Published in final edited form as:

Ann Hum Genet. 2016 March ; 80(2): 123–135. doi:10.1111/ahg.12144.

On efficient and accurate calculation of significance p-values for sequence kernel association testing of variant set

Baolin Wu¹, Weihua Guan¹, and James S. Pankow²

¹Division of Biostatistics, School of Public Health, University of Minnesota

²Division of Epidemiology and Community Health, School of Public Health, University of Minnesota

Summary

The objective of this paper is to discuss and develop alternative computational methods to accurately and efficiently calculate significance p-values for the commonly used sequence kernel association test (SKAT) and adaptive sum of SKAT and burden test (SKAT-O) for variant set association. We show that the existing software can lead to either conservative or inflated type I errors. We develop alternative and efficient computational algorithms that quickly compute the SKAT p-value and have well-controlled type I errors. In addition, we derive an alternative and simplified formula for calculating the significance p-value of SKAT-O, which sheds light on the development of efficient and accurate numerical algorithms. We implement the proposed methods in the publicly available R package that can be readily used or adapted to large-scale sequencing studies. Given that more and more large-scale exome and whole genome sequencing or re-sequencing studies are being conducted, the proposed methods are practically very important. We conduct extensive numerical studies to investigate the performance of the proposed methods. We further illustrate their usefulness with application to associations between rare exonic variants and fasting glucose levels in the Atherosclerosis Risk in Communities (ARIC) Study.

Keywords

GWAS; SKAT; SKAT-O; Sequencing data

Introduction

Recent efforts of large-scale genetic studies have focused on exploring the rare variant contribution to missing heritability (Manolio *et al.*, 2009; Bonnefond *et al.*, 2012; Zhan *et al.*, 2013; Steinthorsdottir *et al.*, 2014; Wang *et al.*, 2014; Estrada *et al.*, 2014). Due to their low minor allele frequencies (MAF), individual rare variant based tests typically lack power. Gene-level based rare variant set (RVS) tests have gained popularity by aggregating rare variant effects. There are typically two classes of RVS tests: the burden test and the variance component score test. The burden test collapses all rare variants or adaptively groups rare variants based on their MAFs into burden scores to be linked to the phenotype

(Morgenthaler and Thilly, 2007; Madsen and Browning, 2009; Morris and Zeggini, 2010; Li and Leal, 2008; Price *et al.*, 2010; Lin and Tang, 2011). The burden test works well for variants with similar effects, but could suffer substantial power loss with both protective and deleterious variants. The sequence kernel association test (SKAT) is based on the variance component score test and works well under various combinations of protective and deleterious variants (Wu *et al.*, 2010; Neale *et al.*, 2011; Wu *et al.*, 2011; Lee *et al.*, 2014). SKAT has been extended to various outcomes and study designs (see, for example, Barnett *et al.*, 2013; Schaid *et al.*, 2013; Ouakacha *et al.*, 2013; Chen *et al.*, 2013; Zhang *et al.*, 2014; Chen *et al.*, 2014). A more flexible approach is SKAT-O, which adaptively combines the burden and the SKAT statistics (Lee *et al.*, 2012b). The SKAT based approach performs well and is widely used in RVS association tests.

The objective of this paper is to discuss and develop alternative computational methods to accurately calculate significance p-values for the commonly used SKAT and SKAT-O tests of variant set association. We show that the existing software can lead to either conservative or inflated type I errors. We develop alternative and efficient computational algorithms that quickly compute the SKAT p-value and have accurate control of type I errors. In addition, we derive a simplified formula for calculating the significance p-value of SKAT-O, which sheds light on the development of efficient and accurate numerical algorithms. We implement the proposed methods in the publicly available R package that can be readily used or adapted to large-scale sequencing studies. Given that more and more large-scale exome and whole genome sequencing or re-sequencing studies are being conducted, the proposed methods are practically very important. We conduct extensive numerical studies to investigate the performance of the proposed methods. We further illustrate their usefulness with application to associations between rare exonic variants and fasting glucose levels in the Atherosclerosis Risk in Communities (ARIC) Study.

The main contributions of this paper are: (1) We note and numerically show the liberal type I errors for the widely used SKAT method, which has been routinely adopted in many large-scale sequencing studies for testing variant set association. We develop and implement alternative significance p-value calculation methods for SKAT that are very efficient and have accurate control of type I errors. (2) We derive alternative and much simplified formula to accurately and efficiently compute the significance p-value for the adaptive SKAT-O test, which also sheds insights on how the minimum p-value under-estimates the actual significance level when ‘adapting to the data’. We develop an efficient algorithm to compute the significance p-value for the adaptive SKAT-O test with well-controlled type I errors. (3) We develop a publicly available R package that implements the proposed methods and can be readily applied or adapted to test variant set association in large-scale sequencing studies.

Methods

When developing the Cox SKAT model for testing survival outcomes, Chen *et al.* (2014) discussed the use of individual variant test statistics to construct the SKAT statistic. In a related article, Wu *et al.* (2015) proposed a similar SKAT statistic based on the marginal variant test statistics for binary traits. In general the original SKAT statistics are directly constructed based on the vector of individual outcomes (or residuals). Both approaches lead

to equivalent test results. Adopting the variant test statistics approach of Chen *et al.* (2014) and Wu *et al.* (2015), we develop alternative and very efficient algorithms to accurately compute significance p-values of SKAT and SKAT-O tests.

Throughout the following discussion, we focus on the SKAT with (weighted) linear kernels and use a logistic regression model for binary outcome as an illustration example. The results can be readily generalized to other outcomes through the use of a generalized linear model. Consider a collection of n individuals. Denote Y as the n -vector of binary outcomes, X a $n \times p$ matrix of covariates (intercept included) to be adjusted, and G the $n \times m$ matrix of genotype scores for m variants measured in a gene region. Denote the n -vector of residuals $Z = Y - \hat{Y}$, where $\hat{Y}$ are the fitted probabilities under the null model: $\log[\Pr(Y = 1)/\Pr(Y = 0)] = X\alpha$. Denote the $n \times n$ null covariance matrix $V = \text{Cov}(Z)$, and a $m \times m$ diagonal weight matrix W , which is typically determined by the variant MAF.

SKAT for variant set association test

The SKAT statistic is defined as $Q = Z'GWWG'Z$. Under the null hypothesis Q asymptotically follows the mixture of one degree of freedom (1-DF) χ^2 distributions with the mixing coefficients being the positive eigen values of $K = V^{1/2}GWWG'V^{1/2}$ (Wu *et al.*, 2010, 2011). Though K is a $n \times n$ matrix, its rank $\leq \min(n, m)$. Therefore K has at most $\min(n, m)$ positive eigen values. It is not hard to verify that for any matrix A , $A'A$ and AA' have the same positive eigen values. Therefore K has the same positive eigen values as matrix $R = WG'V^{1/2}V^{1/2}GW = WG'VGW$, which is of dimension $m \times m$. Alternatively we can define the SKAT statistic based on the variant test statistics. Define a m -vector $U = WG'Z$. We can easily verify its null covariance $\text{Cov}(U) = R$. The SKAT statistic can be equivalently written as $Q = U'U$. Hence under the null hypothesis Q asymptotically follows the mixture of 1-DF χ^2 distributions with mixing coefficients being the positive eigen values of $\text{Cov}(U) = R$.

Therefore whether we define the SKAT statistic based on the n -vector of observation residuals or the m -vector of variant statistics, they will lead to identical results. Using the m -vector of variant statistics could offer some computational advantages since typically $m < n$. This approach has appeared previously in the literature (see, for example, Chen *et al.*, 2014; Jiang and McPeck, 2014; Wu *et al.*, 2015). We further improve the computational efficiency of calculating the asymptotic distribution of the SKAT statistic based on matrix decompositions (see appendix for technical details and the numerical study section for results).

SKAT significance p-value calculation

Various methods have been proposed to compute the tail probability of the mixture of 1-DF χ^2 distributions. For example, we can use the Davies' method (Davies, 1980), the moment-matching based non-central χ^2 approximation method (Liu *et al.*, 2009; Lee *et al.*, 2012b), or the saddlepoint approximation method (Kuonen, 1999). The Davies' method has been implemented in the 'CompQuadForm' R package (Lafaye De Micheaux, 2013). The saddlepoint approximation method has been implemented in the 'survey' R package (Lumley, 2014). It was used in Chen *et al.* (2014) to compute the Cox SKAT significance p-

value for survival outcomes. The ‘SKAT’ R package (version 1.0.7, Lee *et al.*, 2015) uses the Davies’ method (with default 10^{-6} accuracy level and 10^4 maximum number of integration terms), and then switches to the non-central χ^2 approximation method of Lee *et al.* (2012b) when the Davies’ method fails to converge.

In general, the Davies’ method is the most accurate, but calculating small tail probabilities requires a very large number of integration terms and is very computing intensive. The saddlepoint approximation method provides a very close approximation to the Davies’ method. The moment-matching based non-central χ^2 approximation methods are anti-conservative and could lead to inflated type I errors especially for small significance levels (see the numerical study section for results).

We use the Davies’ method with 10^{-9} accuracy level and 10^6 maximum number of integration terms, and then switch to the saddlepoint approximation method when the Davies’ method fails to converge. We will show that this hybrid approach (denoted as ‘SKATh’) yields more accurate results (see the numerical study section).

Adaptive sum of SKAT and burden test

The SKAT-O test computes an optimal average of SKAT and burden tests based on minimizing the p-values (Lee *et al.*, 2012b). SKAT-O is typically defined based on the vector of observation residuals, and can be equivalently computed based on the variant test statistics. Consider the weighted average of the SKAT and burden test, $Q_\rho = (1 - \rho)U'U + \rho U'JU$, where $0 \leq \rho \leq 1$ and J is a $m \times m$ matrix with all elements equal to 1. The SKAT-O test statistic is defined as the minimum of the p-values of Q_ρ , p_ρ , over a finite grid of ρ , $0 = \rho_1 < \dots < \rho_K = 1$: $T = \min_k p_{\rho_k}$. Denote q_ρ as the $(1 - T)$ -th percentile of the distribution of Q_ρ , which can be computed based on moment matching (Liu *et al.*, 2009).

We can analytically compute the significance p-value of SKAT-O using numerical integrations following the approach of Lee *et al.* (2012b) and Wu *et al.* (2015). In this paper, we derive much simplified formulas to eliminate the need of many complicated matrix operations and develop more accurate and efficient numerical algorithms to compute SKAT-O p-values (see Appendix for technical derivations). Define $\gamma_1 = \mathbf{1}'R\mathbf{1}$, $\gamma_2 = \mathbf{1}'R^2\mathbf{1}$, $\gamma_3 = \mathbf{1}'R^3\mathbf{1}$. Here R is the null covariance matrix of U , and $\mathbf{1}$ is m -vector of ones. Denote the eigen

values of $R - RJR/\gamma_1$ as $(\lambda_1, \dots, \lambda_m)$. Define $\mu = \sum_{j=1}^m \lambda_j$, and

$r = \sqrt{\sum_{j=1}^m \lambda_j^2} / \sqrt{\sum_{j=1}^m \lambda_j^2 + 2(\gamma_3/\gamma_1 - \gamma_2^2/\gamma_1^2)}$. Let $f(\cdot|\chi_1^2)$ and $F(\cdot|\chi_1^2)$ denote the 1-DF χ^2 density/distribution functions, and $S(\cdot)$ the survival function (i.e., tail probability) for the mixture of 1-DF χ^2 distributions with coefficients $(\lambda_1, \dots, \lambda_m)$. Define

$\tilde{q}_1 = F^{-1}(1 - T|\chi_1^2)$, and

$$\delta(x) = \min_{k: \rho_k < 1} \frac{q_{\rho_k} - [\rho_k \gamma_1 + (1 - \rho_k) \gamma_2 / \gamma_1] x}{1 - \rho_k} r + \mu (1 - r). \quad (1)$$

The significance of SKAT-O can be computed as

$$P = 1 - \int_0^{\tilde{q}_1} [1 - S(\delta(x))] f(x|\chi_1^2) dx, \quad (2)$$

which can be equivalently written as

$$P = T + \int_0^{\tilde{q}_1} S[\delta(x)] f(x|\chi_1^2) dx. \quad (3)$$

Remarks:

1. The first part is the minimum p-value T , and the second part accounts for the fact that the minimum p-value is obtained over multiple ρ s in SKAT-O.
2. In general the larger the K , the more likely we will find smaller $\delta(x)$ and smaller minimum p-value T (hence larger $\tilde{q}_1$), which generally will lead to larger values for the second part. So it is not necessarily always better to include more ρ s when searching for the minimum p-value in SKAT-O.
3. Compared to formula (2), formula (3) can produce more accurate numerical results especially for small p-values. In formula (2), the numerical integration is on computing $1 - P$, and formula (3) is on computing $P - T$. For small p-values, it is generally more difficult to achieve high accuracy for $1 - P$ compared to $P - T$ using numerical integration.
4. If $\rho_k < 1$, $\forall k$ (i.e., $\rho = 1$ not included), define

$$\bar{q}_1 = \min_k \frac{q_{\rho_k} + (1 - \rho_k)(1/r - 1)\mu}{\rho_k \gamma_1 + (1 - \rho_k)\gamma_2/\gamma_1}.$$

The SKAT-O p-value can be computed as

$$1 - F\left(\bar{q}_1|\chi_1^2\right) + \int_0^{\bar{q}_1} S[\delta(x)] f(x|\chi_1^2) dx. \quad (4)$$

Note that $q_1 = \gamma_1 \tilde{q}_1$. So equation (4) also works when $\rho = 1$ is included in the search for the minimum p-value.

Numerical studies

Numerical computation of SKAT and SKAT-O

We first evaluate computing the significance p-values of SKAT and SKAT-O based on the vector of observation residuals as implemented in the SKAT R package (denoted as ‘SKAT’ and ‘SKATO’) and vector of variant test statistics (denoted as ‘SKATv’ and ‘SKATov’). In order to fairly compare the computing time of SKATv and SKAT, we use the same algorithms as the SKAT R package to compute the tail probabilities of mixture of 1-DF χ^2 distributions. In addition, we also included the proposed SKATh in the comparison.

We generated 10,000 European-like haplotypes of length 1000 kb using a calibrated coalescent model (Schaffner *et al.*, 2005). We randomly pair the haplotypes to simulate a population of 10^6 individuals. We analyze a gene region with m rare variants (MAF < 0.01 and at least two copies of minor alleles). We consider two covariates $Z = (Z_1, Z_2)'$: $Z_1 \in \{0, 1\}$ follows Bernoulli(0.5), and $Z_2 \sim N(0, 1)$. We model the disease risk as

$$1 / \left[1 + \exp \left(- \left(\beta_0 + Z' \beta_Z + \sum_{j=1}^m \beta_j G_j \right) \right) \right].$$

We set $\beta_0 = -3.4$, $\beta_Z = (0.5, 0.5)'$

(corresponding to 5% population disease rate), and all $\beta_j = 0$. We randomly select n individuals with equal number of cases and controls. We assign weight $\text{Beta}(p_j; 1, 25)$ to the j th variant G_j . Here p_j is the MAF of G_j . For SKAT-O, we search over $\rho = \{0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 0.999\}$ for the minimum p-values. Here we have thresholded the ρ at 0.999 in order to reproduce the SKAT-O p-values as implemented in the SKAT R package. We evaluate the average CPU time (sec) used to compute the SKAT significance p-values for 10,000 variant sets with $n = 4000, 10,000, 20,000$ and $m = 20, 50, 100$. We repeated ten times and reported the average computing time. The computation of SKATO takes significantly more time and here we reported the average CPU time used to compute the SKAT-O significance p-values for 1000 variant sets. Numerically we obtained identical p-values between SKAT and SKATv, and between SKATO and SKATov in all simulations. Figure 1 shows the average CPU sec used for SKAT, SKATv and SKATh on a single Linux workstation with 2.5 GHz CPU and 24 GB memory. Overall we can see that all methods roughly scale linearly with the variant set size. The proposed SKATh takes slightly more (less than 10%) time than SKATv. And both SKATv and SKATh are very efficient: they only take around 22% to 25% of the computation time of SKAT on the same data. Figure 2 shows the average CPU sec used for SKATO and SKATov on the same Linux workstation. We can see that both methods roughly scale linearly with the variant set size. SKATO does not scale well with increasing sample size. Compared to SKATO, the SKATov was only slightly affected by the sample size with nearly constant computation time for various sample size. Its computation time was mainly determined by the variant size, and it offered more efficiency gain compared to SKATO for larger sample size. Overall the SKATov is very efficient: it only took around 10% of the computation time of SKATO under large sample size $n = 20,000$.

As we discussed in the Methods section, the numerical computation of the SKAT-O significance p-value using integration (2) is more accurate than using integration (3). Figure 3 shows the computed SKAT-O significance p-value P as a function of the minimum p-value T based on the two integrations. Here we fix R as the computed 9×9 covariance matrix for nine rare variants in the gene *G6PC2* from the ARIC sequencing data (see the Application section for details), and search over a finite grid of $\rho \in \{0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1\}$ for the minimum p-value. First note that for fixed R and grid of ρ s, P is easily verified to be a monotone and smooth function of T . Overall we can see that two integrations have similar results. The p-values computed using (2) have several 'outliers' (negative values) and show more variation. The integration with (3) yields monotone p-values with respect to T and leads to more accurate results.

Type I error of SKAT and SKAT-O

We evaluate the type I errors of various approaches of computing significance p-values for SKAT and SKAT-O. Note that computing the analytical significance p-value for a variant set association typically reduces to evaluating the tail probabilities of mixture of 1-DF χ^2 distributions. We consider a diagonal covariance matrix $R = \text{diag}\{\lambda_1, \dots, \lambda_m\}$ with $m = 20, 50, 100$ variants, which are generated as follows. We follow the previous simulation setup. We select a region with m rare variants and consider two Beta distribution based variant weights, $\text{Beta}(p_j; 1, 25)$ and $\text{Beta}(p_j; 0.5, 0.5)$. We compute the null covariance matrix of the weighted variant test statistics and its eigen values, which are then scaled to have unit mean value. We repeat 100 simulations to average the scaled eigen values as $\{\lambda_1, \dots, \lambda_m\}$. We search over a finite grid of $\rho \in \{0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1\}$ for the minimum p-values in SKAT-O.

We conducted 10^{12} Monte Carlo simulations based on independently generated 1-DF χ^2 random numbers to estimate the type I errors of various approaches. For SKAT, we evaluate the following approaches: the proposed SKATh (see the Methods section); the moment-matching based non-central χ^2 approximation method of Liu *et al.* (2009) (denoted as ‘Liu’); the modified Liu method implemented in the SKAT R package (option ‘Liu.mod’ in the SKAT function; denoted as ‘LM’); the saddlepoint approximation method (denoted as ‘SP’); the saddlepoint approximation based approach as implemented in the skatMeta R package (version 1.4.3, Voorman *et al.*, 2013; denoted as ‘SM’); ‘SKAT’ as implemented in the SKAT R package. For SKAT-O, we compare the proposed approach (denoted as ‘SKATOh’, using SKATh for SKAT p-value) to that implemented in the SKAT R package (denoted as ‘SKATO’), and the method as implemented in the skatMeta R package (denoted as ‘SMO’). Table 1 summarizes the results for SKAT significance p-value calculation.

Overall we can see that the nominal significance levels of the saddlepoint approximation method are close to the true α level, and the SKATh has the most accurate control of type I errors. Both the Liu method and modified Liu method (LM) underestimate the p-values and yield overly optimistic results. The SKAT approach occasionally has very liberal type I errors at very small α level. This is mainly due to the convergence problem of the Davies algorithm implemented in the CompQuadForm R package (Lafaye De Micheaux, 2013). For very small p-values, the SKAT Davies method (with 10^{-6} accuracy and 10^4 limit) could give falsely convergent results and produce aberrant results. Figure 4 shows a comparison of SKAT and LM p-values versus the SKATh p-values for the setting of $m = 20$ variants with $\text{Beta}(0.5, 0.5)$ weights in Table 1. We choose 100 equally spaced p-values for SKATh on the log scale from 10^{-2} to 10^{-9} . We then compute the ratios of corresponding SKAT and LM p-values over SKATh. The SKAT p-values are computed with the Davies method and use the LM method when the Davies method fails to converge (i.e, when SKAT and LM p-values are identical). When SKAT and LM give different results and SKAT deviates largely from SKATh (e.g., many p-values in the range of 10^{-6} to 10^{-8}), the SKAT Davies method gives falsely convergent results.

Table 2 summarizes the results for SKAT-O significance p-value calculations. Overall we can see that the SKATO and SKATOh perform similarly at relatively large α level. The SKATOh in general controlled the type I errors though was conservative at small α level.

The SKATO approach occasionally has significantly inflated type I errors at very small α . The SMO approach has very conservative type I errors at large α level, and very liberal type I errors at very small α level.

Application to ARIC sequencing study

We analyze the rare variants in the gene *G6PC2* using the genetic data from the Atherosclerosis Risk in Communities (ARIC) study samples (The ARIC Investigators, 1989). The ARIC study is a prospective investigation of atherosclerotic disease with a total of 15792 individuals recruited from four U.S. communities participating in the baseline examination in 1987-1989. *G6PC2* is a glucose-6-phosphatase gene. *G6PC2* rare variants have shown significant associations with fasting glucose in large scale sequencing studies (see, for example, Service *et al.*, 2014; Mahajan *et al.*, 2015; Wessel *et al.*, 2015). We study a set of nine *G6PC2* rare variants measured in the exome chip data from 5866 non-diabetic white ARIC participants (Wessel *et al.*, 2015). We analyzed the average fasting glucose levels measured at four visits approximately three years apart and applied an additive genetic model adjusting for age, gender, and study center. We also manually created and analyzed a binary trait by dichotomizing the average fasting glucose level at its 30 percentile. We assign the variant weights based on the Beta distribution function with parameters (1,25).

For the joint association test of nine *G6PC2* rare variants with the continuous and binary outcomes, the SKAT p-values are 4.20×10^{-5} and 4.60×10^{-5} respectively (computed using the SKAT R package, denoted as SKAT), and 4.22×10^{-5} and 4.61×10^{-5} respectively (computed using the skatMeta R package, denoted as SM). The SKATh p-values are 4.22×10^{-5} and 4.61×10^{-5} respectively. For the adaptive SKAT-O test, the reported SKATO p-values are 8.46×10^{-7} and 4.69×10^{-7} (computed using the SKAT R package, denoted as SKATO), and 1.14×10^{-4} and 1.10×10^{-4} (computed using the skatMeta R package, denoted as SMO), while the SKATOh p-values are 1.20×10^{-6} and 8.01×10^{-7} respectively. Overall we can see that SKAT, SM and SKATh yielded very similar results, while SKATO had much smaller p-values than SKATOh, and SMO produced much larger p-values. We also performed 10^{12} Monte Carlo simulations to compute the ‘exact’ p-values. Specifically we simulated multivariate normal random numbers from the computed null covariance matrix of the variant test statistics to compute the Monte Carlo p-values. For SM, SKAT and SKATh, the Monte Carlo ‘exact’ p-values are 4.22×10^{-5} and 4.61×10^{-5} for the continuous and dichotomized traits respectively. For the adaptive SKAT-O test, we note that the SMO, SKATO and SKATOh will lead to different ‘exact’ p-values since they could yield different minimum p-values for the same observed variant statistics. For SKATO, the Monte Carlo ‘exact’ p-values are 1.71×10^{-6} and 1.18×10^{-6} ; for SMO, the Monte Carlo ‘exact’ p-values are 1.39×10^{-6} and 9.25×10^{-7} ; and for SKATOh, the Monte Carlo ‘exact’ p-values are 1.35×10^{-6} and 8.87×10^{-7} respectively. Both the SKATh and SKATOh lead to more accurate significance p-values. The SKATO significantly under-estimated the actual p-values, and SMO significantly over-estimated the actual p-values.

In addition, we also performed 10^{12} numerical simulations as in the previous section using the computed covariance matrix of the variant test statistics. Table 3 summarizes the

estimated type I errors for SKAT. Overall we observed that the SKATh offered the most accurate control of type I errors. The SKAT has accurate control of type I errors at $\alpha = 10^{-4}$, 10^{-5} , slightly inflated type I errors at $\alpha = 10^{-6}$, and increased inflations at small α level. Both the Liu method (Liu) and modified Liu method (LM) have significantly inflated type I errors. The saddlepoint approximation based methods (SP and SM) have slightly conservative type I errors.

Table 4 summarizes the results for SKAT-O significance p-value calculations. The SKATOh in general controlled the type I errors well at all α levels. The SKATO had inflated type I errors at $\alpha = 10^{-6}$, 10^{-7} , 10^{-8} . The SMO approach had inflated type I errors at $\alpha = 10^{-8}$, and very conservative type I errors at $\alpha = 10^{-4}$, 10^{-5} , 10^{-6} , 10^{-7} . These are consistent with the ARIC sequencing data analysis results.

Discussion

As shown in our numerical studies and the original SKAT-O paper (Lee *et al.*, 2012b), accurate and analytical computation of significance p-values for SKAT-O is generally a difficult problem, and the SKAT-O p-values are less reliable compared to the SKAT p-values. The performance of current SKAT-O implementations depends on the size of variant set and the correlation structures among variant statistics (Tables 2 and 4). More research is needed to improve upon the current approaches for SKAT-O significance p-value calculation. In this paper, we have focused on the numerical computation of significance p-values for SKAT and SKAT-O conditional on the computed asymptotic covariance matrix of variant test statistics. It is well-known that for binary traits, the asymptotic distribution does not work well for finite samples, the SKAT p-values tend to be conservative (especially at low significance level), and small-sample adjustment is generally needed to improve the asymptotic approximation (Lee *et al.*, 2012a). The original SKAT (Wu *et al.*, 2010, 2011) has been proposed as a flexible method that can use linear kernels as well as other kernels (e.g., polynomial or IBS kernels). The SKAT-O (Lee *et al.*, 2012b) is primarily based on combining the linear kernel SKAT with the burden test. In this study we have focused on the SKAT with (weighted) linear kernel, which is the most widely used in practice especially for testing rare variant set association. For other kernel matrices, we will face the same problem of computing eigen values of a large kernel matrix since a typical kernel matrix has the same dimension of the sample size, which generally poses an obstacle for their applicability to large-scale sequencing studies. It would be useful to study efficient approaches to computing significance p-values for SKAT with general kernel matrix.

We have implemented the proposed methods in R programs posted at http://www.umn.edu/~baolin/research/skat_Rcode.html. In summary, we advocate using the proposed methods as a complementary approach to more efficiently and accurately analyzing the variant set associations in large-scale sequencing studies.

Acknowledgements

This research was supported in part by NIH grants GM083345 and CA134848. We are grateful to the University of Minnesota Supercomputing Institute for assistance with the computations. We want to thank the reviewers for their constructive comments which have greatly improved the presentation of the paper.

The ARIC Study is carried out as a collaborative study supported by the National Heart, Lung, and Blood Institute (NHLBI) contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C). The authors thank the staff and participants of the ARIC study for their important contributions. Funding support for “Building on GWAS for NHLBI-disease: the U.S. CHARGE consortium” was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419).

APPENDIX

Fast computation of asymptotic distribution of SKAT statistic with weighted linear kernel

We need to compute the covariance matrix $R = WGVGW$, where V is the asymptotic null covariance matrix of the residual vector. Under the logistic regression model, we can check that $V = V_0 - V_0X(XV_0X)^{-1}XV_0$, where $V_0 = \text{diag}(\hat{Y}(1 - \hat{Y}))$. We can directly compute V using matrix inverse and product, which is used, e.g., in the SKAT R package. This direct approach does not scale well to large sample size. We propose an alternative approach based on the matrix decomposition as follows. Denote the singular value decomposition (SVD)

$V_0^{1/2}X = UDV\tilde{V}$, where U is a $n \times p$ sub-orthogonal matrix, D a $p \times p$ diagonal matrix, and $\tilde{V}$ a $p \times p$ orthogonal matrix. Note that we just need to compute U in the SVD. We can check that $V = V_0 - V_0^{1/2}UU'V_0^{1/2}$. Let $\tilde{G} = U'V_0^{1/2}GW$. We then compute $R = WGV_0GW - G\tilde{G}\tilde{G}'$.

SKAT-O significance p-value calculation based on the variant test statistics

Consider the following weighted sum of SKAT and burden test statistics, $Q_\rho = U'LU$, where $L = (1 - \rho)I + \rho J$, and $\text{Cov}(U) = R$. Here I is an $m \times m$ identity matrix and J an $m \times m$ matrix will all elements equal to one. Define $\gamma_1 = \mathbf{1}'R\mathbf{1}$, $\gamma_2 = \mathbf{1}'R^2\mathbf{1}$, $\gamma_3 = \mathbf{1}'R^3\mathbf{1}$. Following Wu *et al.*

(2015), we can check that $L = A_0^2$ with $A_0 = \sqrt{1 - \rho}I + \left[\sqrt{1 - \rho + m\rho} - \sqrt{1 - \rho} \right] J/m$. Q_ρ follows a mixture of 1-DF χ^2 distributions with coefficients being the eigen values of A_0RA_0

$= (1 - \rho)R + c_1(JR + RJ) + c_2J$, where $c_1 = \sqrt{1 - \rho} \left[\sqrt{1 - \rho + m\rho} - \sqrt{1 - \rho} \right] / m$,

$c_2 = \left[\sqrt{1 - \rho + m\rho} - \sqrt{1 - \rho} \right]^2 \frac{\gamma_1}{m^2}$. With this approach, we can quickly compute the p-value p_ρ and corresponding quantile q_ρ for each given ρ .

Many complicated matrix operations are involved to compute SKAT-O p-values in the approach of Lee *et al.* (2012b) and Wu *et al.* (2015). In the following we derive much simplified formulas to eliminate the need of many complicated matrix computations and develop more accurate and efficient numerical algorithms to compute significance p-values of SKAT-O.

Consider the decomposition of covariance matrix $R = R_h^2$, where R_h is a $m \times m$ symmetric matrix. Significance of minimum p-value of Q_ρ can be studied equivalently based on $Q_\rho = Z'R_hLR_hZ$, where Z are vector of iid standard normal random variables.

Define a rank-1 projection matrix $H_1 = R_h \mathbf{1} \mathbf{1}' R_h / \gamma_1$. Note $H_1 R H_1 = \frac{R_h \mathbf{1} \mathbf{1}' R_h^2 \mathbf{1} \mathbf{1}' R_h}{\gamma_1^2} = \frac{\gamma_2}{\gamma_1} H_1$. We can check that $R_h (I - H_1) R_h = R - \frac{R \mathbf{1} (R \mathbf{1})'}{\gamma_1}$, and denote its eigen values as $\{\lambda_1, \dots, \lambda_m\}$. Note that actually we have $\lambda_m = 0$ since rank of $I - H_1$ is $m - 1$.

We can decompose $Q_\rho = (1-\rho)[\psi_1 + \psi_2] + \tau(\rho)\psi_3$, where $\psi_1 = Z(I - H_1)R(I - H_1)Z$, $\psi_2 = 2Z\mathcal{H}_1R(I - H_1)Z$, $\psi_3 = Z\mathcal{H}_1Z$, and $\tau(\rho) = (1 - \rho) \frac{\gamma_2}{\gamma_1} + \gamma_1\rho$. Note that H_1Z and $(I - H_1)Z$ are independent. We can check that $Cov(\psi_2, \psi_1) = Cov(\psi_2, \psi_3) = 0$, ψ_1 is independent of ψ_3 , and $E(\psi_2) = 0$, $Var(\psi_2) = 4tr(R_h(I - H_1)RH_1R_h)$. We can further show that

$Var(\psi_2) = 4\gamma_3/\gamma_1 - 4\gamma_2^2/\gamma_1^2$. Here ψ_1 follows a mixture of 1-DF χ^2 mixture distributions with coefficients being $(\lambda_1, \dots, \lambda_m)$. And ψ_3 follows 1-DF χ^2 distribution. Denote

$\mu = \sum_{i=1}^m \lambda_i$, $\sigma^2 = 2 \sum_{i=1}^m \lambda_i^2$, $\sigma_0^2 = \sigma^2 + 4 \left(\gamma_3/\gamma_1 - \gamma_2^2/\gamma_1^2 \right)$, and $r = \sigma/\sigma_0$. We treat ψ_2 as a negligible term, and approximate the distribution of $(\psi_1 + \psi_2 - \mu)r + \mu$ with the distribution of ψ_1 .

Given a finite grid of ρ , $0 = \rho_1 < \dots < \rho_K = 1$, the significance p-value based on the minimum p-value T is

$$1 - Pr \left(Q_{\rho_1} < q_{\rho_1}, \dots, Q_{\rho_K} < q_{\rho_K} \right) = 1 - E \left[Pr \left(\psi_1 + 2\psi_2 < \min_{k < K} \frac{q_{\rho_k} - \tau_{\rho_k} \psi_3}{1 - \rho_k} \mid \psi_3 \right) I(\gamma_1 \psi_3 < q_1) \right],$$

where ψ_3 follows the 1-DF χ^2 distribution, and $I(\cdot)$ is an indicator function. Note

$q_1 = \gamma_1 F^{-1} \left(1 - T \mid \chi_1^2 \right)$, where $F(\cdot \mid \chi_1^2)$ is the 1-DF χ^2 distribution function. We can easily verify that given the covariance matrix R and the grid of ρ values, the SKAT-O significance p-value is a monotone and smooth function of SKAT-O test statistic T (the minimum p-value). Let

$$\delta(x) = \left(\min_{k < K} \frac{q_{\rho_k} - \tau_{\rho_k} x}{1 - \rho_k} - \mu \right) r + \mu.$$

Let $\tilde{q}_1 = F^{-1} \left(1 - T \mid \chi_1^2 \right)$. We treat ψ_3 and $\psi_1 + \psi_2$ as approximately independent. The p-value can then be computed as

$$P = 1 - \int_0^{\tilde{q}_1} M(\delta(x)) f(x \mid \chi_1^2) dx,$$

where $f(\cdot \mid \chi_1^2)$ is the density of 1-DF χ^2 distribution, and $M(\cdot)$ is the distribution function of 1-DF χ^2 mixture with coefficients $(\lambda_1, \dots, \lambda_m)$. Denote the survival function $S(\cdot) = 1 - M(\cdot)$. We then have

$$P = 1 - \int_0^{\tilde{q}_1} f(x \mid \chi_1^2) dx + \int_0^{\tilde{q}_1} S(\delta(x)) f(x \mid \chi_1^2) dx = T + \int_0^{\tilde{q}_1} S(\delta(x)) f(x \mid \chi_1^2) dx.$$

When $\rho = 1$ is not included in searching for the minimum p-value T , we compute the p-value P as follows. Note that when $\delta(x) = 0$, $M(\delta(x)) = 0$ and $S(\delta(x)) = 1$. For a given ρ_k , denote

$x_k = \frac{q\rho_k + (1-\rho_k)(1/r-1)\mu}{\tau\rho_k}$. Let $q_1^- = \min_k x_k$. Then we have $\delta(x) = 0$, $\forall x \leq q_1^-$, and hence we compute the p-value as

$$P = 1 - \int_0^{q_1^-} M(\delta(x)) f(x|\chi_1^2) dx = 1 - F\left(q_1^-|\chi_1^2\right) + \int_0^{q_1^-} S(\delta(x)) f(x|\chi_1^2) dx.$$

To our knowledge, these two formulas are completely new. They lead to more accurate significance p-values in our numerical studies.

References

- Barnett II, Lee S, Lin X. Detecting rare variant effects using extreme pheno-type sampling in sequencing association studies. *Genet Epidemiol.* 2013; 37(2):142–151. [PubMed: 23184518]
- Bonnefond A, Clement N, Fawcett K, Yengo L, Vaillant E, Guillaume JL, Dechaume A, Payne F, Roussel R, Czernichow S, Hercberg S, Hadjadj S, Balkau B, Marre M, Lantieri O, Langenberg C, Bouatia-Naji N, Meta-Analysis of Glucose and Insulin-Related Traits Consortium (MAGIC). Charpentier G, Vaxillaire M, Rocheleau G, Wareham NJ, Sladek R, McCarthy MI, Dina C, Barroso I, Jockers R, Froguel P. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet.* 2012; 44(3):297–301. [PubMed: 22286214]
- Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol.* 2013; 37(2):196–204. [PubMed: 23280576]
- Chen H, Lumley T, Brody J, Heard-Costa NL, Fox CS, Cupples LA, Dupuis J. Sequence kernel association test for survival traits. *Genet Epidemiol.* 2014; 38(3):191–197. [PubMed: 24464521]
- Davies RB. Algorithm AS 155: the distribution of a linear combination of χ^2 random variables. *J R Stat Soc Ser C Appl Stat.* 1980; 29(3):323–333.
- Estrada K, Aukrust I, Bjorkhaug L, Burt NP, Mercader JM, Garcia-Ortiz H, Huerta-Chagoya A, Moreno-Macias H, Walford G, Flannick J, Williams AL, Gomez-Vazquez MJ, Fernandez-Lopez JC, Martinez-Hernandez A, Centeno-Cruz F, Mendoza-Caamal E, Revilla-Monsalve C, Islas-Andrade S, Cordova EJ, Soberon X, Gonzalez-Villalpando ME, Henderson E, Wilkens LR, Le Marchand L, Arellano-Campos O, Ordonez-Sanchez ML, Rodriguez-Torres M, Rodriguez-Guillen R, Riba L, Najmi LA, Jacobs SB, Fennell T, Gabriel S, Fontanillas P, Hanis CL, Lehman DM, Jenkinson CP, Abboud HE, Bell GI, Cortes ML, Boehnke M, Gonzalez-Villalpando C, Orozco L, Haiman CA, Tusie-Luna T, Aguilar-Salinas CA, Altshuler D, Njolstad PR, Florez JC, MacArthur DG. Association of a low-frequency variant in HNF1A with type 2 diabetes in a latino population. *JAMA.* 2014; 311(22):2305–2314. [PubMed: 24915262]
- Jiang D, McPeck MS. Robust Rare Variant Association Testing for Quantitative Traits in Samples With Related Individuals. *Genet Epidemiol.* 2014; 38(1):10–20. [PubMed: 24248908]
- Kuonen D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika.* 1999; 86(4):929–935.
- Lafaye De Micheaux P. CompQuadForm: Distribution function of quadratic forms in normal variables. R package version 1.4.1. 2013. <http://cran.r-project.org/web/packages/CompQuadForm/index.html>
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project-ESP Lung Project Team. Christiani DC, Wurfel MM, Lin X. Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *Am J Hum Genet.* 2012a; 91(2):224–237. [PubMed: 22863193]
- Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics.* 2012b; 13(4):762–775. [PubMed: 22699862]

- Lee S, Abecasis G, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet.* 2014; 95(1):5–23. [PubMed: 24995866]
- Lee, S.; Miropolsky, L.; Wu, MC. SKAT: SNP-set (Sequence) Kernel Association Test. R package version 1.0.7. 2015. <http://cran.r-project.org/web/packages/SKAT/index.html>
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83(3):311–321. [PubMed: 18691683]
- Lin DY, Tang ZZ. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet.* 2011; 89(3):354–367. [PubMed: 21885029]
- Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput Stat Data Anal.* 2009; 53(4): 853–856.
- Lumley, T. survey: analysis of complex survey samples.. R package version 3.30-3. 2014. <http://cran.r-project.org/web/packages/survey/index.html>
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5(2):e1000384. [PubMed: 19214210]
- Mahajan A, Sim X, Ng HJ, Manning A, Rivas MA, Highland HM, Locke AE, Grarup N, Im HK, Cingolani P, Flannick J, Fontanillas P, Fuchsberger C, Gaulton KJ, Teslovich TM, Rayner NW, Robertson NR, Beer NL, Rundle JK, Bork-Jensen J, Ladenvall C, Blancher C, Buck D, Buck G, Burt NP, Gabriel S, Gjesing AP, Groves CJ, Hollensted M, Huyghe JR, Jackson AU, Jun G, Justesen JM, Mangino M, Murphy J, Neville M, Onofrio R, Small KS, Stringham HM, Syvanen AC, Trakalo J, Abecasis G, Bell GI, Blangero J, Cox NJ, Duggirala R, Hanis CL, Seielstad M, Wilson JG, Christensen C, Brandslund I, Rauramaa R, Surdulescu GL, Doney ASF, Lannfelt L, Linneberg A, Isomaa B, Tuomi T, Jorgensen ME, Jorgensen T, Kuusisto J, Uusitupa M, Salomaa V, Spector TD, Morris AD, Palmer CNA, Collins FS, Mohlke KL, Bergman RN, Ingelsson E, Lind L, Tuomilehto J, Hansen T, Watanabe RM, Prokopenko I, Dupuis J, Karpe F, Groop L, Laakso M, Pedersen O, Florez JC, Morris AP, Altshuler D, Meigs JB, Boehnke M, McCarthy MI, Lindgren CM, Gloyn AL, On Behalf of the T2D-GENES consortium and GoT2D consortium. Identification and functional characterization of G6PC2 coding variants influencing glycemic traits define an effector transcript at the g6pc2-ABCB11 locus. *PLoS Genet.* 2015; 11:e1004876. [PubMed: 25625282]
- Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorf LA, Hunter DJ, Mc-Carthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TFC, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. *Nature.* 2009; 461(7265):747–753. [PubMed: 19812666]
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007; 615(1-2):28–56. [PubMed: 17101154]
- Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol.* 2010; 34(2):188–193. [PubMed: 19810025]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genet.* 2011; 7(3):e1001322. [PubMed: 21408211]
- Ouakacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, Richards JB, Ciampi A, Greenwood CMT. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet Epidemiol.* 2013; 37(4):366–376. [PubMed: 23529756]
- Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86(6): 832–838. [PubMed: 20471002]
- Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 2005; 15(11):1576–1583. [PubMed: 16251467]
- Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet Epidemiol.* 2013; 37(5):409–418. [PubMed: 23650101]

- Service SK, Teslovich TM, Fuchsberger C, Ramensky V, Yajnik P, Koboldt DC, Larson DE, Zhang Q, Lin L, Welch R, Ding L, McLellan MD, O'Laughlin M, Fronick C, Fulton LL, Magrini V, Swift A, Elliott P, Jarvelin MR, Kaakinen M, McCarthy MI, Peltonen L, Pouta A, Bonnycastle LL, Collins FS, Narisu N, Stringham HM, Tuomilehto J, Ripatti S, Fulton RS, Sabatti C, Wilson RK, Boehnke M, Freimer NB. Re-sequencing Expands Our Understanding of the Phenotypic Impact of Variants at GWAS Loci. *PLoS Genet.* 2014; 10:e1004147. [PubMed: 24497850]
- Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, Sigurdsson A, Helgadóttir HT, Johannsdóttir H, Magnusson OT, Gudjonsson SA, Justesen JM, Harder MN, Jørgensen ME, Christensen C, Brandslund I, Sandbak A, Lauritzen T, Vestergaard H, Linneberg A, Jørgensen T, Hansen T, Daneshpour MS, Fallah MS, Hreidarsson AB, Sigurdsson G, Azizi F, Benediktsson R, Masson G, Helgason A, Kong A, Gudbjartsson DF, Pedersen O, Thorsteinsdóttir U, Stefánsson K. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet.* 2014; 46(3):294–298. [PubMed: 24464100]
- The ARIC Investigators. The atherosclerosis risk in communities (aric) study: design and objectives. *Am J Epidemiol.* 1989; 129(4):687–702. [PubMed: 2646917]
- Voorman, A.; Brody, J.; Lumley, T. skatMeta: Efficient meta analysis for the SKAT test.. R package version 1.4.3. 2013. <http://cran.r-project.org/web/packages/skatMeta/index.html>
- Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, Zong X, Laplana M, Wei Y, Han Y, Lloyd A, Delahaye-Sourdeix M, Chubb D, Gaborieau V, Wheeler W, Chatterjee N, Thorleifsson G, Sulem P, Liu G, Kaaks R, Henrion M, Kinnarsley B, Valle M, LeCalvez-Kelm F, Stevens VL, Gapstur SM, Chen WV, Zaridze D, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Krokan HE, Gabrielsen ME, Skorpén F, Vatten L, Njølstad I, Chen C, Goodman G, Benhamou S, Vooder T, Vlk K, Nelis M, Metspalu A, Lener M, Lubiski J, Johansson M, Vineis P, Agudo A, Clavel-Chapelon F, Bueno-de Mesquita HB, Trichopoulos D, Khaw KT, Johansson M, Weiderpass E, Tjønneland A, Riboli E, Lathrop M, Scelo G, Albanes D, Caporaso NE, Ye Y, Gu J, Wu X, Spitz MR, Dienemann H, Rosenberger A, Su L, Matakidou A, Eisen T, Stefánsson K, Risch A, Chanock SJ, Christiani DC, Hung RJ, Brennan P, Landi MT, Houlston RS, Amos CI. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet.* 2014; 46(7):736–741. [PubMed: 24880342]
- Wessel J, Chu AY, Willems SM, Wang S, Yaghootkar H, Brody JA, Dauriz M, Hivert M-F, Raghavan S, Lipovich L, Hidalgo B, Fox K, Huffman JE, An P, Lu Y, Rasmussen-Torvik LJ, Grarup N, Ehm MG, Li L, Baldrige AS, Stancakova A, Abrol R, Besse C, Boland A, Bork-Jensen J, Fornage M, Freitag DF, Garcia ME, Guo X, Hara K, Isaacs A, Jakobsdóttir J, Lange LA, Layton JC, Li M, Hua Zhao J, Meidtnér K, Morrison AC, Nalls MA, Peters MJ, Sabater-Lleal M, Schurmann C, Silveira A, Smith AV, Southam L, Stoiber MH, Strawbridge RJ, Taylor KD, Varga TV, Allin KH, Amin N, Aponte JL, Aung T, Barbieri C, Bihlmeyer NA, Boehnke M, Bombieri C, Bowden DW, Burns SM, Chen Y, Chen Y-D, Cheng C-Y, Correa A, Czajkowski J, Dehghan A, Ehret GB, Eiriksdóttir G, Escher SA, Farmaki A-E, Franberg M, Gambaro G, Giulianini F, Goddard WA, Goel A, Gottesman O, Grove ML, Gustafsson S, Hai Y, Hallmans G, Heo J, Ho mann P, Ikram MK, Jensen RA, Jørgensen ME, Jørgensen T, Karaleftheri M, Khor CC, Kirkpatrick A, Kraja AT, Kuusisto J, Lange EM, Lee IT, Lee W-J, Leong A, Liao J, Liu C, Liu Y, Lindgren CM, Linneberg A, Malerba G, Mamakou V, Marouli E, Maruthur NM, Matchan A, McKean-Cowdin R, McLeod O, Metcalf GA, Mohlke KL, Muzny DM, Ntalla I, Palmer ND, Pasko D, Peter A, Rayner NW, Renstrom F, Rice K, Sala CF, Sennblad B, Serafetinidis I, Smith JA, Soranzo N, Speliotes EK, Stahl EA, Stirrups K, Tentolouris N, Thanopoulou A, Torres M, Traglia M, Tsafantakis E, Javad S, Yanek LR, Zengini E, Becker DM, Bis JC, Brown JB, Adrienne Cupples L, Hansen T, Ingelsson E, Karter AJ, Lorenzo C, Mathias RA, Norris JM, Peloso GM, Sheu WH-H, Toniolo D, Vaidya D, Varma R, Wagenknecht LE, Boeing H, Bottinger EP, Dedoussis G, Deloukas P, Ferrannini E, Franco OH, Franks PW, Gibbs RA, Gudnason V, Hamsten A, Harris TB, Hattersley AT, Hayward C, Hofman A, Jansson J-H, Langenberg C, Launer LJ, Levy D, Oostra BA, O'Donnell CJ, O'Rahilly S, Padmanabhan S, Pankow JS, Polasek O, Province MA, Rich SS, Ridker PM, Rudan I, Schulze MB, Smith BH, Uitterlinden AG, Walker M, Watkins H, Wong TY, Zeggini E, The EPIC-InterAct Consortium. Laakso M, Borecki IB, Chasman DI, Pedersen O, Psaty BM, Shyong Tai E, van Duijn CM, Wareham NJ, Waterworth DM, Boerwinkle E, Linda Kao WH, Florez JC, Loos RJF, Wilson JG, Frayling TM, Siscovick DS, Dupuis J, Rotter JJ, Meigs

- JB, Scott RA, Goodarzi MO. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun.* 2015; 6:5897. [PubMed: 25631608]
- Wu B, Pankow JS, Guan W. Sequence kernel association analysis of rare variant set based on the marginal regression model for binary traits. *Genet Epidemiol.* 2015; 39(6):399–405. [PubMed: 26282996]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. Powerful SNP-Set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010; 86(6):929–942. [PubMed: 20560208]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011; 89(1):82–93. [PubMed: 21737059]
- Zhan X, Larson DE, Wang C, Koboldt DC, Sergeev YV, Fulton RS, Fulton LL, Fronick CC, Branham KE, Bragg-Gresham J, Jun G, Hu Y, Kang HM, Liu D, Othman M, Brooks M, Ratnapriya R, Boleda A, Grassmann F, von Strachwitz C, Olson LM, Buitendijk GHS, Hofman A, van Duijn CM, Cipriani V, Moore AT, Shahid H, Jiang Y, Conley YP, Morgan DJ, Kim IK, Johnson MP, Cantsilieris S, Richardson AJ, Guymer RH, Luo H, Ouyang H, Licht C, Pluthero FG, Zhang MM, Zhang K, Baird PN, Blangero J, Klein ML, Farrer LA, DeAngelis MM, Weeks DE, Gorin MB, Yates JRW, Klaver CCW, Pericak-Vance MA, Haines JL, Weber BHF, Wilson RK, Heckenlively JR, Chew EY, Stambolian D, Mardis ER, Swaroop A, Abecasis GR. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nat Genet.* 2013; 45(11):1375–1379. [PubMed: 24036949]
- Zhang Q, Wang L, Koboldt D, Boreki IB, Province MA. Adjusting family relatedness in data-driven burden test of rare variants. *Genet Epidemiol.* 2014; 38(8):722–727. [PubMed: 25169066]

Figure 1.

Average CPU sec used to analyze 10,000 variant sets for SKAT, SKATv and SKATh: the x-axis shows the variant set size, and the y-axis shows the average CPU sec used.

Figure 2.

Average CPU sec used to analyze 1000 variant sets for SKATO and SKATov: the x-axis shows the variant set size, and the y-axis shows the average CPU sec used.

Figure 3.

SKAT-O significance p-value calculation based on two numerical integrations: the x-axis is the minimum p-value T (on log scale), the y-axis shows the corresponding computed SKAT-O significance p-value P based on the two equivalent numerical integrations (2) and (3) discussed in the Methods section.

Figure 4.

SKAT and LM p-values over SKATh for $m = 20$ variants with Beta(0.5,0.5) weights. The x-axis is the SKATh p-values on the log scale. When SKAT and LM differ from each other and SKAT deviates largely from SKATh, the SKAT Davies method gives falsely convergent results.

Table 1

Estimated type I errors (scaled by the significance level α) for SKAT test: Liu is the method proposed in Liu *et al.* (2009), LM is the modified Liu method implemented in SKAT R package, SP is the saddlepoint approximation method, SM is based on the saddlepoint method implemented in the skatMeta R package, SKAT is based on the Davies method implemented in the SKAT R package, and SKATh is the proposed method.

m	α	Beta(1,25)					Beta(0.5,0.5)						
		Liu	LM	SP	SM	SKAT	SKATh	Liu	LM	SP	SM	SKAT	SKATh
20	10^{-4}	1.69	1.27	0.96	0.96	1.00	1.00	1.09	1.04	1.07	1.07	1.00	1.00
	10^{-5}	2.45	1.65	0.94	0.94	0.99	1.00	1.20	1.11	1.09	1.09	1.00	1.00
	10^{-6}	3.69	2.21	0.93	0.93	0.92	1.00	1.36	1.23	1.10	1.10	1.04	1.00
	10^{-7}	5.75	3.06	0.92	0.92	0.14	1.00	1.60	1.42	1.09	1.09	1.29	1.00
	10^{-8}	9.10	4.30	0.91	0.91	0.66	1.01	1.95	1.70	1.06	1.06	2.96	1.02
50	10^{-4}	2.05	1.38	0.95	0.95	1.00	1.00	1.53	1.17	1.05	1.05	1.00	1.00
	10^{-5}	3.35	1.93	0.93	0.93	1.00	1.00	2.32	1.61	1.00	1.00	1.00	1.00
	10^{-6}	5.74	2.84	0.92	0.92	0.99	1.00	3.94	2.48	0.96	0.96	1.00	1.00
	10^{-7}	10.2	4.30	0.92	0.92	0.90	1.00	7.36	4.20	0.93	0.93	0.97	1.00
	10^{-8}	18.6	6.62	0.91	0.91	0.45	1.00	14.7	7.59	0.94	0.94	0.66	1.02
100	10^{-4}	2.12	1.38	0.97	0.97	1.00	1.00	1.75	1.17	1.04	1.04	1.00	1.00
	10^{-5}	3.62	2.00	0.94	0.94	1.00	1.00	2.91	1.69	0.99	0.99	1.00	1.00
	10^{-6}	6.57	3.09	0.93	0.93	1.00	1.00	5.46	2.74	0.95	0.95	0.96	1.00
	10^{-7}	12.5	4.96	0.92	0.92	0.99	1.00	11.3	4.89	0.93	0.93	0.66	1.00
	10^{-8}	24.6	8.19	0.92	0.92	1.02	1.00	24.9	9.30	0.95	0.95	0.51	1.02

Table 2

Estimated type I errors (scaled by the significance level α) for SKAT-O test: SKATO is the method implemented in the SKAT R package, SMO is from the skatMeta R package, and SKATOh is the proposed method.

m	α	Beta(1,25)			Beta(0.5,0.5)		
		SKATO	SMO	SKATOh	SKATO	SMO	SKATOh
20	10^{-4}	1.14	0.23	1.14	0.79	0.27	0.79
	10^{-5}	1.04	0.10	1.10	0.58	0.11	0.57
	10^{-6}	0.92	0.06	1.02	0.56	0.07	0.40
	10^{-7}	1.12	0.31	0.92	1.87	0.44	0.41
	10^{-8}	4.44	3.01	0.85	3.15	4.32	0.42
50	10^{-4}	1.08	0.38	1.07	0.75	0.42	0.75
	10^{-5}	0.98	0.20	0.97	0.48	0.19	0.49
	10^{-6}	0.82	0.09	0.82	0.34	0.06	0.41
	10^{-7}	0.66	0.09	0.67	0.34	0.03	0.37
	10^{-8}	1.44	0.50	0.53	1.14	0.15	0.40
100	10^{-4}	0.93	0.46	0.94	0.75	0.50	0.75
	10^{-5}	0.75	0.25	0.76	0.46	0.24	0.47
	10^{-6}	0.54	0.12	0.57	0.33	0.09	0.37
	10^{-7}	0.42	0.09	0.41	0.37	0.03	0.38
	10^{-8}	1.42	0.45	0.42	1.54	0.15	0.39

Table 3
Estimated type I errors (scaled by the significance level α) for SKAT test based on the *G6PC2* rare variant data.

α	Continuous trait						Binary trait					
	Liu	LM	SP	SM	SKAT	SKATh	Liu	LM	SP	SM	SKAT	SKATh
10^{-4}	1.48	1.22	0.95	0.95	1.00	1.00	1.50	1.23	0.95	0.95	1.00	1.00
10^{-5}	1.99	1.53	0.93	0.93	1.02	1.00	2.02	1.54	0.93	0.93	1.02	1.00
10^{-6}	2.77	1.96	0.92	0.92	1.16	1.00	2.84	1.99	0.92	0.92	1.16	1.00
10^{-7}	3.94	2.58	0.91	0.91	2.64	1.00	4.06	2.63	0.91	0.91	2.57	1.00
10^{-8}	5.66	3.43	0.90	0.90	4.26	0.99	5.88	3.50	0.90	0.90	4.54	0.99

Table 4

Estimated type I errors (scaled by the significance level α) for the SKAT-O test based on the *G6PC2* rare variant data.

α	Continuous trait			Binary trait		
	SKATO	SMO	SKATOh	SKATO	SMO	SKATOh
10^{-4}	1.17	0.39	1.18	1.17	0.43	1.18
10^{-5}	1.30	0.23	1.19	1.30	0.26	1.20
10^{-6}	1.80	0.13	1.18	1.78	0.16	1.18
10^{-7}	4.29	0.24	1.15	4.20	0.25	1.15
10^{-8}	7.71	1.87	1.11	7.08	1.86	1.13