Title: Powerful Genetic Association Analysis for Common or Rare Variants with High Dimensional Structured Traits

Running Title: DKAT for Genetic Association Studies

Xiang Zhan[1], Ni Zhao[2], Anna Plantinga[3], Timothy A. Thornton[3], Karen N. Conneely[4], Michael P. Epstein[4], Michael C. Wu[1,*]

[1]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109

[2]Department of Biostatistics, The Johns Hopkins University, Baltimore, MD, 21202

[3]Department of Biostatistics, University of Washington, Seattle, WA, 98195

[4]Department of Human Genetics, Emory University, Atlanta, GA, 30322

Address for Correspondence:

Michael C. Wu

Public Health Sciences Division

Fred Hutchinson Cancer Research Center

1100 Fairview Avenue North, M3-C102

Seattle, WA 98109-1024

Phone: (206) 667-6603

Email: mcwu@fhcrc.org

Abstract

Many genetic association studies collect a wide range of complex traits. As these traits may be correlated and share a common genetic mechanism, joint analysis can be statistically more powerful and biologically more meaningful. However, most existing tests for multiple traits cannot be used for high-dimensional and possibly structured traits, such as network-structured transcriptomic pathway expressions. To overcome potential limitations, in this paper we propose the dual kernel-based association test (DKAT) for testing the association between multiple traits and multiple genetic variants, both common and rare. In DKAT, two individual kernels are used to describe the phenotypic and genotypic similarity, respectively, between pairwise subjects. Using kernels allows for capturing structure while accommodating dimensionality. Then, the association between traits and genetic variants is summarized by a coefficient which measures the association between two kernel matrices. Finally, DKAT evaluates the hypothesis of non-association with an analytical p-value calculation without any computationally expensive resampling procedures. By collapsing information in both traits and genetic variants using kernels, the proposed DKAT is shown to have correct type I error rate and higher power than other existing methods in both simulation studies and application to a study of genetic regulation of pathway gene expressions.

Key Words: Dual kernels; Genetic association analysis; High-dimensional traits; Network structure; Pleiotropy.

Introduction

Large scale genome wide association studies and next generation sequencing association studies have resulted in the identification of a wide range of genetic variants, common and rare, related to a host of complex traits and disorders.[1–3] Traditional genetic association analyses have focused on identifying associations between individual genetic variants or groups of genetic variants with a single trait of interest. However, this approach proves inadequate when a single variable does not fully capture the trait or phenotype of interest and further may result in power loss. In many situations, joint analysis of multiple traits, simultaneously, may prove advantageous as compared to single trait analysis for a number of reasons. First, joint analysis tends to be statistically more powerful than the individual trait analysis:[4–6] joint analysis can reduce the multiple testing correction burden associated with individually testing multiple traits and, more importantly, can exploit the correlation structure by borrowing information across multiple, related traits. Second, joint analysis facilitates the elucidation of shared genetic mechanisms and pleiotropic relationships, thus serving as an appropriate means for improving biological understanding.[7–14] Finally, many traits are inherently multi-phenotypic. For example, metabolic syndrome, which increases risk for heart disease, diabetes, and stroke, is defined based on the presence of three out of five conditions;[15,16] information may be gained by using all five conditions as trait measures rather than considering only formal diagnosis of metabolic syndrome.

A wide range of statistical and computational methods have been developed for analyzing multiple phenotypes. Broadly speaking, these methods fall into

three main categories. The first category is based on directly integrating univariate results from analyzing each trait separately. However, these methods can handle at most a moderate number of traits, e.g., less than 20 traits.[17,18] Furthermore, such methods generally do not directly harness correlation and relationships among the traits. The second category of methods are based on applying classical dimension reduction methods, e.g. principal component analysis[5] and canonical correlation analysis,[19] in order to collapse multiple traits into a single score. However, results based on dimension reduction methods are difficult to interpret and lose power when the weights for collapsing the multiple traits are imperfect.[6] The final category is the broadest and is based multivariate regression methods, which often assume a model for the relationships between multiple traits and a single SNP.[20–25] The specific modeling strategies underlying each approach varies with some approaches using strategies such as classical mixed models and others using alternative strategies, e.g., MultiPhen[20] which uses ordinal regression to regress a SNP on multiple traits. These methods often suffer when underlying parametric assumptions are violated.[26] Many of these methods have been extended to allow for accommodation of multiple variants as well as multiple traits,[27–30] with the understanding the multi-variant analysis can oftentimes improve power for the same reasons that multi-trait analysis can improve power.

There are considerable and increasing interests in high-dimensional structured phenotypes, such as imaging traits or other omics data, as they are often inherently interesting and also can serve as intermediate traits which help in elucidating underlying molecular mechanisms while being more directly related to eti-

ology. However, despite interest, phenotypes such as imaging outcomes,[31] and other sources of -omics data such as gene expression, metabolomics intensity[32] and microbiome composition,[33] continue to pose grand challenges. Beyond the intrinsic high-dimensionality and scale of the data, such phenotypes are often statistically complex in that they have underlying structure that needs to be accommodated. Examples of structures include network/pathway relationships in metabolomic data and gene expression data, and phylogenetic relationships in microbiome data. Most existing multivariate-trait association methods do not generally accommodate high-dimensional structured phenotypes. Methods based on univariate analysis and collapsing rapidly lose power as dimensionality increases,[17,18] since they typically suffer from power loss due to heavy multiple testing burden, which comes with the high-dimensional traits. Dimension reduction-based association analysis usually considers surrogate outcomes (e.g., principal components), which breaks down the inherent structures in the original phenotypes. More complicated multivariate regression modeling strategies often become unstable or computationally intractable when dimensionality of traits increases.[27,29] None of the methods directly consider the issue of incorporating high-dimensional structured traits, which leads to potential power loss of detecting existing associations.[34] Thus, new methods are necessary.

A powerful approach in genetic association analysis is the kernel machine regression (KMR) framework, which has proven to be a useful tool for association studies with both common and rare variants.[35–39] Under the original KMR framework, a single phenotype is modeled to be related to a group of genetic variants.

The relationship is captured by way of a kernel function which measures similarity among the risky variants. Then testing proceeds by comparing pair-wise similarity in genetic variant profiles between subjects (measured by the kernel matrix) to pairwise similarity in phenotypes (measured by the cross product matrix of traits), with correspondence in similarity indicative of association.[40,41] By intelligently choosing kernels, structure in the genetic variants can be directly accommodated,[42,43] while dealing with high dimensionality.

Motivated by these kernel-based genetic association tests, we propose the dual kernel-based association test (DKAT) which is designed to assess the association between a high-dimensional, possibly structured, phenotypes of interest with multiple genetic variants, though the approach trivially applies to single genetic variant analysis as well. The idea of DKAT is that we propose to use not only a kernel for the genetic variants but also a kernel for the high dimensional and structured traits. In other words, we replace the cross product matrix for traits in existing KMR framework with a kernel matrix to better capture the high-dimensionality as well the structure of the traits. To associate the traits (now embedded within a kernel) and a group of genetic variants, we again compare similarity in genetic variant profiles to similarity in phenotypic profiles. In particular, the normalized Frobenius inner product between two kernel matrices is used as the statistic to summarize the genotype-phenotype association.

Besides being able to incorporate high-dimensional structured traits in genetic association analysis, another major contribution of DKAT is that we introduce a new test design for genetic association testing. Currently, two most popular

p-value calculation methods for genetic association analysis is either based on large-sample asymptotic theory[29,30,36,37] or via permutations.[28,44] However, the large-sample asymptotic theory-based p-value calculation can lead to conservative test with accumulated estimation error,[45,46] as in studies with small samples or high-dimensional traits. On the other hand, the permutation test is inefficient when a stringent p-value is required, as in many genome-wide association studies. We propose a fast pseudo-permutation technique for DKAT, which approximates the empirical distribution of all $n!$ potential permuted DKAT statistics by moment matching. In this new test design, we only calculate the first three sample moments of permutations without explicitly calculating the permutations themselves. Then, the Pearson type III density with the same moments is used to approximate the empirical distribution of all permutations, where a Pearson type III density is selected in this paper due to its good approximation performance for DKAT-similar statistics.[47–49] Fortunately, first three sample moments of these $n!$ permutations have closed-form expressions.[50] Thus, we can analytically calculate both the Pearson type III density and the DKAT p-value. Our DKAT test design is more efficient and accurate than those currently used for genetic association tests, since it neither requires explicit permutations nor relies on large-sample asymptotic theory.

Material and Methods

Throughout this paper, we assume that we have a study with n unrelated individuals who have been genotyped and phenotyped. For the ith subject ($i = 1, \ldots, n$), let $G_i = (g_{i1}, \ldots, g_{im})$ denote the vector of genotypes, where $g_{ij} = 0$, 1, or 2 representing the number of minor alleles, and $Y_i = (y_{i1}, \ldots, y_{ip})$ denote the set of p traits, e.g. the expression values of p genes in a pathway or the abundances of p metabolites in a pathway. The objective is to test the global association between the group of traits and the group of genetic variants which will be accomplished by using the kernel machine framework. We emphasize that although our focus is on the setting in which we have multiple genetic variants, our method trivially applies to the scenario when $m = 1$, that is, when we are interested in the relationship between a single variant and multiple traits.

Single Kernel-based Association Tests

Before discussing multi-trait association analysis, we first briefly review the KMR framework, which has been widely used to test the association between a set of genetic variants and a single trait.[35–39,51,52] Specifically, the KMR relates the trait (continuous or dichotomous) to the set of genotype values using the following generalized partial linear model:[51,52]

$$g(E(y_i|\mathbf{X}_i, \mathbf{G}_i)) = \mathbf{X}_i\boldsymbol{\alpha} + f(\mathbf{G}_i), \tag{1}$$

where $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_q)'$ is the regression coefficients for the covariates, $f(\cdot)$ is a generally specified function belongs to a space spanned by a kernel function $k_g(\cdot, \cdot)$, $g(\cdot)$ is a link function, such as identity function for continuous traits and logit function for dichotomous traits. The kernel $k_g(\cdot, \cdot)$ is the genotype kernel and has corresponding kernel matrix $\mathbf{K}_G$, where $\mathbf{K}_G(i, j) = k_g(\mathbf{G}_i, \mathbf{G}_j)$, $i, j = 1, \ldots, n$. The key to this KMR framework is usage of a positive semi-definite kernel function $k_g(\mathbf{G}_i, \mathbf{G}_j)$ as a similarity measure between genotypes $\mathbf{G}_i$ and $\mathbf{G}_j$,[42,43] which can facilitate capture of structure and relationships among genetic variants.

In the KMR model (1), the trait is related to the variants through $f(\cdot)$. Hence, testing the hypothesis of no association between the trait and genetic variants after adjusting for covariates is equivalent to testing $f(\cdot) = 0$. Through connections between KMR and generalized linear mixed models,[51,52] we can treat $f(\mathbf{G})$ as a vector of subject specific random effects with mean zero and variance $\tau \mathbf{K}_G$. Then testing $f(\cdot) = 0$ is equivalent to testing whether the variance component τ is equal to zero, which can be easily accomplished using a variance component score test with the following test statistic

$$S := \frac{1}{2\phi}(\mathbf{y} - \hat{\mathbf{y}})' \mathbf{K}_G (\mathbf{y} - \hat{\mathbf{y}}) = \frac{1}{2\phi} tr\left(\mathbf{K}_G (\mathbf{y} - \hat{\mathbf{y}})(\mathbf{y} - \hat{\mathbf{y}})'\right), \tag{2}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\hat{\mathbf{y}}$ is the estimated trait values under the null model of $f(\cdot) = 0$, and $tr(\cdot)$ denotes the trace of a matrix. When the trait is continuous, $\phi = \hat{\sigma}^2$ with $\hat{\sigma}^2$ being estimated under the null model. When the trait is dichotomous, $\phi = 1$. Under the null, Q follows a mixture of χ^2 distributions which can be approximated using exact methods.[53]

Test statistic (2) is essentially the sum of element-wise product of two $n \times n$ matrices. One is $\mathbf{K}_G$ and the other is cross product of the trait residuals $(\mathbf{y}-\hat{\mathbf{y}})(\mathbf{y}-\hat{\mathbf{y}})'$. In genetic association analysis, the kernel matrix $\mathbf{K}_G$ is often used to measure the subject-pairwise similarity in terms of genotypes[35–37] and the cross product of residuals $(\mathbf{y}-\hat{\mathbf{y}})(\mathbf{y}-\hat{\mathbf{y}})'$ is often used to measure subject-pairwise similarity of phenotypes.[40,41] Heuristically speaking, statistic S compares the subject-pairwise similarity in the trait to that in genotypes, where a high correspondence usually leads to a large statistic value and suggests existence of association.

There are two straightforward ways to extend the single kernel-based association test statistic (2) to accommodate multiple traits $\mathbf{Y}$. One is to stack the columns of $\mathbf{Y}$ into a huge column vector $\mathbf{y}^* = vec(\mathbf{Y})$ and apply the statistic (2) to $\mathbf{y}^*$.[27] However, a major limitation is that this approach can be computationally intractable with high-dimensional traits since it needs to eigendecompose a $np \times np$ matrix. The other approach to incorporate multiple traits is simply to replace the univariate trait residuals cross product matrix $(\mathbf{y}-\hat{\mathbf{y}})(\mathbf{y}-\hat{\mathbf{y}})'$ by the multivariate traits residuals cross product matrix $(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})'$, where $\hat{\mathbf{Y}}$ is estimated under the null model $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ assuming all traits are continuous. The second approach typically loses power when traits are highly or even modestly correlated with each other.[29] Furthermore, both approaches fail to capture any complicated structures within traits (e.g., inherent regulatory network structure within transcriptomic pathway expressions), which can further lead to power loss.[34] To address this issue, we propose the DKAT approach in the following section to allow for testing association between a high-dimensional, possibly structured traits and

one or more genetic variants.

A DKAT

To address the aforementioned limitations, we propose to use a phenotype kernel $\mathbf{K}_Y$ to model multiple traits simultaneously. Similar the genotype kernel $\mathbf{K}_G$, the phenotype kernel $\mathbf{K}_Y$ is used to summarize the phenotypic similarity. Compared with the cross product matrix $(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})'$ used in some existing methods, DKAT is able to capture complex structures among the multiple phenotypes by embedding the phenotypes in a kernel.

Like the single kernel-based association tests in KMR, we test the association between multiple traits and multiple genetic variants by comparing the phenotypic similarity matrix and genotypic similarity matrix across pairs of individuals. Motivated by works of relating two matrices from the same individuals,[47–49] we propose the new DKAT statistic as

$$D := \frac{tr(\mathbf{H}\mathbf{K}_G\mathbf{H}\mathbf{K}_Y)}{\sqrt{tr(\mathbf{H}\mathbf{K}_G\mathbf{H}\mathbf{K}_G)tr(\mathbf{H}\mathbf{K}_Y\mathbf{H}\mathbf{K}_Y)}}, \tag{3}$$

where $\mathbf{H} = \mathbf{I}_n - \mathbf{1}\mathbf{1}'/n$ is a centering matrix, $\mathbf{I}_n$ is the nth order identity matrix, and $\mathbf{1}$ is a n-dimensional vector of ones. Since $\mathbf{H}$ is idempotent, the numerator $tr(\mathbf{H}\mathbf{K}_G\mathbf{H}\mathbf{K}_Y)$ is essentially the same as $tr(\mathbf{H}\mathbf{K}_G\mathbf{H}\mathbf{H}\mathbf{K}_Y\mathbf{H})$, which is the element-wise multiplication of centered genotype kernel matrix $\mathbf{H}\mathbf{K}_G\mathbf{H}$ and centered phenotype kernel matrix $\mathbf{H}\mathbf{K}_Y\mathbf{H}$. Hence, our DKAT statistic shares the same spirit of comparing two similarities as the single kernel-based association tests statistic (2).

Moreover, if the phenotype kernel is picked as $\mathbf{K}_Y = (\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})'$, then the DKAT statistic reduces to the form of KMR statistic in (2). Therefore, most existing kernel association tests[27,29,35–37,39,51,52] can be viewed as special forms of DKAT. Alternative to comparing two kernel matrices, there exist some similar statistics either comparing two input matrices[47] or two distance matrices.[48] Kernels have been widely used to capture structures among genotypes.[30,35–38] Following this steam, specific kernels are used in this paper to capture the inherent structures among both genotypes and phenotypes.

Intuitively speaking, the larger the DKAT statistic, the more likely the genotype kernel matrix resembles the phenotype kernel matrix, which further implies that the phenotypes might be associated with the genotypes in a specific way. To calculate the exact critical value of a DKAT under a given significance level, we need to study its distribution under the null hypothesis of no association. Two current standard approaches of calculating the null distribution of a genetic association test statistic are permutation-based resampling methods[28,44] and large sample-based asymptotic methods.[29,30,35–37] However, both methods have potential limitations. On one hand, it is computationally expensive to use permutations to achieve genome-wide significance. On the other hand, it is observed that asymptotic methods can be conservative when the sample size is small or modest.[45,46] To overcome these potential limitations, we calculate the p-value of DKAT using a fast pseudo-permutation method, closely following the strategy being used in the RV coefficient literature,[47–49] where a typical RV coefficient shares the same form of DKAT statistic but uses totally different matrices other than K_G

and K_Y (both introduced in the next section) as used in this paper. Specifically, a Pearson type III distribution is used to approximate the permutation null distribution of DKAT by matching first three moments.[49] The advantages of the new DKAT p-value calculation strategy are two-fold. First, no explicit permutation is required as the finite-sample empirical moments can be analytically calculated. Second, closed-form expression of the Pearson type III density is available, and thus our method allows fast and analytic p-value calculation for genetic association analysis.

Choices of Kernels

A key aspect of DKAT is the kernels, which appropriately summarize the phenotypic and genotypic similarities between pairwise subjects. Even though DKAT is statistically valid in protecting the correct type I error, irrespective of the kernels being used. However, good choice of kernels, which better reflect the unique data features, can improve the test power.[33,34] In this section, we first briefly review some genotype kernels widely used in existing kernel-based association tests and some kernels that could potentially be used for phenotypes. And then, we propose a specific phenotype kernel for the high-dimensional structured phenotypes considered in this paper.

In literature, many kernels have been proposed for genotype data.[42,43] Some popular examples include the linear kernel and the identity-by-state (IBS) kernel:

- Linear Kernel: $k_g(\mathbf{G}_i, \mathbf{G}_j) = \mathbf{G}_i'\mathbf{G}_j = \sum_{l=1}^{m} g_{il}g_{jl}$

- IBS Kernel: $k_g(\mathbf{G}_i, \mathbf{G}_j) = \frac{1}{2m}\sum_{l=1}^{m}(2 - |g_{il} - g_{jl}|)$

The linear kernel assumes a linear association pattern. That is, the function $f(\cdot)$ in model (1) is of a linear form. It is simple and can be powerful when the true underlying association pattern is linear. The IBS kernel measures the similarity between $\mathbf{G}_i$ and $\mathbf{G}_j$ in terms of the number of alleles with IBS sharing by a pair. The IBS kernel is positive definite,[35] however the spanned functional space is less studied. Both the linear kernel and the IBS kernel are additive forms, which makes it easy to incorporate weights $w_l, l = 1, \ldots, m$ for each genetic variant.[37]

On the other hand, few studies have described the use of kernels for the complex multi-dimensional traits as considered in this paper. In general, if all traits are continuous, then the Gaussian kernel and the dth-order polynomial kernel are often used. Also the binary kernel was shown to be a valid kernel function for all multivariate binary traits.

- Gaussian Kernel: $k_y(\mathbf{Y}_i, \mathbf{Y}_j; \rho) = \exp\{-\sum_{l=1}^{p}(y_{il} - y_{jl})^2/\rho\}$
- Polynomial Kernel: $k_y(\mathbf{Y}_i, \mathbf{Y}_j; d) = \sum_{l=1}^{p}(y_{il}y_{jl} + 1)^d$
- Binary Kernel: $k_y(\mathbf{Y}_i, \mathbf{Y}_j) = \sum_{l=1}^{p} I[y_{il} \neq y_{jl}]$

If the traits are mixed (a combination of continuous variables and binary variables), then we can define kernels for both the continuous and binary parts separately and then multiply them together as the final kernel function, which has been shown to be valid for association analysis.[39]

No matter how large the dimension p is, the information in all traits is pooled into a scalar by using the phenotype kernel. In this sense, DKAT is robust a-

gainst high-dimensional phenotypes, which can be a major advantage over most existing multivariate regression-based testing methods.[27,29] Besides the robustness to high-dimensional traits, another major concern of this paper is to address the network-type traits, such as expression of genes belonging to the same pathway. For such gene pathway data, a network-based kernel has been proposed of the form $\mathbf{K}_Y = \mathbf{YNY}'$,[34] where $\mathbf{N}$ is the undirected adjacency matrix, and $\mathbf{N}_{ij} = 1$ represents that gene i and gene j interact with each other in an activating fashion, $\mathbf{N}_{ij} = -1$ represents an inhibition pattern.

In reality, it is difficult to know the functional relationship between each gene pair within the pathway. Hence, we replace the adjacency matrix $\mathbf{N}$ with the precision matrix $\mathbf{\Theta}$ (also called inverse covariance matrix $\mathbf{\Sigma}^{-1}$), which can be estimated from the data without any prior biological knowledge. The precision matrix $\mathbf{\Theta}$ is useful in estimating partial correlations, which incorporates the functional mechanism of the whole pathway. For example, under the Gaussian assumption, $\mathbf{\Theta}_{ij} = 0$ indicates that gene i and gene j are conditionally independent given all other genes in the network/pathway, or equivalently speaking, gene i and j are unconnected in the gene network/pathway.[55] Similar to the undirected adjacency matrix $\mathbf{N}$, $\mathbf{\Theta}$ can also incorporate the underlying network-structure. Thus, we propose the phenotype kernel matrix as $\mathbf{K}_Y = \mathbf{Y}\hat{\mathbf{\Theta}}\mathbf{Y}'$, where $\hat{\mathbf{\Theta}}$ is the estimated precision matrix. A simple estimator is the sample precision matrix $\hat{\mathbf{\Theta}}_s$, and the corresponding phenotype kernel matrix $\mathbf{K}_Y$ is proportional to the so-called projection similarity matrix in literature.[30,56] When the dimension of traits is high, the sample precision matrix $\hat{\mathbf{\Theta}}_s$ is unstable or even not estimable. In such a high-

dimensionality scenario, we estimate the precision matrix via regularization. For example, a graphical lasso estimator $\hat{\mathbf{\Theta}}_{gl}$ can be derived by maximizing the lasso-penalized log-likelihood.[55]

In practice, it is often true that multiple kernels $K_G^1, \ldots, K_G^t$ and $K_Y^1, \ldots, K_Y^s$ are available for testing in DKAT. Without knowing the true underlying association model, it is of importance to accommodate multiple candidate kernels. In general, there are two approaches to tackle this issue. The first average-type strategy is to calculate an omnibus K_G^o which is usually a linear combination of $K_G^1, \ldots, K_G^t$, and another omnibus K_Y^o which is usually a linear combination of $K_Y^1, \ldots, K_Y^s$. Then a final DKAT(K_G^o, K_Y^o) test is applied. The other minimum-type approach to accommodate multiple candidate kernels is to pick the most significant kernel pair. That is, K_G^* and K_Y^* are selected such that DKAT(K_G^*, K_Y^*) has the smallest p-value over all ts kernel pairs (K_G^i, K_Y^j), $i = 1, \ldots, t$, $j = 1, \ldots, s$. However, the minimum p-value is no longer a genuine p-value and permutations are often needed to establish the final significance. Details of these two approaches of accommodating multiple candidate kernels, along with numerical evaluations, can be found in the supplementary materials.

Besides the kernels, another important practical issue is to adjust for the confounding covariates effects, such as age, gender and principal components of genotypes (for adjusting population structures). In genetic association tests, a common strategy of adjusting for covariates is the residual-based approach.[28,30,36,37,40,41] That is, we first fit the null model with covariates only: $g(E(y_i|\mathbf{X}_i)) = \mathbf{X}_i\boldsymbol{\alpha}$ and then calculate the residuals $\boldsymbol{\epsilon}_Y = \mathbf{Y} - \hat{\mathbf{Y}}$ of the null model. Next, one can con-

struct the phenotype kernel on the residuals as the subject-wise trait similarity after adjusting for covariates. That is, the phenotype kernel matrix $\mathbf{K}_Y = (\mathbf{Y} - \hat{\mathbf{Y}})\hat{\mathbf{\Theta}}(\mathbf{Y} - \hat{\mathbf{Y}})'$ is used in DKAT, where $\hat{\mathbf{\Theta}}$ is the estimated precision matrix of residuals. Existing numerical studies have shown that it can have the correct type I error as long as the number of covariates is much smaller than the sample size.[28,30,36,37]

Simulation Studies

We conducted extensive simulation studies under different scenarios to evaluate the performance of DKAT in testing the association between high-dimensional structured traits and genotypes. To mimic a relatively high-dimensional scenario, $p = 200$ traits (e.g., expressions of genes belonging to a pathway) were considered in our simulation. As a comparison, most existing multivariate association tests usually considered less than 20 traits.[5,14,18,25,27,29,30] Two different correlation structures were used in this simulation. One was the compound symmetry covariance structure as commonly used in literature.[14,25,27,29,30] That was, $\mathbf{\Sigma}_{ii} = 1$ and $\mathbf{\Sigma}_{ij} = \rho$ for all $i \neq j$, where $\mathbf{\Sigma}$ was the covariance matrix of the traits. The other correlation structure was the banded inverse covariance (precision) matrix $\mathbf{\Theta}$ with $\mathbf{\Theta}_{i,i} = 1$, $\mathbf{\Theta}_{i,i-1} = \mathbf{\Theta}_{i-1,i} = \rho$, and zero otherwise, where $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ was the precision matrix of traits. Assuming all traits were continuous, then $\mathbf{R}_{ij|-\{i,j\}} = -\mathbf{\Theta}_{ij}/\sqrt{\mathbf{\Theta}_{ii}\mathbf{\Theta}_{jj}}$, where $\mathbf{R}_{ij|-\{i,j\}}$ was the partial correlation between trait i and j given all other traits. Thus, the banded precision matrix $\mathbf{\Theta}$ represented such a pathway that each gene was only related to its nearby genes

conditional on all other genes in the pathway. In contrast to the compound symmetry covariance structure, the banded inverse covariance structure mimicked the complicated functional regulatory mechanisms in a gene pathway. For simplicity, we denoted these two covariance structures as $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2 = \mathbf{\Theta}^{-1}$ in the rest of the simulation section. To guarantee positive definiteness of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, we simply simulated ρ from Uniform (0,0.5) distribution. Finally, we conducted three different simulation studies, where **Simulation I** was for a single SNP, **Simulation II** was for multiple SNPs, and **Simulation III** was for multiple rare variants. Under each simulation scenario, we considered sample size of either 500 or 1000 subjects.

Simulation I This simulation was designed to mimic the pleiotropy effect, where a common SNP affected multiple traits. The data were generated from the model

$$y_{ij} = \beta_j \cdot g_i + \epsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, p, \tag{4}$$

where y_{ij} was the expression value of gene j for subject i and g_i was a single SNP taking values 0, 1 and 2, with a MAF of 0.3. For each i, $\epsilon_{ij}, j = 1, \ldots p$ was distributed as multivariate Gaussian with mean zero and covariance matrix $\mathbf{\Sigma}$, where $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ or $\mathbf{\Sigma}_2$. For simplicity, we did not consider covariates in the model since they could be easily adjusted via the residual-based approach described previously. Under the null model, all $\beta_j = 0$. Under the alternative model, we set a proportion ($\gamma = 10\%, 20\%, 30\%$) of traits to be truly associated with the SNP (with non-zero β-coefficients). Without loss of generality, we set the first $p^* = \gamma p$ traits as relevant ones with coefficients β_j generated from a uniform (0,$\sqrt{30/n}$)

distribution, for $j = 1, \ldots, p^*$, and $\beta_j = 0$ for $j = (p^* + 1), \ldots, p$. The effect sizes (following uniform $(0, \sqrt{30/n})$ distribution) changed with sample size and hence it was meaningless to compare test powers under different sample sizes. These effect sizes were selected to better distinguish different tests under each scenario.

Simulation II In the second simulation scenario, we tested the association between multiple SNPs and multiple traits. The multiple SNPs were generated based on the LD structure of gene ASAH1, acid ceramidase 1, as used previously.[36] A total of 93 HapMap SNPs are located within this gene. Based on the LD structure of the ASAH1 gene, we used HAPGEN[57] to generate SNP genotype data at each of the 93 loci. After the SNPs were simulated, we generated the traits from the following model:

$$y_{ij} = \sum_{k=1}^{93} \beta_{kj} g_{ik} + \epsilon_{ij}, i = 1, \ldots, n, j = 1, \ldots, p, \tag{5}$$

where relevant model parameters (e.g., $\gamma, \mathbf{\Sigma}$) were the same as the previous **Simulation I**. We selected 29 typed SNPs on Affy6 to calculate the genotype kernel K_G in the analysis. Under the null model $\beta_{kj} = 0, k = 1, \ldots, 93, j = 1, \ldots, p$. Under the alternative model, we selected the first $p^* = \gamma p$ traits as causal ones which were truly associated with the SNPs. For each causal trait, we randomly selected three SNPs from the 93 SNPs as the causal SNPs for that trait, and simulated the nonzero β_{kj}-coefficient from uniform $(0, \sqrt{30/n})$ distribution for $k = j_1, j_2, j_3 \in \{1, \ldots, 93\}$ and $j = 1, \ldots, p^*$, where different traits could have different causal SNPs. Finally, to allow for the heterogeneous effect of different loci, we randomly assigned a

sign for the β-coefficient of each SNP with even probability.

Simulation III For simulation of rare variants, we considered the previous design[37] to generate rare variants. We simulated 10000 haplotypes for 1Mb region on the basis of COSI[58] to mimic the LD pattern, local recombination rate and population history of European descent. Only those variants with MAF$< 3\%$ were included in the analysis. After rare variants being simulated, we generated the traits according to model (5). Under the alternative model, we randomly selected 10% of the rare variants as causal ones and simulated the nonzero β-coefficients from uniform $(0, 2\sqrt{30/n}) \times |log_{10}(MAF)|$. Other simulation settings were the same as **Simulation II**.

Competing methods After the data were generated, DKAT was applied to test the association between genotypes and phenotypes. The phenotype kernel used in DKAT was $\mathbf{K}_Y = (\mathbf{Y} - \hat{\mathbf{Y}})\hat{\mathbf{\Theta}}_{gl}(\mathbf{Y} - \hat{\mathbf{Y}})'$, where $\hat{\mathbf{Y}}$ was the phenotypes sample mean and the graphical lasso regularization parameter was set as $\rho^{gl} = 0.1$ in our simulation. The graphical lasso method was used for illustrative purposes of constructing the phenotype kernel incorporating the high dimensionality as well as network structures in traits. An optimal graphical lasso regularization parameter was beyond the scope of this paper.

Along with DKAT, we also evaluated other methods for comparison. Among existing multivariate-trait association tests, both multiple testing adjusted univariate trait methods[17,18] and dimension reduction-based methods[5,19] can be limited with high-dimensional traits. Other multivariate traits-single SNP associa-

tion testing methods[20] suffer from power loss when there are systematic but weak marginal effects for each SNP. To make the comparison fair, we focus on existing methods that test association between multivariate-trait and multiple SNPs/rare variants. Two of such methods are the Gene Association with Multiple Traits (GAMuT) test[30] and the multivariate sequence kernel association tests (MSKAT),[29] which are briefly introduced in the following paragraph.

The GAMuT test statistic[30] is actually the numerator of the DKAT statistic in (3). However, it calculates the p-value differently, using large-sample results.[30] The asymptotic distribution of GAMuT statistic is $\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\xi_j\chi_{ij}^2$, where λ_i, ξ_j are eigenvalues of $\mathbf{HK}_G\mathbf{H}$ and $\mathbf{HK}_Y\mathbf{H}$ respectively, and χ_{ij}^2 are i.i.d. χ^2 distributed with 1 degree of freedom. Then, the GAMuT p-value is calculated based on this asymptotic distribution with quadratic form approximations.[53,54] To make a fair comparison, the same phenotype kernel $\mathbf{K}_Y = (\mathbf{Y}-\hat{\mathbf{Y}})\hat{\mathbf{\Theta}}_{gl}(\mathbf{Y}-\hat{\mathbf{Y}})'$ in DKAT was applied in GAMuT in our simulations. The other method MSKAT assumes a linear model between each individual trait with multiple genetic variants, and considers the score test statistics s_{jk} between the kth trait and jth variant, where $k = 1,\ldots,p$, $j = 1,\ldots,M$. Let $S_j = (s_{j1},\ldots s_{jp})'$ be the score vector between the jth variant and all traits. Ignoring the weights, the MSKAT statistic has been proposed as $Q = \sum_{j=1}^{M} S_j'\hat{\mathbf{\Theta}}_s S_j$,[29] where $\hat{\mathbf{\Theta}}_s$ is the sample precision matrix. Unlike $\hat{\mathbf{\Theta}}_{gl}$-based DKAT and GAMuT, the MSKAT statistic uses $\hat{\mathbf{\Theta}}_s = \hat{\mathbf{\Sigma}}^{-1}$, which requires $n > p$. To avoid this potential limitation, another variant of statistic $Q_2 = \sum_{j=1}^{M} S_j'S_j$ is also considered,[29] which is termed as MSKAT2 in our simulations. MSKAT2 represents a broad class of multivariate-trait association testing

methods that ignore the correlation structures among outcomes (such as DKAT and GAMuT with the linear phenotype kernel $\mathbf{K}'_Y = (\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}})'$). MSKAT and MSKAT2 p-values are calculated in a similar way as GAMuT, which is based on its asymptotic quadratic form approximation.[29] MSKAT and MSKAT2 implicitly used the (weighted) linear kernel for genetic variants. To make the comparison fair, the same linear kernels were used in DKAT and GAMuT. In particular, we used the same linear kernel for the SNPs simulations (**Simulation I and I-I**) and weighted linear kernel for the rare variants simulation (**Simulation III**). The weight for each rare variant was specified as Beta(MAF,1,25) as suggested in SKAT.[37] Finally, under each simulation scenario, we evaluated the type I error of each test with 1,000,000 replicates under the null model, and the power with 1,000 replicates under the alternative model. The empirical type I error rate and power were calculated as the proportion of replicates with a p-value smaller than the nominal significance level.

Results

Type I Error Simulation Results

The empirical type I error rates under **Simulation I** are reported in Table 1. Based on the table, DKAT is always able to protect the correct type I error across different scenarios. On the other hand, GAMuT and MSKAT are conservative under each simulation scenario, especially when the sample size is relatively small ($n = 500$). MSKAT2 seems to be more conservative under $\mathbf{\Sigma}_2$ than $\mathbf{\Sigma}_1$. To further explore the

type I error of all tests at more stringent significance levels, we present the QQ-plots of p-values under the configuration of $n = 500$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_1$ in Figure 1. As we can see, the p-values of DKAT stick with the 45 degree line, which indicates that the type I error of DKAT is well controlled under different significance levels. For GAMuT and MSKAT, we can see a clear departure from the 45 degree line with plots skewing downward, implying these tests are very conservative, which are all consistent with the results from Table 1. QQ-plots under other simulation configurations are qualitatively similar and hence are not reported. Similar empirical type I error results have also been observed in **Simulation II** and **Simulation III**.

It has been observed in single trait kernel association tests that estimation error (due to small sample size) can lead to conservative tests,[45,46] which also explains the conservativeness of GAMuT and MSKAT in this simulation. Taking GAMuT as an example, the asymptotic null distribution of GMAuT depends on the eigenvalues of matrix $\mathbf{H}\mathbf{K}_Y\mathbf{H}$ where $\mathbf{K}_Y = (\mathbf{Y} - \hat{\mathbf{Y}})\hat{\mathbf{\Theta}}_{gl}(\mathbf{Y} - \hat{\mathbf{Y}})'$, which further requires accurate estimation of the precision matrix. Given the high dimensionality of traits, many parameters in the precision matrix need to be estimated. The accumulated estimation errors in GAMuT deteriorate the performance of the test resulting in over-protected (conservative) p-values.[45,46] Unlike GAMuT and MSKAT, which need to estimate the whole precision matrix $\mathbf{\Theta}$, MSKAT2 only needs to estimate the diagonals $\mathbf{\Sigma}_{jj}$, $j = 1, \ldots p$. The accumulated estimation errors in MSKAT2 is much smaller and hence it is less conservative than GAMuT and MSKAT. Finally, the way DKAT calculates its p-value is more robust to these estimation errors,

and hence DKAT is robust to small samples and high-dimensional traits. To summarize, the proposed DKAT always has the correct type I error rate even under a very stringent nominal significance level. On the other hand, GAMuT, MSKAT and MSKAT2 can be conservative especially when the sample size is relatively small or modest.

Power Simulation Results

Without loss of generality, we compare the power of all tests under significance level $\alpha = 2.5 \times 10^{-6}$ (reflecting a genome-wide Bonferroni correction for 20,000 genes). The power under **Simulation I** is presented in Figure 2. It is clear to see that DKAT is always the most powerful test under each scenario. On the other hand, MSKAT2 always tends to be the least powerful test (except for the small sample scenario, where MSKAT can have lower power due to the its conservativeness as seen in the previous type-I error simulation results section). This is because the phenotype kernel $\mathbf{K}_Y = (\mathbf{Y} - \hat{\mathbf{Y}})\hat{\mathbf{\Theta}}_{gl}(\mathbf{Y} - \hat{\mathbf{Y}})'$ used in DKAT and GAMuT (or $\hat{\mathbf{\Theta}}_s$ used in MSKAT) can incorporate the inherent correlation structure among the multivariate traits, while MSKAT2 simply ignores the correlations among traits. The power gain of DKAT/GAMuT/MSKAT over MSKAT2 increases with the (partial) correlation strength among traits (i.e., ρ value in $\mathbf{\Sigma}_1$ or $\mathbf{\Sigma}_2$). For each test considered in this simulation study, the power of test increases as the proportion (γ) of associated traits increases (i.e., as the genes are increasingly pleiotropic). This is because it can further amplify the association signal by including more relevant traits into the multi-trait association analysis. Qualitatively

similar empirical power results are also observed in **Simulation II** and **Simulation III**.

To summarize, DKAT is always more powerful than GAMuT, MSKAT and MSKAT2. The power gain probably comes from two aspects. One is the usage of phenotype kernel to incorporate the complex structure of traits into association analysis (compared to MSKAT2). The other is from the new efficient and robust p-value calculation (compared to GAMuT and MSKAT).

Analysis of the Grady Trauma Project Data

We applied the newly proposed DKAT approach to a Grady Trauma Project data set that was collected as part of a larger study investigating the role of genetic and environmental factors in predicting response to stressful life events.[60] 337 individuals were recruited from the Grady Memorial Hospital in Atlanta, Georgia. Blood samples were collected from these individuals who provided informed consent and participated in a verbal interview. For each individual, both gene expressions and genotypes were measured. Demographic data such as gender, age and race were also collected. Details on data collection and preprocessing can be obtained from previous publications.[60] Previous studies have shown that genetic risk factors may account for up to 30%-40% of the heritability of developing post-traumatic stress disorder (PTSD) following a trauma, and many gene pathways that are associated with PTSD have been identified.[61] In this analysis, we further studied the genetic regulation of expressions of genes belonging to these pathways. In particular, we were particularly interested in the cis-regulation, that

was, whether pathway gene expressions were associated with the SNPs in the same pathway. Expressions of 8,588 genes belonging to 224 pathways (with more than one gene in each pathway) were measured. In each pathway analysis, the phenotypes were the gene expression values and the genotypes were the SNPs in that pathway. A total of 164,503 SNPs were mapped to the 8,588 genes in 224 pathways. The median number of genes in a pathway was 27 with the first and third quantiles being 14 and 48.

Two different sets of association analyses were conducted. In the first set of association analysis, we evaluated the association between the multiple gene expressions in a pathway and all SNPs in the same pathway using DKAT, GAMuT, MSKAT and MSKAT2. In the second set of association analysis, we evaluated the importance of each individual gene for certain pathways that might be of interest based on results of the first analysis. In other words, we examined the association between the pathway gene expressions and SNPs in each individual gene belonging to the pathway. For all association analyses, we adjusted the covariates effects of gender, age, race and the top ten principal components of the genotype data.

To account for multiple testing, we set family-wise significance level of $2.2 \times 10^{-4} = 0.05/224$, which corresponds to a Bonferroni correction based on the number of pathways being tested. Under this significance level, 18, 17, 16 and 1 pathways have been found that their gene expressions were significantly associated with their SNPs by DKAT, GAMuT, MSKAT and MSKAT2 respectively. Compared with MSKAT2, it is clear that incorporating the network-type gene regulatory structure via the precision matrix (as in DKAT/GAMuT/MSKAT) can large-

ly enhance the discovery power of association analysis between pathway gene expressions and SNPs. The DKAT is slightly more powerful than GAMuT and MSKAT, which is probably because the test design of DKAT is more efficient for this data set.

The only significant pathway detected by all DKAT, GAMuT, MSKAT and MSKAT was asthma (KEGG: hsa05310). A further interesting analysis was to test which individual gene regulates the asthma pathway gene expressions. That was, we tested the association between asthma pathway gene expressions and all SNPs in a single gene belonging to that pathway. In this data, a total of 167 SNPs were detected in 10 genes in the asthma pathway. Under the gene-level SNPs and pathway-level expressions association analysis, DKAT, GAMuT and MSKAT all detected four genes (HLA-DRA, HLA-DRB1, HLA-DQA1, HLA-DQB1) which regulated the asthma pathway expressions while MSKAT2 only detected two of them (HLA-DRB1, HLA-DQA1). Further functional study of these genes on asthma may be of biological interest.

Discussion

In this article, we have proposed DKAT for evaluating the association between high-dimensional structured traits and multiple SNPs or rare variants. Compared with most existing kernel association tests (e.g., SKAT), the novelties of DKAT are two-folded. First, an additional phenotype/trait kernel is used, which can incorporate the inherent complex structure of the traits and thereby improve the statis-

tical power for detecting an existing association signal. The numerical studies in this paper are mainly designed to mimic the scenario of high-dimensional, structured traits, where we propose a network-type phenotype kernel by replacing the adjacency matrix[34] with the precision matrix. We emphasize that it is possible to design new appropriate kernels for other data types, which can lead to useful and powerful association analysis. Second, unlike existing association tests, DKAT provides a new robust strategy to compute p-values in genetic association testing. The DKAT p-value is less sensitive to estimating errors in covariance terms compared to other methods (e.g., GAMuT and MSKAT), and is extremely appealing with high-dimensional traits, where it is difficult to accurately estimate the trait covariance matrix given the dimensionality. Thus, DKAT is more robust than most existing methods in testing the association between high-dimensional structured traits and genotypes.

As an association test, DKAT has four advantages. First, DKAT is methodologically flexible in testing association between an arbitrary set of traits and an arbitrary set of genetic variants. It can test the association between multiple traits and either a single/multiple SNPs or multiple rare variants, without making parametric assumptions. On the contrary, many existing multivariate trait association tests can only handle a single SNP.[14,20,23–25] Others often assume that traits are associated with SNPs through a linear model.[27,29] Second, DKAT can evaluate biologically meaningful hypotheses. The phenotype kernel in DKAT can capture pleiotropy effects among the phenotypes and the genotype kernel can capture epistasis effects among SNPs. With prior biological knowledge, it can be of interest to apply

DKAT to test associations between a pre-specified set of traits and a pre-specified region of genetic variants, the results of which may further lead to meaningful biological insights. Third, DKAT is also statistically very powerful. As illustrated previously in the SNP-set association test,[35,36] a SNP-kernel can amplify the association signal by collapsing information across multiple SNPs. Moreover, the phenotype kernel in DKAT can further amplify the association signal by collapsing information across multiple traits. After amplifying twice, DKAT can greatly improve the statistical power to detect any existing association signal. Fourth, DKAT is also computationally scalable. Only matrix multiplication is required in DKAT. However, both GAMuT and MSKAT requires eigendecomposition of $n \times n$ matrices, which can be computationally unstable for large sample size. Furthermore, the asymptotic p-value calculation in GAMuT and MSKAT requires large n or small p, otherwise it can be conservative due to estimation error.[45,46] On the other hand, DKAT is applicable to any sample size n and trait dimension p. In this regard, DKAT is appropriate for the large p small n problems as frequently encountered in modern scientific studies.

The design of Simulation II (SNPs-set) and Simulation III (rare variants) is in vein with previous simulation studies in the literature.[36,37] For example, the same ASAH1 gene/LD structure is used in the paper. Since no relevant assumptions are made, we believe that our method should also work well with other genes/LD structures. As indicated in our numerical studies, including more relevant traits in DKAT increases the power to a large extent. However, when more noise traits (not associated with the SNP-set) are added, it may lead to power loss. In practice,

the true association signal may not be known. Adaptive testing strategies could be used to address this uncertainty.[44,64] Finally, to aid interpretation of which genetic variants or which traits are associated, it is of interest to prioritize individual genetic variants/traits by incorporating variable selection in DKAT.[65] We believe these issues are of importance and warrant further investigation.

References

[1] Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.

[2] McCarthy, M.I., Abecasis, G.R., Cardon, L.R., Goldstein, D.B., Little, J., Ioannidis, J.P., and Hirschhorn, J.N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.

[3] Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42(D1)*, D1001–D1006.

[4] Allison, D. B., Thiel, B., Jean, P. S., Elston, R. C., Infante, M. C., and Schork, N. J. (1998). Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. Am. J. Hum. Genet. *63*, 1190-1201.

[5] Klei, L., Luca, D., Devlin, B., and Roeder, K. (2008). Pleiotropy and principal components of heritability combine to increase power for association analysis. Genet. Epidemiol. *32*, 9–19.

[6] Aschard, H., Vilhjlmsson, B.J., Greliche, N., Morange, P.E., Trgout, D.A., and Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. Am. J. Hum. Genet. *94*, 662–676.

[7] Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., and Threadgill, D.W. (2005). Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. Nat. Genet. *37*, 233–242.

[8] Huang, J., Perlis, R. H., Lee, P. H., Rush, A. J., Fava, M., Sachs, G. S., et al. (2010). Cross-disorder genomewide analysis of schizophrenia, bipolar disorder, and depression. Am. J. Psychiatry *167*, 1254–1263

[9] Huang, J., Johnson, A. D., and O'Donnell, C. J. (2011). PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies. Bioinformatics *27*, 1201-1206.

[10] Cross-Disorder Group of the Psychiatric Genomics Consortium. (2013). Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. Lancet *381*, 1371-1379.

[11] van Vliet-Ostaptchouk, J. V., den Hoed, M., Luan, J., Zhao, J. H., Ong, K. K., Van Der Most, P. J., et al. (2013). Pleiotropic effects of obesity-susceptibility loci on metabolic traits: a meta-analysis of up to 37,874 individuals. Diabetologia *56*, 2134-2146

[12] Kraja, A. T., Chasman, D. I., North, K. E., Reiner, A. P., Yanek, L. R., Kilpelinen, T. O., et al. (2014). Pleiotropic genes for metabolic syndrome and inflammation. Mol. Gen. Metab. *112*, 317-338.

[13] Andreassen, O. A., Harbo, H. F., Wang, Y., Thompson, W. K., Schork, A. J., Mattingsdal, M., et al. (2015). Genetic pleiotropy between multiple sclerosis and schizophrenia but not bipolar disorder: differential involvement of immune-related gene loci. Mol. Psychiatry *20*, 207-214.

[14] Schaid, D. J., Tong, X., Larrabee, B., Kennedy, R. B., Poland, G. A., and Sinnwell, J. P. (2016). Statistical Methods for Testing Genetic Pleiotropy. Genetics *204*, 483–497.

[15] Alberti, K., George, M.M., Zimmet, P., Shaw, J., and IDF Epidemiology Task Force Consensus Group. (2005) The metabolic syndrome — a new worldwide definition. The Lancet *366*, 1059–1062.

[16] Carty, C. L., Bhattacharjee, S., Haessler, J., Cheng, I., Hindorff, L. A., Aroda, V., et al. (2014). Comparative Analysis of Metabolic Syndrome Components in over 15,000 African Americans Identifies Pleiotropic Variants: Results from the PAGE Study. Circulation: Cardiovascular Genetics, CIRCGENETICS-113.

[17] Yang, Q., Wu, H., Guo, C. Y., and Fox, C. S. (2010). Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. Genet. Epidemiol. *34*, 444-454.

[18] van der Sluis, S., Posthuma, D., and Dolan, C. V. (2013). TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. *9*, e1003235

[19] Ferreira, M. A., and Purcell, S. M. (2009). A multivariate test of association. Bioinformatics *25*, 132-133.

[20] O'Reilly, P. F., Hoggart, C. J., Pomyen, Y., Calboli, F. C., Elliott, P., Jarvelin, M. R., and Coin, L. J. (2012). MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. PloS One *7*, e34861.

[21] Schifano, E.D., Li, L., Christiani, D.C., and Lin, X. (2013). Genome-wide association analysis for multiple continuous secondary phenotypes. Am. J. Hum. Genet. *92*, 744–759.

[22] Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. PloS One *8*, e65245.

[23] Zhou, X. and Stephens, M. (2014). Efficient algorithms for multivariate linear mixed models in genome-wide association studies. Nat. Met. *11*, 407.

[24] Wu, B., and Pankow, J. S. (2015). Statistical Methods for Association Tests of Multiple Continuous Traits in Genome-Wide Association Studies. Ann. Hum. Genet. *79*, 282-293.

[25] Ray, D., Pankow, J. S., and Basu, S. (2016). USAT: A Unified Score-Based Association Test for Multiple Phenotype-Genotype Analysis. Genet. Epidemiol. *40*, 20-34.

[26] Galesloot, T. E., Van Steen, K., Kiemeney, L. A., Janss, L. L., and Vermeulen, S. H. (2014). A comparison of multivariate genome-wide association methods. PloS One, *9*, e95923.

[27] Maity, A., Sullivan, P.F. and Tzeng, J.Y. (2012). Multivariate Phenotype Association Analysis by Marker-Set Kernel Machine Regression. Genet. Epidemiol. *36*, 686–695.

[28] Hua, W.Y., and Ghosh, D. (2015). Equivalence of kernel machine regression and kernel distance covariance for multidimensional phenotype association studies. Biometrics *71*, 812–820.

[29] Wu, B., and Pankow, J.S. (2016). Sequence kernel association test of multiple continuous phenotypes. Genet. Epidemiol. *40*, 91–100.

[30] Broadaway, K. A., Cutler, D. J., Duncan, R., Moore, J. L., Ware, E. B., Jhun, M. A., et al. (2016). A Statistical Approach for Testing Cross-Phenotype Effects of Rare Variants. Am. J. Hum. Genet. *98*, 525-540.

[31] Zhang, Y., Xu, Z., Shen, X., Pan, W., and Alzheimer's Disease Neuroimaging Initiative (2014). Testing for association with multiple traits in generalized estimation equations, with application to neuroimaging data. NeuroImage *96*, 309–325.

[32] Zhan, X., Patterson A.D., and Ghosh, D. (2015a). Kernel approaches for differential expression analysis of mass spectrometry-based metabolomics data. BMC Bioinformatics *16*, 77.

[33] Zhao, N., Chen, J., Carroll, I.M., Ringel-Kulka, T., Epstein, M.P., Zhou, H., Zhou, J.J., Ringel, Y., Li, H., and Wu, M.C. (2015). Testing in Microbiome-

Profiling Studies with MiRKAT, the Microbiome Regression-Based Kernel Association Test. Am. J. Hum. Genet. *96*, 797–807.

[34] Freytag, S., Manitz, J., Schlather, M., Kneib, T., Amos, C.I., Risch, A., Chang-Claude, J., Heinrich, J., and Bickeboller, H. (2013). A network-based kernel machine test for the identification of risk pathways in genome-wide association studies. Hum. Hered. *76*, 64–75.

[35] Kwee, L.C., Liu, D., Lin, X., Ghosh, D., and Epstein, M.P. (2008). A powerful and flexible multilocus association test for quantitative traits. Am. J. Hum. Genet. *82*, 386–397.

[36] Wu, M.C., Kraft, P., Epstein, M.P., Taylor, D.M., Chanock, S.J., Hunter, D.J., and Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. Am. J. Hum. Genet. *86*, 929–942.

[37] Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.

[38] Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. Am. J. Hum. Genet. *92*, 841–853.

[39] Zhan, X., Girirajan, S., Zhao, N., Wu, M. C., and Ghosh, D. (2016). A novel copy number variants kernel association test with application to autism spectrum disorders studies. Bioinformatics *32*, 3603–3610.

[40] Tzeng, J. Y., Zhang, D., Chang, S. M., Thomas, D. C., and Davidian, M. (2009). Gene–Trait Similarity Regression for Multimarker–Based Association Analysis. Biometrics *65*, 822–832.

[41] Tzeng, J. Y., Zhang, D., Pongpanich, M., Smith, C., McCarthy, M. I., Sale, M. M., et al. (2011). Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am. J. Hum. Genet. *89*, 277-288

[42] Schaid, D.J. (2010a). Genomic similarity and kernel methods I: advancements by building on mathematical and statistical foundations. Hum. Hered. *70*, 109–131.

[43] Schaid, D.J. (2010b). Genomic similarity and kernel methods II: methods for genomic information. Hum. Hered. *70*, 132–140.

[44] Pan, W., Kwak, I. Y., and Wei, P. (2015). A powerful pathway-based adaptive test for genetic association with common or rare variants. Am. J. Hum. Genet. *97*, 86-98.

[45] Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Christiani, D. C., Wurfel, M. M., and Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am. J. Hum. Genet. *91*, 224–37.

[46] Chen, J., Chen, W., Zhao, N., Wu, M.C., and Schaid, D.J. (2016). Small sample kernel association test for human genetic and microbiome association studies. Genet. Epidemiol. *40*, 5–19.

[47] Josse, J., Pages, J., and Husson, F. (2008) Testing the significance of the RV coefficient. Comput. Stat. Data Anal. *53*, 82–91.

[48] Minas, C., Curry, E., and Montana, G. (2013). A distance-based test of association between paired heterogeneous genomic data. Bioinformatics *29*, 2555–2563.

[49] Zhan, X., Plantinga A., Zhao, N., and Wu, M. C. (2017). A fast small-sample kernel independence test for microbiome community-level association analysis Biometrics DOI: 10.1111/biom.12684.

[50] Kazi-Aoual, F., Hitier, S., Sabatier, R., and Lebreton, J.D. (1995). Refined approximations to permutation tests for multivariate inference. Comput. Stat. Data Anal. *20*, 643–656.

[51] Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics *63*, 1079–88.

[52] Liu, D., Ghosh, D., and Lin, X. (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.

[53] Davies, R. (1980). The distribution of a linear combination of chi-2 random variables. Appl. Stat. *29*, 323–333.

[54] Duchesne, P. and Lafaye de Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the liu-tang-zhang approximation and exact methods. Comput. Stat. Data Anal. *54*, 858–862.

[55] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics *9*, 432–441.

[56] Wessel, J., and Schork, N. J. (2006). Generalized genomic distancebased regression methodology for multilocus association analysis. Am. J. Hum. Genet., *79*, 792–806.

[57] Spencer, C.C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. *5*, e1000477.

[58] Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. *15*, 1576–1583.

[59] Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schlkopf, B., and Smola, A.J. (2008). A kernel statistical test of independence. Advances in Neural Information Processing Systems *21*, 585–592.

[60] Gillespie, C.F., Bradley, B., Mercer, K., Smith, A.K., Conneely, K., Gapen, M., Weiss, T., Schwartz, A.C., Cubells, J.F. and Ressler, K.J. (2009). Trauma expo-

sure and stress-related disorders in inner city primary care patients General Hospital Psychiatry *31*, 505-514.

[61] Almli, L. M., Fani, N., Smith, A. K., and Ressler, K. J. (2014). Genetic approaches to understanding post-traumatic stress disorder. Int. J. Neuropsychopharmacol. *17*, 355–370.

[62] Goodwin, R. D., Fischer, M. E., and Goldberg, J. (2007). A twin study of posttraumatic stress disorder symptoms and asthma. Am. J. Respir. Crit. Care. Med. *176*, 983–987.

[63] Wu, M. C., Maity, A., Lee, S., Simmons, E. M., Harmon, Q. E., Lin, X., et al. (2013). Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels. Genet. Epidemiol. *37*, 267-275.

[64] Zhan, X., Epstein, M.P., and Ghosh, D. (2015b). An adaptive genetic association test using double kernel machines. Stat. Biosci. *7*, 262–281.

[65] He, Q., Cai, T., Liu, Y., Zhao, N., Harmon, Q. E., Almli, L. M., et al. (2016). Prioritizing individual genetic variants after kernel machine testing using variable selection. Genet. Epidemiol. *40*, 722-731.

Figure Legends

Figure 1: **QQ plots**: $-\log_{10}$ QQ plots for DKAT, GAMuT, MSKAT and MSKAT2 under *Simulation I* with 500 samples. X-axis represents $-\log_{10}$ expected p-values and Y-axis represents $-\log_{10}$ observed p-values.

Figure 2: **Power under *Simulation I***: Power for DKAT (black), GAMuT (red), MSKAT (green) and MSKAT2 (blue). X-axis represents proportion of relevant traits ($\gamma = 10\%, 20\%, 30\%$) and Y-axis represents power.

Table 1: Empirical type I error rates (divided by the nominal significance level α) under **Simulation I**.

Σ	n	α	DKAT	GAMuT	MSKAT	MSKAT2
		10^{-3}	1.04	0.09	0.01	0.87
	500	10^{-4}	1.06	0.01	0	0.67
		10^{-5}	0.90	0	0	0.50
Σ_1						
		10^{-3}	0.92	0.31	0.19	0.93
	1000	10^{-4}	1.07	0.13	0.12	0.56
		10^{-5}	1.00	0	0	0.80
		10^{-3}	0.96	0.10	0.01	0.21
	500	10^{-4}	1.03	0.02	0	0.12
		10^{-5}	0.90	0	0	0.10
Σ_2						
		10^{-3}	1.06	0.23	0.27	0.37
	1000	10^{-4}	0.89	0.14	0.10	0.23
		10^{-5}	0.90	0	0	0.30

Figure 1

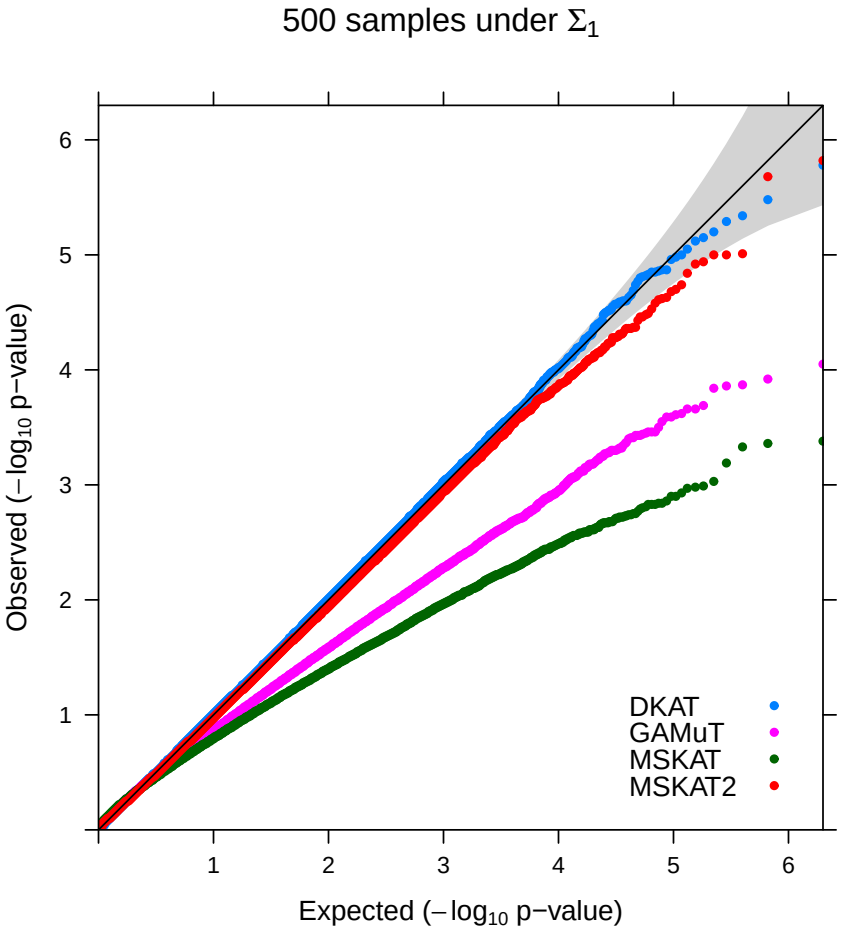
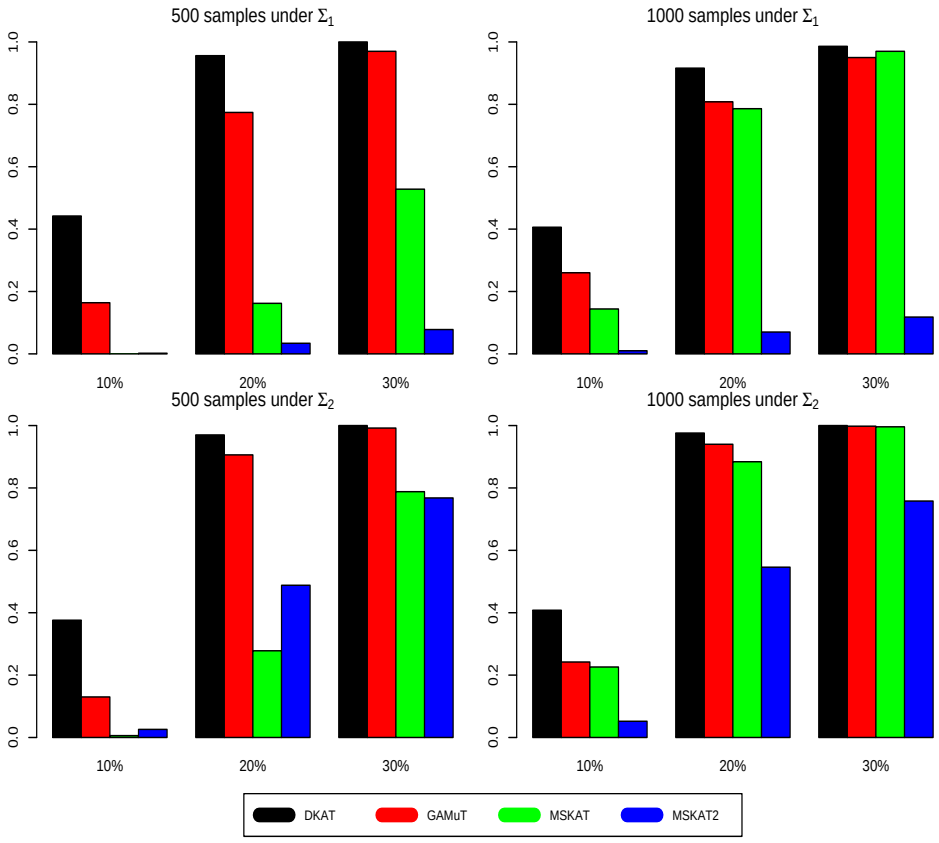

500 samples under Σ_1
Observed ($-\log_{10}$ p−value)
Expected ($-\log_{10}$ p−value)
1
2
3
4
5
6
DKAT
GAMuT
MSKAT
MSKAT2

Figure 2


500 samples under Σ_1
1000 samples under Σ_1
500 samples under Σ_2
1000 samples under Σ_2
0.0
0.2
0.4
0.6
0.8
1.0
10%
20%
30%
DKAT
GAMuT
MSKAT
MSKAT2