Published in final edited form as:

Genet Epidemiol. 2019 March ; 43(2): 122–136. doi:10.1002/gepi.22180.

A Review of Kernel Methods for Genetic Association Studies

Nicholas B. Larson^{#1,†}, Jun Chen^{#1}, and Daniel J. Schaid¹

¹Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

[#] These authors contributed equally to this work.

Abstract

Evaluating the association of multiple genetic variants with a trait of interest by use of kernel-based methods has made a significant impact on how genetic association analyses are conducted. An advantage of kernel methods is that they tend to be robust when the genetic variants have effects that are a mixture of positive and negative effects, as well as when there is a small fraction of causal variants. Another advantage is that kernel methods fit within the framework of mixed models, providing flexible ways to adjust for additional covariates that influence traits. Herein, we review the basic ideas behind the use of kernel methods for genetic association analysis, as well as recent methodological advancements for different types of traits, multivariate traits, pedigree data, and longitudinal data. Finally, we discuss opportunities for future research.

Keywords

genetic association analysis; kernel statistic; mixed model; multivariate; pedigree data

INTRODUCTION

Recent advances in high-throughput next-generation sequencing technologies have led to an explosion of novel genomic assays and new information. A major analytical challenge has been to develop sophisticated and robust statistical methods that can accommodate the high-dimensional and structured nature of the genetic information captured by these assays. Even for large genetic studies with tens of thousands of samples, single-marker association tests can have limited power for low-frequency and rare variants. Alternative methods that analyze the association of a trait with multiple markers within a genomic region have the potential to improve power and lend additional biological interpretability. One strategy is to summarize the multiple genetic markers into a single “burden” score and analyze the association of the trait with this score [Madsen and Browning 2009; Morgenthaler and Thilly 2007]. These approaches, however, have diminished power when the effects of the genetic markers on the trait are in opposing directions, or when only a small fraction of the variants have an effect on the trait [Lee, et al. 2014; Moutsianas, et al. 2015]. In contrast, kernel methods provide a highly flexible framework for aggregative variant testing, and can

[†] Corresponding Author: Nicholas B. Larson, PhD, 200 First St SW, Rochester, MN 55905, Larson.nicholas@mayo.edu, Phone: 507-293-1700, Fax: 507-284-1516.

have greater power than burden tests under a variety of conditions. Kernel methods have intuitive appeal, because they quantify the genetic similarity among pairs of subjects and test whether this genetic similarity is associated with trait similarity. Kernel methods also fit within the framework of regression models, making it straight-forward to adjust for covariates that influence the trait. The flexibility and computational feasibility of kernel methods make them particularly appealing for large-scale genetic studies.

In this review, we discuss the basics of kernel methods for genetic association analysis while highlighting recent advances in the areas of hypothesis testing, extensions to a variety of traits, and underlying kernel design. We draw connections between kernel tests and other statistical methods and offer concluding thoughts on new research directions for kernel methods in genomic data analysis. As a guide for the various methods that will be discussed and software that implements them, we present in Table 1 software tools that are based on kernel statistics for genetic analyses.

KERNEL MACHINES REGRESSION FRAMEWORK

Consider a genetic association study of sample size N for some continuous-valued outcome $y = (y_1, \dots, y_N)'$. We define the measured set of P non-genetic covariates (e.g., age and sex) as the matrix X . For a given variant set of interest of size M , we similarly define the $N \times M$ matrix G to represent codes for the observed (and possibly imputed) variant genotypes. Under an additive genetic model, it follows that G corresponds to minor allele dosage (0, 1, or 2), although alternative representations (e.g., dominant, recessive) are possible without loss of generality.

Simultaneously modeling the linear genetic effects for all M variants in G while adjusting for may be conducted using ordinary multiple linear regression, such that

$$y_i = \alpha_o + X_i' \alpha + G_i' \beta + \varepsilon_i$$

where α_o is an intercept, $\alpha = (\alpha_1, \dots, \alpha_P)'$ are the effects corresponding to the non-genetic covariates, $\beta = (\beta_1, \dots, \beta_M)'$ are the genetic effect parameters, and $\varepsilon_i \sim N(0, \sigma_e^2)$ are the independent and identically distributed (i.i.d.) residual errors. In practice, the assumption of linearity in the effects of G_i may not hold, and model fitting may be under-determined for large M when the sample size N is small. An alternative strategy is to adopt a more flexible semi-parametric regression approach, such that the covariates are modeled parametrically while the genetic markers are modeled in a nonparametric manner. The regression equation can be expressed as

$$y_i = \alpha_o + X_i' \alpha + f(G_i) + \varepsilon_i$$

where $f(\cdot)$ is some unknown smooth function. Before we discuss the mathematical details, a brief intuitive introduction to semi-parametric regression is warranted. Recall that typical regression models require the user to explicitly code the effects of each term in the model,

such as linear or quadratic effects. These “mapping” functions are the basis functions for how the terms are coded in the regression model. More flexible mapping functions include spline functions, which are piecewise polynomials. However, as the number of genetic markers increases, it is difficult to specify all the basis functions that make their joint effects smooth for the function $f(\cdot)$. The advantage of kernel methods is that they do not need to explicitly define these basis functions.

Because the true representation of $f(G_i)$ may be highly non-linear and complex, kernel regression methods can be applied in the form of regularized least squares kernel machines (LSKM) [Liu, et al. 2007]. Instead of the usual regression approach of explicitly defining basis functions to represent the effect of $f(\cdot)$, kernel methods approximate $f(\cdot)$ via an kernel matrix K , such that $K_{ij} = \kappa(G_i, G_j)$ for some kernel function $\kappa(\cdot, \cdot)$. A simple and popular kernel function is the linear-weighted kernel, $\kappa(G_i, G_j) = \sum_{k=1}^M w_k G_{ik} G_{jk}$, for some set of *a priori* defined weights w_k . The resulting regression equation provides a flexible nonparametric model for the joint effects of all the genetic markers, and is based on a regularized regression to avoid over-fitting of the model to the data. The mathematical basis for these types of models is well-developed [Kimeldorf and Wahba 1971; Liu, et al. 2007] and reviewed elsewhere [Schaid 2010]. A key aspect of fitting these types of models, as shown by Liu et al. [Liu, et al. 2007], is that the estimating equations for the regularized LSKM model are equivalent to fitting the mixed-effects model $y_i = \alpha_o + X_i' \alpha + f(G_i) + \varepsilon_i$, where X_i is a fixed effect and $f(G_i)$ is a random effect. To express the model for all N subjects, the subjects' components of this model can be stacked as $y' = (y_1, \dots, y_N)$, $X' = (X_1', \dots, X_N')$, $f = (f(G_1), \dots, f(G_N))$, and $\varepsilon' = (\varepsilon_1, \dots, \varepsilon_N)$. We can represent the model as $y = X\alpha + f + \varepsilon$, where $f \sim MVN(0, \sigma_K^2 K)$ and $V = \sigma_\varepsilon^2 I + \sigma_K^2 K$ is the marginal variance matrix of the vector y . Thus, can be intuited as the covariance structure of the genetic random effects f , and the heritability (i.e., percent of trait variation explained by the kernel matrix) is determined by $h^2 = \sigma_K^2 / (\sigma_K^2 + \sigma_\varepsilon^2)$.

HYPOTHESIS TESTING AND P-VALUES

A common use of the kernel approach is to test whether the genetic variants in $f(G)$, as represented by K , are associated with the outcome of interest. Within the mixed modeling framework of kernel regression, testing the null hypothesis that no genetic effects are present, $H_0: f = 0$, is equivalent to $H_0: \sigma_K^2 = 0$. That is, a high-dimensional test is reduced to testing a single variance component. Liu et al. [Liu, et al. 2007] proposed applying a variance component score test in the context of pathway-level testing of microarray expression data, which was later adapted for aggregative genetic variant testing by Kwee et al. [Kwee, et al. 2008]. The LSKM score test statistic can be derived by taking the first derivative of the restricted maximum likelihood (REML) equation with respect to σ_K^2 and evaluating it under the null hypothesis (see Harville [Harville 1977]). The score statistic takes on the quadratic form

$$Q = \frac{1}{2\sigma_\varepsilon^2}(y - \hat{y})'K(y - \hat{y}), \quad (1)$$

where $\hat{y}$ is the fitted value of y under the null model, $\hat{y}_i = \alpha_o + X_i'\hat{\alpha}$, which is easy to fit with standard regression models for fixed effects (e.g., linear regression for quantitative traits, or logistic regression for binary traits). Note that some authors use σ_ε^2 in the denominator, as in expression (1), others ignore this term [Wu, et al. 2011], and a formal derivation based on derivatives of the log-likelihood function would use σ_ε^4 in the denominator. However, these variations are just different scalings of the quadratic statistic. As long as the scaling is considered when computing p-values, the resulting p-value will be valid.

This testing framework is particularly advantageous when testing many different sets of genetic markers, because the null model needs to be fit only once for all sets of candidate variants G (e.g., all genes in whole-exome sequencing). Because variance components are restricted to be non-negative, the null condition of $\sigma_k^2 = 0$ lies on the boundary of the support $[0, \infty)$. This means that the score test statistic Q does not correspond to the standard χ_1^2 distribution under H_0 .

To calculate p-values for kernel tests under large N , moment-matching methods were initially proposed to approximate the null distribution. This entails computing the first two moments of Q under the null hypothesis, and using these moments to estimate the degrees of freedom and a factor to rescale the chi-square distribution. Liu et al. [Liu, et al. 2007] and Kwee et al. [Kwee, et al. 2008] both applied the same Satterthwaite approximation, although this can be anti-conservative at the extreme tail and higher moment-matching methods have been proposed to improve accuracy [Wu, et al. 2016; Wu and Pankow 2016a]. For the sequence kernel association test (SKAT), Wu et al. [Wu, et al. 2011] considered a more accurate asymptotic approximation of the distribution of Q , recognizing that its distribution is a linear combination of i.i.d. χ_1^2 random variables, and applied the Davies method [Davies 1980a]. This method provides an analytic solution to computing p-values for Q by inverting the characteristic function of a linear combination of χ_1^2 random variables. However, Davies' method requires specifying a numerical accuracy parameter, and for very small p-values, the numerical results can be inaccurate if the accuracy parameter is not specified well. In contrast, a method by Kuonen [Kuonen 1999] that uses a saddlepoint method does not have this complication, and has been recommended to be the better method to compute p-values [Chen, et al. 2013].

To understand why the distribution of Q is a linear combination of χ_1^2 random variables, we provide a brief derivation, because the ideas behind this approach can be used in different contexts when developing kernel statistics. It is helpful to recall that $\hat{\alpha} = (X'X)^{-1}X'y$, so that $(y - \hat{y}) = \hat{\varepsilon} = P_o y$, where $P_o = I - H$, and H is the usual “hat” matrix for linear regression,

$H = X(X'X)^{-1}X'$. This implies that $\hat{\varepsilon} \sim N(0, \sigma_\varepsilon^2 P_0)$. By de-correlating the residuals, $\varepsilon^* = \sigma_\varepsilon^{-1} P_0^{-1/2} \hat{\varepsilon} \sim N(0, I)$. This lets us represent the residuals as $\hat{\varepsilon} = \sigma_\varepsilon P_0^{-1/2} \varepsilon^*$. Substituting this expression of the residuals into Q results in $Q = \varepsilon^{*'} A \varepsilon^*$, where $A = \frac{1}{2} P_o^{1/2} K P_o^{1/2}$. The eigen-decomposition of $A = \sum \lambda_i e_i e_i'$, where λ_i are the eigenvalues and e_i are the orthonormal eigenvectors. This allows us to express $Q = \sum \lambda_i (\varepsilon^{*'} e_i)^2 = \sum \lambda_i z_i^2$, where $z_i \sim N(0, 1)$. Because the vectors are normalized to have length 1, $(\varepsilon^{*'} e_i) = z_i \sim N(0, 1)$, and because they are orthogonal, the z_i 's are independent. This illustrates that the distribution of Q is a linear combination of i.i.d. χ_1^2 random variables, with weights equal to the eigenvalues of matrix $A = \frac{1}{2} P_o^{1/2} K P_o^{1/2}$. For conditions under very large N and M , such as large-scale sequencing studies, calculating these eigenvalues can be computationally demanding since the complexity is $O(\min(N^3, M^3))$ Lumley et al. [Lumley, et al. 2018] recently proposed fastSKAT, which combines the Satterthwaite and saddlepoint strategies to greatly improve computational efficiency. This hybrid approach computes the k leading eigenvalues for a weighted sum of χ_1^2 random variables, and then approximates the “remainder” term of the distribution using moment matching.

Another limitation of using Davies and Kuonen methods to determine p-values is that they can be conservative for small samples. This conservativeness is caused by not accounting for the variability in the estimated σ_ε^2 . To overcome this, Chen et al. [Chen, et al. 2016] proposed an exact p-value that directly incorporates the estimate uncertainty of $\hat{\sigma}_\varepsilon^2$ by recognizing that Q is a ratio of two quadratic statistics. Specifically, the exact p-value was calculated based on the adjusted score statistic

$$Q_a = \frac{(y - \hat{y})' K (y - \hat{y})}{(y - \hat{y})' (y - \hat{y})}, \quad (2)$$

Using the same trick above, where $(y - \hat{y}) = \hat{\varepsilon} = \sigma_\varepsilon P_o^{1/2} \varepsilon^*$, and substituting into Q_a , the p-value can be determined by

$$\begin{aligned} P(Q_a > Q_{a, obs}) &= P\left(\frac{\varepsilon^{*'} P_o^{1/2} K P_o^{1/2} \varepsilon^*}{\varepsilon^{*'} P_o^{1/2} P_o^{1/2} \varepsilon^*} > Q_{a, obs}\right) \\ &= P\left(\varepsilon^{*'} P_o^{1/2} K P_o^{1/2} \varepsilon^* > Q_{obs} [\varepsilon^{*'} P_o \varepsilon^*]\right) \\ &= P\left(\varepsilon^{*'} \left[P_o^{1/2} K P_o^{1/2} \varepsilon^* - Q_{obs} P_o\right] \varepsilon^* > 0\right) \\ &= P(\varepsilon^{*'} A \varepsilon^* > 0) \end{aligned}$$

Hence, by using the eigenvalues of $A = P_o^{1/2} K P_o^{1/2} \varepsilon^* - Q_{obs} P_o$ the Davies or Kuonen method can be used to compute more accurate p-values that account for estimation of σ_ε^2

Analogous small-sample adjustment strategies have been developed for binary-valued traits [Chen, et al. 2016] and for multivariate traits [Zhan, et al. 2017a] and correlated traits [Zhan, et al. 2018]. Although the adjusted score statistic was motivated by small-sample studies, Schweiger et al. [Schweiger, et al. 2017] showed that the adjustment could also increase the statistical power for large sample sizes, with the extent of power improvement depending on the coefficient of variation of the eigenvalues of the kernel matrix. Schweiger et al. [Schweiger, et al. 2017] also implemented a computationally efficient approach for large sample sizes. Recently, Wang et al. [Wang 2016] proposed a modified approach to estimate the coefficients λ under the linear kernel, SKAT+, by using sample subsets to better estimate the null genetic covariance.

Other hypothesis testing procedures based on kernel methods have been proposed beyond score tests. For example, Zeng et al. [Zeng, et al. 2014] outlined a restricted likelihood ratio test approach for aggregative rare variant testing. The authors argued for the advantage of direct parameter estimation, since both null (i.e., reduced) and alternative (i.e., full) models must be fit as a condition of hypothesis testing when using a likelihood ratio statistic, in contrast to score statistics that require only parameter estimates for the null hypothesis model. Simulations by the authors indicated potential power improvements over the score test used by SKAT. Bayesian approaches to kernel-based inference have also been developed using Bayesian linear mixed models. Wen [Wen 2015] derived a computationally efficient approach for approximate Bayes factors for continuous traits and demonstrated comparable performance to counterpart frequentist approaches. The Bayesian framework also affords a higher degree of modeling flexibility, as prior information and hierarchical relationships can be directly accommodated in the analysis. Both of these examples adopt a model comparison framework. Although there is a natural computational advantage of score tests, given that only the null model is required to be fit, there is appeal to the flexibility of Bayesian modeling, which presents opportunities for future research in kernel-based methods for association analysis.

Binary Traits

The LSKM model can be extended to other types of traits with exponential family distributions (e.g., binary disease status, Poisson count data) based on score statistics for the generalized linear mixed model [Lin 1997]. Under the null hypothesis that none of the genetic markers are associated with a trait, the generalized linear mixed model reduces to the usual generalized linear model $g(\mu_i) = \alpha_o + X_i'\alpha + f_i$, where $g(\cdot)$ is an appropriate link function and is the expected value of y_i . The general form of the kernel statistic considers the link function for the expected value, as well as a variance function for how the variance depends on the expected value [Lee, et al. 2012b]. For example, for binary traits and a logit link function without over-dispersion, the regression model reduces to the familiar logistic regression model with variance function $v(\mu_i) = \mu_i(1 - \mu_i)$. When using canonical links, the kernel statistic reduces to $Q = (y - \hat{y})'K(y - \hat{y})$ for binary and Poisson data [Wu, et al. 2010; Wu, et al. 2011]. Hence, the kernel statistic for binary traits, such as case-control data, has this simple form, making it computationally rapid to compute since the covariate-adjusted fitted values only need to be compute once. Although the Davies and Kounen methods can

be used to compute p-values in the same as approach for quantitative traits, for small samples the p-values can be conservative. Small-sample moment corrections for variance and kurtosis for binomial for case-control data have been developed to improve kernel methods [Lee, et al. 2012a]. Although such correction improves the accuracy at small α levels, anti-conservativeness has been observed at traditional α levels (e.g., $\alpha = 0.05, 0.1$) [Chen, et al. 2016].

Censored Survival Data

Kernel statistics for the association of a set of genetic markers with censored survival data have been developed by a number of investigators for different types of genetic data and different types of kernels [Cai, et al. 2011; Chen, et al. 2014; Li and Luan 2003; Lin, et al. 2011; Plantinga, et al. 2017]. The general ideas are similar to the approach used for linear regression kernel machines, but instead of using linear regression residuals we now use estimated martingale residuals that result from fitting a Cox regression model to adjusting covariates. If we let M denote a vector of martingale residuals for N subjects, and K the $N \times N$ kernel matrix, the kernel statistic take the familiar quadratic form, $Q = M'KM$. Under the null hypothesis, the test statistic follows a mixture of χ^2 distributions which can be approximated by the Satterthwaite method based on a rescaled χ^2 distribution or by resampling methods [Cai, et al. 2011]. However, for modest sample sizes and complicated kernels, the test statistic can be conservative. Methods that account for over-dispersion, like those outlined for expression (2), provide more robust p-values [Plantinga, et al. 2017].

Interaction Testing

In addition to the effects of the variants included in a given variant set, kernel methods have also been extended to evaluating statistical interactions with environmental exposures (GxE) or other genetic variants (GxG) in an aggregative fashion. For GxG associations with a continuous outcome, Li and Cui [Li and Cui 2012] proposed a smoothing spline ANOVA approach, SPA3G, for modeling marginal and interaction variant set effects using kernels. Larson and Schaid [Larson and Schaid 2013] later extended this approach to binary traits using generalized linear mixed models via penalized quasi-likelihood. Lin et al. [Lin, et al. 2013] proposed aggregative GxE testing by treating the main genetic and environmental effects as fixed and testing the significance of random GxE interaction effects using a score test. Broadaway et al. [Broadaway, et al. 2015] provided an alternative formulation of GESAT, which jointly tests the genetic, environmental, and GxE interaction effects using the joint weighted two-way interaction kernel. Kernel methods outside of the kernel regression framework, such as kernel canonical correlation analysis [Larson, et al. 2014; Yuan, et al. 2012], have also been developed for GxG association analyses in case-control studies.

Multivariate Traits

To extend testing to multiple quantitative traits, it can be more powerful to directly account for the trait correlations rather than perform multiple single-trait analyses and correct for multiple testing. To simultaneously analyze all P traits, we move from the linear mixed model described above to the multivariate linear mixed model. By stacking vectors for each trait on top of each other, we can express the model as $y = X\alpha + f(G) + \epsilon$, where

$y' = (y'_1, \dots, y'_p)$, $f' = (f'_1, \dots, f'_p)$, $X = \text{diag}(X)$ (i.e., the matrix of covariates as a block-diagonal matrix), $\alpha' = (\alpha'_1, \dots, \alpha'_p)$, and $\varepsilon' = (\varepsilon'_1, \dots, \varepsilon'_p)$. For simplicity we assume that the same covariate matrix, X , applies to all traits, although this can be altered. The random genetic effect $f \sim N(0, T \otimes K)$, where $\otimes$ is the Kronecker product operator, T is a $P \times P$ matrix of the genetic variance and covariance components, τ_{ij} , and K is the $N \times N$ kernel matrix of genetic similarities. The error term $\varepsilon \sim N(0, \Sigma \otimes I_N)$, where Σ is a $P \times P$ covariance matrix for the within-subject covariances of the errors, σ_{ij} , and I_N is an $N \times N$ identity matrix. From this, it can be seen that the covariance matrix of the vector Y is $V = T \otimes K + \Sigma \otimes I$. This illustrates that a main feature of the multivariate linear mixed model is that it decomposes the covariance matrix of the vector Y into genetic variances and covariances, T , and residual variances and covariances, Σ .

To test the null hypothesis that none of the traits is associated with a set of genetic markers, Maity et al. [Maity, et al. 2012] derived the REML score statistic for multivariate kernel machine regression (MVKM), based on derivatives of the log-likelihood for a restricted multivariate normal likelihood equation [Harville 1977]. The null hypothesis that none of the traits are associated with a set of genetic markers is $H_0: \tau_{11} = \tau_{22} = \dots = \tau_{pp} = 0$. Note that the off-diagonal genetic covariances (τ_{ij}) do not exist when genetic variances (τ_{ii}) are zero. This means we can specify a null hypothesis multivariate kernel matrix for p traits as block-diagonal, $M_o = I_p \otimes K$, where I_p is a $P \times P$ identity matrix. From this setup, the score statistic derived by Maity et al. [Maity, et al. 2012] is

$$Q_{\text{Maity}} = (y - X\hat{\alpha})' V_o^{-1} M_o V_o^{-1} (y - X\hat{\alpha}),$$

where $V_o = \Sigma \otimes I_N$. The asymptotic distribution of the REML score statistic is a linear combination of χ^2_1 random variables, which can be evaluated by Davies' [Davies 1980b] or Kuonen's [Kuonen 1999] method. To account for estimation of $\hat{\Sigma}$, the statistic Q_{Maity} can be scaled by an estimate of the null residual variances in order to avoid conservative p-values [Zhan, et al. 2018]. Details are provided in the Appendix.

An alternative strategy to analyze multivariate traits is to create a kernel matrix for traits (denoted K_y), a measure of similarity for pairs of subjects across all traits, as well as a kernel matrix for genetic markers (denoted K_g). Then, a test of independence between these two kernel matrices can be expressed as $T_{\text{GAMuT}} = \text{trace}(K_y K_g) / N$. By taking the trace of the matrix product, the statistic is a sum over all products of the elements of these two matrices, much like a score statistic for the association of a trait y with a genetic marker g depends on their product $y_i g_j$. When a linear kernel is used to create the matrix K_y , it can be shown that the REML score statistic of Maity et al. [Maity, et al. 2012] is the same as the GAMuT statistic (genome association with multiple traits) derived by Broadaway et al. [Broadaway, et al. 2016]. The asymptotic distribution of T_{GAMuT} depends on the eigenvalues of both

matrices K_y and K_g . An advantage of this approach is that it can be used for both binary and quantitative traits.

A caution regarding the asymptotic distribution for both the mixed model approach of Maity [Maity, et al. 2012] and the kernel matrix product approach of Broadaway [Broadaway, et al. 2016] is that it can be conservative when the sample size is small relative to the number of traits [Chen, et al. 2016; Lee, et al. 2012a]. Permutation methods, which permute indices of subjects to randomly distribute their multivariate traits, can overcome this limitation, but can be computationally time consuming. An alternative is to use the kernel matrix product approach of Broadaway [Broadaway, et al. 2016], or similar approaches [Zhan, et al. 2017b], to calculate the association between these two kernel matrices, but compute p-values as follows. Instead of actually permuting subjects, one can compute the finite sample moments of the statistic under the null hypothesis of no association. The finite sample moments are the moments of the distribution if all $N!$ permutations were enumerated, yet the moments can be calculated without actually enumerating all permutations. P-values [Zhan, et al. 2017b] can then be efficiently computed by matching the first three moments with a Pearson type III density [Zhan, et al. 2017b]. This approach can be especially useful for high-dimensional traits.

Multivariate binary traits can be complicated to model with kernel methods, because likelihoods for multivariate binary data are notoriously challenging, both theoretically and computationally. Nonetheless, Davenport et al. [Davenport, et al. 2017] developed a novel solution by combining kernel machine logistic regression models with the framework of generalized estimating equations. By fitting a marginal logistic regression model for each binary trait (e.g., adjusting each trait for covariates), they were able to obtain residuals $(y - \hat{y})$ from the fitted models. Recall that the usual kernel score statistics have the general form $Q = (y - \hat{y})'K(y - \hat{y})$. For a single trait, this type of quadratic statistic can be viewed as a test of goodness-of-fit of the model residuals using the kernel matrix to define the dependency among the residuals [le Cessie and van Houwelingen 1995]. If the multivariate traits are stacked on top of each other, the residuals $(y - \hat{y})$ would be correlated between the different traits, and accounting for this correlation is important in order to avoid inflated Type-I error rates. Davenport et al. [Davenport, et al. 2017] accounted for between-trait correlations using the framework of generalized estimating equations, relying on empirical estimates of the cross-trait correlations. Although the authors did not consider a mixture of different types of traits (e.g., binary and quantitative), their approach should be easily generalized to a mixture of different types of traits that can be modeled marginally by generalized linear models. Further research to understand the strengths and limitations of model-based approaches by Maity et al. [Maity, et al. 2012] and Davenport et al. [Davenport, et al. 2017] versus the kernel matrix product approach of Broadaway [Broadaway, et al. 2016] would be beneficial.

PEDIGREE DATA

When using the kernel association test for pedigree data, it is important to account for the correlation of regression model residuals among members of a pedigree. General approaches based on fitting likelihood mixed models for pedigree data have been used [Almeida, et al.

2014], and computationally efficient score tests for quantitative traits have become popular. Score statistics can be constructed by including an additional random effect in the regression model, denoted by vector b_i for the i^{th} pedigree, so that the resulting linear mixed model for a pedigree looks like $y_i = \alpha_o + X_i' \alpha + f(G_i) + b_i + \varepsilon_i$ [Chen, et al. 2013; Ouakacha, et al. 2013; Schifano, et al. 2012]. By assuming the random effect has a multivariate normal distribution with mean 0 and variance matrix $V(b_i) = \sigma_b^2 \Phi_i$, residual within-pedigree correlations can be accounted for. The matrix Φ_i represents the genetic similarity within a pedigree based on pedigree relationships. This is accomplished by the diagonal elements of Φ_i having elements $\Phi_{i,jj} = 1 + h_{j,j}$, where $h_{j,j}$ is the inbreeding coefficient for subject j , and the off-diagonal elements $\Phi_{i,jk} = 2\varphi_{i,jk}$, where $\varphi_{i,jk}$ is the kinship coefficient between individuals j and k . Recall that the kinship coefficient is the probability that a randomly chosen allele at a given locus from individual j is identical by descent to a randomly chosen allele from individual k , conditional on their ancestral relationship. For autosomes, the genetic correlation between a pair of subjects is twice the kinship coefficient. For known pedigree relationships, the kinship coefficients can be efficiently computed by a recursive algorithm [Lange 2002], as implemented in the R package *kinship2* [Sinnwell, et al. 2014]. With this additional variance component, the null variance matrix for all subjects for a quantitative trait is $V_o = \sigma_b^2 \text{diag}\{\Phi_i\} + \sigma_\varepsilon^2 I$, where $\text{diag}\{\Phi_i\}$ is a block-diagonal matrix, implying that residuals are assumed to be uncorrelated between subjects not in the same pedigree.

The score statistic to test the association of a quantitative trait with a set of genetic markers is similar to that in expression (1) for unrelated subjects, but extended to account for within-pedigree correlations. To test whether the variance component for the kernel matrix is zero, $H_o: \sigma_K^2 = 0$, one can take the derivative of the log-likelihood for a multivariate mixed model to derive the score statistic

$$Q = \frac{1}{2}(y - \hat{y})' V_o^{-1} K V_o^{-1} (y - \hat{y}). \quad (3)$$

Note that the null variance matrix requires estimation of both σ_b^2 and σ_ε^2 , and these are also used to estimate the maximum likelihood estimate $\hat{\alpha} = (X' V_o^{-1} X)^{-1} X' V_o^{-1} y$. Furthermore, if $\sigma_b^2 = 0$, then $V_o^{-1} = (1/\sigma_\varepsilon^2) I$, and the score statistic reduces to $Q = \frac{1}{2}(y - \hat{y})' K (y - \hat{y}) / \sigma_\varepsilon^4$, which is just a rescaled version of expression (1) for unrelated subjects. To compute p-values, Schifano et al. [Schifano, et al. 2012] suggested either moment-matching or Davies' method. Chen et al. [Chen, et al. 2013] derived a similar score statistic, and suggested use of Kuonen's method to compute p-values. Although both Schifano [Schifano, et al. 2012] and Chen [Chen, et al. 2013] showed that their methods controlled the Type-I error rates for large sample sizes, their methods have been improved for small sample sizes by scaling the score statistic [Zhan, et al. 2018], similar to the approach in expression (2). Details are described in the Appendix. A limitation of the derivation of both Schifano [Schifano, et al.

2012] and Chen [Chen, et al. 2013] is that they did not extend their methods to the X chromosome, which would have a different matrix of kinship coefficients compared to that for autosomes.

A potential source of bias in association studies is population stratification. A way to overcome this source of bias is to focus the test of association on only within-family associations to avoid the bias that can arise from between-family associations. Although these types of transmission disequilibrium tests (TDT) were developed years ago for single genetic markers, only recently were they extended to kernel methods that test the association of a set of genetic markers with a single trait using trios of child-parents or nuclear families [Jiang, et al. 2014]. Further extensions to test the association of multiple traits with multiple genetic variants for child-parent trios have recently been developed [Fischer, et al. 2018]. This approach offers promise for testing rare variants in this type of data.

Another generalization for pedigree data was developed by Yan et al. [Yan, et al. 2015a] to analyze multivariate quantitative traits. Their approach is similar to the multivariate linear mixed model, but includes an extra random effect for each pedigree. The linear model for the vector of the j^{th} trait for all members of the i^{th} model is $y_{ij} = \alpha_o + X'_{ij}\alpha + f(G_{ij}) + b_{ij} + \varepsilon_{ij}$, similar to that for a single trait for pedigree data. By stacking all the traits into a single vector across all pedigrees, the model can be expressed as $y = \alpha_o + X\alpha + f(G) + b + \varepsilon$, where X is block-diagonal with blocks for the different traits, and the distributions of the random effects are structured as

$$\begin{aligned} f(G) &\sim N(0, \Sigma_K \otimes K), \\ b &\sim N(0, \Sigma_b \otimes \Phi), \\ \varepsilon &\sim N(0, \Sigma_\varepsilon \otimes I_n) \end{aligned}$$

Each of the variance components matrices (e.g., Σ) is a $P \times P$ matrix. Under the null hypothesis that none of the traits is associated with the set of genetic markers that are used to create kernel matrix K , the score statistic has the same form as expression (3). The main distinction is the estimated null variance matrix, $V_o = \hat{\Sigma}_b \otimes \Phi + \hat{\Sigma}_\varepsilon \otimes I$. The pedigree variance components matrix, $\hat{\Sigma}_b$, accounts for correlations induced by the pedigrees, both between traits and between subjects in the same pedigree. A computational challenge for a large number of traits is estimating the large number of variance components. In addition, further extensions for longitudinal family data are conceptually straight-forward, adding additional random effects to account for correlations for longitudinal data [Yan, et al. 2015b], with the caveat of additional computational challenges.

Retrospective Likelihood for Binary Traits

The above approaches to analyze quantitative traits in pedigrees assume the traits have a multivariate normal distribution such that the score statistic for the random effect can be used to test associations. When traits are binary, and particularly when pedigrees are enriched to have multiple affected members, a multivariate normal distribution is unrealistic.

For example, in extreme over-sampling of many affected per pedigree, the likelihood methods of Schifano and Chen [Chen, et al. 2013; Schifano, et al. 2012] would over-estimate the null variance component for shared pedigree relationships, σ_b^2 , which would in turn cause the test to lose power. To avoid these complications, as well as the need to correct for how pedigrees were ascertained, Schaid et al. [Schaid, et al. 2013] derived kernel and burden statistics for a retrospective likelihood that treated the genotypes as random and the traits as fixed. By this approach, they were able to account for complex correlations among genetic markers: correlations arising from linkage disequilibrium among the set of markers, and correlations arising from pedigree relationships. A key step in this approach was recognizing that the covariance between genetic markers k and l , for subjects i and j in the same pedigree, is $Cov(g_{ik}, g_{jl}) = 2R_{kl}\sqrt{p_k(1-p_k)p_l(1-p_l)}\Phi_{ij}$. The term R_{kl} accounts for the correlation between markers k and l , with minor allele frequencies p_k and p_l , and the kinship coefficient Φ_{ij} accounts for their pedigree relationship. Some advantages of their approach are that case-control studies with related subjects can be analyzed, and they allowed for analysis of markers on the X chromosome, by appropriate calculation of the kinship coefficients. Another feature of their retrospective likelihood is that it can be used for arbitrary traits, because the retrospective approach conditions on the trait. Although this was not discussed in Schaid et al. [Schaid, et al. 2013], later simulations found that the power of the retrospective approach for quantitative traits was very close to the power of the methods by Schifano et al. [Schifano, et al. 2012] and Chen et al. [Chen, et al. 2013] when the trait heritability was not large, less than 25% (unpublished data). Software for these methods is available as the R package *pedgene* in the CRAN public repository.

An approach similar to the methods by Schaid et al. [Schaid, et al. 2013] was developed by Choi et al. [Choi, et al. 2014], a method they call family-based rare variant associate test (FARVAT). Simulations have shown that *pedgene* and FARVAT have similar power [Wang, et al. 2016a], and more power than methods based on generalized estimating equations [Wang, et al. 2016b]. Because FARVAT was written C++, it was computationally faster than *pedgene* written in R.

To analyze censored survival traits for pedigree data, Chien et al. [Chien, et al. 2017] also used a retrospective likelihood by treating the censored traits as fixed and the genotypes as random. The advantage of this strategy is that it avoids having to estimate a variance component for the polygenic background, similar to that required for the methods of Schifano et al. [Schifano, et al. 2012] and Chen et al. [Chen, et al. 2013] for quantitative traits. However, the methods are only well-developed for weighted linear kernels.

KERNEL DESIGNS AND MULTIPLE KERNELS

The validity of the score test (i.e., controlling the type I error to a specified level) only depends on the correct specification of the null model. Any choice of kernel can lead to a valid test. However, the power of the test depends strongly on the choice of kernel. In the Appendix, we show that the distribution of the kernel statistic under a specified alternative hypothesis is a linear combination of independent χ_1^2 random variables with weights that

depend in the eigenvalues of a matrix. This matrix depends on the kernel matrix K and the ratio $\sigma_K^2/\sigma_\epsilon^2$. This ratio can be expressed as $h^2/(1-h^2)$, where h^2 represents the percent of the trait variation explained by the kernel matrix – the “kernel heritability”. An ideal kernel would maximize the variance component σ_K^2 , so that the kernel heritability is maximized.

Although generic kernels such as linear and quadratic kernels could be used, a kernel that captures the prior biological knowledge about the underlying signal structure is expected to be more powerful. To embed domain knowledge into the kernel, two approaches are possible: a variance-component approach and a distance-based approach.

Variance Component Kernels

The connection between kernel machines regression models and mixed-effects models allows us to define the kernel by imposing a covariance structure on the random effects part. By the variance components approach, concepts of similarity are used, such as similarity of genetic marker profiles leading to similarity of traits. To illustrate by example, consider a regression model ignoring intercept and covariates, to focus on the matrix of genetic markers, G . The model is $y = G\beta + \epsilon$, and $\beta \sim MVN(0, \sigma_\beta^2 \Sigma_\beta)$, where Σ_β is assumed to have a known structure, and σ_β^2 is the parameter to be tested. Then, the variance matrix of y is $\sigma_\beta^2 G \Sigma_\beta G' + \sigma_\epsilon^2 I$. Therefore, we can define

$$K = G \Sigma_\beta G',$$

where the assumed covariance structure Σ_β can be used to capture the plausible signal structure (variance and correlation). The linear weighted kernel can be seen as a special case of such a design [Wu, et al. 2011]. It assumes that Σ_β is a diagonal matrix with the diagonal elements being the weights assigned to each genetic marker. These reflect the anticipated effect size of the variants. If we let β have the exchangeable correlation structure $R^\rho = (1-\rho)I + \rho 11'$, which assumes that the effects of the causal variants tend to be in the same direction, the resulting weighted kernel matrix is

$$K_\beta^\rho = GWR^\rho WG'.$$

This is the composite kernel used in the optimal sequence kernel association test (SKAT-O) [Lee, et al. 2012a; Lee, et al. 2012b]. The parameter ρ provides the flexibility to capture signals with varying degrees of correlation. When $\rho = 1$, the test reduces to a burden test, which assumes all genetic markers have the same effect on a trait. When $\rho = 0$, the test reduces to the linear weighted kernel [Wu, et al. 2011]. A challenge is computing reliable p-values for SKAT-O when evaluating a range of values for the parameter ρ . It has been shown that the SKAT-O p-values are less reliable than the SKAT p-values, and the performance of current SKAT-O implementations depends on the set size of genetic markers and their correlation [Wu, et al. 2016].

Similar to the ideas of SKAT-O, Sun et al. [Sun, et al. 2016] developed structured kernels for multivariate traits, attempting to improve power by reducing the number of parameters. For P traits and M genetic markers, the regression coefficients can be represented as $P \times M$ matrix, $\beta \sim MVN(0, \sigma_\beta^2 \Sigma_\beta)$, where now we have $\Sigma_\beta = R \otimes W$. The $P \times P$ matrix R represents the assumed correlation structure among the rows of β and the $M \times M$ matrix W represents the assumed correlation structure among the columns of β . Sun et al. [Sun, et al. 2016] assumed a common correlation for the effects of a given genetic marker on different traits and that the effects of different genetic markers are uncorrelated. These assumptions led to $R = \rho 1_P 1_P' + (1 - \rho)I_P$, and W is a diagonal matrix for the weights. By these assumptions, they were able to reduce the assumed kernel to have only one parameter, ρ .

Although the linear (weighted) kernels are still most widely applied in genetic association studies, owing to its biological interpretability and computational efficiency, there have been recent attempts to design kernels that adapt to more complex signal structures by taking into account prior biological knowledge. For example, new variance component kernels have been proposed to accommodate functional annotations of noncoding variations in the mixed effects model framework [Hao, et al. 2018; He, et al. 2017]. Using these new kernels, more genetic associations have been recovered. Moving from genetics applications to -omics applications, more structural relationships are available among the -omics features, and they can be readily incorporated into kernels. For example, a phylogeny-based kernel for microbiome data was designed to capture the evolutionary relationship among bacteria species [Xiao and Chen 2017].

Distance-based Kernels

In contrast to using concepts of similarity, dissimilarity might be more intuitive in some scientific domains. Concepts of distance measures can quantify the dissimilarities between samples while reducing multiple measures of trait features to a single distance value. These distance measures take into account the data characteristics as well as domain-specific knowledge and are instrumental in revealing the biological signals through distance-based multivariate methods [Zapala and Schork 2012]. For example, the UniFrac distance has been widely used to quantify the distance between two metagenomic samples, accounting for the phylogenetic relationship among taxa [Chen, et al. 2012; Lozupone and Knight 2005]. The UniFrac distance embeds the idea that the difference in distantly related bacteria has more weight than that in closely related bacteria when measuring an overall distance. We can thus take advantage of biological knowledge to create kernels based on distance measures. Both multidimensional scaling (MDS) [Tzeng, et al. 2008] and radial basis function (RBF) [Cai, et al. 2011] approaches can be used to achieve this task. In the MDS-based approach, K_D is created by double centering the squared distance matrix D ,

$$K_D = -\frac{1}{2}(I - 11')D^2(I - 11'),$$

where $\mathbf{1}$ is a vector of 1's. K_D is equivalent to the linear kernel defined by the sample coordinates after embedding the samples into a Euclidean space through MDS. In the RBF-based approach, the kernel is formed by

$$K_D^\rho = \exp\left(-\frac{D^2}{2\rho}\right),$$

where the bandwidth parameter ρ defines how far the influence of a sample reaches. It has been shown that if the distance is metric, both types of kernels are positive definite [Micchelli 1984]. A metric distance satisfies the following properties: 1) it is greater or equal to 0; 2) a distance of 0 implies that two items are equivalent; 3) it is symmetric; 4) it follows the triangle inequality. By this we mean that if $d(x, z)$ is a distance from x to z , then $d(x, z) \leq d(x, y) + d(y, z)$. However, generic distance measures are not necessarily positive definite. In such situations, positive-definiteness correction may be needed to improve the behavior of the kernel-based score test. The distance-based kernel design has been widely used for microbiome association studies such as MiRKAT [Chen and Li 2013; Zhao, et al. 2015], MMiRKAT [Zhan, et al. 2017a] and OMiAT [Koh, et al. 2017].

Combining Multiple Kernels

In many applications, a kernel such as K_D^ρ and K_β^ρ may have a parameter ρ , which allows modeling a wide range of potential signal structures. If the value of ρ coincides with the real signal structure in the data, the power of the score test is maximized. In practice, we may not know *a priori* the best value of ρ for a particular dataset. An omnibus test, which searches for the best value of ρ over a grid in a data-adaptive manner, could potentially improve the robustness and power of the test [Lee, et al. 2012a; Lee, et al. 2012b; Zhao, et al. 2015; Zhao, et al. 2017]. The best value of ρ can be determined by minimizing the p-value of the score test instead of maximizing the score statistic, since the scale of the score statistic may depend on the parameter ρ . Appropriate statistical procedures are needed to control the type I error rate, to account for searching for a value of ρ that minimizes the p-value. To achieve this, the minimum of p-values can be used as a new test statistic and the distribution of the test statistic under the null can be derived either by an analytical approach, if the kernel has a special form such as that in SKAT-O [Lee, et al. 2012a; Lee, et al. 2012b], or by a permutation [Zhao, et al. 2015] or perturbation-based [Wu, et al. 2013] approach. In the permutation-based approach, the residuals estimated from the null model are permuted, added to the fitted outcome vector to generate a new outcome, and the null model is then refitted to calculate the score statistic. The perturbation approach takes advantage of the fact that the standardized residuals $(y - \hat{y})/\hat{\sigma}$ are asymptotically distributed as standard normal when the null hypothesis is true. Therefore, a standard normal vector can be generated to calculate the score statistic. The perturbation-based approach is computationally more efficient but requires a large sample size. If sample sizes are more modest, the residual permutation approach is preferred to obtain the empirical null distribution of the score statistic [Zhao, et al. 2015].

In other applications, we may have several distinct (and possibly correlated) candidate kernels, each of which is expected to be the most powerful to detect a specific effect [Zhao, et al. 2015; Zhao, et al. 2017]. In genetic association studies, we may have two distinct kernels defined based on rare and common variants [Ionita-Laza, et al. 2013], and wish to test the combined effect of rare and common variants. In microbiome association studies, various distance-based kernels for microbiome data can be defined based on traditional ecological distances [Zhao, et al. 2015]. Each distance-based kernel is powered to detect a specific type of microbiome change. Therefore, it is beneficial to consider multiple kernels to improve the power in these association studies. Several approaches for combining multiple kernels have been proposed. In the kernel averaging approach, a composite kernel is created based on the sum or weighted average of the candidate kernels and treated as a new kernel in the kernel machine test [Wu, et al. 2013]. Ideally, the weights (fixed *a priori*) should reflect the power of respective kernels based on our prior knowledge. If no prior knowledge is available, a simple scaling that ensures the kernels are on the same scale can be used. In the case of a weighted average of two kernels, the best weight can be searched on a grid and an analytic p-value can be calculated after orthogonalization of the two kernels [Ionita-Laza, et al. 2013; Zhao, et al. 2017]. However, this approach is difficult to extend to more than two kernels. An alternative to the kernel averaging is the minP approach, where we compute a p-value for each candidate kernel, take the minimum, and then evaluate significance via the residual permutation or perturbation-based approach.

KERNEL WEIGHTS AND FEATURE SELECTION

An important consideration for kernel functions that incorporate feature weights, such as the linear-weighted kernel, is how one should define the weights. For rare variants, inverse relationships between the minor allele frequency (MAF) and effect size are often assumed, such as Madsen-Browning weights [Madsen and Browning 2009], or weights defined by the beta density function as implemented in SKAT [Wu, et al. 2011]. Weights can also be informed by putative functional impact from *in silico* prediction tools that provide scores.

A limitation of the aggregative variant testing implemented by kernel methods is that there is no explicit indication of which variants may be “driving” a positive association signal. He et al. [He, et al. 2016] proposed the use of kernel iterative feature extraction (KNIFE) as a two-step procedure to integrate feature selection into the association analysis. Specifically, a separate set of non-negative feature weights $c = (c_1, \dots, c_M)'$ are defined and integrated into existing kernel functions. For example, the weighted linear kernel is adapted to be $K = GW^{1/2}CW^{1/2}G'$, where $C = \text{diag}(c)$. Feature selection on G is then performed in an iterative fashion by applying a penalized model to c (e.g., non-negative garrote). In a different approach, Ionita-Laza et al. [Ionita-Laza, et al. 2014] proposed a backwards-elimination strategy for SKAT, such that variants in G are iteratively removed until the score statistic p-value from the reduced variant set can no longer be reduced. The authors additionally proposed a resampling strategy to quantify the relative importance of each variant.

CONNECTIONS BETWEEN KERNEL METHODS AND OTHER APPROACHES

Kernel machines regression (KMR) methods, which basically relate the outcome to the covariates in terms of pairwise similarities, is directly related to distance-based multivariate regression (DMR) methods (equivalent to PERMANOVA [McArdle and Anderson 2001] and GDBR [Wessel and Schork 2006] under different contexts) [Pan 2011]. However, the roles of outcome and covariate are switched between KMR and DMR. While the p-value for the KMR method is usually calculated analytically, the p-value for the DMR method is mainly calculated by permutation methods. Compared to the traditional multivariate regression model, the multivariate outcome in DMR is implicitly specified in the form of a distance matrix, and the multivariate F-statistic is expressed directly based on the distance matrix D as

$$F = \frac{\text{tr}(HGH)}{\text{tr}[(I-H)G(I-H)]},$$

where $\text{tr}(\cdot)$ is the trace of a matrix, H is the “hat” matrix ($H = X(X'X)^{-1}X'$) that accounts for covariate effects, and $G = -\frac{1}{2}(I - 11')D^2(I - 11')$. Clearly, G is exactly the same as the distance-based kernel K_D . Pan [Pan 2011] showed that, under the condition that there are no other covariate and $K_D(G)$ is used in KMR, the F-test statistic in DMR and the score test statistic in KMR are equal, up to some constants, for both continuous and binary outcomes. The power of DMR and KMR is almost equivalent under such conditions [Pan 2011]. However, KMR is computationally more efficient and easy to adjust for other covariates. The KMR score test is also closely related to SSU test [Pan 2011; Pan, et al. 2014; Pan, et al. 2015]. The SSU statistic

$$T_{SSU} = \sum_{j=1}^M U_j^2, \quad U_j = \sum_{i=1}^N G_{ij}(y_i - \hat{y}_i)$$

is the sum of the squared score statistics over all genetic variants. Thus, T_{SSU} is equivalent to the KMR score statistic Q up to a scaling constant, when the linear kernel is used. Finally, the KMR test was also shown to be related to the kernel-based adaptive cluster test (KBAC), which computes a weighted integral representing the difference in risk between variants [Zhang, et al. 2017]. For multivariate traits, the score test based on multivariate kernel machine regression (MVKM) is equivalent to the generalized estimating equation-based sum of squared score test (GEE-SPU(2)), when the linear kernel is used in MVKM and the trait correlation matrix is used as the working correlation matrix in GEE-SPU(2). Moreover, GEE-SPU(2) is equivalent to the DMR if the Euclidean distance is used. More details can be found in [Kim, et al. 2016; Zhang, et al. 2014].

CONCLUSIONS

Kernel methods represent a family of highly flexible approaches for performing aggregative variant association analysis, and recent advances have extended these methods to a diverse

array of outcome types and study designs. Additionally, kernel methods naturally fit within the familiar mixed models regression framework, allowing for covariate adjustment and accommodating complex correlated data structures. Because kernel statistics are derived as score statistics under the null hypothesis of no associations, they are rapid to compute. However, there are some notable limitations that warrant further research to enhance the utility of kernel methods. Identifying which variants may be driving identified associations is complicated by the aggregative process, and the idea of replicating positive findings using kernel methods is not straight-forward. Improved ways to choose a kernel that provides optimal power for a given situation would avoid arbitrary choices to create kernels. Kernel methods can also be very computationally demanding under large sample sizes, particularly when direct variance component estimation is part of the analysis procedure. Additional opportunities for future research in kernel tests for genetic association studies include strategies for leveraging the growing amount of publicly available functional annotation, integration of multiple different datatypes (e.g., multi-omics), and further improvements to Bayesian applications.

ACKNOWLEDGEMENTS

This research was supported by the U.S. Public Health Service, National Institutes of Health, contract grants numbers GM065450 (DJS as principal investigator), and Mayo Clinic Center for Individualized Medicine. The authors have no conflicts of interest to declare.

APPENDIX

Scaled Quadratic Statistic: Ratio of Two Quadratic Forms

Assume that under the null hypothesis $H_o: \sigma_K^2 = 0, y \sim N(X\alpha, V_o)$. For example, for pedigree data $V_o = \sigma_b^2 \text{diag}\{\Phi_i\} + \sigma_e^2 I$. By the first derivative of the log-likelihood of a mixed model, the REML score statistic can be shown to be $Q = (y - \hat{y})' V_o^{-1} K V_o^{-1} (y - \hat{y})/2$ [Harville 1977]. To account for estimation of V_o , we follow the approach of Zhan et al. [Zhan, et al. 2018] by scaling Q by an estimate of the null residual variances, thereby creating a ratio of two quadratic forms of normal random variables. The resulting statistic is

$$Q = \frac{(y - X\hat{\alpha})' V_o^{-1} K V_o^{-1} (y - X\hat{\alpha})}{(y - X\hat{\alpha})' V_o^{-1} (y - X\hat{\alpha})}$$

To compute p-values, we need to express the statistic as $\epsilon^{*'} A \epsilon^*$, where $\epsilon^* \sim N(0, I)$, and the eigenvalues of matrix A are used by either Davies' or Kuonen's method. To see that the distribution of Q can be expressed in this manner, it is helpful to know that $\hat{\alpha} = (X' V_o^{-1} X)^{-1} X' V_o^{-1} y$, so that $(y - X\hat{\alpha}) = \hat{\epsilon} = (I - P_o)y$, where $P_o = X(X' V_o^{-1} X)^{-1} X' V_o^{-1}$. This implies that $\hat{\epsilon} \sim N(0, V_{\hat{\epsilon}})$, where $V_{\hat{\epsilon}} = V_o - X(X' V_o^{-1} X)^{-1} X'$. By de-correlating the residuals, $\epsilon^* = V_{\hat{\epsilon}}^{-1/2} \hat{\epsilon} \sim N(0, I)$. This lets us represent the residuals as $\hat{\epsilon} = V_{\hat{\epsilon}}^{1/2} \epsilon^*$. Substituting this expression of the residuals into Q results in

$$Q = \frac{\varepsilon^{*'} V_{\hat{\varepsilon}}^{-1/2} V_o^{-1} K V_o^{-1} V_{\hat{\varepsilon}}^{-1/2} \varepsilon^*}{\varepsilon^{*'} V_{\hat{\varepsilon}}^{-1/2} V_o^{-1} V_{\hat{\varepsilon}}^{-1/2} \varepsilon^*}$$

This illustrates that the statistic can be expressed as the ratio of two quadratic forms of standard normal random variables. To compute p-values, we need the probability $P(Q > Q_{obs})$, where Q_{obs} is the observed value of the statistic. This can be expressed as

$$P(Q > Q_{obs}) = P\left(\frac{\varepsilon^{*'} V_{\hat{\varepsilon}}^{-1/2} V_o^{-1} K V_o^{-1} V_{\hat{\varepsilon}}^{-1/2} \varepsilon^*}{\varepsilon^{*'} V_{\hat{\varepsilon}}^{-1/2} V_o^{-1} V_{\hat{\varepsilon}}^{-1/2} \varepsilon^*} > Q_{obs}\right) = P(\varepsilon^{*'} A \varepsilon^* > 0)$$

where $A = V_{\hat{\varepsilon}}^{-1/2} V_o^{-1} K V_o^{-1} V_{\hat{\varepsilon}}^{-1/2} - Q_{obs} V_{\hat{\varepsilon}}^{-1/2} V_o^{-1} V_{\hat{\varepsilon}}^{-1/2}$. Hence, the eigenvalues of matrix A are needed to compute exact p-values by either Davies' or Kuonen's method.

Distribution of Quadratic Statistic Under the Alternative Hypothesis

Assume that $y \sim N(X\alpha, V_A)$, where the variance of y under an alternative hypothesis $H_A: \sigma_K^2 \neq 0$ is $V_A = \sigma_\varepsilon^2 I + \sigma_K^2 K$. The REML score statistic to test the null hypothesis $H_o: \sigma_K^2 = 0$ is $Q = (y - \hat{y})' V_o^{-1} K V_o^{-1} (y - \hat{y})/2$ [Harville 1977], where $V_o = \sigma_\varepsilon^2 I$. Using the null projection matrix, $P_o = X(X'V_o^{-1}X)^{-1}X'V_o^{-1}$, $(y - X\hat{\alpha}) = \hat{\varepsilon} = [I - P_o]y$, so the variance of $\hat{\varepsilon}$ under an alternative hypothesis is $V_{\hat{\varepsilon}} = (I - P_o)V_A(I - P_o)$. When $V_o = \sigma_\varepsilon^2 I$, $P_o = X(X'X)^{-1}X' = H$, the "hat" matrix, so $V_{\hat{\varepsilon}} = (I - H)V_A(I - H)$. By de-correlating, $V_A^{-1/2}\hat{\varepsilon} = \varepsilon^* \sim N(0, I)$, so we can express $\hat{\varepsilon} = V_A^{1/2}\varepsilon^*$. Substituting this expression into Q results in $Q = \varepsilon^{*'} [V_\varepsilon^{1/2} V_o^{-1} K V_o^{-1} V_\varepsilon^{1/2}/2] \varepsilon^*$. This implies that the distribution of Q is a linear combination of independent χ_1^2 random variables with weights that are the eigenvalues of the matrix $A = V_\varepsilon^{1/2} V_o^{-1} K V_o^{-1} V_\varepsilon^{1/2}/2$, which are the same as the eigenvalues of the matrix $A = V_o^{-1} K V_o^{-1} V_\varepsilon/2$. After substituting $V_o = \sigma_\varepsilon^2 I$ and $V_A = \sigma_\varepsilon^2 I + \sigma_K^2 K$ into A and simplifying, the matrix A can be expressed as

$$A = \frac{1}{2\sigma_\varepsilon^2} K \left[(I - H) + \frac{\sigma_K^2}{\sigma_\varepsilon^2} (I - H)K(I - H) \right].$$

This illustrates that the distribution of the statistic under an alternative hypothesis, the power of the statistic, depends on the ratio $\sigma_K^2/\sigma_\varepsilon^2$.

REFERENCES

- Almeida M, Peralta JM, Farook V, Puppala S, Kent JW, Jr., Duggirala R, Blangero J. 2014 Pedigree-based random effect tests to screen gene pathways. *BMC proceedings* 8(Suppl 1 Genetic Analysis Workshop 18Vanessa Olmo):S100. [PubMed: 25519354]
- Broadaway KA, Cutler DJ, Duncan R, Moore JL, Ware EB, Jhun MA, Bielak LF, Zhao W, Smith JA, Peyser PA and others. 2016 A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. *American journal of human genetics* 98(3):525–40. [PubMed: 26942286]
- Broadaway KA, Duncan R, Conneely KN, Almli LM, Bradley B, Ressler KJ, Epstein MP. 2015 Kernel Approach for Modeling Interaction Effects in Genetic Association Studies of Complex Quantitative Traits. *Genet Epidemiol* 39(5):366–75. [PubMed: 25885490]
- Cai T, Tonini G, LIN X. 2011 Kernel Machine Approach to Testing the Significance of Multiple Genetic Markers for Risk Prediction. *Biometrics* 67(3):975–986. [PubMed: 21281275]
- Chen H, Lumley T, Brody J, Heard-Costa NL, Fox CS, Cupples LA, Dupuis J. 2014 Sequence Kernel Association Test for Survival Traits. *Genetic Epidemiology* 38(3):191–197. [PubMed: 24464521]
- Chen H, Meigs JB, Dupuis J. 2013 Sequence Kernel Association Test for Quantitative Traits in Family Samples. *Genetic Epidemiology* 37(2):196–204. [PubMed: 23280576]
- Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. 2012 Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28(16):2106–13. [PubMed: 22711789]
- Chen J, Chen W, Zhao N, Wu MC, Schaid DJ. 2016 Small Sample Kernel Association Tests for Human Genetic and Microbiome Association Studies. *Genetic Epidemiology* 40(1):5–19. [PubMed: 26643881]
- Chen J, Li H. 2013 *Kernel Methods for Regression Analysis of Microbiome Compositional Data* New York, NY: Springer New York p 191–201.
- Chien LC, Bowden DW, Chiu YF. 2017 Region-based association tests for sequencing data on survival traits. *Genetic epidemiology* 41(6):511–522. [PubMed: 28580640]
- Choi S, Lee S, Cichon S, Nothen MM, Lange C, Park T, Won S. 2014 FARVAT: a family-based rare variant association test. *Bioinformatics* 30(22):3197–205. [PubMed: 25075118]
- Davenport CA, Maity A, Sullivan PF, Tzeng J-Y. 2017 A Powerful Test for SNP Effects on Multivariate Binary Outcomes Using Kernel Machine Regression. *Statistics in Biosciences* 55:1–22.
- Davies RB. 1980a Algorithm AS 155: The distribution of a linear combination of χ^2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29(3):323–333.
- Davies RB. 1980b The distribution of a linear combination of chi-square random variables. *Applied Statistics* 29(3):323–333.
- Fischer ST, Jiang Y, Broadaway KA, Conneely KN, Epstein MP. 2018 Powerful and robust cross-phenotype association test for case-parent trios. *Genetic epidemiology* 42(5):447–458. [PubMed: 29460449]
- Hao X, Zeng P, Zhang S, Zhou X. 2018 Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. *PLoS Genet* 14(1):e1007186. [PubMed: 29377896]
- Harville DA. 1977 Maximum likelihood approaches to variance component estimation and to related problems. *J Am Statistical Association* 72(358):320–340.
- He Q, Cai T, Liu Y, Zhao N, Harmon QE, Almli LM, Binder EB, Engel SM, Ressler KJ, Conneely KN and others. 2016 Prioritizing individual genetic variants after kernel machine testing using variable selection. *Genet Epidemiol* 40(8):722–731. [PubMed: 27488097]
- He Z, Xu B, Lee S, Ionita-Laza I. 2017 Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in MetaboChip Data. *Am J Hum Genet* 101(3):340–352. [PubMed: 28844485]
- Ionita-Laza I, Capanu M, De Rubeis S, McCallum K, Buxbaum JD. 2014 Identification of rare causal variants in sequence-based studies: methods and applications to VPS13B, a gene involved in Cohen syndrome and autism. *PLoS Genet* 10(12):e1004729. [PubMed: 25502226]

- Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, LIN X. 2013 Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. *The American Journal of Human Genetics* 92(6):841–853. [PubMed: 23684009]
- Jiang Y, Conneely KN, Epstein MP. 2014 Flexible and robust methods for rare-variant testing of quantitative traits in trios and nuclear families. *Genetic epidemiology* 38(6):542–51. [PubMed: 25044337]
- Kim J, Zhang Y, Pan W, Alzheimer's Disease Neuroimaging I. 2016 Powerful and Adaptive Testing for Multi-trait and Multi-SNP Associations with GWAS and Sequencing Data. *Genetics* 203(2): 715–31. [PubMed: 27075728]
- Kimeldorf G, Wahba G. 1971 Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications* 33(1):82–95.
- Koh H, Blaser MJ, Li H. 2017 A powerful microbiome-based association test and a microbial taxa discovery framework for comprehensive association mapping. *Microbiome* 5(1):45. [PubMed: 28438217]
- Kuonen D 1999 Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* 86(4):929–935.
- Kwee LC, Liu D, LIN X, Ghosh D, Epstein MP. 2008 A Powerful and Flexible Multilocus Association Test for Quantitative Traits. *The American Journal of Human Genetics* 82(2):386–397. [PubMed: 18252219]
- Lange K 2002 *Mathematical and statistical methods for genetic analysis* New York: Springer.
- Larson NB, Jenkins GD, Larson MC, Vierkant RA, Sellers TA, Phelan CM, Schildkraut JM, Sutphen R, Pharoah PP, Gayther SA and others. 2014 Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. *Eur J Hum Genet* 22(1):126–31. [PubMed: 23591404]
- Larson NB, Schaid DJ. 2013 A kernel regression approach to gene-gene interaction detection for case-control studies. *Genet Epidemiol* 37(7):695–703. [PubMed: 23868214]
- le Cessie S, van Houwelingen HC. 1995 Testing the fit of a regression model via score tests in random effects models. *Biometrics* 51(2):600–14. [PubMed: 7662848]
- Lee S, Abecasis GR, Boehnke M, LIN X. 2014 Rare-Variant Association Analysis: Study Designs and Statistical Tests. *The American Journal of Human Genetics* 95(1):5–23. [PubMed: 24995866]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, Christiani DC, Wurfel MM, LIN X. 2012a 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(2):224–237. [PubMed: 22863193]
- Lee S, Wu MC, LIN X. 2012b Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13(4):762–775. [PubMed: 22699862]
- Li H, Luan Y. 2003 Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*:65–76. [PubMed: 12603018]
- Li SY, Cui YH. 2012 Gene-Centric Gene-Gene Interaction: A Model-Based Kernel Machine Method. *Annals of Applied Statistics* 6(3):1134–1161.
- Lin X 1997 Variance component testing in generalised linear models with random effects. *Biometrika* 84:309–326.
- Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, LIN X. 2011 Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology* 35(7): 620–631. [PubMed: 21818772]
- Lin X, Lee S, Christiani DC, LIN X. 2013 Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 14(4):667–681. [PubMed: 23462021]
- Liu D, Lin X, Ghosh D. 2007 Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. *Biometrics* 63(4):1079–88. [PubMed: 18078480]
- Lozupone C, Knight R. 2005 UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12):8228–35. [PubMed: 16332807]

- Lumley T, Brody J, Peloso G, Morrison A, Rice K. 2018 FastSKAT: Sequence kernel association tests for very large sets of markers. *Genet Epidemiol*
- Madsen BE, Browning SR. 2009 A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5(2):e1000384. [PubMed: 19214210]
- Maity A, Sullivan PF, Tzeng Ji. 2012 Multivariate Phenotype Association Analysis by Marker-Set Kernel Machine Regression. *Genetic Epidemiology* 36(7):686–695. [PubMed: 22899176]
- McArdle BH, Anderson MJ. 2001 Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology* 82(1):290–297.
- Micchelli CA. 1984 Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Approximation theory and spline functions*: Springer p 143–145.
- Morgenthaler S, Thilly WG. 2007 A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res* 615(1–2):28–56. [PubMed: 17101154]
- Moutsianas L, Agarwala V, Fuchsberger C, Flannick J, Rivas MA, Gaulton KJ, Albers PK, Go TDC, McVean G, Boehnke M and others. 2015 The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet* 11(4):e1005165. [PubMed: 25906071]
- Ouakacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, Richards JB, Ciampi A, Greenwood CMT. 2013 Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology* 37(4):366–376. [PubMed: 23529756]
- Pan W 2011 Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic Epidemiology* 35(4):211–216. [PubMed: 21308765]
- Pan W, Kim J, Zhang Y, Shen X, Wei P. 2014 A Powerful and Adaptive Association Test for Rare Variants. *Genetics* 197(4):1081–1095. [PubMed: 24831820]
- Pan W, Kwak IY, Wei P. 2015 A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. *Am J Hum Genet* 97(1):86–98. [PubMed: 26119817]
- Plantinga A, Zhan X, Zhao N, Chen J, Jenq RR, Wu MC. 2017 MiRKAT-S: a community-level test of association between the microbiota and survival times. *Microbiome* 5(1):17. [PubMed: 28179014]
- Saad M, Wijsman EM. 2014 Combining family- and population-based imputation data for association analysis of rare and common variants in large pedigrees. *Genet Epidemiol* 38(7):579–90. [PubMed: 25132070]
- Schaid D 2010 Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. *Human Heredity* 70:109–131. [PubMed: 20610906]
- Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. 2013 Multiple Genetic Variant Association Testing by Collapsing and Kernel Methods With Pedigree or Population Structured Data. *Genetic Epidemiology* 37(5):409–418. [PubMed: 23650101]
- Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, LIN X. 2012 SNP Set Association Analysis for Familial Data. *Genetic Epidemiology* 36(8):797–810. [PubMed: 22968922]
- Schweiger R, Weissbrod O, Rahmani E, Müller-Nurasyid M, Kunze S, Gieger C, Waldenberger M, Rosset S, Halperin E. 2017 RL-SKAT: An Exact and Efficient Score Test for Heritability and Set Tests. *Genetics* 207(4):genetics.300395.2017–1283.
- Sinnwell JP, Therneau TM, Schaid DJ. 2014 The kinship2 R package for pedigree data. *Human heredity* 78(2):91–3. [PubMed: 25074474]
- Sun J, Ouakacha K, Forgetta V, Zheng HF, Brent Richards J, Ciampi A, Greenwood CM. 2016 A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects. *European journal of human genetics : EJHG* 24(9): 1344–51. [PubMed: 26860061]
- Tzeng J, Lu HH, Li WH. 2008 Multidimensional scaling for large genomic data sets. *BMC Bioinformatics* 9:179. [PubMed: 18394154]
- Wang K 2016 Boosting the Power of the Sequence Kernel Association Test by Properly Estimating Its Null Distribution. *American journal of human genetics* 99(1):104–114. [PubMed: 27292111]
- Wang L, Choi S, Lee S, Park T, Won S. 2016a Comparing family-based rare variant association tests for dichotomous phenotypes. *BMC proceedings* 10(Suppl 7):181–186. [PubMed: 27980633]

- Wang X, Zhang Z, Morris N, Cai T, Lee S, Wang C, Yu TW, Walsh CA, LIN X. 2016b Rare variant association test in family-based sequencing studies. - PubMed - NCBI. Briefings in bioinformatics 18(6):bbw083–961.
- Wen X 2015 Bayesian model comparison in genetic association analysis: linear mixed modeling and SNP set testing. Biostatistics 16(4):701–12. [PubMed: 25796429]
- Wessel J, Schork NJ. 2006 Generalized genomic distance-based regression methodology for multilocus association analysis. Am J Hum Genet 79(5):792–806. [PubMed: 17033957]
- Wu B, Guan W, Pankow JS. 2016 On Efficient and Accurate Calculation of Significance P-Values for Sequence Kernel Association Testing of Variant Set. Annals of Human Genetics 80(2):123–135. [PubMed: 26757198]
- Wu B, Pankow JS. 2016a On Sample Size and Power Calculation for Variant Set-Based Association Tests. Annals of Human Genetics 80(2):136–143. [PubMed: 26831402]
- Wu B, Pankow JS. 2016b Sequence Kernel Association Test of Multiple Continuous Phenotypes. Genetic Epidemiology 40(2):91–100. [PubMed: 26782911]
- Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, LIN X. 2010 Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. The American Journal of Human Genetics 86(6):929–942. [PubMed: 20560208]
- Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011 Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93. [PubMed: 21737059]
- Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, Engel SM, Molldrem JJ, Armistead PM. 2013 Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels. Genetic Epidemiology 37(3):267–275. [PubMed: 23471868]
- Xiao J, Chen J. 2017 Phylogeny-Based Kernels with Application to Microbiome Association Studies. New Advances in Statistics and Data Science: Springer p 217–237.
- Yan Q, Fang Z, Chen W. 2018 KMgene: a unified R package for gene-based association analysis for complex traits. Bioinformatics 34(12):2144–2146. [PubMed: 29438558]
- Yan Q, Weeks DE, Celedon JC, Tiwari HK, Li B, Wang X, Lin WY, Lou XY, Gao G, Chen W and others. 2015a Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples with a Novel Kernel Machine Regression Method. Genetics 201(4):1329–39. [PubMed: 26482791]
- Yan Q, Weeks DE, Tiwari HK, YI N, Zhang K, Gao G, Lin W-Y, Lou X-Y, Chen W, Liu N. 2015b Rare-Variant Kernel Machine Test for Longitudinal Data from Population and Family Samples. Human Heredity 80(3):126–138. [PubMed: 27161037]
- Yuan Z, Gao Q, He Y, Zhang X, Li F, Zhao J, Xue F. 2012 Detection for gene-gene co-association via kernel canonical correlation analysis. BMC Genet 13:83. [PubMed: 23039928]
- Zapala MA, Schork NJ. 2012 Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. Front Genet 3:190. [PubMed: 23060897]
- Zeng P, Zhao Y, Zhang L, Huang S, Chen F. 2014 Rare variants detection with kernel machine learning based on likelihood ratio test. PLoS One 9(3):e93355. [PubMed: 24675868]
- Zhan X, Tong X, Zhao N, Maity A, Wu MC, Chen J. 2017a A small-sample multivariate kernel machine test for microbiome association studies. Genetic Epidemiology 41(3):210–220. [PubMed: 28019040]
- Zhan X, Xue L, Zheng H, Pantinga A, Wu MC, Schaid DJ, Zhao N, Chen J. 2018 A small-sample kernel machine test for correlated data with application to microbiome association studies. Genetic Epidemiology(In Press.)
- Zhan X, Zhao N, Plantinga A, Thornton TA, Conneely KN, Epstein MP, Wu MC. 2017b Powerful Genetic Association Analysis for Common or Rare Variants with High-Dimensional Structured Traits. Genetics 206(4):1779–1790. [PubMed: 28642271]
- Zhang W, Epstein MP, Fingerlin TE, Ghosh D. 2017 Links Between the Sequence Kernel Association and the Kernel-Based Adaptive Cluster Tests. Statistics in Biosciences 9(1):246–258.
- Zhang Y, Xu Z, Shen X, Pan W, Alzheimer's Disease Neuroimaging I. 2014 Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. Neuroimage 96:309–25. [PubMed: 24704269]

- Zhao N, Chen J, Carroll IM, Ringel-Kulka T, Epstein MP, Zhou H, Zhou JJ, Ringel Y, Li H, Wu MC. 2015 Testing in Microbiome-Profilng Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. *American journal of human genetics* 96(5):797–807. [PubMed: 25957468]
- Zhao N, Zhan X, Huang Y-T, Almli LM, Smith A, Epstein MP, Conneely K, Wu MC. 2017 Kernel machine methods for integrative analysis of genome-wide methylation and genotyping studies. *Genetic Epidemiology* 22(4):623.

Table 1.

Software tools based on kernel statistic useful for genetic analyses.

Tools	Description	Source	Reference(s)
SKAT	SNP-Set (Sequence) Kernel Association Test including Burden test, SKAT and SKAT-O.	CRAN*	[Lee, et al. 2012a; Wu, et al. 2011]
fastSKAT	Computationally efficient calculation of SKAT p-values for large data	https://github.com/tslumley/bigQF	[Lumley, et al. 2018]
SSKAT	A suite of tools implementing small-sample adjusted kernel machine association tests for univariate, multivariate, and correlated outcomes.	https://github.com/jchen1981/SSKAT	[Chen, et al. 2016; Zhan, et al. 2017a; Zhan, et al. 2018]
RL-SKAT	Recalibrated Lightweight Sequence Kernel Association Test with small-sample adjustment. It is optimized for large-scale genetic data.	https://github.com/cozygene/RL-SKAT	[Schweiger, et al. 2017]
SPA3G	Gene-gene interaction analysis for continuous phenotypes using kernels.	CRAN*	[Li and Cui 2012]
iSKAT	Implementation of GESAT for GxE interaction kernel testing.	CRAN*	[Lin, et al. 2013]
iGasso	The package implements SKAT+ for enhanced power over SKAT by properly estimating the null distribution of SKAT.	CRAN*	[Wang 2016]
pedgene	Compute kernel and burden statistics for pedigree data based on retrospective likelihood.	CRAN*	[Schaid, et al. 2013]
FARVAT	Family-based Rare Variant Association Test for gene-based association test with related samples. Implemented in C++ for computational efficiency.	http://healthstat.snu.ac.kr/software/farvat/	[Choi, et al. 2014]
famSKATRC	Family-based association kernel test for both rare and common variants.	CRAN*	[Saad and Wijsman 2014]
MSKAT	A suite of tools implementing various statistical methods for testing multi-trait variant-set association.	https://github.com/tslumley/bigQF ; https://github.com/baolinwu/MSKAT	[Wu and Pankow 2016b]
KMgene	Gene-based association test between a group of SNPs and traits including familial (both continuous and binary) traits, multivariate continuous (both independent and familial) traits, longitudinal continuous traits and survival traits.	CRAN*	[Yan, et al. 2018]
MIRKAT	Test for association between microbiome composition and a continuous, dichotomous, multivariate, survival, and structured outcome.	CRAN*	[Plantinga, et al. 2017; Zhan, et al. 2017a; Zhao, et al. 2015]

* These are available as R packages hosted on The Comprehensive R Archive Network (CRAN, <https://cran.r-project.org/>) under the provided tool name.