Genetics: Early Online, published on October 12, 2017 as 10.1534/genetics.117.300395

RL-SKAT: An exact and efficient score test for heritability and set tests

Regev Schweiger[*,1], **Omer Weissbrod**[†], **Elior Rahmani**[*], **Martina Müller-Nurasyid**[‡,§,**], **Sonja Kunze**[††,‡‡], **Christian Gieger**[††,‡‡], **Melanie Waldenberger**[**,††,‡‡], **Saharon Rosset**[§§] **and Eran Halperin**[***,†††]

[*]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel, [†]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, [‡]Institute of Genetic Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, Neuherberg, Germany, [§]Department of Medicine I, Ludwig-Maximilians-Universität, Munich, Germany, [**]DZHK (German Centre for Cardiovascular Research), partner site Munich Heart Alliance, Munich, Germany, [††]Institute of Epidemiology II, Helmholtz Zentrum München - German Research Center for Environmental Health, 85764 Neuherberg, Germany, [‡‡]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München - German Research Center for Environmental Health, 85764 Neuherberg, Germany, [§§]School of Mathematical Sciences, Department of Statistics, Tel Aviv University, Tel Aviv, Israel, [***]Department of Computer Science, University of California, Los Angeles, CA, USA, [†††]Department of Anesthesiology and Perioperative Medicine, University of California, Los Angeles, CA, USA

ABSTRACT Testing for the existence of variance components in linear mixed models is a fundamental task in many applicative fields. In statistical genetics, the score test has recently become instrumental in the task of testing an association between a set of genetic markers and a phenotype. With few markers, this amounts to set-based variance component tests, which attempt to increase power in association studies by aggregating weak individual effects. When the entire genome is considered, it allows testing for the heritability of a phenotype, defined as the proportion of phenotypic variance explained by genetics. In the popular score-based Sequence Kernel Association Test (SKAT) method, the assumed distribution of the score test statistic is uncalibrated in small samples, with a correction being computationally expensive. This may cause severe inflation or deflation of p-values, even when the null hypothesis is true. Here, we characterize the conditions under which this discrepancy holds, and show it may occur also in large real datasets, such as a dataset from the Wellcome Trust Case Control Consortium 2 (n=13,950) study, and in particular when the individuals in the sample are unrelated. In these cases the SKAT approximation tends to be highly over-conservative and therefore underpowered. To address this limitation, we suggest an efficient method to calculate exact p-values for the score test in the case of a single variance component and a continuous response vector, which can speed up the analysis by orders of magnitude. Our results enable fast and accurate application of the score test in heritability and in set-based association tests. Our method is available in http://github.com/cozygene/RL-SKAT.

KEYWORDS Statistical genetics; SKAT; Heritability; Set-tests

The variance component model is a well established statistical framework used in many scientific fields. Testing for an association between several explanatory variables and a univariate response produces a variety of useful applications. For example, in metagenomics, an association is tested between a phenotype (e.g, body mass index, blood glucose levels, blood lipid levels, etc.) and the relative abundance counts of the measured species (Zhao *et al.* 2015).

doi: 10.1534/genetics.XXX.XXXXXX
Manuscript compiled: Friday 29th September, 2017
[1]Corresponding author: Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel. E-mail: schweiger@post.tau.ac.il

In statistical genetics, testing for an association between a set of genetic markers and a phenotype, such as a disease or a trait, is a fundamental task. Since studies to detect genetic signals are often underpowered, even with large datasets becoming available, the common approach to help alleviate this issue is grouping together genetic markers and testing them jointly. Grouping genetic markers is commonly implemented under the framework of variance component models. In addition to association testing, this framework can be used to answer several questions, such as estimation of the underying heritability of a phenotype (Kang *et al.* 2010); estimating the uncertainty of such estimation (Furlotte *et al.* 2014; Schweiger *et al.* 2016, 2017); phenotype prediction (Hayes *et al.* 2001), and more.

We consider two main scenarios in which such tests are performed: (i) a single phenotype, many sets of markers; (ii) many phenotypes, a single set of markers. Scenario (i) is common in set-testing, where relatively few markers are tested jointly. This is particularly useful in the case of rare variants, which are increasingly available for study using sequencing technologies, and which consistute a large part of human genetic variability. In such studies, a single phenotype is often tested against several sets of markers (for example, all rare variants in a single gene), because single-marker tests are often underpowered. Scenario (ii) occurs when studying heritability, defined as the proportion of phenotypic variance explained by genetics. Here, the tested markers are commonly the entire set of genotyped or sequenced single-nucleotide polymorphism (SNP) variants, or large portions of the genome (defined by, e.g., chromosome or functional annotation), and they are often tested against many (e.g., thousands) of phenotypes. Such phenotypes could be expression profiles of genes (Price *et al.* 2011; Wright *et al.* 2014; Lloyd-Jones *et al.* 2017), methlyation levels across of various methylation sites in the DNA (Quon *et al.* 2013; Van Dongen *et al.* 2016) or neuroimaging measurements (Ganjgahi *et al.* 2015; Ge *et al.* 2015).

Within the variance components framework, a common approach for association testing is the score test. It is used, for example, for testing the heritability of morphometric measurements derived from brain structural MRI scans (Ge *et al.* 2015) and on fractional anisotropy measures in subjects from the Genetics of Brain Structure study (Ganjgahi *et al.* 2015).

The main popular alternative to the score test is the generalized likelihood ratio (LR) test, e.g. as implemented by GCTA, a popular software package for heritability estimation (Yang *et al.* 2011). Both the score test and the LR test are based on properties of the likelihood function. The LR test statistic is calculated from the likelihood of the best fitting model across different heritability values, and from the likelihood of the model corresponding to no heritability. Conversely, the score test is based on the derivative of the likelihood function at the point corresponding to zero association, and testing if it is significantly nonzero. Compared with the LR test, the score test is often advantageous as it requires parameter estimation only for the null model, whereas the LR test requires parameter estimation for both the null and the alternative model. Additionally, the score test is the locally most powerful test; see (Lippert *et al.* 2014) for a thorough comparison between the two tests, mainly in the context of set testing.

The Sequence Kernel Association Test (SKAT) (Wu *et al.* 2011) has become the standard score-based test in statistical genetics and in metagenomics (Zhao *et al.* 2015), in large part due to its computational tractability. One of its merits is that it does not rely on the asymptotic distribution of the score test statistic, instead specifying a non-asymptotic distribution for the statistic under the null hypothesis of no association. However, it has been observed that this distribution may be inaccurate. In the SKAT-O extension (Lee *et al.* 2012), a resampling-based moment-matching correction is suggested. An adaptive permutation testing procedure is suggested in (Hasegawa *et al.* 2016). Chen et al. provide a method for calculating exact p-values (Chen *et al.* 2016); however, their method may be significantly slower than that of SKAT, as it requires the eigendecomposition of a full rank square matrix, whose computational complexity is typically cubic in the sample size, for each distinct response variable (e.g., phenotype) or each set of explanatory variables (e.g., SNP set). Finally, in these works, it is reported that this discrepancy occurs mainly in studies having a small sample size, and it is currently unclear to which extent the p-values of SKAT are calibrated for large sample sizes.

Here, we undertake a thorough analysis of the null distribution of the score test statistic, and its discrepancy under the SKAT approximation. We suggest a practical way to quantify this discrepancy, and show that such discrepancies may occur even at large sample sizes. We show that a discrepancy is expected when the number of markers is comparable to or larger than the number of individuals, and when the individuals are relatively unrelated. In particular, in addition to such inaccuracies occuring in tests of sets of rare-variants in small samples, we conclude that they may also occur in large scale heritability studies. We further suggest a computational method, Recalibrated Lightweight SKAT (RL-SKAT), that allows exact p-value computation while maintaining computation time as in SKAT; in particular, for multiple phenotypes tested against the same marker set, only a single eigendecomposition is required. Finally, we demonstrate and validate our results on two real datasets, a large dataset from the Wellcome Trust Case Control Consortium 2 (Consortium *et al.* 2011) (WTCCC2) study and the Cooperative health research in the Region of Augsburg (KORA) study (Holle *et al.* 2005) dataset.

Materials and Methods

We begin by reviewing the score test, as defined by the SKAT method (Wu *et al.* 2011) (see also the Supplemetary Information of (Lippert *et al.* 2014) for an excellent review). We focus here on continuous phenotypes, and on the case of a single variance component; for other cases, see the discussion below.

The variance components model

We consider the following standard variance components model (see (Searle *et al.* 2009) for a detailed review). Let n be the number of observations and $\mathbf{y}$ be a $n \times 1$ vector of responses. Let $\mathbf{X}$ be a $n \times p$ design matrix of p covariates, associated with fixed effects (possibly including an intercept vector $\mathbf{1}_n$ as a first column, as well as other covariates) and let $\boldsymbol{\beta}$ be a $p \times 1$ vector of fixed effects. Finally, let $\mathbf{K}$ be a kernel matrix, which, in a kernel-based method such as SKAT, can be taken to be any symmetric positive-definite matrix that encodes similarity between individuals. Then, $\mathbf{y}$ is assumed to follow:

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma_g^2\mathbf{K} + \sigma_e^2\mathbf{I}_n\right), \tag{1}$$

The fixed effects $\boldsymbol{\beta}$ and the coefficients σ_g^2 and σ_e^2 are the parameters of the model.

In the context of statistical genetics, $\mathbf{y}$ is a vector of phenotype measurements for each individual and $\mathbf{X}$ is a matrix of covariates (often including an intercept, sex, age, etc.). Let $\mathbf{Z}$ be a $n \times m$ standardized (i.e., columns have zero mean and unit variance) genotype matrix containing the m SNPs we test. The common choice for $\mathbf{K}$ is a weighted dot product of the genetic markers (Yang *et al.* 2010); formally, define $\mathbf{K} = \mathbf{ZWZ}^\top$, where $\mathbf{W}$ is a non-negative $m \times m$ diagonal matrix assigning a weight per SNP. A standard choice is the uniform $\mathbf{W}_{i,i} = 1/m$ (see (Wu *et al.* 2011) for a discussion). The narrow-sense heritability due to genotyped common SNPs is defined as the proportion of total variance explained by genetic factors (Visscher *et al.* 2008):

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}. \tag{2}$$

The score test

Under the above model, evaluating whether the tested covariates influence the response, while adjusting for additional covariates, corresponds to testing the null hypothesis $\sigma_g^2 = 0$. SKAT tests this hypothesis with a variance component score test in the corresponding mixed model. Specifically, the score statistic in the single-kernel case is obtained from the derivative of the restricted likelihood, discarding terms which are constant with respect to $\mathbf{y}$ (Lippert *et al.* 2014):

$$Q(\mathbf{y}) = \mathbf{y}^\top \mathbf{SKSy} \tag{3}$$

where $\mathbf{S} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ is the projection matrix to the subspace orthogonal to the covariates $\mathbf{X}$. For clarity of presentation, we will divide the statistic by σ_e^2. Then,

Proposition 1. *Let $\{\phi_i\}$ be the eigenvalues of $\mathbf{SKS}^\top$ and be $\chi^2_{1,i}$ are i.i.d. random variables distributed chi-square with one degree of freedom. Then,*

$$Q/\sigma_e^2 \sim \sum_{i=1}^{n} \phi_i \chi^2_{1,i}\,. \tag{4}$$

The proof of Proposition 1, as well as all proofs below, are deferred to the Supplemental Material.

The exact distribution of the score test statistic

The above derivation is exact whenever σ_e^2 is known. However, in practice, σ_e^2 is not known and needs to be estimated from the data; most often, from the single response vector we are testing. In practice, σ_e^2 is replaced with its restricted maximum likelihood (REML) estimate. The REML estimate is simply the corrected mean of the squared entries of the phenotype, after regressing out the covariates and using $\mathbf{S}^\top\mathbf{S} = \mathbf{S}$:

$$\hat{\sigma}_e^2(\mathbf{y}) = \frac{\|\mathbf{Sy}\|^2}{n-p} = \frac{\mathbf{y}^\top\mathbf{Sy}}{n-p}\,. \tag{5}$$

We note that sometimes the ML estimate $\mathbf{y}^\top\mathbf{Sy}/n$ is used, or just $\mathbf{y}^\top\mathbf{Sy}$; as this only introduces a multiplicative constant, we use the unbiased REML estimate for simplicity of presentation later. The statistic Q and $\hat{\sigma}_e^2$, are in fact dependent random variables. Therefore, the assumed distribution of $Q/\hat{\sigma}_e^2$ (described in Proposition (1)) does not hold when substituting σ_e^2 with its estimate, $\hat{\sigma}_e^2$. In (Zhang and Lin 2003; Liu *et al.* 2007, 2008; Wu *et al.* 2011), this subtitution is justified by the claim that the (restricted) ML estimator $\hat{\sigma}_e^2$ is consistent, and may therefore be substituted by its true value for a sample size n large enough. However, this argument does not take into consideration the dependency between Q and $\hat{\sigma}_e^2$. Also, as shown below, this distribution might not hold in realistic settings. In (Chen *et al.* 2016), this discrepancy is reported for small samples, and an exact distribution is derived for the statistic $Q/\hat{\sigma}_e^2$, and for any n, $\mathbf{K}$ and $\mathbf{X}$, which we review here:

Proposition 2. *The distribution of $Q/\hat{\sigma}_e^2$ may be modeled as a ratio of quadratic forms of normal variables. In particular, if $\mathbf{z} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{I}_n)$, then*

$$\frac{Q}{\hat{\sigma}_e^2} \overset{d}{=} (n-p)\cdot\frac{\mathbf{z}^\top\mathbf{SKSz}}{\mathbf{z}^\top\mathbf{Sz}} \tag{6}$$

Assessing the discrepancy

While noted in the literature (Zhao *et al.* 2015; Chen *et al.* 2016), the above discrepancy is reported for small samples only. However, as we show now, it may occur also when the number of individuals is large. We give a qualitative measure for when to expect large discrepancies between the asymptotic approximation of a weighted mixture of chi-squares and the exact distribution.

In the Supplemental Material, it is shown that the distributions of Q/σ_e^2 and $Q/\hat{\sigma}_e^2$ have the same means, but that $\mathrm{Var}(Q/\sigma_e^2) > \mathrm{Var}(Q/\hat{\sigma}_e^2)$, i.e. the latter having a smaller variance. We can further quantify the ratio between the variances as an indicator to the discrepancy between the distributions.

Proposition 3. *Denote the eigenvalues of $\mathbf{SKS}$ by $\phi_1, \ldots, \phi_n$ and note that there are at most $n-p$ non-zero eigenvalues ϕ_i. Denote the first two sample moments of the eigenvalues by $\bar{\phi} = \sum_{i=1}^n \phi_i/(n-p)$ and $\overline{\phi^2} = \sum_{i=1}^n \phi_i^2/(n-p)$. Denote the empirical variance of the eigenvalues by $\sigma^2(\phi) = \overline{\phi^2} - (\bar{\phi})^2$. Then,*

$$R := \frac{\mathrm{Var}(Q/\sigma_e^2)}{\mathrm{Var}(Q/\hat{\sigma}_e^2)} = \frac{n-p+2}{n-p}\cdot\left(\left(\frac{\sigma(\phi)}{\bar{\phi}}\right)^{-2} + 1\right) \tag{7}$$

The expression $\sigma(\phi)/\bar{\phi}$ is the (sample) coefficient of variation (CV) of the eigenvalues – a unitless, relative measure of their dispersion. Therefore, the ratio becomes larger when the CV is smaller. Also, as noted above, since the approximation wrongly ignores the dependency between the statistic Q and $\hat{\sigma}_e^2$, we expect the discrepancy to grow larger as the correlation between Q and $\hat{\sigma}_e^2$ increases. We therefore examine this correlation as an additional measure of this discrepancy.

Proposition 4. *Let $\sigma(\phi)/\bar{\phi}$ be the coefficient of variation (CV) of the eigenvalues as above. Then,*

$$\mathrm{Corr}(Q, \hat{\sigma}_e^2) = \left(\left(\frac{\sigma(\phi)}{\bar{\phi}}\right)^2 + 1\right)^{-1/2}\,. \tag{8}$$

This again demonstrates that CV affects discrepancy – the correlation becomes stronger when the CV is smaller. When $CV \ll 1$, for example when $\mathbf{K} \approx \mathbf{I}_n$, we have $R \gg 1$ and $\mathrm{Corr}(Q, \hat{\sigma}_e^2) \approx 1$. Conversely, when $CV \gg 1$, we have $R \approx 1$ and $\mathrm{Corr}(Q, \hat{\sigma}_e^2) \approx 1/CV$. This also gives the variance ratio as the function of the correlation as

$$R = \frac{n-p+2}{n-p}\cdot\frac{1}{1-\mathrm{Corr}^2(Q, \hat{\sigma}_e^2)}\,. \tag{9}$$

To summarize, the discrepancy is strong when the eigenvalues are more uniformly dispersed, and is weak when they have large variability. The dispersion of the eigenvalues of a kinship matrix has been previously shown to be related to the uncertainty in estimation of heritability: In (Visscher and Goddard 2015), it is shown that the asymptotic variance of the heritability REML estimator decreases with the variance of the entries of the kinship matrix, and with the variance of the eigenvalues. In (Schweiger *et al.* 2016), this result is shown without assumptions of asymptotics.

Examples. We now employ Propositions 3 and 4 to analyze several interesting examples in a genetic context. For simplicity, in the following, we use $\mathbf{X} = \mathbf{0}$, so that $p = 0$ and $\mathbf{S} = \mathbf{I}_n$.

- **Completely unrelated cohort.** Suppose the cohort contains completely unrelated individuals; then, $\mathbf{K} = \mathbf{I}_n$. Thus, $\phi_1 = \ldots = \phi_n = 1$, so $R = \infty$, $\mathrm{Corr}(Q, \hat{\sigma}_e^2) = 1$, and $Q/\hat{\sigma}_e^2$ is the constant n. Compare this to the case where σ_e^2 is known; then, it can be easily seen that $Q/\sigma_e^2 \sim \chi^2_n$. Therefore, the mean is the same but the variance vanishes completely.

- **Rank-one kinship matrix.** Consider the case of a simple burden test (Lee *et al.* 2012): If we assume the random effects $\mathbf{s}$ of all SNPs are identical, the burden test becomes equivalent to the score test with $\mathbf{K} = \mathbf{u}\mathbf{u}^\top$, where $\mathbf{u} = \mathbf{Z}\mathbf{1}_m$. Alternatively, consider the extreme case, where all the individuals are identical - $\mathbf{K} = \mathbf{1}\mathbf{1}^\top$ (while unlikely in human, this could be approximately true in studies of plants, yeast, etc.). In both these cases, there is a single nonzero eigenvalue: $\phi_2 = \ldots = \phi_n = 0$, which gives $R \approx 1$ and $\mathrm{Corr}(Q, \hat{\sigma}_e^2) = (\phi_1/n)/\sqrt{\phi_1^2/n} = 1/\sqrt{n}$; that is, with large enough sample size, we expect the correlation to be effectively zero, and the SKAT mixture approximation to hold well.
- **A full rank kinship matrix.** Assume the matrix $\mathbf{Z}$ contains $m > n$ SNPs in linkage equilibrium, where each column was mean-centered and normalized to have unit variance. Choosing the linear kernel $\mathbf{K} = \mathbf{Z}\mathbf{Z}^\top/m$, we follow (Patterson *et al.* 2006) in modeling $\mathbf{Z}$ as a matrix of random standard normal variables, from which it follows that $\mathbf{K}$ is a Wishart matrix. The limit distribution of the density of the eigenvalues of $\mathbf{K}$ is specified by the Marčhenko-Pastur distribution (Marcenko and Pastur 1967), with its first two moments known to be 1 and $1 + n/m$. Under this approximation, $\bar{\phi} \approx 1$, $\overline{\phi^2} \approx 1 + n/m$, $\sigma^2(\phi) \approx n/m$, $R \approx (n-p+2)/(n-p) \cdot (1+n/m)/(n/m)$ and $\mathrm{Corr}(Q, \hat{\sigma}_e^2) \approx 1/\sqrt{1+n/m}$. When $m \gg n$, as is often the case, $R \gg 1$ and $\mathrm{Corr}(Q, \hat{\sigma}_e^2) \approx 1$. This shows that for a large class of kinship matrices, we would expect the SKAT mixture approximation to hold poorly.
- **A SNP set.** Now, consider the case of set-testing, where $\mathbf{Z}$ is a normalized matrix of $m < n$ SNPs in linkage equilibrium. Following the modeling above, we have again $R \approx (n-p+2)/(n-p) \cdot (1+n/m)/(n/m)$ and $\mathrm{Corr}(Q, \hat{\sigma}_e^2) \approx 1/\sqrt{1+n/m}$; when $m \ll n$, $R \approx 1$ and $\mathrm{Corr}(Q, \hat{\sigma}_e^2) \approx \sqrt{m/n} \ll 1$, and thus expecting a good approximation by the mixture. This perhaps shows why the SKAT mixture approximation was considered good in the context of set-tests, when few variants or a large sample is considered. This also shows why, in small samples, the mixture is expected to be a poor approximation.

Calculating p-values

We now describe how to efficiently calculate p-values for the distribution of the statistic $r = Q(\mathbf{y})/\hat{\sigma}_e^2(\mathbf{y})$ calculated from the data; that is, given an observed statistic r, what is $\Pr(Q/\hat{\sigma}_e^2 > r)$ under the null? We review the result in (Chen *et al.* 2016):

Proposition 5. *Let r be the observed value of the statistic. Denote by $\alpha_1^{(r)}, \ldots, \alpha_n^{(r)}$ the eigenvalues of $\mathbf{SKS} - r/(n-p) \cdot \mathbf{S}$. Then,*

$$\Pr\left(\frac{Q}{\hat{\sigma}_e^2} > r\right) = \Pr\left(\sum_{i=1}^{n} \alpha_i^{(r)} \chi_{1,i}^2 > 0\right) \tag{10}$$

where $\chi_{1,i}^2$ are i.i.d. random variables distributed chi-square with one degree of freedom.

However, this condition requires us to calculate the eigenvalues of $\mathbf{SKS} - r/(n-p) \cdot \mathbf{S}$ for each new value r, which, naively, has a complexity of $O(n^3)$. We consider two scenarios where this is problematic. First, in many heritability studies, we wish to test the heritability of many (e.g., thousands) of phenotypes, all relative to the same kernel or kinship matrix (see above). For each phenotype $\mathbf{y}_1, \ldots, \mathbf{y}_N$, we calculate its score test statistic r_i. For p-value calculation, we need to compute the eigendecomposition of $\mathbf{SKS} - r_i/(n-p) \cdot \mathbf{S}$ for each observed statistic r_i, which is a significant computational burden.

A second problematic scenario is of an association study of a single phenotype with many sets of SNPs, e.g. rare variants. Choosing a weighted linear kernel as in SKAT (Wu *et al.* 2011), we have $\mathbf{K}_i = \mathbf{Z}_i\mathbf{W}_i\mathbf{Z}_i^\top$ for each set. As $\mathbf{K}_i$ changes with each test, in principle, we need to perform a costly $O(n^3)$ eigendecomposition for each matrix $\mathbf{K}_i$. However, a significant computational saving is gained due to the fact that the nonzero eigenvalues of $\mathbf{SK}_i\mathbf{S} = \mathbf{SZ}_i\mathbf{W}_i\mathbf{Z}_i^\top\mathbf{S}$ are the same as those of $\mathbf{W}_i^{1/2}\mathbf{Z}_i^\top\mathbf{S}\mathbf{Z}_i\mathbf{W}_i^{1/2}$, which is an $m \times m$ matrix (Lippert *et al.* 2014). As the number of tested SNPs m is often small, calculating the eigenvalues of this matrix instead is significantly faster, taking only $O(m^3)$, with matrix construction taking only $O(n(m+p)^2)$ (see (Lippert *et al.* 2014)). However, with the exact approach, we need to calculate the eigenvalues of $\mathbf{SK}_i\mathbf{S} - r_i/(n-p) \cdot \mathbf{S}$ instead of $\mathbf{SK}_i\mathbf{S}$. Even when $\mathbf{K}_i$ is low rank, the matrix $\mathbf{SK}_i\mathbf{S} - r_i/(n-p) \cdot \mathbf{S}$ may be close to full rank, so another approach is needed.

The following characterizes the eigenvalues of $\mathbf{SKS} - r/(n-p) \cdot \mathbf{S}$ given the eigenvalues of $\mathbf{SKS}$:

Proposition 6. *Let r be the observed score test statistic. Denote by $\phi_1, \ldots, \phi_n$ the eigenvalues of $\mathbf{SKS}$. Denote the column space of a matrix $\mathbf{A}$ by $\mathrm{col}(\mathbf{A})$, its null space by $\ker(\mathbf{A})$. Then,*

$$\Pr\left(\frac{Q}{\hat{\sigma}_e^2} > r\right) = \Pr\left(\sum_{i=1}^{k}\left(\phi_i - \frac{r}{n-p}\right)\chi_{i,1}^2 - \sum_{i=k+1}^{k+q}\frac{r}{n-p} \cdot \chi_{i,1}^2 > 0\right) \tag{11}$$

where $k = \mathrm{rank}(\mathbf{SKS})$ is the number of nonzero eigenvalues ϕ_i, $q = \dim(\ker(\mathbf{SKS}) \cap \mathrm{col}(\mathbf{S}))$, and $\chi_{1,i}^2$ are i.i.d. random variables distributed chi-square with one degree of freedom, $i = 1, \ldots, k+q$.

Proposition 6 shows that calculating the p-value amounts to evaluating the cumulative distribution function (cdf) of a certain weighted mixture of chi-square distribution at 0. This can be done rapidly using the Davies method (Davies 1980), which is based on the numerical inversion of the characteristic function and runs in $O(n)$ complexity, or using other methods (Duchesne and De Micheaux 2010).

It remains to calculate k and q. Naively, this can be done in $O(n^3)$, for example by calculating the singular value decomposition (SVD) of $\mathbf{SKS}$ and $\mathbf{S}$ to get k and to obtain vector bases for $\ker(\mathbf{SKS})$ and $\mathrm{col}(\mathbf{S})$, and by calculating the SVD of a matrix whose columns are the two vector bases to obtain q. When the same kernel is used with many phenotypes, it is a single preprocessing step. However, when the number of SNPs used to construct the kernel and the number of covariates are small, these quantities can be calculated much faster:

Proposition 7. *Suppose $\mathbf{K} = \mathbf{ZWZ}^\top$, and let $k = \mathrm{rank}(\mathbf{SKS})$ and $q = \dim(\ker(\mathbf{SKS}) \cap \mathrm{col}(\mathbf{S}))$. Then, k and q can be calculated in complexity $O(n(m+p)^2)$.*

Most commonly, $k = \min(m, n) - 1$. When the number of SNPs m and the number of covariates p are small, the computational saving is substantial.

Data Availability

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from

www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. The KORA study was initiated and financed by the Helmholtz Zentrum München German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research (BMBF) and by the State of Bavaria. Furthermore, KORA research was supported within the Munich Center of Health Sciences (MC-Health), Ludwig-Maximilians-Universität, as part of LMUinnovativ. The data used in this manuscript were obtained via KORA.PASST (https://epi.helmholtz-muenchen.de/) with the following variables: KORA F4 Illumina HumanMethylation450K BeadChip array, BMIQ normalization KORA F4 Affymetrix 6.0 SNP Array; imputed (HapMap2 reference panel). Access to the data may be obtained by request to KORA.

Results and Discussion

Performance summary

We summarize the results described in above in Table 1 and in Algorithms 1 and 2. We compare our method, RL-SKAT, with the SKAT formulation and the correction of (Chen *et al.* 2016) using the naive implementation of Proposition 5, as implemented by the MiRKAT software package (Zhao *et al.* 2015). The two scenarios discussed are those of a heritability study (same $\mathbf{K}$ with many responses $\mathbf{y}_i$) and SNP set-testing (many low rank $\mathbf{K}_i$). In all methods, a preprocessing step of calculating $\mathbf{X}^\dagger$ and $\{\phi_i\}$ is required. In a heritability study, calculating the statistic $Q/\hat{\sigma}_e^2$ amounts to evaluating two quadratic forms in $O(n^2)$. Compared to RL-SKAT, MiRKAT requires a full $O(n^3)$ eigendecomposition for each $\mathbf{y}_i$. For a set-testing study, these quadratic forms can be calculated in $O(n(m+p))$ due to the low rank of $\mathbf{K}_i$. Again, MiRKAT requires a full $O(n^3)$ eigendecomposition, compared to the $O(n(m+p)^2)$ procedure described in Proposition 7.

We now demonstrate our results on two datasets: a dataset from the Wellcome Trust Case Control Consortium 2 (Consortium *et al.* 2011) (WTCCC2) study and the Cooperative health research in the Region of Augsburg (KORA) study (Holle *et al.* 2005). A full description of data preprocessing is given in the Supplemental Material.

A simulation study using WTCCC2 data

We first analyze data with real genotypes from the WTCCC2 Multiple Sclerosis dataset, and simulated phenotypes. We used the same data processing described in (Yang *et al.* 2014), resulting in m=360,556 SNPs for n=13,950 individuals. We constructed the kinship matrix by a standard, uniformly weighted linear kernel. We sought to demonstrate the discrepancy between the true null distribution and the chi-square weighted mixture distribution. Following Proposition 4, we calculated the correlation to be 0.886 and variance ratio to be $R = 4.69$, indicating that a large discrepancy is possibly expected. To verify this, we simulated 10,000 random phenotypes, where each phenotype is a vector of i.i.d. standard normal variables. We tested whether the variance component is significantly greater than zero, and calculated their p-values under the assumption of either of the two distributions. In Figure 1, we show the quantile-quantile plots for the two sets of p-values. As evidenced, using the SKAT mixture distribution results in a severe deflation of small p-values, while using the correct distribution as in Proposition 1 results in an accurate p-value distribution. This shows that even for large sample sizes (n=13,950), such a discrepancy is possible.

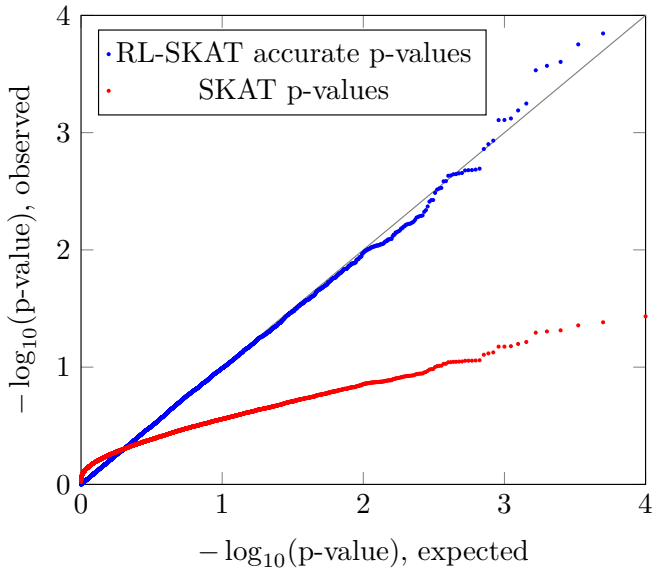

Figure 1 Statistic distribution. Results of the WTCCC2 data analysis, presented by quantile-quantile plots of the $-\log_{10}(p)$-values for heritability significance of 10,000 random phenotypes drawn under the null distribution. Significant deviation from the black line indicates a deflation arising from an inaccurate null distribution. Calculation under the assumption of a weighted mixture of chi-square distributions, gives deflated p-values and potentially creating false negatives. Using the correct distribution, as implemented in RL-SKAT, results in calibrated p-values.

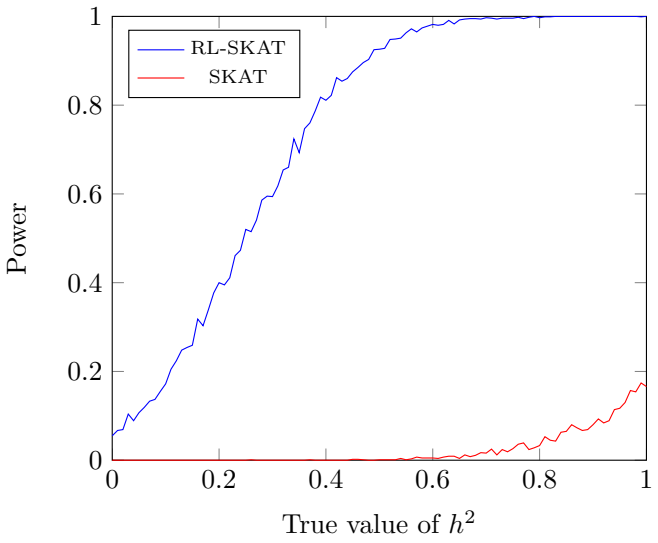

Figure 2 Power study. The power of the accurate approach and SKAT is shown for p-value threshold of $p = 0.05$, for the KORA dataset, on 10,000 simulated phenotypes with varying degrees of true underlying heritability. SKAT is seen to be severely underpowered.

Algorithm 1 RL-SKAT for heritability

procedure PREPROCESSING($\mathbf{X}, \mathbf{K}$) ▷ Preprocessing step, done once
- Calculate $\mathbf{X}^\dagger = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ ▷ $O(np^2)$
- Calculate $\mathbf{SKS}$ using $\mathbf{S} = \mathbf{I} - \mathbf{X}\mathbf{X}^\dagger$ ▷ $O(n^2p)$
- Calculate $\phi_1, \ldots, \phi_n$, the eigenvalues of $\mathbf{SKS}$ ▷ $O(n^3)$
- Extract $k = \text{rank}(\mathbf{SKS})$ ▷ $O(1)$
- Calculate $q = \dim(\ker(\mathbf{SKS}) \cap \text{col}(\mathbf{S}))$ using Proposition 7 ▷ $O(n(n+p)^2)$

procedure TEST($\mathbf{y}$) ▷ Calculate p-value for a single phenotype $\mathbf{y}$
- Calculate the score: $r := Q/\hat{\sigma}_e^2 = (n-p) \cdot \mathbf{y}^\top\mathbf{SKSy}/\mathbf{y}^\top\mathbf{Sy}$ ▷ $O(n^2)$
- Calculate $\{\alpha_i^{(r)}\}$ as in Propositions 5 and 6 ▷ $O(n)$
- Calculate the p-value $p = \Pr\left(\sum_{i=1}^n \alpha_i^{(r)}\chi^2_{1,i} > 0\right)$ using the Davies method ▷ $O(n)$

Algorithm 2 RL-SKAT for set-tests

procedure PREPROCESSING($\mathbf{X}, \mathbf{ZW}^{1/2}$) ▷ Preprocessing step, done once
- Calculate $\mathbf{X}^\dagger = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ ▷ $O(np^2)$
- Calculate $\mathbf{SZW}^{1/2}$ using $\mathbf{S} = \mathbf{I} - \mathbf{X}\mathbf{X}^\dagger$ ▷ $O(nmp)$
- Calculate $\phi_1, \ldots, \phi_n$ as the squares of the singular values of $\mathbf{SZW}^{1/2}$ ▷ $O(nm^2)$
- Extract $k = \text{rank}(\mathbf{SKS})$ ▷ $O(1)$
- Calculate $q = \dim(\ker(\mathbf{SKS}) \cap \text{col}(\mathbf{S}))$ using Proposition 7 ▷ $O(n(m+p)^2)$

procedure TEST($\mathbf{y}$) ▷ Calculate p-value for a single phenotype $\mathbf{y}$
- Calculate the score: $r := Q/\hat{\sigma}_e^2 = (n-p) \cdot \mathbf{y}^\top\mathbf{SKSy}/\mathbf{y}^\top\mathbf{Sy}$, using $\mathbf{K} = \mathbf{ZWZ}^\top$ ▷ $O(n(m+p))$
- Calculate $\{\alpha_i^{(r)}\}$ as in Propositions 5 and 6 ▷ $O(n)$
- Calculate the p-value $p = \Pr\left(\sum_{i=1}^n \alpha_i^{(r)}\chi^2_{1,i} > 0\right)$ using the Davies method ▷ $O(n)$

Table 1 Performance summary. Comparison of the different approaches for p-value calculation discussed. RL-SKAT achieves accuracy while remaining computationally efficient.

Scenario	Algorithm	Exact?	Preprocessing	Calculating $Q/\hat{\sigma}_e^2$	Calculating p-value
Heritability	SKAT	Approximate	$O(np^2 + n^2p + n^3)$	$O(n^2)$	$O(n)$
	MiRKAT	Exact	$O(np^2 + n^2p)$	$O(n^2)$	$O(n^3)$
	RL-SKAT	Exact	$O(np^2 + n^2p + n^3)$	$O(n^2)$	$O(n)$
Set-testing	SKAT	Approximate	$O(np^2 + nmp + nm^2)$	$O(n(m+p))$	$O(n)$
	MiRKAT	Exact	$O(np^2 + nmp)$	$O(n(m+p))$	$O(n^3)$
	RL-SKAT	Exact	$O(np^2 + nmp + n(m+p)^2)$	$O(n(m+p))$	$O(n)$

Testing for heritable methylation sites in the KORA dataset

The longitudinal KORA study consists of whole-blood methylation levels and genotypes of n=1,799 individuals. The phenotype is the proportion of methylated samples at a specific site, averaged across DNA samples of an individual. The study consists of independent population-based subjects from the general population living in the region of Augsburg, southern Germany (Holle *et al.* 2005). Whole-blood samples of the KORA F4 study were used as described elsewhere (Pfeifferm *et al.* 2015). In summary, a total of 431,366 methylation site phenotypes, and 657,103 SNPs, were available for analysis. The correlation as in Proposition 4 is 0.976 and the variance ratio is $R = 22.01$, indicating again that a large discrepancy is expected. We performed a heritability study of multiple phenotypes with the same kinship matrix, by testing the heritability of the N=43,140 methylation sites on chromosome 1. As it is common for a methylation site to be correlated with its surrounding SNPs (Gibbs *et al.* 2010; Zhang *et al.* 2010; Bell *et al.* 2011), we avoided such *cis* effects by using a kinship matrix constructed from the m=604,170 SNPs on all chromosomes other than 1. The kinship matrix is constructed by a standard, uniformly weighted linear kernel. For covariates, we used $\mathbf{X}$ consisting only of an intercept vector. Again, we calculated p-values under the assumption of the two distributions. We note that it has been shown that some methylation site profiles often display significant heritability, while others do not; thus, both significant and insignificant p-values are expected (Rahmani *et al.* 2017).

In Figure 3 we show the histograms of the $\log_{10}$ of the p-value of all the considered phenotypes. The two histograms are indeed very different; p-values calculated using the inaccurate SKAT mixture distribution indicate that the heritability of almost all sites is considered insignificant; for example, using a Bonferroni threshold of $0.05 \cdot 1/43140 \approx 10^{-6}$, only 8/43,140 sites are significant. In light of the results above, it is reasonable to suspect that p-values of many heritable phenotypes are deflated, thus causing false negatives. The p-values distribution has a peak around 0.5, likely an artifact of the inaccurate calculation method. In comparison, p-values calculated by RL-SKAT do not exhibit such a peak. They are significantly smaller, and using the same Bonferroni threshold, we now find 319/43,140 significant sites. Indeed, a simulated power study of both approaches under varying degrees of true underlying heritability validates that the inaccurate approach results in a severe decrease in power (Figure 2), which has been reported in the literature (Uemoto *et al.* 2013). As a point of reference, we compared the power of RL-SKAT with that of the popular LR test approach, and found the have similar power (see the Supplemental Material). We conclude that in this dataset, using the SKAT distribution for p-value calculation is highly problematic.

Benchmarks

Finally, we benchmarked the methods discussed here on the KORA dataset under the two above scenarios, on a 64G, 2.2GHz Linux workstation, using our implementation in the Python language. We verified that the relevant part of our implementation is equivalent to MiRKAT and has a very similar running time. For the scenario of heritability testing, we calculated the p-values of 1000 phenotypes with the kinship matrix. For the scenario of set testing, we used 1000 sets of 100 SNPs each. The results are summarized in Table 2; as expected, the computational savings are very significant, achieving a speedup of more than two orders of magnitude. We expect the speedup to be even more significant for larger datasets.

Discussion

In summary, we have shown that the distribution suggested by SKAT to the score test statistic may be very inaccurate. Unlike previous studies, which have noted this discrepancy only in small sample sizes, we have shown that it might occur in large studies as well. We have proposed a computational method to accurately calculate p-values without compromising computational time. Finally, we demonstrated our findings in two datasets.

The exact calculation of p-values can be applied to other variants of the score test; for example, the SKAT-O (Lee *et al.* 2012) seeks to find an optimal combination of burden tests and non-burden tests, which amounts to the score test with a certain kernel.

In this work, we focused on the case of a single kernel, and on a continuous phenotype. The extension of this work to multiple kernels (e.g., corresponding to several sets of SNPs) or to binary phenotypes (e.g., case/control studies) is nontrivial, as the null distribution cannot be modeled as a ratio of quadratic forms; see, e.g., (Wu *et al.* 2016; Wang 2016). It therefore remains a subject for future work.

We believe that the prominence of likelihood-ratio based tests in heritability studies might stem from the statistical issues discussed above; see for example (Uemoto *et al.* 2013), where SKAT was found to be significantly less powerful. It is our hope that this paper would facilitate the use of score tests in heritability studies in the future.

Acknowledgements

R.S. is supported by the Colton Family Foundation. This study was supported in part by a fellowship from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University to R.S and E.R.

Literature Cited

Bell, J. T., A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, *et al.*, 2011 DNA methylation patterns associate with genetic and gene expression variation in hapmap cell lines. Genome biology **12**: R10.

Chen, J., W. Chen, N. Zhao, M. C. Wu, and D. J. Schaid, 2016 Small sample kernel association tests for human genetic and microbiome association studies. Genetic epidemiology **40**: 5–19.

Consortium, I. M. S. G., W. T. C. C. C. 2, *et al.*, 2011 Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature **476**: 214–219.

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**: 323–333.

Duchesne, P. and P. L. De Micheaux, 2010 Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. Computational Statistics & Data Analysis **54**: 858–862.

Furlotte, N. A., D. Heckerman, and C. Lippert, 2014 Quantifying the uncertainty in heritability. Journal of human genetics **59**: 269–275.

Ganjgahi, H., A. M. Winkler, D. C. Glahn, J. Blangero, P. Kochunov, *et al.*, 2015 Fast and powerful heritability inference for family-based neuroimaging studies. NeuroImage **115**: 256–268.

Ge, T., T. E. Nichols, P. H. Lee, A. J. Holmes, J. L. Roffman, *et al.*, 2015 Massively expedited genome-wide heritability analysis

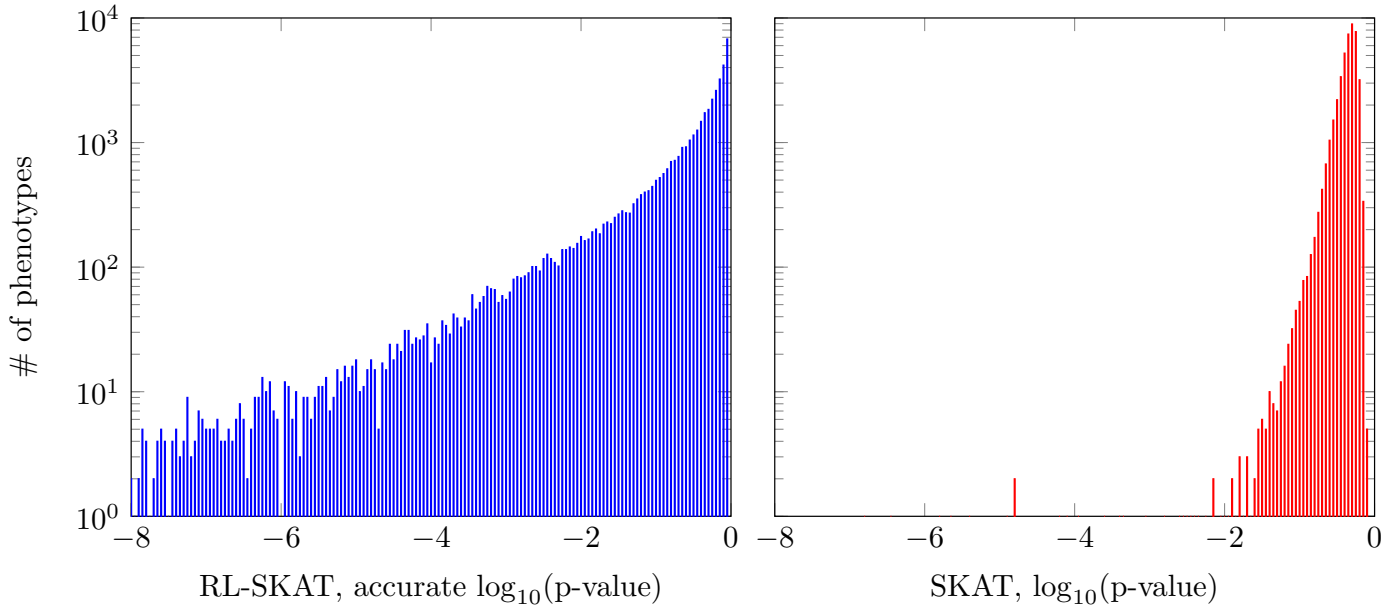

Figure 3 Heritability study. Histograms of the p-values of the studied phenotypes in the KORA dataset, as calculated by the accurate method (left) and the inaccurate method (right). Histograms are shown in log-scale, and are capped at $p = 10^{-8}$ for clarity of presentation. SKAT tends to severely deflate p-values which are small according to the accurate calculation, leading to a severe loss of power.

Table 2 Benchmarks. Benchmark of the performance of different approaches for p-value calculation, applied to the KORA dataset.

Scenario	Algorithm	Exact?	Preprocessing	Calculating $Q/\hat{\sigma}_e^2$	Calculating p-value
Heritability	SKAT	Approximate	3 sec	0.3 sec	5 sec
	MiRKAT	Exact	0.2 sec	0.3 sec	37 minutes
	RL-SKAT	Exact	9 sec	0.3 sec	5 sec
Set-testing	SKAT	Approximate	45 sec	2 sec	1 sec
	MiRKAT	Exact	5 sec	2 sec	43 minutes
	RL-SKAT	Exact	50 sec	2 sec	4 sec

(MEGHA). Proceedings of the National Academy of Sciences **112**: 2479–2484.

Gibbs, J. R., M. P. van der Brug, D. G. Hernandez, B. J. Traynor, M. A. Nalls, *et al.*, 2010 Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet **6**: e1000952.

Hasegawa, T., K. Kojima, Y. Kawai, K. Misawa, T. Mimori, *et al.*, 2016 AP-SKAT: highly-efficient genome-wide rare variant association test. BMC genomics **17**: 745.

Hayes, B., M. Goddard, *et al.*, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Holle, R., M. Happich, H. Löwel, H. Wichmann, M. study group, *et al.*, 2005 KORA - a research platform for population based health research. Das Gesundheitswesen **67**: 19–25.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S.-y. Kong, *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. Nature genetics **42**: 348–354.

Lee, S., M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, *et al.*, 2012 Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. The American Journal of Human Genetics **91**: 224–237.

Lippert, C., J. Xiang, D. Horta, C. Widmer, C. Kadie, *et al.*, 2014 Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. Bioinformatics p. btu504.

Liu, D., D. Ghosh, and X. Lin, 2008 Estimation and testing for the effect of a genetic pathway on a disease outcome using logistic kernel machine regression via logistic mixed models. BMC bioinformatics **9**: 292.

Liu, D., X. Lin, and D. Ghosh, 2007 Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics **63**: 1079–1088.

Lloyd-Jones, L. R., A. Holloway, A. McRae, J. Yang, K. Small, *et al.*, 2017 The genetic architecture of gene expression in peripheral blood. The American Journal of Human Genetics **100**: 228–237.

Marcenko, V. A. and L. A. Pastur, 1967 Distribution of eigenvalues for some sets of random matrices. Mathematics of the USSR-Sbornik **1**: 457.

Patterson, N., A. L. Price, and D. Reich, 2006 Population structure and eigenanalysis. PLoS genet **2**: e190.

Pfeifferm, L., S. Wahl, L. C. Pilling, E. Reischl, J. K. Sandling, *et al.*, 2015 DNA methylation of lipid-related genes affects blood lipid levels. Circulation: Cardiovascular Genetics pp. CIRCGENETICS–114.

Price, A. L., A. Helgason, G. Thorleifsson, S. A. McCarroll, A. Kong, *et al.*, 2011 Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. PLoS Genet **7**: e1001317.

Quon, G., C. Lippert, D. Heckerman, and J. Listgarten, 2013 Patterns of methylation heritability in a genome-wide analysis of four brain regions. Nucleic acids research p. gks1449.

Rahmani, E., L. Shenhav, R. Schweiger, P. Yousefi, K. Huen, *et al.*, 2017 Genome-wide methylation data mirror ancestry information. Epigenetics & Chromatin **10**: 1.

Schweiger, R., E. Fisher, E. Rahmani, L. Shenhav, S. Rosset, *et al.*, 2017 Using stochastic approximation techniques to efficiently construct confidence intervals for heritability. In *International Conference on Research in Computational Molecular Biology*, pp. 241–256, Springer.

Schweiger, R., S. Kaufman, R. Laaksonen, M. E. Kleber, W. März, *et al.*, 2016 Fast and accurate construction of confidence intervals for heritability. The American Journal of Human Genetics **98**: 1181–1192.

Searle, S. R., G. Casella, and C. E. McCulloch, 2009 *Variance components*, volume 391. John Wiley & Sons.

Uemoto, Y., R. Pong-Wong, P. Navarro, V. Vitart, C. Hayward, *et al.*, 2013 The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. Frontiers in genetics **4**.

Van Dongen, J., M. G. Nivard, G. Willemsen, J.-J. Hottenga, Q. Helmer, *et al.*, 2016 Genetic and environmental influences interact with age and sex in shaping the human methylome. Nature communications **7**.

Visscher, P. M. and M. E. Goddard, 2015 A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. Genetics **199**: 223–232.

Visscher, P. M., W. G. Hill, and N. R. Wray, 2008 Heritability in the genomics era – concepts and misconceptions. Nature Reviews Genetics **9**: 255–266.

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

Wright, F. A., P. F. Sullivan, A. I. Brooks, F. Zou, W. Sun, *et al.*, 2014 Heritability and genomics of gene expression in peripheral blood. Nature genetics **46**: 430–437.

Wu, B., W. Guan, and J. S. Pankow, 2016 On efficient and accurate calculation of significance p-values for sequence kernel association testing of variant set. Annals of human genetics **80**: 123–135.

Wu, M. C., S. Lee, T. Cai, Y. Li, M. Boehnke, *et al.*, 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics **89**: 82–93.

Yang, J., B. Benyamin, B. P. McEvoy, S. Gordon, A. K. Henders, *et al.*, 2010 Common SNPs explain a large proportion of the heritability for human height. Nature Genetics **42**: 565–9.

Yang, J., S. H. Lee, M. E. Goddard, and P. M. Visscher, 2011 GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics **88**: 76–82.

Yang, J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, 2014 Advantages and pitfalls in the application of mixed-model association methods. Nature genetics **46**: 100–106.

Zhang, D., L. Cheng, J. A. Badner, C. Chen, Q. Chen, *et al.*, 2010 Genetic control of individual differences in gene-specific methylation in human brain. The American Journal of Human Genetics **86**: 411–419.

Zhang, D. and X. Lin, 2003 Hypothesis testing in semiparametric additive mixed models. Biostatistics **4**: 57–74.

Zhao, N., J. Chen, I. M. Carroll, T. Ringel-Kulka, M. P. Epstein, *et al.*, 2015 Testing in microbiome-profiling studies with MiRKAT, the microbiome regression-based kernel association test. The American Journal of Human Genetics **96**: 797–807.