Received: 19 June 2017 | Accepted: 22 October 2017

DOI: 10.1002/gepi.22098

RESEARCH ARTICLE

WILEY Genetic Epidemiology

OFFICIAL JOURNAL INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY www.geneticepi.org

Gene-based genetic association test with adaptive optimal weights

Zhongxue Chen[1] | Yan Lu[2] | Tong Lin[3] | Qingzhong Liu[4] | Kai Wang[5]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, Bloomington, Indiana, United States of America

[2]Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, United States of America

[3]The Key Laboratory of Machine Perception (Ministry of Education), School of EECS, Peking University, Beijing, China

[4]Department of Computer Science, Sam Houston State University, Huntsville, Texas, United States of America

[5]Department of Biostatistics, College of Public Health, University of Iowa, Iowa City, Iowa, United States of America

Correspondence
Zhongxue Chen, Department of Epidemiology and Biostatistics, School of Public Health, Indiana University Bloomington, 1025 E. 7th street, Bloomington, IN 47405, USA.
Email: zc3@indiana.edu

ABSTRACT

It is well known that using proper weights for genetic variants is crucial in enhancing the power of gene- or pathway-based association tests. To increase the power, we propose a general approach that adaptively selects weights among a class of weight families and apply it to the popular sequencing kernel association test. Through comprehensive simulation studies, we demonstrate that the proposed method can substantially increase power under some conditions. Applications to real data are also presented. This general approach can be extended to all current set-based rare variant association tests whose performances depend on variant's weight assignment.

KEYWORDS

burden test, gene set, genetic association, SKAT, weighting

1 | INTRODUCTION

Genome-wide association studies (GWASs) have successfully identified numerous genetic variants that are significantly associated with many common diseases (Manolio et al., 2009). However, these known variants explain only a small portion of disease heritability. Many genetic variants, especially those with small effect sizes, remain to be identified (Manolio et al., 2009). In order to increase power of genetic association test, one approach is to conduct gene-based tests, where genotyped single nucleotide polymorphisms (SNPs) within a gene are used simultaneously.

Many statistical methods for GWASs have been developed in the literature (Chen, 2011b, 2013, 2014, 2017; Chen & Ng, 2012; Chen & Wang, 2017; Chen et al., 2014; Chen, Huang, & Ng, 2014; Chen, Huang, & Ng, 2016; Chen, Lin, & Wang, 2017; Chen, Ng, Li, Liu, & Huang, 2017; Wang, 2012; Zang & Fung, 2011; Zheng & Ng, 2008). Other methods designed specifically for rare variant association test were also proposed. For instance, the burden test (Li & Leal, 2008) combines multiple variables (e.g., loci) as a single one on which an association test will be performed. The burden test performs well if all SNPs have the same effect direction and similar effect sizes. However, if causal SNPs within a gene have different effect directions, the burden test may have little or low power. If the effect sizes are not similar, one may make them similar by assigning appropriate weights to SNPs. When causal SNPs have different directions, methods more robust than the burden test should be considered. Many such methods have been developed, for instance, C-alpha test (Neale et al., 2011), sequencing kernel association test (SKAT) (Wu et al., 2011), optimal SKAT (SKAT-O) (Lee, Wu, & Lin, 2012), and variations of SKAT (Chen, Han, & Wang, 2017; Sun, Zheng, & Hsu, 2013; Wang, 2016; Wu, Pankow, & Guan, 2015). In addition, some adaptive tests based on multivariate analysis were also proposed in the literature (Han & Pan, 2010; Pan, Kim, Zhang, Shen, & Wei, 2014).

SKAT is one of the most popular tests in this area. SKAT assigns weights to SNPs based on their minor allele frequencies (MAFs). Typically, the weight w is set as a function of MAF. The weighing function in SKAT is $w = dbeta(MAF, a, b)$, where $dbeta(\cdot, a, b)$ is the probability

density function of a beta distribution with two shape parameters a and b. The authors of SKAT suggested to use 1 and 25 as the default values for a and b, respectively, and applied it to the Dallas heart study (DHS) data (Romeo et al., 2009). Based on our own experience with the DHS data, using other values of a and b than their default values results in lowered power. That is, these default values for a and b seem to be chosen for optimizing the power performance of SKAT on the DHS data.

Other weighting functions have also been introduced in the literature. For instance, for resequencing data, Madsen and Browning suggested to use the weight as the inverse of $\hat{w} = \sqrt{nq(1-q)}$, where $q = \frac{m^U+1}{2n^U+2}$, n is the sample size, m^U is the number of mutant alleles observed in the unaffected individuals, n^U is the number of unaffected individuals for each variant (Madsen & Browning, 2009). In other words, the above weight is the estimated standard deviation of the total number of mutations in the sample under the null hypothesis of no frequency differences between affected and unaffected individuals.

In this study, we show that SKAT is sensitive to the weight assignment. We propose an approach to improve its detecting power through searching for the optimal weights. The organization of the paper is as follows. In Section 2, we present the new approach. In Section 3, a comprehensive simulation study is conducted to demonstrate the improvements of the proposed test. Real data applications are performed in Section 4. In Section 5 we give some discussions and conclude this study.

2 | METHOD

2.1 | Notations and SKAT

In an association study, the phenotype can be binary (e.g., case-control study) or quantitative. In general, a generalized linear model can be used to adjust for some covariates. In this paper, without loss of generality, we use $y = (y_1, y_2, \cdots, y_n)^T$ to denote the phenotypes after adjusting for some covariate and being standardized (e.g., each component divided by its estimated standard deviation), where the superscript T denotes the transpose operation, and n the number of subjects included in the study. We use m to denote the number of SNPs in a set (e.g., gene) from which an association test will be performed. We use an $n \times m$ matrix G, $G = (g_{ij})$, to denote the genotype data with its (i,j)th component, g_{ij}, equals the number of minor allele of SNP j from subject i. Therefore, $g_{ij} = 0, 1$ or $2 (i = 1, 2, \ldots, n; j = 1, 2, \ldots, m)$.

The SKAT test statistic can be written as:

$$SKAT = y^T GWWG^T y, \tag{1}$$

where $W = diag(w_1, w_2, \cdots, w_m)$, $w_i > 0$ is the weight assigned to the ith variant.

Let $G_W = GW$, and the k^{th} eigenvalue and its associated eigenvector of $G_W^T G_W$ be $\lambda_{W,k}$ and $\mu_{W,k}$, respectively. Without loss of generality, suppose $\lambda_{W,1} \geq \lambda_{W,2} \geq \cdots \geq \lambda_{W,m} > 0$. Then, $G_W^T G_W \mu_{W,k} = \lambda_{W,k}\mu_{W,k}$, $\mu_{W,k}^T \mu_{W,k} = 1$, $k = 1, 2, \ldots, m$.

Let $\upsilon_{W,k} = G_W \mu_{W,k} / \sqrt{\lambda_{W,k}}$. It is easy to verify that $G_W G_W^T \upsilon_{W,k} = \lambda_{W,k}\upsilon_{W,k}$, $\upsilon_{W,k}^T \upsilon_{W,k} = 1$, $k = 1, 2, \cdots, m$. That is, $\{\upsilon_{W,k},\ k = 1, 2, \ldots, m\}$ are the eigenvectors of $G_W G_W^T$ with $\lambda_{W,k}{}'s$ the corresponding eigenvalues. Because for conformable matrices **AB** and **BA**, they have the same set of none-zero eigenvalues, $\{\lambda_{W,k}\}$ are also the set of non-zero eigenvalues of $G_W G_W^T$. Therefore, SKAT statistic can be rewritten as:

$$SKAT = y^T GWWG^T y = \sum_{k=1}^{m} \lambda_{W,k} \left(y^T \upsilon_{W,k}\right)^2. \tag{2}$$

It can be shown that under the null hypothesis, the above statistic in (2) asymptotically follows a linear combination of chi-square distributions (Chen, Han et al., 2017),

$$SKAT \sim \sum_{k=1}^{m} \lambda_{W,k} \chi_{k,1}^2, \tag{3}$$

where $\chi_{k,1}^2$ is independently and identically distributed chi-square distribution with degree of freedom (*df*) 1. Therefore, the P-value for SKAT can be estimated by the Davis' approach or other methods (Davies, 1980; Wu, Guan, & Pankow, 2016). From (3), we know that the performance of SKAT depends on the weight matrix W.

The burden test statistic can be written as:

$$B = y^T GW\mathbf{1}\mathbf{1}^T WG^T y, \tag{4}$$

where 1 is the $m \times 1$ vector with all elements equal to 1.

For any $0 \leq \rho \leq 1$,we can define a new statistic which is a linear combination of the burden and SKAT statistics:

$$S_\rho = \rho B + (1-\rho)\, SKAT. \tag{5}$$

It should be pointed out that under the null hypothesis, the two test statistics, S and B, in (5) are correlated (Lee et al., 2012). In general they are not asymptotically independent either.

The SKAT-O test statistic (Lee et al., 2012) is defined as:

$$S_o = \min_{0 \leq \rho \leq 1} p_\rho, \tag{6}$$

where p_ρ is the P-value from the test S_ρ in (5). Its practical version is:

$$S_o = \min \left\{ p_{\rho_1}, p_{\rho_2}, \ldots, p_{\rho_b} \right\}. \tag{7}$$

The distribution of S_o is usually unknown, numerical methods are used to estimate its P-value.

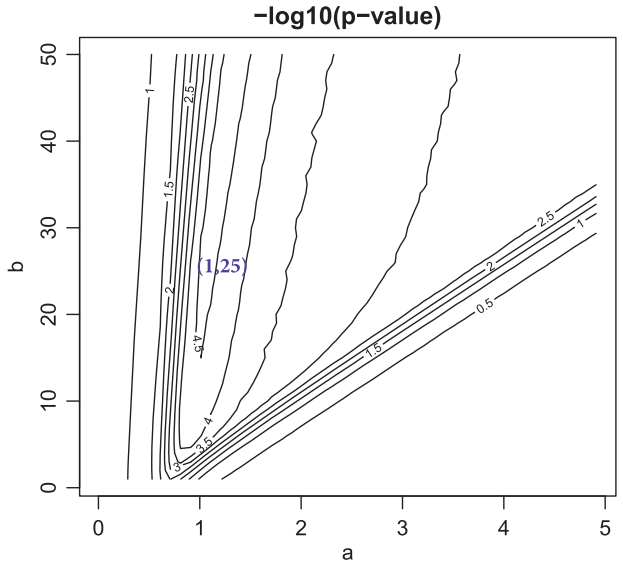

FIGURE 1 Contour plot of −log10(P values) from the Dallas heart study data by SKAT with different a and b values in the weighting function, $w_i = dbeta(MAF_i, a, b)$

Note: The default values for a and b are (1, 25) in SKAT.

To investigate the influence of the weights, we apply SKAT with different a and b values to the DHS data. Figure 1 is the contour plot of –log10(P-value) from this data set by SKAT with different a and b values in the weighting function. It clearly shows that for this specific data set, $a = 1$ and $b = 25$ are nearly the optimal values, in terms of obtaining the smallest P-value, for SKAT using $dbeta(MAF_i, a, b)$ as the weighting function. However, it should be pointed out that, first, the parameters in the weighting function should be determined before we see the data, otherwise, we need consider the multiple comparison issue as many values of a and b have been tested in finding the optimal values. Second, those default values in SKAT were obtained only based on one single data set; different data may have very different optimal weights. Therefore, the SKAT can be improved by searching for optimal weights.

More generally, for an association test, we want to test the null hypothesis, H_0, that all of the SNPs are independent of the phenotype versus the alternative hypothesis, H_1, that at least one of the SNPs is associated. Let $f : R^m \to R^k$ be a real-valued function such that $f(G_i) = H_i$, where G_i is the genotype data from the ith subject, that is, the ith row of matrix G, and $H_i = (h_{i1}, h_{i2}, \cdots, h_{ik})$, the aforementioned association test can be conducted based on the newly defined "genotype" H. Each of the current association tests can be viewed as a special case of this more general test with a specific function f. For instance, for SKAT, $f(G_i) = G_i W$; and for the burden test, $f(G_i) = 1^T G_i$, where 1 is a vector with all elements equal to 1. For an appropriately chosen function f, the association test based on H can potentially increase the detecting power (Fan et al., 2015).

2.2 | Proposed method

In this paper, we will focus on SKAT test with the weighting function, $w_i = dbeta(MAF_i, 1, b)$; we will search for the optimal value b to improve the detecting power. For a given set of possible values, B, for b, the test statistic of the proposed method is defined as:

$$S = \min_{b \in B} p_b, \tag{8}$$

where p_b is the P-value from SKAT with weighting function, $w_i = dbeta(MAF_i, 1, b)$.

There is no simple expression for the null distribution of the above test statistic S. Its P-value can be estimated based on permutation. Specifically, we permute the phenotype data T times, for each permutation, $1 \le t \le T$, we calculate the statistic s_t using (8). The P-value will be estimated as:

$$\hat{p} = \frac{\sum_{t=1}^{T} I_{s_t < s}}{T}, \tag{9}$$

where I is the indicator function.

One limitation of the permutation-based tests, such as the above one, is the computational burden, especially when there are hundreds of thousands of genes and very small significance level is used. We develop a fast algorithm for the permutation, which will dramatically reduce the computation burden. This algorithm will reduce the precision of the estimate when the true P-value is large. We first define the following parameters:

p_0= constant × significance level (e.g., 10×10^{-6});
T_{max} = maximal number of permutations (e.g., 10^6);
T_0 = minimal number of permutations (e.g., 10);
M = multiplying increment for the number of permutation (e.g., 10).

The fast algorithm works as follows:

Step 0. Calculate the test statistic s using (8) for a given set B;
Step 1. Set initial values: $p_0 = 10^{-5}$, $T_{max} = 10^6$, $T_0 = 10$, $M = 10, T = T_0$;
Step 2. Use (8) and (9) to estimate the P-value, $\hat{p}$. Set $T \leftarrow T \times M$; and
Step 3. If $\hat{p} > p_0$ or $T > T_{max}$, report $\hat{p}$ and stop; otherwise go to Step 2.

In the above algorithm, T_0 can be assigned a larger number, for example, 100, to increase the precision of the P values for those nonsignificant genes, if the estimated P values will be used in a follow-up analysis, for instance, gene ranking based on their P values.

3 | SIMULATION STUDY

3.1 | Simulation settings

To evaluate the performance of the proposed test, in this section, we conduct a simulation study to compare it with some existing methods. In the simulation study, we mainly focus on comparing the proposed test (new) with the SKAT, the SKAT-O, and the burden test. For the proposed test, we use $B = (1, 6, 11, 16, \ldots, 46)$; for other tests, we use the default weights as in SKAT, that is, $w_i = dbeta(MAF_i, 1, 25)$. We use the program, simRareSNP (https://www.biostat.umn.edu/~weip/), provided by Dr. Wei Pan to generate case-control rare-variant SNP data. For the genotype data, we use a latent multivariate Gaussian variable with a compound symmetry (CS) covariance structure. The correlation coefficient (ρ) in the CS takes different values, ranging from −0.8 to 0.8, in the simulation study. We simulate SNPs with MAFs ranging from 0.001 to 0.05.

To investigate how the new method controls type I error rate, we simulate 10 null SNPs, 1,000 cases, and 1,000 controls. Using significance level 0.05, we obtain the empirical type I error rate based on 1,000 replicates. To estimate the power value, we randomly select a proportion (θ) of 10 variants as causal SNPs, where θ takes values 0.2, 0.4, 0.6, 0.8, and 1.0. Following the simulation settings as described in the SKAT paper (Wu et al., 2011), we assume the effect size of each causal SNP is a function of its MAF. Specifically, we assume the magnitude of logarithmic relative risk (RR) of heterozygous to homozygous major genotypes is $d \times \log_{10}(MAF)$, with various values for d, −0.3, −0.2, and −0.15. The logarithmic RR is very close to the logarithmic odds ratio (OR), which was used with similar magnitudes for simulation study in the SKAT paper (Wu et al., 2011), if the disease prevalence is low. Of those causal SNPs, we randomly assign 20%, 50%, and 80% as protective variants. The commonly used log-additive genetic model is assumed in the simulation. The genotype frequencies of cases can be determined by those of controls and the relative risks of heterozygous and homozygous minor to homozygous major (Chen, 2014; Chen & Ng, 2012; Chen, Huang et al., 2014; Chen, Huang et al., 2016; Chen, Huang, & Ng, 2012a). Specifically, if the genotype frequencies of homozygous minor, heterozygous, and homozygous major are p_0, p_1, and p_2 (q_0, q_1, and q_2), respectively, for controls (cases), and the relative risks of heterozygous and homozygous minor to homozygous major are r_1 and r_2, then we have the following relationships:

$$\begin{cases} q_0 = \frac{p_0}{p_0 + r_1 p_1 + r_2 p_2} \\ q_1 = \frac{r_1 p_1}{p_0 + r_1 p_1 + r_2 p_2} \\ q_2 = \frac{r_2 p_2}{p_0 + r_1 p_1 + r_2 p_2} \end{cases}$$

TABLE 1 Empirical type I error rate for each method using significance levels $\alpha = 0.05$ and 1,000 replicates when there are 1,000 cases, 1,000 controls, and 10 SNPs

ρ	SKAT	SKAT-O	Burden	New
0	0.059	0.053	0.052	0.062
0.2	0.044	0.055	0.067	0.063
0.4	0.041	0.050	0.049	0.062
0.6	0.055	0.050	0.053	0.064
0.8	0.045	0.040	0.044	0.055
−0.2	0.046	0.036	0.035	0.050
−0.4	0.039	0.049	0.054	0.052
−0.6	0.058	0.055	0.051	0.054
−0.8	0.050	0.055	0.054	0.063

TABLE 2 Empirical power of each method using significance levels $\alpha = 0.05$ and 1,000 replicates when there are 1,000 cases, 1,000 controls, and 10 SNPs with 20% of those 10θ causal SNPs are protective

ρ	$(\theta, -d)$	SKAT	SKAT-O	Burden	New
0	(0.2,0.3)	0.46	0.44	0.20	0.73
	(0.4,0.2)	0.38	0.30	0.21	0.62
	(0.6,0.15)	0.33	0.43	0.36	0.48
	(0.8,0.15)	0.45	0.51	0.43	0.58
	(1.0,0.15)	0.57	0.70	0.58	0.71
0.2	(0.2,0.3)	0.40	0.34	0.09	0.67
	(0.4,0.2)	0.46	0.40	0.24	0.63
	(0.6,0.15)	0.27	0.34	0.32	0.39
	(0.8,0.15)	0.56	0.59	0.45	0.66
	(1.0,0.15)	0.56	0.66	0.53	0.78
−0.2	(0.2,0.3)	0.44	0.42	0.19	0.72
	(0.4,0.2)	0.35	0.38	0.29	0.60
	(0.6,0.15)	0.35	0.37	0.26	0.48
	(0.8,0.15)	0.50	0.57	0.41	0.71
	(1.0,0.15)	0.55	0.64	0.55	0.82

3.2 | Simulation results

Table 1 reports the empirical type I error rates for all methods included in the comparison. It shows that under various conditions, all method controlled type I error rate well. Tables 2–5 present the empirical power values (the highest power value is highlighted for each comparison) from each test when 1,000 cases and 1,000 controls were simulated, with the proportion of protective causal variants being 20%, 50%, 80%, and 100%, respectively. From the simulation results, we have the following observations. First, as expected, the burden test is less powerful than others in most situations; it has comparable detecting power as SKAT and SKAT-O only when θ is large (e.g., $\theta = 1$) and most of the effects have the same direction (e.g., in Tables 2 and 4). Second, when the burden test has reasonable power, SKAT-O is more powerful than SKAT; otherwise, SKAT performs better than SKAT-O. Third, for many of

TABLE 3 Empirical power of each method using significance levels $\alpha = 0.05$ and 1,000 replicates when there are 1,000 cases, 1,000 controls, and 10 SNPs with 50% of those 10θ causal SNPs are protective

ρ	$(\theta, -d)$	SKAT	SKAT-O	Burden	New
0	(0.2,0.3)	0.50	0.44	0.18	0.71
	(0.4,0.2)	0.50	0.40	0.16	0.59
	(0.6,0.15)	0.25	0.17	0.12	0.38
	(0.8,0.15)	0.41	0.35	0.13	0.51
	(1.0,0.15)	0.61	0.50	0.20	0.70
0.2	(0.2,0.3)	0.49	0.43	0.14	0.74
	(0.4,0.2)	0.34	0.29	0.08	0.45
	(0.6,0.15)	0.36	0.27	0.10	0.46
	(0.8,0.15)	0.48	0.40	0.12	0.59
	(1.0,0.15)	0.48	0.45	0.19	0.69
−0.2	(0.2,0.3)	0.46	0.42	0.22	0.77
	(0.4,0.2)	0.41	0.39	0.17	0.61
	(0.6,0.15)	0.30	0.23	0.13	0.47
	(0.8,0.15)	0.48	0.41	0.14	0.68
	(1.0,0.15)	0.50	0.51	0.25	0.81

TABLE 4 Empirical power of each method using significance levels $\alpha = 0.05$ and 1,000 replicates when there are 1,000 cases, 1,000 controls, and 10 SNPs with 80% of those 10θ causal SNPs are protective

ρ	$(\theta, -d)$	SKAT	SKAT-O	Burden	New
0	(0.2,0.3)	0.46	0.40	0.22	0.65
	(0.4,0.2)	0.34	0.36	0.29	0.50
	(0.6,0.15)	0.31	0.35	0.22	0.33
	(0.8,0.15)	0.38	0.50	0.46	0.56
	(1.0,0.15)	0.62	0.73	0.59	0.68
0.2	(0.2,0.3)	0.47	0.39	0.26	0.61
	(0.4,0.2)	0.39	0.37	0.20	0.55
	(0.6,0.15)	0.26	0.30	0.28	0.36
	(0.8,0.15)	0.36	0.39	0.30	0.49
	(1.0,0.15)	0.50	0.61	0.54	0.61
−0.2	(0.2,0.3)	0.40	0.39	0.22	0.63
	(0.4,0.2)	0.45	0.43	0.30	0.53
	(0.6,0.15)	0.24	0.39	0.32	0.36
	(0.8,0.15)	0.38	0.52	0.45	0.56
	(1.0,0.15)	0.61	0.76	0.66	0.76

the situations considered in the simulation study, the proposed test has the highest power values. Forth, the proposed test performs comparable to or better than SKAT-O under many situations. Finally, for all situations considered, the proposed test is more powerful than the original SKAT. This indicates that the new test, like SKAT-O, is more robust than SKAT.

TABLE 5 Empirical power of each method using significance levels $\alpha = 0.05$ and 1,000 replicates when there are 1,000 cases, 1,000 controls, and 10 SNPs with 100% of those 10θ causal SNPs are protective

ρ	$(\theta, -d)$	SKAT	SKAT-O	Burden	New
0	(0.2,0.3)	0.57	0.51	0.28	0.82
	(0.4,0.2)	0.40	0.48	0.48	0.63
	(0.6,0.15)	0.28	0.49	0.54	0.45
	(0.8,0.15)	0.37	0.80	0.83	0.58
	(1.0,0.15)	0.57	0.96	0.98	0.81
0.2	(0.2,0.3)	0.53	0.50	0.34	0.79
	(0.4,0.2)	0.44	0.53	0.49	0.66
	(0.6,0.15)	0.31	0.41	0.47	0.44
	(0.8,0.15)	0.38	0.71	0.84	0.60
	(1.0,0.15)	0.54	0.95	0.97	0.65
−0.2	(0.2,0.3)	0.42	0.42	0.31	0.78
	(0.4,0.2)	0.37	0.52	0.55	0.65
	(0.6,0.15)	0.25	0.52	0.61	0.51
	(0.8,0.15)	0.42	0.79	0.85	0.62
	(1.0,0.15)	0.50	0.94	0.96	0.77

3.3 | GAW17 data

The Genetic Analysis Workshop 17 (GAW17) data set (Almasy et al., 2011) uses the genotypes of a subset of genes whose sequencing data are available in the 1000 Genomes Project. It includes SNPs from gene ELAVL4 that influences the simulated quantitative phenotype Q1, and gene VNN1 which is associated with the simulated quantitative phenotype Q2. Except for the genetic risk factors, both Q1 and Q2 were also assumed to be associated with some covariates, such as age, gender, and smoking status. For each gene, a total of 200 simulated Q1 and Q2 values were included in the GAW17 data set. To account for the effects of those nongenetic factors, we use a linear regression model, then applied our proposed test using the residual as the phenotype, along with SKAT, SKAT-O, and the burden test, to the standardized residuals from the regression. For the proposed test, we use $B = (1, 6, 11, 16, 21, 26, 31)$ with 10^5 permutations to estimate its P-value.

Because the estimated P values from the new test are 0 for some cases, a small number of 10^{-5} is added to all P values before log transformation. Figures 2 and 3 plot the –log10(P values + 10^{-5}) with P values obtained by those methods from genes ELAVL4, and VNN1, respectively. Those plots clearly show that the proposed test produced smaller P values compared to SKAT, SKAT-O, and the burden test, for most of the cases. This indicates that the new test is more powerful than its competitors. For some situations, the improvements of the new method over others were substantial.

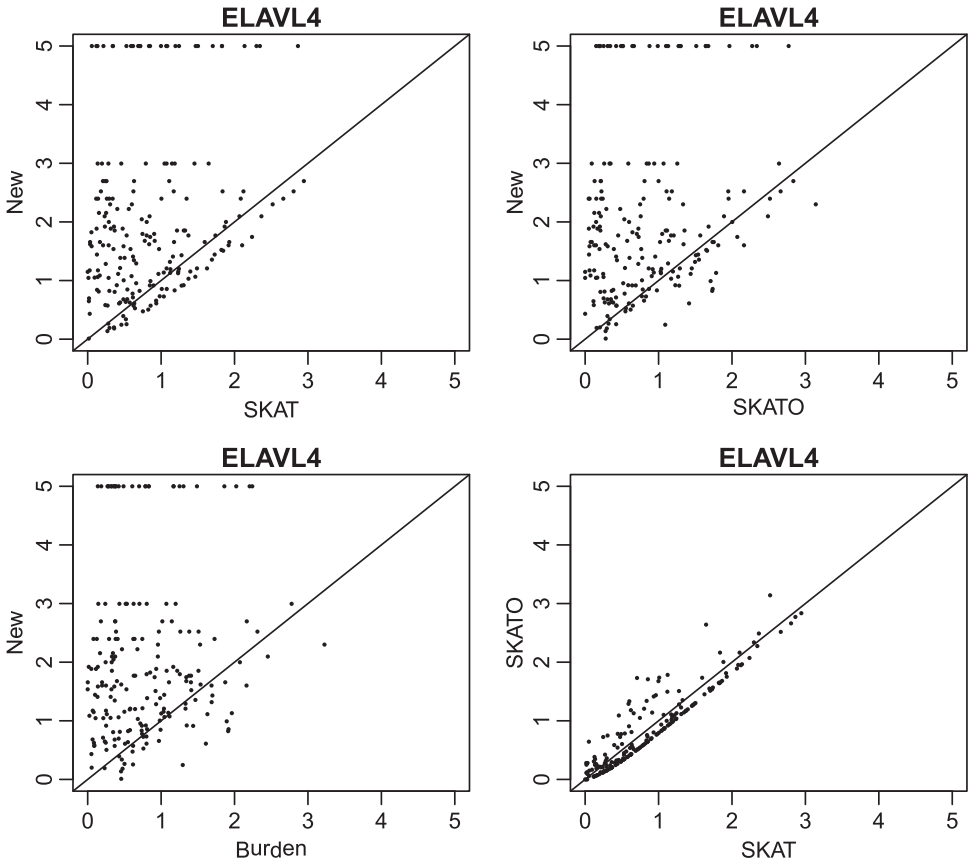

FIGURE 2 Pair-wise comparison of the –log10(P-value + 10^{-5}) obtained by the proposed test, SKAT, SKAT-O, and the burden test from gene ELAVL4 in the GAW 17 study

4 | REAL DATA APPLICATION

In this section, we apply the new method, along with some existing ones, to the ocular hypertension treatment study (OHTS) data (Gordon & Kass, 1999). OHTS is a National Eye Institute sponsored multicenter, randomized clinical trial. Its goal is to investigate the efficacy of medical treatment in delaying or preventing the onset of primary open angle glaucoma (POAG) in individuals with elevated intraocular pressure. Two hundred forty-nine non-Hispanic black individuals between 40 and 80 years old with both genotype and phenotype data available in this data set are used for this application. Data for this genetic study is available at Database of Genotypes and Phenotypes (dbGaP, Study Accession phs000240.v1.p1). There were 1,051,295 genotyped SNPs. There HGNC gene symbols were obtained using the R/Bioconductor package biomaRt (version 2.26.1). There are 30,562 autosomal genes.

In this application, we want to detect the association between each gene and the outcome central corneal thickness (CCT), which is used to assess POAG in this study. After adjusting for covariates age and gender using a linear regression, the standardized residues from the regression analysis are used for the association tests. For the proposed test, we use the fast algorithm described in Section 2.2 with $\boldsymbol{B} = (1, 6, 11, 16, 21, 26, 31)$, $p_0 = 3.0 \times 10^{-4}$, $T_{max} = 10^6$, $T_0 = 10$, and $M = 10$. Table 6 reports the P values obtained by SKAT, SKAT-O, the burden test, and the proposed method for genes whose smallest P values from the four tests are less than 5.0×10^{-5}. For all of the 11 listed genes, two have the smallest P values from SKAT, another two from SKAT-O, and the rest of them from the proposed test. Except for one gene, the P values obtained by the new method in Table 6 are all less than 10^{-4}, while SKAT, SKAT-O, and the burden test obtained very large P values for some genes. It should be pointed out that given the small sample size, the P values obtained by SKAT, SKAT-O, and the burden test may not be reliable as they are estimated based on their respective asymptotic distributions. On the other hand, the P value based on the permutation procedure from the proposed test is more accurate. However, to confirm the true association, those listed genes warrant further investigation.

We also use this real data set to compare the computational speed for those methods. All of the tests were implemented in R. Table 7 reports their running times when they were run in a computer with 3.60 GHz Intel Core processors and 8 GB of RAM memory. Both of the burden and SKAT required much less computations. The computational burden for SKAT-O

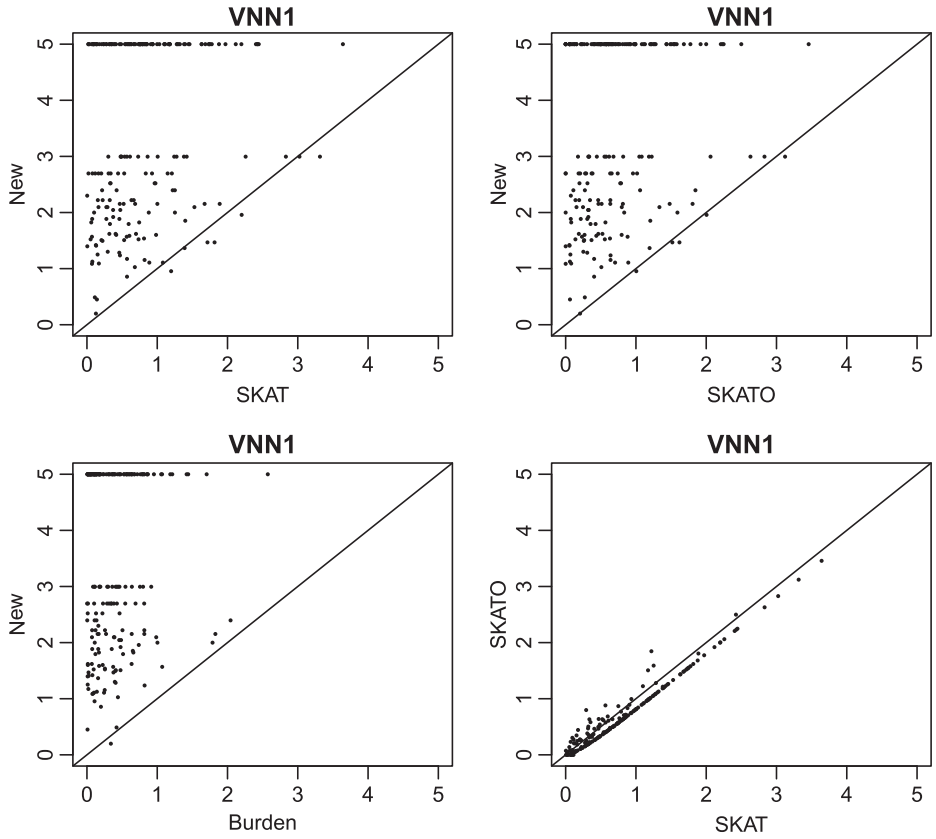

FIGURE 3 Pair-wise comparison of the –log10(P-value + 10^{-5}) obtained by the proposed test, SKAT, SKAT-O, and the burden test from gene VNN1 in the GAW 17 study

TABLE 6 Genes in the OHTS data with smallest P-value less than 5.0×10^{-5} from the four methods

Chr	Gene	SKAT	SKAT-O	Burden	New
3	CASR	3.4×10^{-4}	$\mathbf{3.8 \times 10^{-5}}$	8.7×10^{-5}	1.0×10^{-3}
3	MTCO1P29	2.7×10^{-5}	$\mathbf{2.6 \times 10^{-5}}$	2.7×10^{-5}	5.1×10^{-5}
11	DCHS1	$\mathbf{2.8 \times 10^{-5}}$	4.5×10^{-5}	9.5×10^{-2}	3.7×10^{-5}
16	SYT17	$\mathbf{4.9 \times 10^{-5}}$	5.5×10^{-5}	2.6×10^{-3}	7.7×10^{-5}
17	SENP3	6.1×10^{-2}	6.1×10^{-2}	4.1×10^{-2}	$\mathbf{4.6 \times 10^{-5}}$
17	SENP3-EIF4A1	1.2×10^{-1}	2.1×10^{-1}	5.4×10^{-1}	$\mathbf{2.0 \times 10^{-5}}$
17	MIR6779	4.0×10^{-4}	3.7×10^{-4}	6.4×10^{-4}	$\mathbf{2.4 \times 10^{-5}}$
17	KCNH4	2.7×10^{-1}	1.0×10^{-1}	6.9×10^{-2}	$\mathbf{3.0 \times 10^{-6}}$
17	HCRT	7.8×10^{-2}	6.6×10^{-2}	6.6×10^{-2}	$\mathbf{2.0 \times 10^{-6}}$
17	GHDC	7.8×10^{-2}	7.8×10^{-2}	7.8×10^{-2}	$\mathbf{2.0 \times 10^{-6}}$
17	STAT5B	1.6×10^{-1}	2.5×10^{-1}	6.3×10^{-1}	$\mathbf{2.0 \times 10^{-5}}$

TABLE 7 Computational time for each method based on the OHTS data

Method	SKAT	SKAT-O	Burden	New, 10^3 perm	New, 10^6 perm (est.)	New, fast algorithm
Time (hr)	0.47	3.62	0.45	115	481 days	78.2

was about 7–8 times of that for SKAT. For the proposed test, it can only be applied when the number of permutations for each gene was small. However, if the fast algorithm was used, the proposed method can be completed in about 3 days.

5 | DISCUSSION AND CONCLUSION

In this paper, we propose an improved version of a given association test by adaptively searching for optimal weights.

Using the popular test SKAT, we show how to improve its power by finding the optimal b value in the weighting function, $w_i = dbeta(MAF_i, 1, b)$ from a set of preset values B. This approach can be easily extended for other tests, such as SKAT-O and the burden test. Furthermore, this strategy can be extended to find an optimal function f described in Section 2.1. The optimal function can be searched from a set of meaningful candidate functions. However, it is challenging to preset those functions; more research is needed in this area.

Due to the fact that the distribution of the proposed test in (8) is intractable, a permutation-based approach should be used to estimate its P value. To reduce the computational burden, we develop a fast algorithm that gives lower estimation precisions to those genes with large P values. On the other hand, there are some advantages associated with the permutation-based approach. First, it is a distribution-free method; it can be applied to any situation without assuming a specific distribution of the test statistic. Second, its P value estimation can be more reliable than other methods based on the asymptotic properties of the test statistics; this is especially true when the sample sizes are small. Our simulation study and real data application reveal that the proposed approach can potentially increase the detecting power. The new approach provides alternative statistical tool in genetic association studies.

There are some limitations associated with the proposed test. First, only relatively simple weighting functions are considered. Second, compared to other methods, the computational burden for the proposed test can be high even if the fast algorithm is used. Third, in this report, we propose the optimally weighted SKAT test. From the simulation results, we find that under some situations (e.g., the assumptions for the burden test are valid) the proposed test may have lower power than the burden test and SKAT-O. In other words, it is less robust than SKAT-O. One possible remedy is to base on SKAT-O, instead of SKAT, to find the optimal weighting functions. However, this will increase the computational burden. In the future, we will consider more gene-based association tests and more weighting functions with robust P value combination methods (Chen, 2011a; Chen, 2013; Chen & Nadarajah, 2014; Chen et al., 2014; Chen, Huang, & Qiu, 2016; Chen, Liu, & Nadarajah, 2012). We will compare their performances through simulation studies and real data applications.

ACKNOWLEDGMENT

The authors would like to thank the editor, the associate editor, and two anonymous referees for their insightful comments that resulted in an improved presentation of the paper. TL was supported by the Natural Science Foundation of China (NSFC grant 61375051).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Zhongxue Chen http://orcid.org/0000-0003-2537-7843

REFERENCES

Almasy, L., Dyer, T. D., Peralta, J. M., Kent, J. W., Charlesworth, J. C., Curran, J. E., & Blangero, J. (2011). Genetic analysis workshop 17 mini-exome simulation. *BMC proceedings*, *5*(9), S2, 1–9.

Chen, Z. (2011a). Is the weighted z-test the best method for combining probabilities from independent tests? *Journal of Evolutionary Biology*, *24*(4), 926–930.

Chen, Z. (2011b). A new association test based on chi-square partition for case-control GWA studies. *Genetic Epidemiology*, *35*(7), 658–663.

Chen, Z. (2013). Association tests through combining *P*-values for case control genome-wide association studies. *Statistics and Probability Letters*, *83*(8), 1854–1862.

Chen, Z. (2014). A new association test based on disease allele selection for case-control genome-wide association studies. *BMC Genomics*, *15*, 358, 1–7.

Chen, Z. (2017). Testing for gene-gene interaction in case-control GWAS. *Statistics and Its interface*, *10*(2), 267–277.

Chen, Z., Han, S., & Wang, K. (2017). Genetic association test based on principal component analysis. *Applications in Genetics and Molecular Biology*, *16*(3), 189–198.

Chen, Z., Huang, H., & Ng, H. K. T. (2012). Design and analysis of multiple diseases genome-wide association studies without controls. *Gene*, *510*(1), 87–92.

Chen, Z., Huang, H., & Ng, H. K. T. (2014). An improved robust association test for GWAS with multiple diseases. *Statistics & Probability Letters*, *91*, 153–161.

Chen, Z., Huang, H., & Ng, H. K. T. (2016). Testing for association in case-control genome-wide association studies with shared controls. *Statistical Methods in Medical Research*, *25*(2), 954–967.

Chen, Z., Huang, H., & Qiu, P. (2016). Comparison of multiple hazard rate functions. *Biometrics*, *72*, 39–45.

Chen, Z., Lin, T., & Wang, K. (2017). A powerful variant-set association test based on chi-square distribution. *Genetics*, *207*(3), 903–910.

Chen, Z., Liu, Q., & Nadarajah, S. (2012). A new statistical approach to detecting differentially methylated loci for case control Illumina array methylation data. *Bioinformatics*, *28*(8), 1109–1113.

Chen, Z., & Nadarajah, S. (2014). On the optimally weighted z-test for combining probabilities from independent studies. *Computational Statistics & Data Analysis*, *70*, 387–394.

Chen, Z., & Ng, H. K. T. (2012). A robust method for testing association in genome-wide association studies. *Human Heredity*, *73*(1), 26–34.

Chen, Z., Ng, H. K. T., Li, J., Liu, Q., & Huang, H. (2017). Detecting associated single-nucleotide polymorphisms on the X chromosome in case control genome-wide association studies. *Statistical Methods in Medical Research*, *26*(2), 567–582.

Chen, Z., & Wang, K. (2017). A gene-based test of association through an orthogonal decomposition of genotype scores. *Human Genetics*, *136*(10), 1385–1394.

Chen, Z., Yang, W., Liu, Q., Yang, J. Y., Li, J., & Yang, M. Q. (2014). A new statistical approach to combining *P*-values using gamma distribution and its application to genome-wide association study. *BMC Bioinformatics*, *15*(Suppl 17), S3.

Davies, R. B. (1980). Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society Series C (Applied Statistics)*, *29*(3), 323–333.

Fan, R., Wang, Y., Boehnke, M., Chen, W., Li, Y., Ren, H., … Xiong, M. (2015). Gene level meta-analysis of quantitative traits by functional linear models. *Genetics*, *200*(4), 1089–1104.

Gordon, M. O., & Kass, M. A. (1999). The ocular hypertension treatment study: Design and baseline description of the participants. *Archives of Ophthalmology*, *117*(5), 573–583.

Han, F., & Pan, W. (2010). A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, *70*(1), 42–54.

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, *13*(4), 762–775.

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

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

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Chakravarti, A. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., … Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet*, *7*(3), e1001322.

Pan, W., Kim, J., Zhang, Y., Shen, X., & Wei, P. (2014). A powerful and adaptive association test for rare variants. *Genetics*, *197*(4), 1081–1095.

Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H., & Cohen, J. C. (2009). Rare loss-of-function mutations in ANGPTL family members contribute to plasma triglyceride levels in humans. *Journal of clinical investigation*, *119*(1), 70–79.

Sun, J., Zheng, Y., & Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*, *37*(4), 334–344.

Wang, K. (2012). Statistical tests of genetic association for case–control study designs. *Biostatistics*, *13*(4), 724–733.

Wang, K. (2016). Boosting the power of the sequence kernel association test by properly estimating its null distribution. *American Journal of Human Genetics*, *99*(1), 104–114.

Wu, B., Guan, W., & Pankow, J. S. (2016). On efficient and accurate calculation of significance *P*-values for sequence kernel association testing of variant set. *Annals of Human Genetics*, *80*(2), 123–135.

Wu, B., Pankow, J. S., & Guan, W. (2015). Sequence kernel association analysis of rare variant set based on the marginal regression model for binary traits. *Genetic Epidemiology*, *39*(6), 399–405.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American Journal of Human Genetics*, *89*(1), 82–93.

Zang, Y., & Fung, W. K. (2011). Robust Mantel-Haenszel test under genetic model uncertainty allowing for covariates in case-control association studies. *Genetic Epidemiology*, *35*(7), 695–705.

Zheng, G., & Ng, H. K. T. (2008). Genetic model selection in two-phase analysis for case-control association studies. *Biostatistics*, *9*(3), 391–399.

How to cite this article: Chen Z, Lu Y, Lin T, Liu Q, Wang K. Gene-based genetic association test with adaptive optimal weights. *Genet Epidemiol*. 2017;1–9. https://doi.org/10.1002/gepi.22098