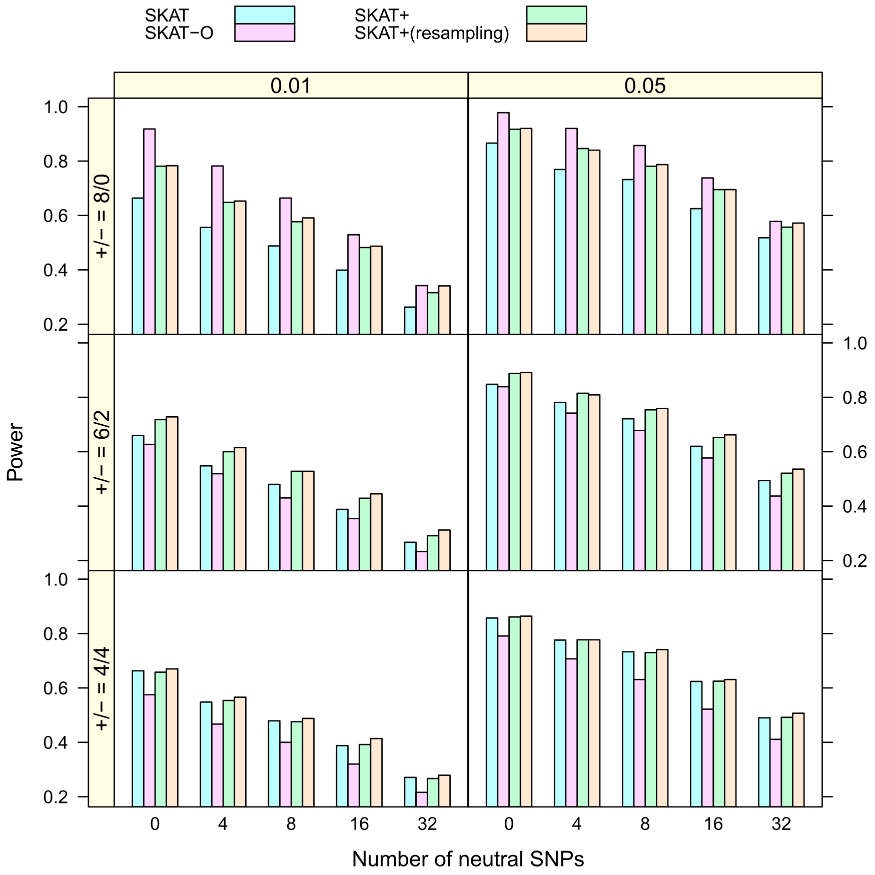

Figure 3. Simulated Power of Continuous Traits over 1,000 Replicates
There are two covariates. The third quartile of $\{y_i - \widehat{y}_i\}_{i=1,\ldots,1{,}000}$ was used for selecting subjects for SKAT+.

Q4 was not influenced by any of the genotyped SNPs. The latent liability was influenced by 51 SNPs in 15 genes. As a result, the dichotomous trait was influenced by 162 genotyped SNPs in 36 genes (Q1 and the latent liability shared one common gene). Age, gender, and smoking status were confounders to Q1, Q4, and the latent liability, but not to Q2. I compared the proportion of times SKAT+, SKAT, and SKAT-O were significant at level 0.05 at each causal gene across 200 simulated datasets. I focused on the continuous traits Q1, Q2, and Q4. The following linear null model was considered:

$$y = \alpha_0 + \alpha_1 \text{ age} + \alpha_2 \text{ smoke} + \alpha_3 \text{ gender}.$$

Note that covariates were controlled for although they were known to have no effect on Q2. Q4 was used here for confirming that the type I error rate of SKAT+ was under control. The effects of the genotyped causal SNPs were all in the same direction of increasing the continuous traits. Only SNPs whose MAFs were no larger than 1% were used. Genes with only one genotyped SNP were excluded. Because the majority of the MAFs at the causal genes were extremely rare at a magnitude of 1×10^{-4}, the 95th percentile of the residuals was used for selecting subjects for SKAT+. The proportion of rejecting the null hypothesis at level 0.05 out of the 200 simulated replicates is presented in Table 3. Given that the power was not high for most causal genes even at level 5% for all methods, only those genes at which the maximum power of the three statistics was greater than 10% are presented. It is clear that SKAT+ has higher power than SKAT. Its power is less than SKAT-O in most situations. This is very possibly because the direction of the causal variants in these genes is simulated to be the same, a situation SKAT-O is optimized for. The results for trait Q4 suggest that the type I error rates for all three methods are under control, especially given that there were only 200 replicates.

Application to OHTS

Primary open-angle glaucoma (POAG [MIM: 137760]) is a leading cause of irreversible blindness. Although a genetic basis has been established for a substantial fraction of POAG, no risk alleles of major effect have been identified.[32] The etiology of POAG is likely to be complex. Because POAG is assessed through quantitative measures such as central corneal thickness (CCT), intraocular pressure, and cup-to-disc ratio, one promising research direction is to map genes underlying these quantitative measures. Indeed, large-scale GWASs have identified genes that affect CCT.[33–35] Using data from the OHTS,[27] I applied the methods SKAT+, SKAT, and SKAT-O to a gene-based GWAS of CCT.

OHTS is a multi-center, randomized clinical trial sponsored by the National Eye Institute. Its goal is to investigate the efficacy of medical treatment in delaying or preventing the onset of POAG in individuals with elevated intraocular pressure. A total of 1,636 individuals between 40 and 80 years old were enrolled, and 1,077 of them were genotyped in a subsequent study. Data for this genetic study are available from dbGaP (study accession number dbGaP: phs000240.v1.p1). Both genotype data and baseline phenotype data are available for 1,057 subjects. The vast majority of these subjects are non-Hispanic white (752) and black (249).

There were 1,051,295 genotyped SNPs and 30,562 autosomal genes. The HGNC gene symbols were obtained with the R/Bioconductor package biomaRt (version 2.26.1). As in Lee et al.,[28] genes that contained fewer than three SNPs were excluded from further consideration. This reduced the number of genes to 23,778.

Table 3. Proportion of Significant Test Results at Level 0.05 out of the 200 Simulated GAW17 Datasets

	Q1			Q2			Q4		
Gene	SKAT+	SKAT	SKAT-O	SKAT+	SKAT	SKAT-O	SKAT+	SKAT	SKAT-O
FLT1 (MIM: 165070)	1.000	1.000	1.000	–	–	–	0.045	0.045	0.035
FLT4 (MIM: 136352)	0.500	0.430	0.570	–	–	–	0.055	0.065	0.045
HIF1A (MIM: 603348)	0.110	0.100	0.195	–	–	–	0.070	0.060	0.065
KDR (MIM: 191306)	0.995	0.995	1.000	–	–	–	0.075	0.065	0.075
BCHE (MIM: 177400)	–	–	–	0.400	0.340	0.410	0.050	0.040	0.055
PDGFD (MIM: 609673)	–	–	–	0.710	0.680	0.715	0.035	0.025	0.035
RARB (MIM: 180220)	–	–	–	0.440	0.420	0.345	0.065	0.055	0.050
SIRT1 (MIM: 604479)	–	–	–	0.625	0.600	0.640	0.080	0.055	0.075
SREBF1 (MIM: 184756)	–	–	–	0.425	0.375	0.670	0.065	0.065	0.075
VLDLR (MIM: 192977)	–	–	–	0.185	0.175	0.180	0.035	0.030	0.040
VNN1 (MIM: 603570)	–	–	–	0.195	0.185	0.150	0.075	0.065	0.085
VWF (MIM: 613160)	–	–	–	0.335	0.300	0.300	0.025	0.030	0.035

Because a thinner CCT increases the risk of POAG,[32] subjects with higher CCT measurements were used as control subjects in calculations of the null distribution for SKAT+. In particular, two inclusion criteria were used: the first and second quartiles. CCT measurements from both eyes were averaged and used as the response, and age and gender were used as covariates. The analysis was done on non-Hispanic white samples only, given that this group is much larger than other ethnic groups. A scatterplot matrix of the genome-wide gene-based p values from SKAT+ (with two different inclusion criteria), SKAT, and SKAT-O is shown in Figure 4. SKAT-O behaved rather differently from the others. The other three behaved pretty similarly to each other, but they did differ, especially SKAT+ with 50% of the subjects, which is not surprising. Most importantly, there doesn't seem to exist a systematic bias in favor of SKAT+. Those genes at which at least one of SKAT+, SKAT, and SKAT-O had a p value less than 0.0001 are listed in Table 4. These genes warrant further investigation. Unfortunately, they do not overlap *ZNF469* (MIM: 612078), *COL5A1* (MIM: 120215), *COL8A2* (MIM: 120252), *AKAP13* (MIM: 604686), or *AVGR8*, genes for which association with CCT has been reported previously.[33–35] I also conducted a simulation study to confirm that the type I error rate of SKAT+ is under control at a much higher level. For this purpose, genes *TRERF1* (MIM: 610322) and *IQUB* in Table 4 were arbitrarily selected. A continuous trait following a standard normal distribution and a dichotomous trait with disease probability 0.3 were each generated with 5×10^6 replicates. For the continuous trait, 75% of the subjects were used as control subjects. For the dichotomous trait, all simulated control subjects were used for the SKAT+ method. The expected proportion of control subjects is equal to $1 - 0.3 = 0.7$. The type I error rates are presented in Table 5. They are clearly under control.

Discussion

I have proposed the SKAT+ method for gene-based association testing. This method uses the same test statistic as SKAT but estimates the null distribution differently. By using a properly selected subset of subjects, this estimation method leads to a more powerful testing procedure. The selection is based on the residuals of the phenotype after the effect of covariates has been removed. The null distribution depends only on the second moments of the phenotype and the genotypes controlling for the effect of covariates. Selection based on phenotype residuals does not affect the validity of the test but has an effect on the power.

The current estimation method has a valid type I error rate, but the power is compromised because the estimated distribution does not correspond to the desired one when the null hypothesis is not true. It is contaminated by the distribution of SKAT under the alternative hypothesis. Because of this, the proposed SKAT+ method is almost surely destined to be more powerful than SKAT as sample size increases. No other SKAT competitors can make such a statement given that they perform better in only certain situations. To demonstrate this point, I simulated 10,000 datasets from the covariate-free dichotomous-trait model (Simulation Studies) used in the simulation studies with ORs = (2, 2, 2, 2, 2, 2, 2, 2) and no neutral SNPs. The $-\log_{10}$-transformed 10,000 p values for SKAT+ were plotted against SKAT p values transformed in the same way (Figure 5). The vast majority of the 10,000 p values from SKAT+ were much more significant than those

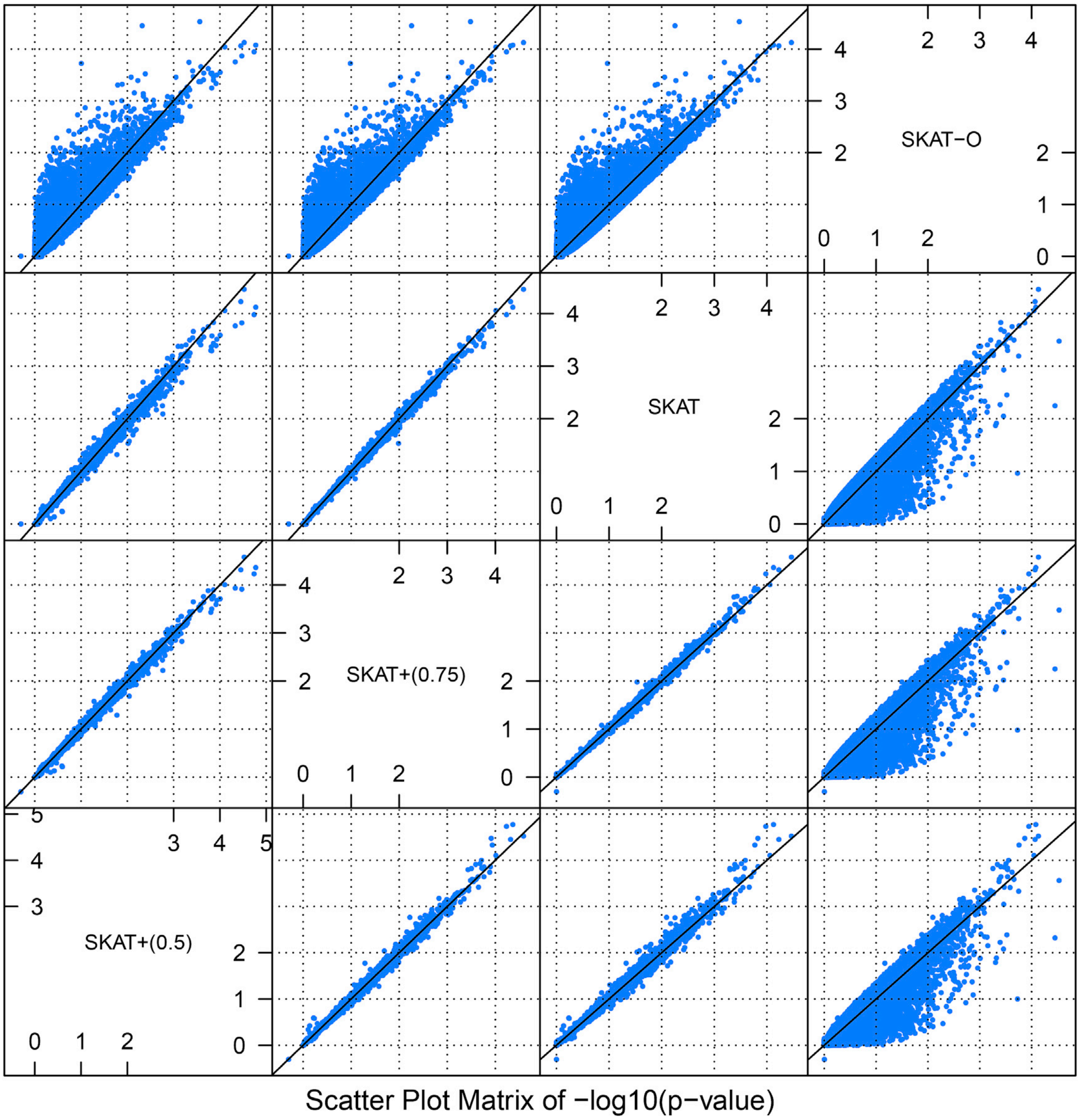

Figure 4. Scatterplots of Genome-wide Gene-Based p Values for the OHTS
SKAT+ (0.5) refers to the SKAT+ method using 50% of the subjects for its p value calculation. SKAT+ (0.75) uses 75% of the subjects.

from SKAT, and almost none of them were apparently worse.

Selection of subjects for the SKAT+ method is critical for power improvement. For case-control studies without covariates, the selection seems to be obvious. When there are covariates or the trait is continuous, the selection is less so. For the simulation studies, I selected subjects whose trait values were not higher than the third quartile after the effect of covariates had been removed. I also tried other thresholds, such as the median and the 80% quantile. The results are not shown, but they were consistent with the intuition that SKAT+ becomes more similar to SKAT as more subjects are selected. Generally, a lower threshold leads to fewer subjects and less accuracy in p value computation. The larger the sample size, the lower the threshold one can afford while still having enough selected subjects for estimating the null distribution with certain accuracy. SKAT+ contains SKAT as a special case. If small sample size is a concern, one is recommended to use SKAT instead of SKAT+. Furthermore, one can use the recently proposed small-sample method.[36] A previous method[28] for small-sample situations "can overcorrect and lead to inflated type I error."[36]

A related issue in selecting control subjects is the allele frequencies of the variants. The rarer they are, the more sensitive the sample variance matrix of genotype scores is to the selected control subjects. Hence, a larger number of control subjects should be used. For an MAF of 1%, there are on average two copies of the minor allele per 100 subjects. So, 500 subjects would have ten copies of the minor allele. If the MAFs are rarer, having a more stable sample variance matrix would require more subjects for the control group.

Table 4. A Summary of Gene-Based Association p Values with Data from the OHTS

Chromosome	Gene	SKAT+ (0.5)	SKAT+ (0.75)	SKAT	SKAT-O
2	*HAGLROS*	3.525×10^{-5}	4.821×10^{-5}	5.918×10^{-5}	8.619×10^{-5}
6	*EXOC2* (MIM: 615329)	2.725×10^{-4}	3.285×10^{-4}	3.327×10^{-4}	2.951×10^{-5}
	TRERF1 (MIM: 610322)	2.976×10^{-5}	2.586×10^{-5}	3.427×10^{-5}	7.390×10^{-5}
7	*NDUFA5* (MIM: 601677)	1.822×10^{-5}	5.848×10^{-5}	1.042×10^{-4}	1.129×10^{-4}
	IQUB	1.680×10^{-5}	4.322×10^{-5}	7.473×10^{-5}	8.301×10^{-5}
15	*MTMR10* (MIM: 208500)	4.767×10^{-3}	5.528×10^{-3}	5.592×10^{-3}	3.536×10^{-5}
	SNAP23 (MIM: 602534)	3.355×10^{-5}	1.217×10^{-4}	1.486×10^{-4}	1.359×10^{-4}
	HAUS2 (MIM: 613429)	9.859×10^{-5}	1.930×10^{-4}	2.581×10^{-4}	2.784×10^{-4}
	MYL12BP1	4.672×10^{-5}	1.155×10^{-4}	1.693×10^{-4}	1.748×10^{-4}
16	*TXNDC11*	7.867×10^{-5}	9.746×10^{-5}	8.708×10^{-5}	9.013×10^{-5}

Genes were selected if any statistic had a p value less than 0.0001 in the non-Hispanic white-only sample. SKAT+ (0.5) refers to SKAT+ method using 50% of the subjects for its p value calculation. SKAT+ (0.75) uses 75% of the subjects.

Note that when there are no covariates, it is possible to estimate the unknown null-distribution parameters by using data from a reference database, such as the 1000 Genomes Project or the International HapMap Project. One can compute the variance matrix for the SNPs used in the study and use it in place of the matrix $\mathbf{S}_0$. In this way, there is no need to select subjects for the SKAT+ method, but one needs to use an appropriate matching data sample. When there are covariates, generally it is not possible to use a reference database this way.

Also note that the basic principle behind SKAT+ is applicable to extensions of SKAT, as mentioned in the Introduction. It can also be generalized to phenotype with a distribution in the exponential family if the diagonal matrix $\mathbf{V}$ is replaced with a diagonal matrix of proper estimates of $\{\text{var}(y_i)\}_{i=1,\ldots,n}$.[14]

An implementation of the SKAT+ method is provided in the R package iGasso.

Acknowledgments

This work is dedicated to the memory of the late Dr. Leon F. Burmeister. Leon was a nice gentleman, a great colleague, and a wise mentor. I thank Leon for his long-time support and inspiration. Preparation of Genetic Analysis Workshop 17 (GAW17) data was partly supported by NIH grant R01 MH059490 and used data from the 1000 Genomes Project. The GAW17 was supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. I also thank the handling editor and three anonymous reviewers for their useful comments.

Received: January 21, 2016
Accepted: May 2, 2016
Published: June 9, 2016

Table 5. Simulated Type I Error Rates of SKAT+ for Two Top Genes in Table 4 over 5×10^6 Replicates

		Nominal Level		
Trait Type	Gene	10^{-4}	10^{-5}	2.5×10^{-6}
Continuous	*TRERF1*	8.1800×10^{-5}	2.8000×10^{-6}	1.2000×10^{-6}
	IQUB	1.0120×10^{-4}	7.6000×10^{-6}	2.0000×10^{-6}
Dichotomous	*TRERF1*	9.7600×10^{-5}	4.6000×10^{-6}	3.0000×10^{-6}
	IQUB	1.1128×10^{-4}	9.6000×10^{-6}	2.0000×10^{-6}

Only non-Hispanic white subjects were used. The continuous trait followed a standard normal distribution, and the dichotomous trait had a disease probability equal to 0.3. 75% of subjects were used as controls subjects.

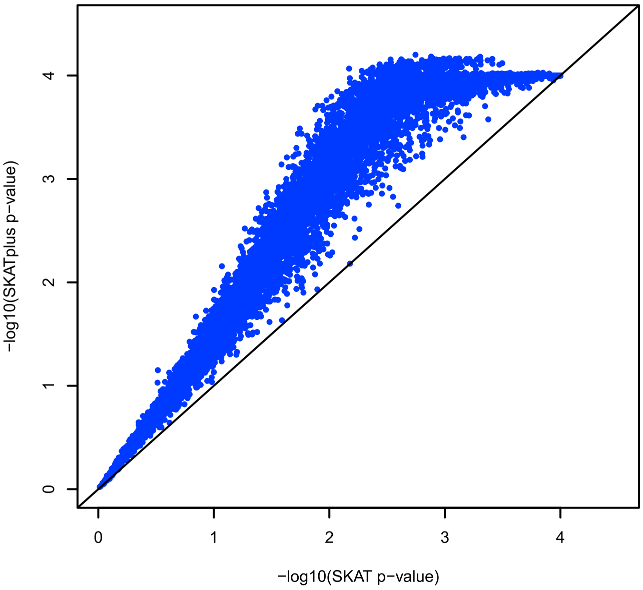

Figure 5. The p Values from SKAT+ versus the p Values from SKAT over 10,000 Replicates
The vast majority of the 10,000 p values from SKAT+ are much more significant than those from SKAT, given that the points are almost all above the 45° line. The simulation model is 1, with ORs = (2, 2, 2, 2, 2, 2, 2, 2) and no neutral SNPs.

Web Resources

1000 Genomes Project, http://www.1000genomes.org/
Dr. Wei Pan's website, http://www.biostat.umn.edu/~weip/prog
International HapMap Project, https://hapmap.ncbi.nlm.nih.gov/
OMIM, http://www.omim.org
R package iGasso, http://cran.r-project.org/web/packages/iGasso/index.html

References

1. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., and Parkinson, H. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.
2. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.
3. Zuk, O., Hechter, E., Sunyaev, S.R., and Lander, E.S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. Proc. Natl. Acad. Sci. USA *109*, 1193–1198.
4. Lee, S.H., Wray, N.R., Goddard, M.E., and Visscher, P.M. (2011). Estimating missing heritability for disease from genome-wide association studies. Am. J. Hum. Genet. *88*, 294–305.
5. Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H., and Nadeau, J.H. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. Nat. Rev. Genet. *11*, 446–450.
6. Morgenthaler, S., and Thilly, W.G. (2007). A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutat. Res. *615*, 28–56.
7. Madsen, B.E., and Browning, S.R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. *5*, e1000384.
8. Price, A.L., Kryukov, G.V., de Bakker, P.I., Purcell, S.M., Staples, J., Wei, L.-J., and Sunyaev, S.R. (2010). Pooled association tests for rare variants in exon-resequencing studies. Am. J. Hum. Genet. *86*, 832–838.
9. Li, B., and Leal, S.M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet. *83*, 311–321.
10. Morris, A.P., and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol. *34*, 188–193.
11. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.
12. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
13. Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics *63*, 1079–1088.
14. Liu, D., Ghosh, D., and Lin, X. (2008). 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*, 292.
15. Lee, S., Emond, M.J., Bamshad, M.J., Barnes, K.C., Rieder, M.J., Nickerson, D.A., Christiani, D.C., Wurfel, M.M., Lin, X., et al.; NHLBI GO Exome Sequencing Project—ESP Lung Project Team (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–237.
16. Lee, S., Teslovich, T.M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. Am. J. Hum. Genet. *93*, 42–53.
17. Barnett, I.J., Lee, S., and Lin, X. (2013). Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. Genet. Epidemiol. *37*, 142–151.
18. Lin, X., Cai, T., Wu, M.C., Zhou, Q., Liu, G., Christiani, D.C., and Lin, X. (2011). Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet. Epidemiol. *35*, 620–631.
19. Cai, T., Tonini, G., and Lin, X. (2011). Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics *67*, 975–986.
20. Chen, H., Meigs, J.B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. Genet. Epidemiol. *37*, 196–204.
21. Wang, X., Lee, S., Zhu, X., Redline, S., and Lin, X. (2013). GEE-based SNP set association test for continuous and discrete traits in family-based association studies. Genet. Epidemiol. *37*, 778–786.
22. Lin, X., Lee, S., Christiani, D.C., and Lin, X. (2013). Test for interactions between a genetic marker set and environment in generalized linear models. Biostatistics *14*, 667–681.
23. Lin, X., Lee, S., Wu, M.C., Wang, C., Chen, H., Li, Z., and Lin, X. (2016). Test for rare variants by environment interactions in sequencing association studies. Biometrics *72*, 156–164.
24. Davies, R. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. J. R. Stat. Soc. Ser. C Appl. Stat. *29*, 323–333.
25. Imhof, J. (1961). Computing the distribution of quadratic forms in normal variables. Biometrika *48*, 419–426.
26. Almasy, L., Dyer, T.D., Peralta, J.M., Kent, J.W., Jr., Charlesworth, J.C., Curran, J.E., and Blangero, J. (2011). Genetic Analysis Workshop 17 mini-exome simulation. BMC Proc. *5* (*Suppl 9*), S2.
27. Gordon, M.O., and Kass, M.A. (1999). The Ocular Hypertension Treatment Study: design and baseline description of the participants. Arch. Ophthalmol. *117*, 573–583.
28. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics *13*, 762–775.
29. Pan, W., Kim, J., Zhang, Y., Shen, X., and Wei, P. (2014). A powerful and adaptive association test for rare variants. Genetics *197*, 1081–1095.
30. Basu, S., and Pan, W. (2011). Comparison of statistical tests for disease association with rare variants. Genet. Epidemiol. *35*, 606–619.
31. Wang, T., and Elston, R.C. (2007). Improved power by use of a weighted score test for linkage disequilibrium mapping. Am. J. Hum. Genet. *80*, 353–360.

32. Fingert, J.H. (2011). Primary open-angle glaucoma genes. Eye (Lond.) *25*, 587–595.

33. Lu, Y., Dimasi, D.P., Hysi, P.G., Hewitt, A.W., Burdon, K.P., Toh, T., Ruddle, J.B., Li, Y.J., Mitchell, P., Healey, P.R., et al. (2010). Common genetic variants near the Brittle Cornea Syndrome locus ZNF469 influence the blinding disease risk factor central corneal thickness. PLoS Genet. *6*, e1000947.

34. Vitart, V., Bencić, G., Hayward, C., Skunca Herman, J., Huffman, J., Campbell, S., Bućan, K., Navarro, P., Gunjaca, G., Marin, J., et al. (2010). New loci associated with central cornea thickness include COL5A1, AKAP13 and AVGR8. Hum. Mol. Genet. *19*, 4304–4311.

35. Vithana, E.N., Aung, T., Khor, C.C., Cornes, B.K., Tay, W.-T., Sim, X., Lavanya, R., Wu, R., Zheng, Y., Hibberd, M.L., et al. (2011). Collagen-related genes influence the glaucoma risk factor, central corneal thickness. Hum. Mol. Genet. *20*, 649–658.

36. Chen, J., Chen, W., Zhao, N., Wu, M.C., and Schaid, D.J. (2016). Small sample kernel association tests for human genetic and microbiome association studies. Genet. Epidemiol. *40*, 5–19.