Kernel Machine SNP-set Testing under Multiple Candidate Kernels

Michael C. Wu[1], Arnab Maity[2], Seunggeun Lee[3], Elizabeth M. Simmons[1], Quaker E. Harmon[4], Xinyi Lin[3], Stephanie M. Engel[4], Jeffrey J. Molldrem[5], Paul M. Armistead[6]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC
[2]Department of Statistics, North Carolina State University, Raleigh, NC
[3]Department of Biostatistics, Harvard School of Public Health, Boston, MA
[4]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC
[5]Department of Stem Cell Transplantation and Cellular Therapy, MD Anderson Cancer Center, Houston, TX
[6]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC

December 31, 2012

Address for Correspondence:
Michael C. Wu
Department of Biostatistics
The University of North Carolina at Chapel Hill
4115C McGavran-Greenberg Hall, CB# 7420
Chapel Hill, NC 27599-7420
Phone: (919) 843-3656
Email: mwu@bios.unc.edu

Abstract

Joint testing for the cumulative effect of multiple single nucleotide polymorphisms grouped on the basis of prior biological knowledge has become a popular and powerful strategy for the analysis of large scale genetic association studies. The kernel machine (KM) testing framework is a useful approach that has been proposed for testing associations between multiple genetic variants and many different types of complex traits by comparing pairwise similarity in phenotype between subjects to pairwise similarity in genotype, with similarity in genotype defined via a kernel function. An advantage of the KM framework is its flexibility: choosing different kernel functions allows for different assumptions concerning the underlying model and can allow for improved power. In practice, it is difficult to know which kernel to use *a priori* since this depends on the unknown underlying trait architecture and selecting the kernel which gives the lowest p-value can lead to inflated type I error. Therefore, we propose practical strategies for KM testing when multiple candidate kernels are present based on constructing composite kernels and based on efficient perturbation procedures. We demonstrate through simulations and real data applications that the procedures protect the type I error rate and can lead to substantially improved power over poor choices of kernels and only modest differences in power versus using the best candidate kernel.

Key Words: Genetic association studies; kernel machines; multi-SNP analysis; similarity based testing; SNP sets.

1 Introduction

Advances in high-throughput biotechnology over the last decade have culminated in large scale genetic association studies which have facilitated discovery of over 1000 single nucleotide polymorphisms (SNPs) [Hindorff et al., 2009] associated with a range of complex traits. Typical analysis of genetic association studies involves single SNP analysis wherein individual SNPs are tested, one-by-one, for association with the trait while adjusting for confounders, such as the top principal components of genetic variability for population stratification [Price et al., 2006]. Standard procedures are applied to control for multiple comparisons. However, single SNP analysis is often underpowered due to the large number of individual variants, inability to capture the effect of ungenotyped SNPs that are tagged by genotyped variants, and difficulties in identifying multi-SNP and SNP-SNP interaction effects. This is often exacerbated by the limited availability of samples, modest effect sizes for most individual SNPs, and poor understanding of the genetic architecture underlying disease and complex trait etiology. To overcome many of these limitations, multi-SNP based analyses of genetic association studies, wherein multiple related (through proximity to a gene, pathway, functional group, etc.) SNPs are grouped into an *SNP set* and jointly tested using a global test, have emerged as powerful approaches for identification of gene variants that are associated with complex traits. SNP set analysis can offer many advantages over single SNP analysis due to its ability to capture the effect of ungenotyped SNPs that are tagged by the genotyped variants, to identify multi-marker effects, to reduce the number of multiple comparisons (ameliorating the stringent genome wide significance threshold), to allow for epistatic effects, and to make inference on biologically meaningful units.

Kernel machine testing [Liu et al., 2007, 2008] is a useful and operationally simple means for SNP set testing that has been successfully applied to identify SNP sets associated a range of disorders and traits [Liu et al., 2010, Lindstrom et al., 2010, Locke et al., 2010,

Monsees et al., 2011, Wu et al., 2011a, Shui et al., 2012, Meyer et al., 2012]. The principle behind the kernel machine test is that it defines genetic similarity through the use of a kernel function, a tool often seen within the framework of support vector machines [Cristianini and Shawe-Taylor, 2000]. The kernel function is a pairwise similarity metric that operates on the genotype values for every pair of individuals in the study. Then, like other similarity based approaches [Reiss et al., 2010, Schaid, 2010a,b, Wessel and Schork, 2006, Mukhopadhyay et al., 2010, Tzeng et al., 2009], the kernel machine test essentially compares pairwise similarity in genotype (of the SNPs in the SNP set) between individuals to pairwise similarity in trait value between individuals. High correspondence suggests association. We note that although our focus is on kernel machine based testing, many other other multi-marker tests for rare and common variants can be shown to be closely related to the kernel machine test [Pan, 2011] such that our approach generalizes to other similarity based tests as well.

The choice of kernel (similarity metric) can significantly impact the power to identify a significant SNP set. For example, when epistasis is present, kernel functions that accommodate nonlinearity such as the IBS kernel [Wessel and Schork, 2006] can sometimes offer improved power, but if no epistasis is present, using the linear kernel is often more powerful [Wu et al., 2010, Lin et al., 2011]. In practice however, information on the underlying genetic architecture is unknown — knowledge on the trait architecture would already preclude the need for conducting an analysis — and one needs to specify the kernel *a priori*. Tempting solutions such as picking the most significant p-value across the candidate kernels can lead to inflated type I error. Using permutation to correct for taking the minimum p-value is computationally expensive (see Supplemental Text) and will break possible correlations between the genotype values and the covariates, which is present in some studies, violating the exhangeability condition and resulting in incorrect type I error [Brown and Maritz, 1982, Anderson and Robinson, 2002, Good, 2004, Huang et al., 2006]. While residual based and parametric bootstrapping procedures can sometimes overcome the difficulties associated

with covariate adjustment, the computational expense is still high. Hence, it is an open problem of considerable practical interest to know, given a set of candidate kernels, how one can perform kernel machine testing.

We propose two simple, but effective, omnibus strategies for SNP set testing when multiple candidate kernels are applicable. First, we propose the use of composite kernels constructed as the simple weighted average of multiple kernels, with weights specified *a priori*. Second, we introduce a computationally efficient test based on perturbation of the score statistic. Both strategies still allow for easy covariate adjustment and simulations show that both correctly protect the type I error rate while maintaining high power in the omnibus. The methods differ in that the first operates by averaging candidate kernels while the second considers each kernel separately and takes the best. Emphasizing that our work does not seek to identify or evaluate the conditions under which particular kernels are most powerful (since this depends on quantities that are never known in reality), the main contribution of this project is to address a key challenge in application of kernel machine testing, as well as other related similarity based tests, by providing practical strategies when one has a set of candidate kernel functions.

2 Methods

2.1 Kernel Machine Test Under a Single Kernel

Kernel machine testing was first proposed within the gene expression framework [Liu et al., 2007, 2008] and was extended and adapted for testing assoications between multiple SNPs and individual complex traits [Kwee et al., 2008, Wu et al., 2010]. The approach has been extended to analysis of censored survival data [Cai et al., 2011, Lin et al., 2011] and multivariate data [Maity et al., 2012]. Recently, kernel machine based methods have been proposed for analysis of rare variants [Shriner and Vaughan, 2011, Basu and Pan, 2011, Wu et al.,

2011b, Lee et al., 2012b,a]. For simplicity, in this article we focus attention on testing the association of SNP sets comprised of common genetic variants with quantitative and dichotomous traits, but we note that our results are directly applicable to the rare variant analysis methods and the extended kernel machine methods as well.

Under the kernel machine regression framework, continuous (quantitative) traits can be related to the genotypes and any additional covariates through the semiparametric model:

$$y_i = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + h(\mathbf{Z}_i) + \varepsilon_i \tag{1}$$

where y_i denotes the trait value for the i^{th} person in the sample, $\mathbf{X}_i$ is a set of covariates for which we would like to control, and $\mathbf{Z}_i = [Z_{i1}, Z_{i2}, \ldots, Z_{ip}]'$ is the vector of genotype values for the p SNPs in the SNP set. Under the commonly used additive genetic model, each Z_{ij} is trinary variable equal to 0, 1, or 2 for non-carriers, heterozygotes, and homozygous carriers of the minor allele. Each ε_i is an error term with mean zero and variance σ^2, β_0 is an intercept, and $\boldsymbol{\beta}$ is the vector of regression coefficients for the covariates. Similarly, for case-control data, the model for risk of the dichotomous trait is given by:

$$\text{logit}P(y_i = 1|\mathbf{X}_i, \mathbf{Z}_i) = \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i + h(\mathbf{Z}_i) \tag{2}$$

where $\mathbf{X}_i$, $\mathbf{Z}_i$, β_0, and $\boldsymbol{\beta}$ are as before, but y_i is now a case-control indicator (0=control/1=case).

For both models $h(\cdot)$ is a function that has form defined only by a positive definite kernel function $K(\cdot, \cdot)$. The $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$ is a measure of the similarity between subjects i and i' based on the genotypes of the SNPs in the SNP set, and importantly, the kernel function fully specifies the relationship between the trait and the SNPs in the SNP set, and vice versa. For example, it can be shown that if $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \mathbf{Z}_i'\mathbf{Z}_{i'}$, called the linear kernel, then this implies that $h(\mathbf{Z}_i) = \boldsymbol{\alpha}'\mathbf{Z}_i$ for some vector of constants $\boldsymbol{\alpha}$, i.e. $h(\mathbf{Z}_i)$ is a linear function of

the SNPs in the SNP set. The converse is also true: setting $h(\mathbf{Z}_i) = \boldsymbol{\alpha}'\mathbf{Z}_i$ also implies that the kernel function is equal to the linear kernel. Hence, by selecting and changing the kernel function, one is implicitly selecting and changing the model being used.

Some examples of commonly used kernel functions for genotype data include:

- *Linear Kernel*: $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = \mathbf{Z}_i'\mathbf{Z}_{i'}$
- *Quadratic Kernel*: $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = (\mathbf{Z}_i'\mathbf{Z}_{i'} + 1)^2$
- *IBS Kernel*: $K(\mathbf{Z}_i, \mathbf{Z}_{i'}) = (2p)^{-1} \sum_{j=1}^{p} IBS(Z_{ij}, Z_{i'j}) = (2p)^{-1} \sum_{j=1}^{p} (2 - |Z_{ij} - Z_{i'j}|)$

Other kernels are possible with the sole condition that they need to satisfy Mercer's theorem [Cristianini and Shawe-Taylor, 2000].

The goal is to test whether the SNPs in the SNP set, $\mathbf{Z}$, are associated with the trait values, $\mathbf{y}$. Since the trait depends on the $\mathbf{Z}$ only through the function $h(\mathbf{Z})$, to test the null hypothesis that no variants in the SNP set are associated with the trait corresponds to testing whether $h(\mathbf{Z}) = 0$. For quantitative traits, we can test this by constructing the score-type statistic

$$Q = \frac{(\mathbf{y} - \widehat{\mathbf{y}}_0)'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{y}}_0)}{\widehat{\sigma}_0^2}$$

where $\widehat{\mathbf{y}}_0 = \widehat{\beta}_0 + \mathbf{X}\widehat{\boldsymbol{\beta}}$ with $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ estimated under the null hypothesis, i.e. under the model where $\mathbf{h} = 0$. Similarly, for dichotomous traits, the kernel machine test operates using the score-type statistic

$$Q = (\mathbf{y} - \widehat{\mathbf{y}}_0)'\mathbf{K}(\mathbf{y} - \widehat{\mathbf{y}}_0)$$

where $\widehat{y} = logit^{-1}(\widehat{\beta}_0 + \mathbf{X}\widehat{\boldsymbol{\beta}})$ with $\widehat{\beta}_0$ and $\widehat{\boldsymbol{\beta}}$ again estimated under the null hypothesis. Since all estimation is under the null, standard software for least squares and logistic regression may be used to estimate all parameters. $\mathbf{K}$ is the kernel matrix and has $(i, i')^{th}$ component $K(\mathbf{Z}_i, \mathbf{Z}_{i'})$. In fact, instead of specifying a particular kernel function, it is sufficient for $\mathbf{K}$ to be a positive semidefinite matrix.

In order to obtain a p-value for significance, it is straightforward to see that Q asymptotically follows an unknown mixture of χ^2_1 distributions. Specifically, we define $\widetilde{\mathbf{X}} = [\mathbf{1}, \mathbf{X}]$, $\mathbf{P}_0 = \mathbf{I} - \widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'$ for quantitative traits, and $\mathbf{P}_0 = \mathbf{D}_0 - \mathbf{D}_0\widetilde{\mathbf{X}}(\widetilde{\mathbf{X}}'\mathbf{D}_0\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\mathbf{D}_0$, where $\mathbf{D}_0 = \text{diag}(\widehat{p}_{0_i}(1 - \widehat{p}_{0_i}))$, for dichotomous traits. Then $Q \sim \sum \kappa_j \chi^2_1$ where the κ_j are the eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{K}\mathbf{P}_0^{1/2}$. We can approximate this distribution using a range of different methods such as moment matching [Liu et al., 2007, 2009] or exact methods based on inversion of the characteristic function [Davies, 1980, Duchesne and Lafaye De Micheaux, 2010].

2.2 Testing Under Multiple Candidate Kernels

The kernel machine based test requires specification of a kernel function or kernel matrix *a priori*. A number of kernels, have been successfully used in real data applications. However, in practice, it is often unclear which kernel to use. Here, we assume that a set of L candidate kernel functions $K_1(\cdot,\cdot), K_2(\cdot,\cdot), \ldots, K_L(\cdot,\cdot)$ with corresponding kernel matrices $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_L$ are under consideration. For instance, $K_1(\cdot,\cdot)$ could be the linear kernel, $K_2(\cdot,\cdot)$ the IBS, and $K_3(\cdot,\cdot)$ the quadratic. Then we develop two strategies for testing under this setting. The first method uses the simple weighted average of multiple kernels, with weights specified *a priori*. Using this composite kernel approach, a p-value can be calculated analytically. The second method uses the minimum p-value from the different candidate kernels and we provide a computationally efficient perturbation approach to control for having taken the minimum. Both methods allow for easy covariate adjustment.

2.2.1 Composite Kernels: Simple Kernel Averaging

Within the prediction based statistical learning literature, when ambiguity in the choice of kernel is present, composite kernels have been proposed as a reasonable compromise [Joachims et al., 2001, Szafranski et al., 2010]. In particular, given L kernel functions,

$K_1(\cdot,\cdot)$, $K_2(\cdot,\cdot)$, ..., $K_L(\cdot,\cdot)$, the composite kernel function evaluated at the genotypes for the i^{th} and i'^{th} subjects is given by:

$$K_C(\mathbf{Z}_i, \mathbf{Z}_{i'}) = w_1 K_1(\mathbf{Z}_i, \mathbf{Z}_{i'}) + w_2 K_2(\mathbf{Z}_i, \mathbf{Z}_{i'}) + \cdots + w_L K_L(\mathbf{Z}_i, \mathbf{Z}_{i'})$$

and the corresponding kernel matrix is found as

$$\mathbf{K}_C = \sum_{d=1}^{L} w_d \mathbf{K}_d$$

for some set of nonnegative weights $w_1, \ldots, w_L$. $\mathbf{K}_C$ is a valid kernel as long as $\mathbf{K}_1, \ldots, \mathbf{K}_L$ are valid. Note that the sum of the weights is not constrained.

Although considerable study has been devoted to estimation and prediction using composite kernels, limited work exists on how to test using composite kernels. For a set of fixed weights one may directly apply the kernel machine test treating the composite kernel as just another single kernel, but the challenge lies in selecting the weights. In the prediction setting, the weights are generally estimated, sometimes sparsely, from the data. Unfortunately, supervised estimation of the weights from the data will lead to inflated type I error rate. Hence, we propose to apply a simple scaling to ensure that the kernel functions are on the same scale and then use a simple average of the candidate kernels. In particular, we let $\gamma_d = \text{tr}\left\{\mathbf{P}_0^{1/2}\mathbf{K}_d\mathbf{P}_0^{1/2}\right\}$ where $\mathbf{P}_0$ is defined as before. Then we set $w_d = \frac{1}{\gamma_d}$. The scaling is necessary to ensure that no single kernel entirely dominates the composite metric. Alternative normalizing constants and choices for γ_d are possible.

2.2.2 Perturbation Based Inference

An alternative to composite kernel testing and averaging across a range of kernels is to compute a p-value under each candidate kernel, take the minimum, and then evaluate significance via permutation. However, as noted earlier, permutation creates difficulties in covariate ad-

justment since it requires the covariates to be uncorrelated with the SNPs in the SNP set which is known to be untrue in many situations: principal components of genetic variability are necessarily correlated and environmental risk factors may also be correlated with genotype. For example, association studies for lung cancer usually require adjustment for smoking, but smoking is also known to be associated with variants in many genes [Furberg et al., 2010]. Beyond the statistical challenges, permutation tends tends to be computationally expensive since it requires recomputing many estimated quantities. Therefore, we propose to take advantage of our knowledge of the asymptotic distribution of the quadratic forms for $Q_1, \ldots, Q_k$ and develop a strategy based on perturbation of the score statistic. Perturbation procedures, sometimes referred to as resampling or simply monte-carlo approaches, have been previously proposed by others [Lin, 2005, Conneely and Boehnke, 2007, Chapman and Whittaker, 2008, Pan et al., 2010]. Recently, Cai et al. [2012] apply the approach within the context of multivariable kernel machine SNP set testing. Our work is closely related to these procedures but differs in the need to accommodate the correlation that necessarily arises from simultaneous consideration of multiple kernels built on the same data which is not necessary for methods that are based on the marginal score statistics (such as minimum p-value based approaches) since the construction of the statistic naturally captures this.

The intuition behind our approach lies in the following. For quantitative traits, with large n, under H_0 the $(\mathbf{y} - \widehat{\mathbf{y}}_0)/\widehat{\sigma}$ are approximately standard normal. Then each $Q_d = (\mathbf{y} - \widehat{\mathbf{y}}_0)'\mathbf{K}_d(\mathbf{y} - \widehat{\mathbf{y}})/\widehat{\sigma}^2$ is essentially comprised of a vector of standard normal variables sandwiching a square matrix. The vectors of normals are the same across all $Q_1, \ldots, Q_k$. Thus, we can perturb each Q_d by replacing $(\mathbf{y} - \widehat{\mathbf{y}}_0)/\widehat{\sigma}$ with a new, common vector of normals to generate new score statistics. Since the vector of normals are the same and we are still using the same central kernel matrices as before, then we are essentially generating new data sets that are capturing the correlation between different kernels built on the same data without using permutation. We can generate a large number of new data sets by changing

the vector of normal variables and use this perturbation distribution to obtain a p-value across all the candidate kernels. The intuition is similar for the logistic scenario except we are using the working linear model.

Formally, to obtain a p-value for a SNP set using perturbation, we propose the following procedure:

1. For each candidate kernel, $\mathbf{K}_d$, we obtain the corresponding score statistic Q_d and p-value p_d.

2. Then find the minimum p-value $p^o = \min_{1 \leq d \leq k} p_d$.

3. For $d \in 1, \ldots, L$, compute $\mathbf{\Lambda}_d = diag(\lambda_{d,1}, \ldots, \lambda_{d,m_d})$, and $\mathbf{V}_d = [\mathbf{v}_{d,1}, \mathbf{v}_{d,2}, \ldots, \mathbf{v}_{d,m_d}]$ where $\lambda_{d,1} \geq \lambda_{d,2} \geq \ldots \geq \lambda_{d,m_d}$ are the m_d positive eigenvalues of $\mathbf{P}_0^{1/2}\mathbf{K}_d\mathbf{P}_0^{1/2}$ with corresponding eigenvectors $\mathbf{v}_{d,1}, \mathbf{v}_{d,2}, \ldots, \mathbf{v}_{d,m_d}$. Note that these quantities may already be calculated in step 1 to obtain a p-value based on each Q_d.

4. Obtain
$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{I} & \mathbf{V}_1'\mathbf{V}_2 & \cdots & \mathbf{V}_1'\mathbf{V}_L \\ \mathbf{V}_2'\mathbf{V}_1 & \mathbf{I} & \cdots & \mathbf{V}_2'\mathbf{V}_L \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{V}_L'\mathbf{V}_1 & \mathbf{V}_L'\mathbf{V}_2 & \cdots & \mathbf{I} \end{bmatrix}$$
and conduct a Cholesky decomposition on $\mathbf{\Sigma} = \mathbf{R}\mathbf{R}'$ where $\mathbf{L}$ is an $n \times m$ matrix with m being the rank of $\mathbf{\Sigma}$.

5. Generate $\mathbf{r} = [r_1, r_2, \ldots, r_m]'$ with each $r_j \sim N(0,1)$. We then obtain $\mathbf{r}^* = \mathbf{R}\mathbf{r}$. Then for the d^{th} kernel, we assign
$$\mathbf{r}_d^* = [r_a^*, r_{a+1}^*, \ldots, r_{a+m_d}^*]'$$
where $a = \sum_{j=1}^{d-1} m_j + 1$.

6. We can then compute $Q_d^* = \mathbf{r}_d^{*\prime} \mathbf{\Lambda}_d \mathbf{r}_d^*$ for each d and obtain a corresponding p-value, p_d^*, using parameters estimated for the original Q_d. We set $p^* = \min_{1 \leq d \leq k} p_d^*$.

7. We repeat steps (5)-(6) B times to obtain $p_{(1)}^*, p_{(2)}^*, \ldots, p_{(B)}^*$ for some large number B.

8. The final p-value for significance is estimated as

$$p = B^{-1} \sum_{b=1}^{B} I(p_{(b)}^* \leq p^o)$$

It is important to note that direct use of the p-value is necessary rather than using the maximum score statistic: testing under different kernels can yield tests that are dramatically different in terms of degrees of freedom, i.e. the raw statistics are often on completely different scales. On the other hand, p-values are scale free.

Although this strategy also generates a monte carlo p-value, the advantage over permutation is, first, our procedure retains any possible correlation between covariates and SNPs, and second, the procedure is far more computationally efficient. The latter is true because the computation now relies only on generating and then rotating m normal random variables. All other parameters remain the same. In contrast, permutation requires complete re-estimation of the kernel (or projection) matrices, eigendecompositions, and/or moment matching parameters. Detailed comparisons are described within the Supplemental Text.

The perturbation strategy is general and can be combined with the composite kernels to generate a p-value. For instance, one could construct candidate composite kernels using different choices of weights and then use perturbation to test across all of the choices of weights.

2.3 Simulations

2.3.1 Type I Error Rate

To demonstrate that the proposed methods are valid tests, in terms of protecting type I error, we conducted a series of simulations under null models for both continuous and dichotomous traits. Specifically, for all simulation settings, we generated 1,000,000 data sets comprised of n individuals under the null hypothesis that the SNP genotypes are not associated with the trait values. Genetic data were simulated by pairing haplotypes based on Gene I of Sha et al. [2005] which contains 10 individual common variants, all of which are treated as genotyped.

For simulations based on continuous traits, we generate (independent) covariates, $\mathbf{X}$, where $X_{i1} \sim N(29.2, 21.1)$ and $X_{i2} \sim bern(0.506)$. These covariate distributions are based on the same model used by Kwee et al. [2008]. The continuous traits are generated under the null model

$$y_i = 0.03X_{i1} + 0.5X_{i2} + \varepsilon_i$$

where $\varepsilon_i \sim N(0, 1)$. For simulations based on dichotomous traits, we generate data sets comprised of $n/2$ "cases" and $n/2$ "controls" using the model:

$$\text{logit}P(y_i = 1|\mathbf{X}_i) = -1.35 + 0.5X_{i1} + 0.3X_{i2}$$

where $X_{i1} \sim N(0, 1)$ and $X_{i2} \sim unif(0, 1)$. For both continuous and dichotomous trait simulations, we let the sample size vary as $n = 500, 1000, 2000$.

For each data set, we tested for an association between genotype and the trait while adjusting for the covariates $\mathbf{X}$. We apply kernel machine testing under the linear, IBS, and quadratic kernel using the Davies exact method [Davies, 1980] to obtain p-values for significance. We also apply the kernel machine test using composite kernels formed by averaging the linear and IBS kernels, and also composite kernels formed by averaging the

linear, IBS, and quadratic kernels. Finally, we also used perturbation (with $B = 100,000$) to determine the p-value across all three candidate kernels. The type I error rate was estimated as the proportion of p-values less than $\alpha = 0.05$ or 0.0005 across the 1,000,000 simulations.

2.3.2 Power

We conducted simulations under alternative models in order to assess the empirical power of the proposed approaches. Specifically, we considered four different settings and for each settings, we use 500 simulated data sets to estimate the empirical power. We focus, here, on simulations with quantitative traits and relegate similar simulations with dichotomous traits to the Supplemental Material.

For each setting considered, we simulated a set of genotypes, covariates, and quantitative traits. For the i^{th} individual in the data set, we denote the simulated genotype information as $\mathbf{S}_i$. Then we simulated continuous traits for $n = 500$ or $n = 1000$ individuals by again setting $X_{i1} \sim N(29.2, 21.1)$ and $X_{i2} \sim bern(0.506)$ as before, but then we generate the trait value using the alternative model:

$$y_i = 0.03X_{i1} + 0.5X_{i2} + \beta_G S_{ia} + \beta_G S_{ib} + \beta_I S_{ia} S_{ib} + \varepsilon_i \tag{3}$$

where S_{ia} and S_{ib} are the genotypes for the a^{th} and b^{th} SNPs in the SNP set , β_G and β_I are the coefficient values for the main genetic effects and the interaction effect.

For Settings 1 and 2, $\mathbf{S}_i$ was generated using the same method as was done for the type I error simulations by sampling and pairing haplotypes based on the 10 SNPs in Gene I of Sha et al. [2005]. Under Setting 1, we set $a = 1$, $b = 3$, $\beta_G = 0$ and $\beta_I = 0.1$, i.e. the first and third SNPs in the SNP set are causal with an interactive epistatic effect, but no separate main genetic effect. Setting 2 is similar to Setting 1, except we set $a = 1$, $b = 10$, and $\beta_G = \beta_I = 0.2$ such that the underlying model depends on the first and tenth SNPs,

but both interactive and main SNP effects are present. As with the type I error simulations, we let $\mathbf{Z} = \mathbf{S}$ which implies that all 10 SNPs in the SNP set are genotyped and grouped to form the SNP set.

In many candidate gene and GWAS studies, only a few tag SNPs are genotyped. Consequently, we followed the strategy of Wu et al. [2010], and in Settings 3 and 4 we conduct simulations by generating data based on 86 HAPMAP SNPs from the *ASAH1* gene using the HAPGEN software [Marchini et al., 2007]. We then generate the quantitative traits under the Model (3) with $\mathbf{S}$ denoting the genotypes for the 86 SNPs and $\mathbf{X}$ defined as before. Under Setting 3, we let $a = 1$, $b = 50$, $\beta_G = 0.1$ and $\beta_I = 0.2$. For Setting 4, we let $a = 1$, $b = 11$, $\beta_G = 0.1$ and $\beta_I = 0.1$. Different from earlier settings, we assume that only 14 of the 86 SNPs in the *ASAH1* gene are genotyped (those on the Illumina 550 SNP chip), so in contrast to the earlier power simulations, for Settings 3 and 4, we restrict $\mathbf{Z}$ to be the 14 genotyped variants. This better reflects real data since in many genotyping studies only some of the causal variants are likely to be observed. The first SNP among the 86 SNPs (rs7508) is genotyped but neither the eleventh nor the fiftieth SNPs (rs425010 and rs2299606, respectively) are genotyped.

For each of the four settings, we simulated 500 data sets. We then tested for an association between the SNPs in $\mathbf{Z}$ (all of the variants for Settings 1 and 2 and only the 14 genotyped variants for Settings 3 and 4) and the trait $\mathbf{y}$ while adjusting for the covariates $\mathbf{X}$. Specifically, for the continuous trait simulations, we applied the linear kernel machine test using the linear, IBS, and quadratic kernels. We also applied the proposed composite kernel methods based on direct kernel averaging of the linear, IBS, and quadratic kernels. Perturbation based on $B = 1000$ was also applied to search over the three candidate kernels. For each method under each setting, the power was estimated as the proportion of p-values less than $\alpha = 0.05$.

2.4 Candidate Gene Study of Pre-term Birth

We illustrate our methods via application to a sub-study of a broader genetic association study examining the role of a mother's (common) genetic polymorphisms on risk of pre-term birth, here defined as a live birth before 37 complete weeks of gestation. Specifically, in our sub-study, we examined the association between pre-term birth and genetic variability at 47 candidate genes (comprised of 735 tagSNPs) in the inflammation and apoptosis pathways within a sample of 863 singleton, live births from women of European ancestry in the Pregnancy, Infection and Nutrition Cohort [Savitz et al., 2001]. 153 of the infants were born pre-term and the remaining infants were born at term. To assess whether the the 47 candidate genes were associated with risk of pre-term birth, we conducted a gene level analysis by grouping the SNPs in and near each individual gene into SNP sets. Then we tested each SNP set for association with pre-term birth, adjusting for smoking and parity. Since we have little understanding of the underlying true model, we considered the linear, IBS, and quadratic kernels as candidate kernels. Specifically, we applied the logistic kernel machine test using the linear kernel, IBS kernel, quadratic kernel, and composite kernels with all three kernels averaged. We also used the perturbation method (based on $B = 10^4$). Significance was determined at the recommended FDR = 0.20 level [Efron, 2007].

We note that the kernel machine testing framework is closely related to many other multi-SNP tests such that using kernel machine testing under single kernels already allows for comparisons with some "alternative methods". For example, the kernel machine test under the IBS kernel is essentially a generalization of the Wessel and Schork approach [Wessel and Schork, 2006, Pan, 2011] that allows for covariate adjustment and analytic p-value computation. Similarly, the kernel machine test under the linear kernel is related to the Global Test [Goeman et al., 2004] and other variance component tests [Pan, 2009]. However, we also include further comparisons with two alternative multi-SNP tests that are used in practice. First, we also applied a variation of the PCA approach [Gauderman et al., 2007,

Zhao et al., 2012] to test for the association between the dichotomous outcome and the SNP set. Under the PCA approach, we collapse the SNPs in each SNP set into the first principal component and use logistic regression to regress the dichotomous outcome on the collapsed value to obtain a p-value. Second, we compared the kernel methods to the minimum p-value approach (MinP) [Chapman and Whittaker, 2008, Pan et al., 2010] wherein we compute the minimum single SNP analysis p-value for the SNPs in the SNP set and use permutation to adjust for having taken the minimum p-value. We applied both the PCA method and the MinP method (with 1000 permutations) to each of the SNP sets and again called significance at the FDR = 0.20 level.

3 Results

3.1 Simulation Results

The type I error rate simulations results are presented in Table i. For both continuous and dichotomous traits, the test size is correctly controlled for each of the methods at each of the considered sample sizes and α-levels.

The power simulation results for continuous traits are presented in Table ii. For each setting and sample size we present the power of each method relative to the optimal method. For example, under Setting 1, for a sample size of 500, the kernel machine testing with the IBS kernel leads to a power loss of 22% relative to kernel machine testing with the quadratic kernel. The absolute powers can be found in the Supplemental Table 1.

Results were generally consistent across sample sizes, though the absolute power improved as the sample size increased. The quadratic kernel had the highest power under Settings 1 and 2, followed by the linear kernel and using the IBS kernel yielded lowest power. For these two settings, averaging the three candidate kernels yielded improved power that was between the power from using the linear and quadratic kernels. Using of our perturbation

based method yielded power between the linear and quadratic kernels.

Under Setting 3, the IBS kernel had nearly double the power of both the linear and quadratic kernels. Using composite kernels by direct averaging of the IBS and linear kernel yielded power that was between the IBS and linear, though compared to Settings 1 and 2 the loss in power was considerably greater. Averaging across all three kernels resulted in lower power than averaging only the IBS and linear kernels. However, the power from using perturbation remained high and close to the power from using the IBS kernel.

In Setting 4, all three kernels performed similarly and using composite kernels based on averaging and perturbation did not change the power.

From these simulations, it is evident that depending on the underlying trait architecture, use of different kernels can yield differential power. Using composite kernels and perturbation allows for power that is intermediate among the candidate kernels. Restricting the composite kernels to a more focused set of kernels can yield improved power. Overall, using perturbation resulted in the higher power than using composite kernels. In fact, the power from using perturbation was generally closer to the power from using the optimal kernel (out of the candidates). We reiterate that these simulations are not meant to identify the situations under which individual methods are most powerful (since such knowledge is unavailable *a priori*) but rather to demonstrate the usefulness of the proposed methods within practical situations.

In the Supplemental Text, we also include simulations using dichotomous traits and also comparisons of the kernel methods with the PCA and MinP methods.

3.2 Data Analysis Results

Table iii presents the number of SNP sets significantly associated with pre-term birth using each of the methods as well as the number of overlapping significant SNP sets across methods. While the specific genes will be reported elsewhere, it is evident from Table iii that using the

Table i: Type I error rate results. Linear, IBS, and Quadratic correspond to the linear, IBS, and quadratic kernels. Average 2 corresponds to using the composite kernel generated as a direct average of the linear and IBS kernels while Average 3 denotes the composite kernel generated as the average of the linear, IBS, and quadratic kernel. Perturb denotes the proposed perturbation procedure using all three candidate kernels.

		Continuous Traits			Dichotomous Traits		
	n	500	1000	2000	500	1000	2000
$\alpha = 0.05$	Linear	0.049	0.050	0.050	0.051	0.050	0.049
	IBS	0.049	0.050	0.050	0.051	0.050	0.050
	Quadratic	0.049	0.050	0.050	0.050	0.050	0.050
	Average 2	0.051	0.051	0.051	0.051	0.051	0.053
	Average 3	0.049	0.050	0.050	0.050	0.049	0.051
	Perturb.	0.047	0.048	0.048	0.048	0.048	0.047
$\alpha = 0.0005$	Linear	0.00039	0.00050	0.00048	0.00046	0.00051	0.00053
	IBS	0.00040	0.00048	0.00049	0.00051	0.00050	0.00049
	Quadratic	0.00046	0.00054	0.00048	0.00046	0.00048	0.00053
	Average 2	0.00046	0.00050	0.00054	0.00048	0.00045	0.00046
	Average 3	0.00048	0.00044	0.00048	0.00051	0.00046	0.00052
	Perturb.	0.00041	0.00051	0.00050	0.00052	0.00046	0.00051

IBS kernel led to marginally more SNP sets being called significant than either the linear and quadratic quadratic kernels. However, only three of the four SNP sets significant using the IBS kernel were significant using the linear kernel and only none were significant using the quadratic kernel. Using the perturbation approach, we identified the same four SNP sets to be significant as the IBS, which is the optimal candidate kernel in this case. Using kernel averaging resulted in some power loss and only a single SNP set was called significant, though this was still better than using the worst kernel, in this case the quadratic kernel. Noting our data analysis is primarily illustrative, we also present results considered alternative significance criteria based on alternative nominal and FDR adjusted levels. These results are presented in Supplemental Tables 4-7. Overall, results are qualitatively similar, but we note that if the higher FDR = 0.25 level was used, individual kernels started yielding very different results (Supplemental Figure 5). However, under this scenario, any SNP set called

Table ii: Power simulation results based on the four configurations for continuous trait values. Linear, IBS, and Quadratic correspond to the linear, IBS, and quadratic kernels. Average corresponds to using the composite kernel generated as the average of the linear, IBS, and quadratic kernel. Perturb denotes the proposed perturbation procedure using all three candidate kernels. Power is expressed as the power relative to the best (of the three individual kernels).

n	Setting	Linear	IBS	Quadratic	Average	Perturb
500	1	0.93	0.78	1.00	0.96	0.93
	2	0.74	0.59	1.00	0.82	0.88
	3	0.56	1.00	0.56	0.66	0.94
	4	1.00	0.93	1.00	1.00	1.00
1000	1	0.96	0.90	1.00	0.97	0.98
	2	0.80	0.70	1.00	0.88	0.94
	3	0.66	1.00	0.69	0.80	0.96
	4	1.00	0.96	1.00	1.00	1.00

significant under individual kernel testing was also called significant using the perturbation approach.

Also included in Table iii are the comparisons with the PCA and MinP approaches. The MinP approach fails to identify any SNP sets as significant and the PCA methods identifies only a single SNP set.

Of the four SNP sets significant at the FDR = 0.20 level, only one has been previously shown to be associated with pre-term birth. Within our data set, the gene containing 26 SNPs was significant using the linear kernel ($p = 0.013$, FDR = 0.20), the IBS kernel ($p = 0.004$, FDR = 0.11), and perturbation analysis ($p = 0.004$, FDR = 0.09). The gene was not significant using the quadratic kernel ($p = 0.022$, FDR = 0.25), PCA ($p = 0.903$, FDR = 0.99), nor MinP ($p = 0.036$, FDR = 0.42). If we were to increase the FDR level to 0.25, then then an additional gene previously shown to be associated with pre-term birth would also have been called significant by using the linear kernel, the quadratic kernel, kernel averaging, perturbation analysis, and PCA analysis, but not using the IBS kernel nor the

Table iii: Pre-term Birth Data Analysis Results. Each number represents the number of SNP sets called significant at the FDR = 0.20 level by both the method along the top and the side of the table. Diagonal elements represent the number of SNP sets called significant by each individual method.

	Linear	IBS	Quadratic	Average	Perturb.	PCA	MinP
Linear	3	3	0	1	3	1	0
IBS		4	0	1	4	1	0
Quadratic			0	0	0	0	0
Average				1	1	1	0
Perturb.					4	1	0
PCA						1	0
MinP							0

MinP methods.

The difference in the results between different individual kernels was in some cases quite large. For example, the most significant gene in the apoptosis pathway contained 22 SNPs and had a marginally significant p-value (at the nominal level) of 0.034 or 0.010 using the linear or IBS kernels, respectively. On the other hand, using the quadratic kernel yielded a clearly nonsignificant p-value of 0.206. Using a composite kernel would lead to a p-value of 0.042 and using the perturbation procedure leads to a p-value of 0.016 which is close to the p-value from the optimal candidate kernel in this scenario. Collectively, our data analysis results indicate that although there is some sacrifice in power for looking across multiple kernels, our proposed omnibus testing procedure can still maintain high power across scenarios.

4 Discussion

Our work addresses a key gap in the statistical genetics literature concerning the choice of similarity metric in similarity based testing. Focusing on the choice of kernel in kernel machine testing, we proposed two simple strategies for inference when multiple candidate

kernels are available. Our solution is based purely on practical considerations. In particular, we make no claims as to understanding which kernel works best under which scenario since this depends on the (unknown) true state of nature. Instead, we acknowledge the limited availability of prior knowledge concerning genetic architecture and develop omnibus methods to consider a range of kernels, each of which functions best under different scenarios.

In general, both strategies we propose offer improved power over weaker choices of kernel but only slightly lower power than using the optimal kernel, chosen from the set of considered candidates. However, using perturbation appears to offer somewhat improved power. On the other hand, though perturbation is considerably faster than permutation since it does not require re-estimation of numerous matrices and parameters, it is still computationally more expensive than kernel averaging. Hence if significance needs to be rapidly estimated at a low α-level, then using composite kernels may be the better choice. If the results are less urgent, then perturbation may be the better choice. Generally, the computational burden may be reduced by initially using a modest value for B, estimating a p-value, and refining the p-value with a larger B if the initial estimate suggests need for more accurate estimation.

Both of our proposed strategies are valid and protect the type I error, even if many (or even all) of the considered candidate kernels do not reflect the underlying model. However, inclusion of many suboptimal kernels can lead to loss in power. The kernel averaging approach will lose power if many of the candidate kernels are far from the optimal kernel due to contamination of the signal. It may be possible to identify weights that use information on the similarity between kernels or that optimally weight different kernels using the outcome, but such a solution remains elusive. The perturbation approach is more robust in that if a few suboptimal kernels are included and all very similar, then the loss in power from using a few suboptimal kernels will be small since the perturbation approach accommodates the correlation between the kernels. However, perturbation will still lose power if too many suboptimal kernels are included and particularly if they are all very different. Therefore,

although one can search across a wide range of kernels, we recommend identifying a few reasonable kernels prior to using either approach.

Recently, Kim et al. [2012] explored the semiparametric model underlying the kernel machine approach within a bayesian framework and explored inferential procedures. While the work centered around gene expression analysis, they showed that bayesian views of the problem were very natural and could result in improved performance over frequentist takes on the work. Furthermore, they developed a strategy for selecting individual kernels using bayes factors, though further work is needed to explore how to conduct inference with multiple kernels. Consequently, exploration of relationships between our approach and possible bayesian approaches is an area of potential future research.

Although we restricted our work to studying kernel machine testing, our work has implications for other similarity based tests as well. In particular, the haplotype based similarity test of Tzeng and Zhang [2007], the MDMR statistic of Wessel and Schork [2006] and distance test of Reiss et al. [2010], as well as others are equivalent to the kernel machine based tests under certain conditions. Therefore, when there is uncertainty in the choice of similarity metrics for these tests, our proposed methods can also be used suggesting that our general approach is widely applicable to a range of tests.

Acknowledgements

This research was supported in part by NIH grants P30 ES010126, R00 ES017744, KL2 RR025746, P50 CA100632, R01 HD058008, R21 HD060207, T32 ES07018, KL2 TR000084, and Leukemia & Lymphoma Society grant SCOR 7262-08. Additional research support was provided by an ASCO Young Investigator Award and Rattay Advanced Research Fellowship. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors report no conflict of interests.

References

M.J. Anderson and J. Robinson. Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43(1):75–88, 2002.

S. Basu and W. Pan. Comparison of statistical tests for disease association with rare variants. *Genetic Epidemiology*, 35(7):606–619, 2011.

BM Brown and JS Maritz. Distribution-free methods in regression. *Australian Journal of Statistics*, 24(3):318–331, 1982.

T. Cai, G. Tonini, and X. Lin. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*, 67(3):975–986, 2011.

T. Cai, X. Lin, and R.J. Carroll. Identifying genetic marker sets associated with phenotypes via an efficient adaptive score test. *Biostatistics*, 13(4):776–790, 2012.

J. Chapman and J. Whittaker. Analysis of multiple snps in a candidate gene or region. *Genetic Epidemiology*, 32(6):560–566, 2008.

K.N. Conneely and M. Boehnke. So many correlated tests, so little time! rapid adjustment of p-values for multiple correlated tests. *The American Journal of Human Genetics*, 81 (6):1158–1168, 2007.

N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge Univ Pr, 2000. ISBN 0521780195.

R. B. Davies. Algorithm AS 155: The distribution of a linear combination of chi-square random variables. *Applied Statistics*, 29:323–333, 1980.

P. Duchesne and P. Lafaye De Micheaux. Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics & Data Analysis*, 54(4):858–862, 2010.

B. Efron. Correlation and large-scale simultaneous significance testing. *Journal of the American Statistical Association*, 102(477):93–103, 2007.

H. Furberg, Y. Kim, J. Dackor, E. Boerwinkle, N. Franceschini, D. Ardissino, L. Bernardinelli, P.M. Mannucci, F. Mauri, P.A. Merlini, et al. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature Genetics*, 42(5):441–447, 2010.

W.J. Gauderman, C. Murcray, F. Gilliland, and D.V. Conti. Testing association between disease and multiple snps in a candidate gene. *Genetic Epidemiology*, 31(5):383–395, 2007.

J.J. Goeman, S.A. Van De Geer, F. De Kort, and H.C. Van Houwelingen. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99, 2004.

P.I. Good. *Permutation, parametric, and bootstrap tests of hypotheses*. Springer, 2004.

L.A. Hindorff, P. Sethupathy, H.A. Junkins, E.M. Ramos, J.P. Mehta, F.S. Collins, and T.A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23): 9362, 2009. ISSN 0027-8424.

Y. Huang, H. Xu, V. Calian, and J.C. Hsu. To permute or not to permute. *Bioinformatics*, 22(18):2244–2248, 2006.

T. Joachims, N. Cristianini, and J. Shawe-Taylor. Composite kernels for hypertext categorisation. In *Proceedings of the International Conference on Machine Learning, ICML'01*, pages 250–257. Morgan Kaufmann, 2001.

I. Kim, H. Pang, and H. Zhao. Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes. *Statistics in Medicine*, 31 (15):1633–51, 2012.

L.C. Kwee, D. Liu, X. Lin, D. Ghosh, and M.P. Epstein. A powerful and flexible multilocus association test for quantitative traits. *The American Journal of Human Genetics*, 82(2): 386–397, 2008. ISSN 0002-9297.

S. Lee, M.J. Emond, M.J. Bamshad, K.C. Barnes, M.J. Rieder, D.A. Nickerson, D.C. Christiani, M.M. Wurfel, and X. Lin. 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*, 2012a.

S. Lee, M.C. Wu, and X. Lin. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 2012b.

DY Lin. An efficient monte carlo approach to assessing statistical significance in genomic studies. *Bioinformatics*, 21(6):781–787, 2005.

X. Lin, T. Cai, M.C. Wu, Q. Zhou, G. Liu, D.C. Christiani, and X. Lin. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. *Genetic Epidemiology*, pages 82–93, 2011.

S. Lindstrom, D.J. Hunter, H. Gronberg, P. Stattin, F. Wiklund, J. Xu, S.J. Chanock, R. Hayes, and P. Kraft. Sequence Variants in the TLR4 and TLR6-1-10 Genes and Prostate Cancer Risk. Results Based on Pooled Analysis from Three Independent Studies. *Cancer Epidemiology Biomarkers & Prevention*, 19(3):873, 2010. ISSN 1055-9965.

C. Liu, M.C. Wu, F. Chen, M. Ter-Minassian, K. Asomaning, R. Zhai, Z. Wang, L. Su, R.S. Heist, M.H. Kulke, et al. A Large-scale genetic association study of esophageal adenocarcinoma risk. *Carcinogenesis*, 31(7):1259–1263, 2010. ISSN 0143-3334.

D. Liu, X. Lin, and D. Ghosh. Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models. *Biometrics*, 63 (4):1079–1088, 2007. ISSN 1541-0420.

D. Liu, D. Ghosh, and X. Lin. 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(1):292, 2008. ISSN 1471-2105.

H. Liu, Y. Tang, and H.H. Zhang. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis*, 53(4):853–856, 2009. ISSN 0167-9473.

A.E. Locke, K.J. Dooley, S.W. Tinker, S.Y. Cheong, E. Feingold, E.G. Allen, S.B. Freeman, C.P. Torfs, C.L. Cua, M.P. Epstein, et al. Variation in folate pathway genes contributes to risk of congenital heart defects among individuals with Down syndrome. *Genetic Epidemiology*, 34(6):613–623, 2010.

A. Maity, P.F. Sullivan, and J. Tzeng. Multivariate phenotype association analysis by marker-set kernel machine regression. *Genetic Epidemiology*, 2012.

J. Marchini, B. Howie, S. Myers, G. McVean, and P. Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7): 906–913, 2007. ISSN 1061-4036.

N.J. Meyer, Z.J. Daye, M. Rushefski, R. Aplenc, P.N. Lanken, M.G.S. Shashaty, J.D. Christie, and R. Feng. Snp-set analysis replicates acute lung injury genetic risk factors. *BMC Medical Genetics*, 13(1):52, 2012.

G.M. Monsees, P. Kraft, S.J. Chanock, D.J. Hunter, and J. Han. Comprehensive screen of genetic variation in dna repair pathway genes and postmenopausal breast cancer risk. *Breast Cancer Research and Treatment*, 125(1):207–214, 2011.

I. Mukhopadhyay, E. Feingold, D.E. Weeks, and A. Thalamuthu. Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genetic Epidemiology*, 34(3):213–221, 2010. ISSN 1098-2272.

W. Pan. Asymptotic tests of association with multiple snps in linkage disequilibrium. *Genetic epidemiology*, 33(6):497–507, 2009.

W. Pan. Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genetic Epidemiology*, 2011. ISSN 1098-2272.

W. Pan, F. Han, and X. Shen. Test selection with application to detecting disease association with multiple snps. *Human Heredity*, 69(2):120–130, 2010.

A.L. Price, N.J. Patterson, R.M. Plenge, M.E. Weinblatt, N.A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

P.T. Reiss, M.H.H. Stevens, Z. Shehzad, E. Petkova, and M.P. Milham. On Distance-Based Permutation Tests for Between-Group Comparisons. *Biometrics*, 66(2):636–643, 2010. ISSN 1541-0420.

D.A. Savitz, N. Dole, J.W. Terry Jr, H. Zhou, and J.M. Thorp Jr. Smoking and pregnancy outcome among african-american and white women in central north carolina. *Epidemiology*, 12(6):636, 2001.

D.J. Schaid. Genomic similarity and kernel methods I: Advancements by building on mathematical and statistical foundations. *Human Heredity*, 70(2):109–131, 2010a. ISSN 0001-5652.

D.J. Schaid. Genomic similarity and kernel methods II: methods for genomic information. *Human Heredity*, 70(2):132–140, 2010b. ISSN 0001-5652.

Q. Sha, J. Dong, R. Jiang, and S. Zhang. Tests of association between quantitative traits and haplotypes in a reduced-dimensional space. *Annals of Human Genetics*, 69(6):715–732, 2005. ISSN 1469-1809.

D. Shriner and L.K. Vaughan. A unified framework for multi-locus association analysis of both common and rare variants. *BMC Genomics*, 12(1):89, 2011. ISSN 1471-2164.

I.M. Shui, L.A. Mucci, P. Kraft, R.M. Tamimi, S. Lindstrom, K.L. Penney, K. Nimptsch, B.W. Hollis, N. DuPre, E.A. Platz, et al. Vitamin d–related genetic variation, plasma vitamin d, and risk of lethal prostate cancer: A prospective nested case–control study. *Journal of the National Cancer Institute*, 104(9):690–699, 2012.

M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. *Machine Learning*, 79(1):73–103, 2010. ISSN 0885-6125.

J.Y. Tzeng and D. Zhang. Haplotype-based association analysis via variance-components score test. *The American Journal of Human Genetics*, 81(5):927–938, 2007. ISSN 0002-9297.

J.Y. Tzeng, D. Zhang, S.M. Chang, D.C. Thomas, and M. Davidian. Gene-Trait Similarity Regression for Multimarker-Based Association Analysis. *Biometrics*, 65(3):822–832, 2009. ISSN 1541-0420.

J. Wessel and N.J. Schork. Generalized genomic distance-based regression methodology for multilocus association analysis. *The American Journal of Human Genetics*, 79(5):792–806, 2006. ISSN 0002-9297.

I.C. Wu, Y. Zhao, R. Zhai, G. Liu, M. Ter-Minassian, K. Asomaning, L. Su, C. Liu, F. Chen, M.H. Kulke, et al. Association between polymorphisms in cancer-related genes and early onset of esophageal adenocarcinoma. *Neoplasia (New York, NY)*, 13(4):386, 2011a.

M.C. Wu, P. Kraft, M.P. Epstein, D.M. Taylor, S.J. Chanock, D.J. Hunter, and X. Lin. Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. *The American Journal of Human Genetics*, 86(6):929–942, 2010.

M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin. Rare variant association testing for sequencing data with the sequence kernel association test (SKAT). *The American Journal of Human Genetics*, 89:82–93, 2011b.

Y. Zhao, F. Chen, R. Zhai, X. Lin, N. Diao, and D.C. Christiani. Association test based on snp set: Logistic kernel machine based test vs. principal component analysis. *PLOS ONE*, 7(9):e44978, 2012.

Kernel Machine SNP-set Testing under Multiple Candidate Kernels Supplemental Material

Michael C. Wu[1], Arnab Maity[2], Seunggeun Lee[3], Elizabeth M. Simmons[1], Quaker E. Harmon[4], Xinyi Lin[3], Stephanie M. Engel[4], Jeffrey J. Molldrem[5], Paul M. Armistead[6]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC
[2]Department of Statistics, North Carolina State University, Raleigh, NC
[3]Department of Biostatistics, Harvard School of Public Health, Boston, MA
[4]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC
[5]Department of Stem Cell Transplantation and Cellular Therapy, MD Anderson Cancer Center, Houston, TX
[6]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC

December 31, 2012

Address for Correspondence:
Michael C. Wu
Department of Biostatistics
The University of North Carolina at Chapel Hill
4115C McGavran-Greenberg Hall, CB# 7420
Chapel Hill, NC 27599-7420
Phone: (919) 843-3656
Email: mwu@bios.unc.edu

Power Simulations for Dichotomous Traits

Simulation Scenarios

We conducted simulations under alternative models in order to assess the empirical power of the proposed approaches with dichotomous traits. As with quantitative trait simulations, we considered four different settings and for each settings, we use 500 simulated data sets to estimate the empirical power.

For each setting considered, we simulated a set of genotypes, covariates, and dichotomous outcome traits. For the i^{th} individual in the data set, we denote the simulated genotype information as $\mathbf{S}_i$. Then we simulated continuous traits for $n = 1000$ or $n = 2000$ individuals by again setting $X_{i1} \sim N(29.2, 21.1)$ and $X_{i2} \sim bern(0.506)$ as before, but then we sampled half $(n/2)$ of the individuals to have dichotomous trait $y = 1$ and $n/2$ individuals to have $y = 0$ using the alternative model:

$$\text{logit}P(y_i = 1|\mathbf{X}_i, \mathbf{S}_i) = -1.35 + 0.5X_{i1} + 0.3X_{i2} + \beta_G S_{ia} + \beta_G S_{ib} + \beta_I S_{ia} S_{ib}. \qquad (1)$$

where S_{ia} and S_{ib} are the genotypes for the a^{th} and b^{th} SNPs in the SNP set, β_G and β_I are the coefficient values for the main genetic effects and the interaction effect.

For Settings 1 and 2, $\mathbf{S}_i$ was generated using the same method as was done for the type I error simulations by sampling and pairing haplotypes based on the 10 SNPs in Gene I of Sha et al. (2005). Under Setting 1, we set $a = 1$, $b = 10$, $\beta_G = 0.2$ and $\beta_I = 0$, i.e. the first and tenth SNPs in the SNP set are causal with only main effects. Setting 2 is similar to Setting 1, except we set $a = 3$, $b = 10$, and $\beta_G = 0$ and $\beta_I = 0.2$ such that the underlying model depends on the interaction between the third and tenth SNPs but not the main SNP effects. As with the type I error simulations, we let $\mathbf{Z} = \mathbf{S}$ which implies that all 10 SNPs in the SNP set are genotyped and grouped to form the SNP set.

Settings 3 and 4 follow a similar rationale to that used for quantitative traits and we conduct simulations by generating data based on 86 HAPMAP SNPs from the *ASAH1* gene using the HAPGEN software. We then generate the dichotomous traits under the Model (1) with $\mathbf{S}$ denoting the genotypes for the 86 SNPs and $\mathbf{X}$ defined as before. Under Setting 3, we let $a = 1$, $b = 50$, $\beta_G = 0.1$ and $\beta_I = 0.3$. For Setting 4, we let $a = 1$, $b = 11$, $\beta_G = 0.2$ and $\beta_I = 0.3$. We again assume that only 14 of the 86 SNPs in the *ASAH1* gene are genotyped (those on the Illumina 550 SNP chip), so we restrict $\mathbf{Z}$ to be the 14 genotyped variants.

For each of the four settings, we simulated 500 data sets. We then tested for an association between the SNPs in $\mathbf{Z}$ (all of the variants for Settings 1 and 2 and only the 14 genotyped variants for Settings 3 and 4) and the trait $\mathbf{y}$ while adjusting for the covariates $\mathbf{X}$. Specifically, for the continuous trait simulations, we applied the linear kernel machine test using the linear, IBS, and quadratic kernels. We also applied the proposed composite kernel methods based on direct kernel averaging of the linear, IBS, and quadratic kernels. Perturbation based on $B = 1000$ was also applied to search over the three candidate kernels. For each method under each setting, the power was estimated as the proportion of p-values less than $\alpha = 0.05$.

Simulation Results

The power simulation results for dichotomous traits are presented in Supplemental Tables 2 and 3. Supplemental Table 2 shows the power of each method relative to the optimal method across the different settings and sample sizes considered. The absolute powers are found in the Supplemental Table 3. Generally, results are qualitatively similar to the results using quantitative traits in that: (1) different kernels can yield very different power; (2) the proposed methods, particularly using perturbation, allows for good power in the omnibus, sacrificing a just little bit of power relative to the optimal method, but dominating poor choices of kernels.

Comparisons with Competing Methods

Within the power simulations for dichotomous traits, we also compared the power of the proposed kernel averaging and perturbation methods to the PCA and MinP methods (described in the main text). In particular, for each of the simulated data sets in each of the four simulation scenarios considered, we also applied the PCA and the MinP method (with 1000 permutations) to assess the association between the the dichotomous outcome and the SNP set while adjusting for covariates. Note that permutation is valid in this case since the covariates were simulated independently of the SNP information. We assessed the power again at the $\alpha = 0.05$ level.

The results of the PCA and MinP methods are also presented in Supplemental Tables 2 and 3. In general, under the considered simulation scenarios, the kernel based approaches tended to offer improved power over the PCA and MinP methods. As a supervised approach, the MinP method outperformed the PCA approach under Settings 1-3 but lost power in Setting 4. The higher power of the kernel methods is expected since the simulation settings involve multiple causal variants. We emphasize, however, that the purpose of this study is not to systematically compare methods, but rather to focus on the kernel machine setting and illustrate an omnibus testing approach that simultaneously considers multiple methods.

In addition to the PCA and MinP methods, we further note that due to the close relationship between kernel machine methods and other multi-SNP tests, the kernel machine tests under individual kernels correspond already to a wide range of alternative tests. In particular, the kernel machine test under the linear kernel is closely related to the variance component tests of Goeman et al. and Pan et al. Similarly, the distance based test of Wessel and Schork under their allele sharing kernel is essentially equivalent to the kernel machine test under the IBS kernel. Consequently, by considering the single kernel tests, our simulations simultaneously encompass several other well known multi-SNP testing approaches due to the equivalence between different approaches and the kernel machine tests.

Comparing Perturbation with Permutation

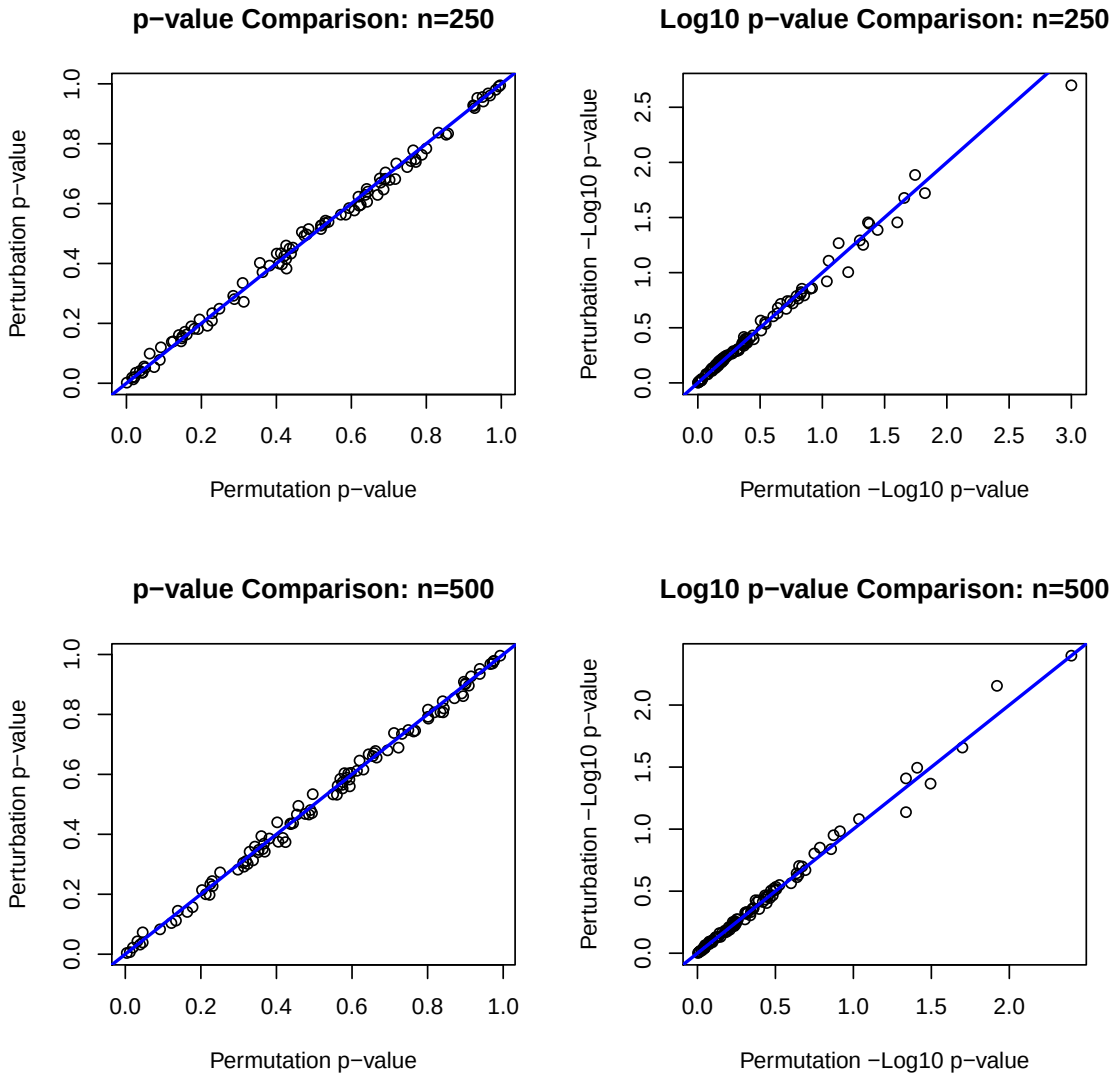
We briefly compared the use of permutation to the use of perturbation. Within the context of quantitative traits, we simulated 100 data sets null data sets (as in the type I error

simulations) for $n = 250$ or 500 individuals. Then we analyzed each data set using both perturbation with the linear, IBS, and quadratic kernels and also permutation with the linear, IBS, and quadratic kernels. We compared the p-values obtained under perturbation and permutation. We also estimated the average runtime for computing a p-value using 1000 perturbations vs. 1000 permutations. Covariates in the simulations were independent of the genotype data such that permutation is valid in this case.

The p-values from using permutation are plotted against the p-values from using perturbation in Supplemental Figure 1. As expected, the p-values are very close with differences mainly being due to monte-carlo error. In terms of run time, on a single processor of a Linux cluster with 4 Gb of reserved memory, the average time to analyze a single data set for our proposed perturbation procedure is 0.59 seconds and 0.70 seconds for $n = 250$ and 500, respectively. In contrast, use of permutations takes approximately 210.02 seconds and 1734.93 seconds for $n = 250$ and 500, respectively. The advantage of the perturbation procedure results from our exploitation of the distribution of the kernel machine score statistic.

Supplemental Figures

Supplemental Figure 1. Comparisons of p-values from perturbation vs. permutation for sample sizes of 250 and 500.



Supplemental Tables

Supplemental Table 1: Absolute power simulation results based on the four configurations for continuous trait values. Linear, IBS, and Quadratic correspond to the linear, IBS, and quadratic kernels. Average corresponds to using the composite kernel generated as the average of the linear, IBS, and quadratic kernel. Perturb denotes the proposed perturbation procedure using all three candidate kernels.

n	Setting	Linear	IBS	Quadratic	Average	Perturb
500	1	0.51	0.43	0.55	0.53	0.51
	2	0.25	0.20	0.34	0.28	0.30
	3	0.18	0.32	0.18	0.21	0.30
	4	0.28	0.26	0.28	0.28	0.28
1000	1	0.87	0.82	0.91	0.88	0.89
	2	0.55	0.48	0.69	0.61	0.65
	3	0.46	0.70	0.48	0.56	0.67
	4	0.50	0.48	0.50	0.50	0.50

Supplemental Table 2: Relative power simulation results based on the four configurations for dichotomous trait values. Linear, IBS, and Quadratic correspond to the linear, IBS, and quadratic kernels. Average corresponds to using the composite kernel generated as the average of the linear, IBS, and quadratic kernel. Perturb denotes the proposed perturbation procedure using all three candidate kernels. Power is expressed as the power relative to the best (of the three individual kernels). Also included are the power comparisons with two competing approaches for multi-SNP analysis using the PCA or minimum p-value methods.

n	Setting	Linear	IBS	Quadratic	Average	Perturb	PCA	MinP
1000	1	0.96	0.81	1.00	0.96	0.91	0.49	0.73
	2	0.83	0.70	1.00	0.89	0.93	0.44	0.67
	3	0.53	1.00	0.53	0.63	0.93	0.23	0.27
	4	1.00	0.96	1.00	1.00	0.98	0.98	0.25
2000	1	1.00	0.98	0.99	1.00	0.98	0.59	0.93
	2	0.97	0.94	1.00	0.97	0.98	0.54	0.87
	3	0.54	1.00	0.58	0.72	0.96	0.09	0.33
	4	1.00	1.00	1.00	1.00	1.00	0.96	0.32

Supplemental Table 3: Absolute power simulation results based on the four configurations for dichotomous trait values. Linear, IBS, and Quadratic correspond to the linear, IBS, and quadratic kernels. Average corresponds to using the composite kernel generated as the average of the linear, IBS, and quadratic kernel. Perturb denotes the proposed perturbation procedure using all three candidate kernels. Also included are the power comparisons with two competing approaches for multi-SNP analysis using the PCA or minimum p-value methods.

n	Setting	Linear	IBS	Quadratic	Average	Perturb	PCA	MinP
1000	1	0.67	0.57	0.70	0.67	0.64	0.34	0.51
	2	0.58	0.49	0.70	0.62	0.65	0.31	0.47
	3	0.16	0.30	0.16	0.19	0.28	0.07	0.08
	4	0.51	0.49	0.51	0.51	0.50	0.50	0.13
2000	1	0.90	0.88	0.89	0.90	0.88	0.53	0.84
	2	0.94	0.91	0.97	0.94	0.95	0.52	0.84
	3	0.36	0.67	0.39	0.48	0.64	0.06	0.22
	4	0.72	0.72	0.72	0.72	0.72	0.69	0.23

Supplemental Table 4: Pre-term Birth Data Analysis Results. Each number represents the number of SNP sets called significant at the FDR = 0.10 level by both the method along the top and the side of the table. Diagonal elements represent the number of SNP sets called significant by each individual method.

	Linear	IBS	Quadratic	Average	Perturb.	PCA	MinP
Linear	0	0	0	0	0	0	0
IBS		1	0	0	1	1	0
Quadratic			0	0	0	0	0
Average				0	0	0	0
Perturb.					2	1	0
PCA						1	0
MinP							0

Supplemental Table 5: Pre-term Birth Data Analysis Results. Each number represents the number of SNP sets called significant at the FDR = 0.25 level by both the method along the top and the side of the table. Diagonal elements represent the number of SNP sets called significant by each individual method.

	Linear	IBS	Quadratic	Average	Perturb.	PCA	MinP
Linear	4	3	3	4	4	3	0
IBS		5	2	3	5	2	0
Quadratic			4	3	4	2	0
Average				4	4	3	0
Perturb.					7	3	0
PCA						4	0
MinP							0

Supplemental Table 6: Pre-term Birth Data Analysis Results. Each number represents the number of SNP sets called significant at the nominal $\alpha = 0.01$ level by both the method along the top and the side of the table. Diagonal elements represent the number of SNP sets called significant by each individual method.

	Linear	IBS	Quadratic	Average	Perturb.	PCA	MinP
Linear	1	1	1	1	1	1	0
IBS		2	1	1	2	1	0
Quadratic			1	1	1	1	0
Average				1	1	1	0
Perturb.					2	1	0
PCA						1	0
MinP							0

Supplemental Table 7: Pre-term Birth Data Analysis Results. Each number represents the number of SNP sets called significant at the nominal $\alpha = 0.05$ level by both the method along the top and the side of the table. Diagonal elements represent the number of SNP sets called significant by each individual method.

	Linear	IBS	Quadratic	Average	Perturb.	PCA	MinP
Linear	8	7	5	8	7	6	5
IBS		8	4	7	6	5	4
Quadratic			6	5	5	5	4
Average				8	7	6	5
Perturb.					7	5	5
PCA						8	3
MinP							5