RARE-VARIANT ASSOCIATION TESTING FOR SEQUENCING DATA WITH THE SKAT

Michael C. Wu, Seunggeun Lee, Tianxi Cai,
Yun Li, Michael Boehnke, Xihong Lin

MATERIALS & METHODS

Sequencing Kernel Association Test.

SKAT is a supervised test for the joint effects of multiple variants in a region on a phenotype. Regions can be defined by genes (in candidate-gene or whole-exome studies) or moving windows across the genome (in whole-genome studies). For each region, SKAT analytically calculates a p value for association while adjusting for covariates. Adjustments for multiple comparisons are necessary for analyzing multiple regions, for example with the Bonferroni correction or FDR control.

Notation.

Assume n subjects are sequenced in a region with p variant sites observed. Covariates might include age, gender, and top principal components of genetic variation for controlling population stratification.[22] For the i-th subject, y_i denotes the phenotype variable, $X_i = (X_{ij}, X_{i2}, ..., X_{im})$ denotes the covariates, and $G_i = (G_{ij}, G_{i2}, ..., G_{ip})$ denotes the genotypes for the p variants within the region. Typically, we assume an additive genetic model and let $G_{ij} = 0, 1, \ or\ 2$ represent the number of copies of the minor allele. Dominant and recessive models can also be considered.

SKAT Model and Test for Linear SNP Effects.

For a simple illustration of SKAT, we focus here on testing for a relationship between the variants and the phenotype by using classical multiple linear and logistic regression. We describe how the SKAT can incorporate epistatic effects later. To relate the sequence variants in a region to the phenotype, consider the linear model.

$$y_i = \alpha_0 + \alpha' X_i + \beta' G_i + \varepsilon i, \qquad \text{(Equation 1)}$$

when the phenotypes are continuous traits, and the logistic model.

$$logit\ P(y_i = 1) = \alpha_0 + \alpha' X_i + \beta' G_i, \qquad \text{(Equation 2)}$$

when the phenotypes are dichotomous (e.g., y = 0/1 for case or control). Here α_0 is an intercept term, $\alpha_0 = [\alpha_1,\ ...,\ \alpha_m]'$ is the vector of regression coefficients for the *m* covariates, $\beta = [\beta_1,\ ...,\ \beta_p]'$ is the vector of regression coefficients for the *p* observed gene variants in the region, and for continuous phenotypes ε_i is an error term with a mean of zero and a variance of σ^2. Under both linear and logistic models, and evaluating whether the gene variants influence the phenotype, adjusting for covariates, corresponds to testing the null hypothesis $H_0 : \beta = 0$, that is, $\beta_1 = \beta_2 =\ ...\ = \beta_p = 0$. The standard p-DF likelihood ratio test has little power, especially for rare variants. To increase the power, SKAT tests H_0 by assuming each β_j follows an arbitrary distribution with a mean of zero and a variance of $w_j\tau$, where τ is a variance component and w_j is a prespecified weight for variant *j*. One can easily see that $H_0 : \beta = 0$ is equivalent to testing $H_0 : \tau = 0$, which can be conveniently tested with a variance-component score test in the corresponding mixed model; this is known to be a locally most powerful test.[25] A key advantage of the score test is that it only requires fitting the null model $y_i = \alpha_0 + \alpha_1' X_i + \varepsilon_i$ for continuous traits and the $logit\ P(y_i = 1) = \alpha_0 + \alpha_1' X_i$ for dichotomous traits.

Specifically, the variance-component score statistic is.

$$Q = (y - \hat{\mu})' K (y - \hat{\mu}), \qquad \text{(Equation 3)}$$

where $K = GWG'$, $\hat{\mu}$ is the predicted mean of y under H_0, that is $\hat{\mu} = \hat{\alpha}_0 + X\hat{\alpha}$ for continuous traits and $\hat{\mu} = logit^{-1}(\hat{\alpha}_0 + X\hat{\alpha})$ for dichotomous traits; $\hat{\alpha}_0$ and $\hat{\alpha}$ and are estimated under the null model by regressing **y** on only the covariates **X**. Here **G** is an *n x p* matrix with the (i, j)-th element being the genotype of variant *j* of subject *i*, and **W** = diag$(w_j, ..., w_p)$ contains the weights of the *p* variants.

In fact, **K** is an *n x n* matrix with the (*i*, *i'*)-th element equal to $K(G_i, G_{i'}) = \Sigma_{j=1}^{p} w_j G_{ij} G_{i'j}.K(\bullet, \bullet)$ is called the kernel function, and $K(G_i, G_{i'})$ measures the genetic similarity between subjects *i* and *i'* in the region via the p markers. This particular form of $K(\bullet, \bullet)$ is called the weighted linear kernel function. We later discuss other choices of the kernel to model epistatic effects.

Good choices of weights can improve power. Each weight w_j is prespecified, with only the genotypes, covariates and external biological information, that is estimated without using the outcome, and reflects the relative contribution of the j-th variant to the score statistic: if w_j is close to zero, then the j-th variant makes only a small contribution to Q. Thus, decreasing the weight of noncausal variants and increasing the weight of causal variants can yield improved power. Because in practice we do not know which variants are causal, we propose to set $\sqrt{w_j} = Beta(MAF_j;\ a_1, a_2)$, the beta distribution density function with prespecified parameters a_1 and a_2 evaluated at the sample minor-allele frequency (MAF) (across cases and controls combined) for the j-th variant in the data. The beta density is flexible and can accommodate a broad range of scenarios. For example, if rarer variants are expected to be more likely to have larger effects, then setting $0 < a_1 \leq 1$ and $a_2 \geq 1$ allows for increasing the weight of rarer variants and decreasing the weight of common weights.We suggest setting $a_1 = 1$ and $a_2 = 25$ because it increases the weight of rare variants while still putting decent nonzero weights for variants with MAF 1%–5%. All simulations were conducted with this default choice unless stated otherwise. Note that a smaller a_1 results in more strongly increasing the weight of rarer variants. Examples of weights across a range of a_1 and a_2 values are presented in Figure S1, available online. Note that $a_1 = a_2 = 1$ corresponds to $w_j = 1$, that is all variants are weighted equally, and $a_1 = a_2 = 0.5$ corresponds to $\sqrt{w_j} = 1/\sqrt{MAF_j(1 - MAF_j)}$, that is w_j is the inverse of the variance of the genotype of marker j, which puts almost zero weight for MAFs ¿ 1% and can be used if one believes only variants with MAF ¡ 1% are likely to be causal. Note that SKAT calculated with this weight is identical to the unweighted SKAT test with the standardized genotypes in Equations 1 and 2. Other forms of the weight as a function of MAF can also be used. Because SKAT is a score test, the type I error is protected for any choice of prechosen weights. Note that the weights used in the weighted sum test13 involve phenotype information and will therefore alter the null distribution of SKAT if such weights are used.

Under the null hypothesis, Q follows a mixture of chi-square distributions, which can be closely approximated with the computationally efficient Davies method.[26] See Appendix A for details.

A special case of SKAT arises when the outcome is dichotomous, no covariates are included, and all $w_j = 1$. Under these conditions, we show in Appendix A that the SKAT test statistic Q is equivalent to the C-alpha test statistic T. Hence, the C-alpha test can be seen as a special case of SKAT, or alternatively, SKAT can be seen as a generalized C-alpha test that does not require permutation but calculates the p value analytically, allows for covariate adjustment, and accommodates either dichotomous or continuous phenotypes. Because SKAT under flat weights is also equivalent to the kernel machine regression test[23,24] and because the kernel machine regression test is in turn related to the SSU test,[27] it follows transitively that SKAT under flat weights, the kernel machine regression test, the SSU test, and the C-alpha test are all equivalent and special cases of SKAT. Note that the null distribution is calculated differently via these methods, and SKAT gives more accurate analytic p values, especially in the extreme tail, when sample sizes are sufficient.

Relationship between Linear SKAT and Individual Variant Test Statistics.

One can efficiently compute the test statistic Q by exploiting a close connection between the SKAT score test statistic Q and the individual variant test statistics. In particular, Q is a weighted sum of the individual score statistics for testing for individual variant effects. Hence, by letting $g_j = [G_{ij}, G_{1j}, ..., G_{nj}]'$ denote the *n x 1* vector containing the genotypes of the n subjects for variant j, it is straightforward to see that $Q = \Sigma_{j=1}^{p} w_j S_j^2$, where $S_j = g_j'(y - \hat{\mu_0})$ is the individual score statistic for testing the marginal effect of the j-th marker ($H_0 : \beta_j = 0$) under the individual linear or logistic regression model of y_i on X_i and only the j-th variant G_{ij}:

$$y_i = \alpha_0 + X_i'\alpha + \beta_j G_{ij} + \varepsilon_i$$

for continuous phenotypes and.

$$logit\ P(y_i) = \alpha_0 + X_i'\alpha + \beta_j G_{ij}$$

for dichotomous phenotypes. $\hat{\mu_0}$ is estimated as $\hat{\mu_0} = \hat{\alpha_0} + Xi'\hat{\alpha}$ for continuous traits and $\hat{\mu_0} = \ logit^{-1}(\hat{\alpha_0} + Xi'\hat{\alpha})$ for dichotomous traits. As a score test, one needs to fit the null model only a single time to be able to compute the S_j for all individual variants j as well as all regions to be tested. Similarly, if multiple regions are under consideration, then the same 0 can be used to compute the SKAT Q statistics for each region.

Accommodating Epistatic Effects and Prior Information under the SKAT. An attractive feature of SKAT is the ability to model the epistatic effects of sequence variants on the phenotype within the flexible kernel machine regression framework.[28–30] To do so, we replace $G_i'\beta$ by a more flexible function $f(G_i)$ in the linear and logistic models (1) and (2) where $f(G_i)$ allows for rare variant by rare variant and common variant by rare-variant interactions. Specifically, for continuous traits we use the semiparametric linear model[23,29].

$$y_i = \alpha_0 + \alpha' X_i + f(G_i) + \varepsilon_i, \qquad \text{(Equation 4)}$$

and for dichotomous traits, we use the semiparametric logistic model[24,30].

$$logit\ P(y_i = 1) = \alpha_0 + \alpha' X_i + f(G_i), \qquad \text{(Equation 5)}$$

Here the variants, G_i, are related to the phenotype through a possibly nonparametric function $f(\bullet)$, which is assumed to lie in a functional space generated by a positive semidefinite kernel function $K(\bullet,\bullet)$. Models (1) and (2) assume linear genetic effects and are specified by $K(G_i, G_{i'}) = \Sigma_{j=1}^{p} w_j G_{ij} G_{i'j.} K(\bullet,\bullet)$. By changing $K(\bullet,\bullet)$, one can allow for more complex models. Intuitively, $K(G_i, G_{i'})$ is a function that measures genetic similarity between the *i*-th and *i'*-th subjects via the p variants in the region, and any positive semidefinite function $K(G_i, G_{i'})$ can be used as a kernel function. We tailored several useful and commonly used kernels specifically for the purpose of rare-variant analysis: the weighted linear kernel, the weighted quadratic kernel, and the weighted identity by state (IBS) kernel.

The weighted linear kernel function $K(G_i, G_{i'}) = \Sigma_{j=1}^{p} w_j G_{ij} G_{i'j}$ implies that the trait depends on the variants in a linear fashion and is equivalent to the classical linear and logistic model presented in Equations 1 and 2. The weighted quadratic kernel $K(G_i, G_{i'}) = (1 + \Sigma_{j=1}^{p} w_j G_{ij} G_{i'j})^2$ implicitly assumes that the model depends on the main effects and quadratic terms for the gene variants and the first-order variant by variant interactions. The weighted IBS kernel $K(G_i, G_{i'}) = \Sigma_{j=1}^{p} w_j G_{ij} G_{i'j}$, defines similarity between individuals as the number of alleles that share IBS. For additively coded autosomal genotype data, $K(G_i, G_{i'}) = \Sigma_{j=1}^{p} w_j (2 - |G_{ij} G_{i'j}|)$. The model implied by the weighted IBS kernel models the SNP effects nonparametrically.[3]1 Consequently, this allows for epistatic effects because the function f($\bullet$) does not assume linearity or interactions of a particular order (e.g., the second order), Using the weighted IBS kernel removes the assumption of additivity because the number of alleles that are identical by state is a physical quantity that does not change on the basis of different genotype encodings.

We note that a kernel function that better captures both the similarity between individuals and the causal variant effects will increase power. In particular, if relationships are linear and no interactions are present, then the weighted linear kernel will have highest power. If interactions are present, the weighted quadratic and weighted IBS kernels can increase power. Our experience suggests using the IBS kernel when the number of interacting variants within the region is modest. As our understanding of genetic architecture improves so too will our knowledge of which kernel to use.

In each of the above kernels, w_j is an allele specific weight that controls the relative importance of the j^{th} variant and might be a function of factors such as allele frequency or anticipated functionality. Without prior information, we suggest the use of the $\sqrt{w_j} = Beta(MAF_j; 1,\ 25)$ suggested earlier. However, if prior information is available, for example some variants are predicted as functional or damaging via Polyphen[32] or Sift,[33] weights can be selected to increase the weight for likely functionality.

To test for the effects of gene variants in a region on a phenotype, one tests the null hypothesis $H_0 : f(G) = 0$. SKAT tests for this null hypothesis by assuming the *n x 1* vector $f = [f(G_1), ..., f(G_n)]'$ for the genetic effects of n subjects follows a distribution with mean zero and covariance τK, where τ is a variance component that indexes the effects of the variants.[29,30] Hence, we can test the null hypothesis that corresponds to testing $H_0 : \tau = 0$ by a variance-component score test. In particular, we simply replace **K** in Equation 3 by using the **K** discussed in this section, for example, the weighted IBS kernel, for epistatic effect. All subsequent calculations for computing a p value remain the same.

Because the SKAT evaluates significance via a score test, which operates under the null hypothesis, the SKAT is valid (in terms of protecting type I error) irrespective of the kernel and the weights used. Good choices of the kernel and the weights simply increase power.