Multivariate Phenotype Association Analysis by Marker-Set Kernel Machine Regression

Arnab Maity,[1] Patrick F. Sullivan,[2] and Jung-Ying Tzeng[1,3]*

[1] *Department of Statistics, North Carolina State University, Raleigh, North Carolina*
[2] *Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina*
[3] *Bioinformatics Research Center, North Carolina State University, Raleigh, North Carolina*

Genetic studies of complex diseases often collect multiple phenotypes relevant to the disorders. As these phenotypes can be correlated and share common genetic mechanisms, jointly analyzing these traits may bring more power to detect genes influencing individual or multiple phenotypes. Given the advancement brought by the multivariate phenotype approaches and the multimarker kernel machine regression, we construct a multivariate regression based on kernel machine to facilitate the joint evaluation of multimarker effects on multiple phenotypes. The kernel machine serves as a powerful dimension-reduction tool to capture complex effects among markers. The multivariate framework incorporates the potentially correlated multidimensional phenotypic information and accommodates common or different environmental covariates for each trait. We derive the multivariate kernel machine test based on a score-like statistic, and conduct simulations to evaluate the validity and efficacy of the method. We also study the performance of the commonly adapted strategies for kernel machine analysis on multiple phenotypes, including the multiple univariate kernel machine tests with original phenotypes or with their principal components. Our results suggest that none of these approaches has the uniformly best power, and the optimal test depends on the magnitude of the phenotype correlation and the effect patterns. However, the multivariate test retains to be a reasonable approach when the multiple phenotypes have none or mild correlations, and gives the best power once the correlation becomes stronger or when there exist genes that affect more than one phenotype. We illustrate the utility of the multivariate kernel machine method through the Clinical Antipsychotic Trails of Intervention Effectiveness antibody study. *Genet. Epidemiol.* 36:686–695, 2012. © 2012 Wiley Periodicals, Inc.

Key words: kernel machine regression; multivariate regression; multivariate phenotypes; score-based test

*Correspondence to: Jung-Ying Tzeng, Department of Statistics and Bioinformatics Research Center, North Carolina State University, Campus Box 7566, Raleigh, NC 27695. E-mail: jytzeng@stat.ncsu.edu

Received 1 April 2012; Revised 23 May 2012; Accepted 18 June 2012
Published online 16 August 2012 in Wiley Online Library (wileyonlinelibrary.com/journal/gepi).
DOI: 10.1002/gepi.21663

INTRODUCTION

Genetic studies of complex diseases commonly collect multiple phenotypes that are relevant to the disorder under study. The multidimensional phenotypic information can consist of traits that relate to the risk of the diseases, such as body mass index (BMI) and blood pressure to cardiovascular disease. It may consist of subclinical phenotypes that underlie the disease syndromes, such as lung function indices for asthma and the endophenotypes proposed to assist in delineating the causes of psychiatric disorders like schizophrenia (e.g., brain imaging or neurocognition) [Gottesman and Gould, 2003]. It may also be the biochemical measurements that reflect the physiological state of the diseases, such as metabolite concentrations [Suhre et al., 2011]. These multiple phenotypes better reflect the underlying molecular mechanisms and are more directly related to the etiology of the diseases. They are also thought to be more heritable and have a less complicated genetic basis than the final disease diagnosis [Klei et al., 2008]. As a result, genetic modeling of these intermediate phenotypes can have larger strengths of association and be more informative to infer the potentially affected pathways and disease-causing process.

The typical strategy of analyzing multiple phenotypes is to conduct separate analysis with one phenotype vs. one marker at a time and then correct for multiple testing. Though straightforward and computationally efficient, univariate phenotype analysis ignores the correlation among phenotypes and only captures one aspect of the phenotypes. As a result, the tests can be inefficient due to the penalty for multiple testing when phenotypes are correlated. An alternative strategy is to model the principal components of the original phenotypes. This principal component (PC) approach reduces dimensionality and yields statistically independent PC-phenotypes. However, these PC-phenotypes can have low heritability and the association results can be difficult to interpret [Klei et al., 2008].

Several recent papers have developed methods for multivariate association analysis of multiple phenotypes [e.g., Lange et al., 2002; Liu et al., 2009; Verzilli et al., 2005; Zapala and Schork, 2006; Zhang et al., 2010]. Compared to univariate phenotype analysis, the multivariate approaches model the joint distribution of the multiple traits and improve the statistical power to detect associated genetic variants. By taking into account the correlation structure of multiple traits and collectively analyzing the multidimensional information, multivariate approaches enhance the ability to identify genes that affect multiple traits, especially when traits are genetically correlated due to pleiotropy [Zhu and Zhang, 2009].

Current multivariate approaches focus mainly on single marker analyses. As with univariate analysis, multivariate phenotype analysis can also benefit from marker-set analysis, including the ability to handle high-dimensional markers and the ability to amplify the association signals via information collapsing [e.g., Wu et al., 2010; Tzeng et al., 2011]. As shown in previous work, kernel machine is an attractive dimension-reduction tool to model the linear or nonlinear effects of multiple markers. It can account for epistatic effects among markers, has been demonstrated to be more powerful over other marker-set approaches, and is applicable to the detection of both common and rare variants [Kwee et al., 2007; Wu et al., 2010, 2011]. To conduct kernel machine analyses with multiple phenotypes, the typical strategy is to perform multiple univariate kernel machine (UV-KM) regressions coupled with Bonferroni correction for multiple testing. The UV-KM regression can take responses as the original phenotypes or the PCs of the phenotypes. However, these strategies may share the same power concerns as observed in the single marker analysis.

Given the advancement brought by the multivariate approaches for single marker analysis, in this work, we construct a multivariate regression framework based on kernel machine to facilitate the joint evaluation of multimarker effects on multiple phenotypes. Through simulations, we assess the validity and efficacy of the proposed multivariate method. We also study the performance of the commonly adapted strategies for kernel machine analysis on multiple phenotypes: the UV-KM tests with original phenotypes and with their PCs. Our results suggest that none of these approaches has the uniformly best power, and the optimal test depends on the magnitude of the correlation among phenotypes and the effect gene patterns. However, the multivariate test retains to be a reasonable approach when the multiple phenotypes have none or mild correlations, and gives the best power once the correlation becomes stronger.

The rest of the article is organized as follows. First, we introduce the multivariate kernel machine (MV-KM) regression model, derive a score test to evaluate the multimarker effects on the multivariate phenotypes, and show that the test statistic follows a weighted chi-squared distribution under the null hypothesis. Next, we conduct simulations studies to evaluate the performance of the multivariate tests and the typical strategies under a variety of scenarios. We then illustrate the utility of the proposed method through the antibody study of the Clinical Antipsychotic Trails of Intervention Effectiveness (CATIE) samples. Finally, we give some concluding remarks in the last section.

MATERIAL AND METHODS

MV-KM REGRESSION

Suppose we observe for each individual $i = 1, \ldots, n$, the response vector $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{pi})^{\mathrm{T}}$, covariates $\mathbf{X}_i$ such as age, gender, etc., and a set of single nucleotide polymorphisms (SNPs) $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{iM})^{\mathrm{T}}$ with $Z_{im} \in \{0, 1, 2\}$, $m = 1, \ldots, m$, recording the number of minor alleles. We assume the following model to relate the health outcome $\mathbf{Y}_i$ to the genetic covariates $\mathbf{Z}_i$ and the clinical covariates $\mathbf{X}_i$: for $k = 1, \ldots, p$ and $i = 1, \ldots, n$,

$$Y_{ki} = \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta}_k + h_k(\mathbf{Z}_i) + \epsilon_{ki},$$

with $(\epsilon_{1i}, \ldots, \epsilon_{pi})^{\mathrm{T}} = \text{Normal}(0, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = \{\sigma_{k\ell}\}$ and $\sigma_{k\ell}$ reflects the correlation between traits Y_k and Y_ℓ of the same individual. To fix the ideas, here we let the covariates $\mathbf{Z}_i$ and $\mathbf{X}_i$ be common to all the phenotypes. However, our methodology readily allows for a general case where these can be different for individual outcomes. In the model, $\boldsymbol{\beta}_k$ are unknown coefficient vectors corresponding to the effect of $\mathbf{X}$ and $h_k(\cdot)$ are unknown functions corresponding to the effect of SNP set of interest. Our goal is to test whether the SNP set has any effect of the outcome. In other words, we are interested in testing the null hypothesis

$$H_0 : h_1(\cdot) = \cdots = h_p(\cdot) = 0.$$

In this paper, we use kernel machine framework to allow $h(\cdot)$ to be specified parametrically and nonparametrically. This approach is more convenient and powerful for multidimensional data. Specifically, we specify $h_\ell(\cdot)$ using a kernel function $K_\ell(\cdot, \cdot)$. Mercer's theorem [Cristianini and Shawe-Taylor, 2000] guarantees that under some regularity conditions, the kernel function $K_\ell(\cdot, \cdot)$ implicitly specifies a unique function space, say $\mathcal{H}_\ell$, spanned by a particular set of orthonormal basis functions $\{\phi_{\ell j}(z), j = 1, \ldots, J_\ell\}$. Here orthogonality is defined with respect to the L_2 norm. Hence, the function space $\mathcal{H}_\ell$ has the property that any function $h_\ell(\cdot) \in \mathcal{H}_\ell$ can be represented in two ways: using a set of basis functions as $h(z) = \sum_{j=1}^{J_\ell} \phi_{\ell j}(z)\eta_{\ell j}$ known as the primal or basis representation; or equivalently using the kernel function as $h(z) = \sum_{k=1}^{n} K_\ell(\mathbf{Z}_i, z)\alpha_{\ell k}$ for some constants $\alpha_{\ell 1}, \ldots, \alpha_{\ell n}$. The later representation is called the dual representation.

In theory, given any basis functions $\phi_\ell(z) = \{\phi_{\ell 1}(z), \ldots, \phi_{\ell J_\ell}(z)\}^{\mathrm{T}}$ in the primal representation, one can construct the corresponding kernel $K_\ell(z_1, z_2) = \sum_{j=1}^{J_\ell} \phi_{\ell j}(z_1)\phi_{\ell j}$ (z_2) to facilitate the dual representation, and vice versa. For multidimensional data, it is more convenient to work with the dual representation for $h_\ell(\cdot)$ using the kernel function $K_\ell(\cdot, \cdot)$, as will be done in this paper. This approach has two main advantages, namely, it can easily deal with high-dimensional data and it can capture potentially complex interaction between SNPs via the specified kernel function. Two most commonly used kernels for SNP data are the dth order polynomial kernel and the identical by state (IBS) kernel. The dth order polynomial kernel $K(z_1, z_2) = (1 + z_1^{\mathrm{T}} z_2)^d$ corresponds to the models with dth-order polynomials including the cross-product terms. For example, the first-order polynomial kernel ($d = 1$) corresponds to the model with only main effects $h(\mathbf{Z}_i) = \mathbf{Z}_i^{\mathrm{T}} \boldsymbol{\eta}$, and the second-order polynomial kernel ($d = 2$) corresponds to the model with linear and quadratic main effects and two-way interactions $h(\mathbf{Z}_i) = \sum_{m=1}^{M} Z_{im}\eta_{1m} + \sum_{j<m} Z_{ij} Z_{im}\gamma_{jm} + \sum_{m=1}^{M} Z_{im}^2 \eta_{2m}$. The IBS kernel is $K(z_1, z_2) = \sum_{m=1}^{M}(2 - |z_{1m} - z_{2m}|)/(2M)$. Both dth polynomial kernel ($d > 1$) and IBS kernels allow for interactions between SNPs.

SCORE TEST FOR THE MARKER-SET EFFECT

We develop a score-based marker-set testing procedure in this section. Define

$$\boldsymbol{Y} = (Y_{11}, \ldots, Y_{1n}, \ldots, Y_{p1}, \ldots, Y_{pn})^{\mathrm{T}},$$
$$\boldsymbol{h} = \{h_1(\mathbf{Z}_1), \ldots, h_1(\mathbf{Z}_n), \ldots, h_p(\mathbf{Z}_1), \ldots, h_p(\mathbf{Z}_n)\}^{\mathrm{T}}$$

and similarly $\boldsymbol{\epsilon}$. Also define $\boldsymbol{X} = diag(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_p^{\mathrm{T}})^{\mathrm{T}}$. We can rewrite the model in matrix form as

$$\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = \text{Normal}(0, \widetilde{\boldsymbol{\Sigma}})$ with $\widetilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} \otimes I_n$. That is, $\widetilde{\boldsymbol{\Sigma}}$ is a $p \times p$ block matrix, and each block is a diagonal matrix of $\sigma_{k\ell}\mathbf{1}_n$ for $k = 1, \ldots, p$ and $\ell = 1, \ldots, p$. In other words, the correlation is not zero between different phenotypes for the same individuals.

Under the full model, for a fixed covariance matrix $\boldsymbol{\Sigma}$, we write the penalized log-likelihood as

$$L(\cdot) = -\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{h}\}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{h}\}/2$$
$$- \sum_{\ell=1}^{p} \tau_\ell^{-1} \, ||h_\ell(\cdot)||^2_{\mathcal{H}_\ell}/2, \qquad (1)$$

where $||h_\ell(\cdot)||_{\mathcal{H}_\ell}$ denotes the function norm of h_ℓ and τ's are penalty parameters. The function norm is defined as $||h_\ell(\cdot)||_{\mathcal{H}_\ell} = \int h_\ell^2(z)\,dz = \sum_{j=1}^{J_\ell} \eta_j^2 \equiv \boldsymbol{\eta}_\ell^{\mathrm{T}}\boldsymbol{\eta}_\ell$, where $\boldsymbol{\eta}_\ell = (\eta_{\ell 1}, \ldots, \eta_{\ell, J_\ell})^{\mathrm{T}}$. The penalized likelihood is needed to perform the inference on the genetic effect $\boldsymbol{h}$, whose dimension is the same as the sample size n. Note that each τ_ℓ controls the smoothness of the corresponding function h_ℓ so that for small values of τ_ℓ, the penalty term for h_ℓ becomes large and as a result h_ℓ becomes flat. On the other hand, larger values of τ_ℓ implies rougher h_ℓ.

RELATIONSHIP TO MULTIVARIATE LINEAR MIXED MODEL

Recall that we have assumed for $\ell = 1, \ldots, p$, h_ℓ belongs to a function space $\mathcal{H}_\ell$ with orthonormal basis functions $\{\phi_{\ell j}, j = 1 \ldots, J_\ell\}$ and can be expressed in the primal form $h_\ell(\mathbf{Z}_i) = \sum_{j=1}^{J_\ell} \phi_{\ell j}(\mathbf{Z}_i)\,\eta_{\ell j} = \boldsymbol{\phi}_\ell^{\mathrm{T}}(\mathbf{Z}_i)\boldsymbol{\eta}_\ell$, where $\boldsymbol{\phi}_\ell = (\phi_{\ell 1}, \ldots, \phi_{\ell, J_\ell})^{\mathrm{T}}$. Define $\boldsymbol{\Phi}_\ell = [\boldsymbol{\phi}_\ell(\mathbf{Z}_1), \ldots, \boldsymbol{\phi}_\ell(\mathbf{Z}_n)]$. Then it is easy to see that the penalized log-likelihood in (1) can be written as

$$L(\cdot) = -\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\eta}\}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\eta}\}/2$$
$$- \boldsymbol{\eta}^{\mathrm{T}}\widetilde{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\eta}/2, \qquad (2)$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\eta}_p^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Phi} = diag\{\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_p\}$ and $\widetilde{\boldsymbol{\Lambda}} = diag(\tau_1 I_{J_1}, \ldots, \tau_p I_{J_p})$.

Differentiating (2) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, we obtain the estimating equations

$$0 = \boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\eta}\}$$
$$0 = \widetilde{\boldsymbol{\Lambda}}\,\boldsymbol{\Phi}\widetilde{\boldsymbol{\Sigma}}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\eta}\} - \boldsymbol{\eta}, \quad \ell = 1, \ldots, p.$$

Define $\boldsymbol{K} = diag(\boldsymbol{K}_1, \ldots, \boldsymbol{K}_p)$ and multiply $\boldsymbol{\Phi}^{\mathrm{T}}$ to the second equation above. Using the facts that $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} = \boldsymbol{K}$ and $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\eta} = \boldsymbol{h}$, we obtain the estimating equation for $\boldsymbol{\beta}$ and $\boldsymbol{h}$ as

$$0 = \boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{h}\}$$
$$0 = \boldsymbol{\Lambda}\boldsymbol{K}\widetilde{\boldsymbol{\Sigma}}^{-1}\{\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{h}\} - \boldsymbol{h}, \quad \ell = 1, \ldots, p,$$

where $\boldsymbol{\Lambda} = diag(\tau_1, \ldots, \tau_p) \otimes I_n$. Equivalently, we have the normal equation

$$\begin{bmatrix} \boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}^{\mathrm{T}} & \boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1} \\ \widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{X}^{\mathrm{T}} & \widetilde{\boldsymbol{\Sigma}}^{-1} + (\boldsymbol{K}\boldsymbol{\Lambda})^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{h} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}\widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y} \\ \widetilde{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y} \end{bmatrix}.$$

From a computational point of view, these normal equations are exactly identical to those of the mixed effects model

$$\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{h} + \boldsymbol{\epsilon},$$

where $\boldsymbol{h} = \text{Normal}(0, \boldsymbol{K}\boldsymbol{\Lambda})$ and $\boldsymbol{\epsilon} = \text{Normal}(0, \widetilde{\boldsymbol{\Sigma}})$ [see, e.g., Harville, 1977]. Hence, one can think of the penalty parameters τ's as the variance components. Hence, testing for $H_0 : h_1(\cdot) = \cdots = h_p(\cdot) = 0$ is equivalent to testing for the variance components $H_0' : \tau_1 = \cdots = \tau_p = 0$.

THE REML SCORE TEST STATISTIC

A straightforward way to test H_0' is to use a likelihood-based score test. However, a major disadvantage of this maximum likelihood approach is that it does not take into account the loss of degrees of freedom due to estimation of $\boldsymbol{\beta}$ and hence the resulting test would suffer from loss of power. Instead, we use the restricted maximum likelihood (REML) estimation procedure [see, e.g., Maity and Lin, 2011 and Tzeng and Zhang, 2007] to derive a score test. We write the REML of (2) as

$$L_{REML} = -\log|\boldsymbol{V}|/2 - \log|\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^{\mathrm{T}}|/2$$
$$- (\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})/2,$$

where $\boldsymbol{V} = \widetilde{\boldsymbol{\Sigma}} + \boldsymbol{K}\boldsymbol{\Lambda}$. Define $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{V}^{-1}\boldsymbol{X}^{\mathrm{T}})^{-1}$ $\boldsymbol{X}\boldsymbol{V}^{-1}$ and denote $\boldsymbol{P}_0$ to be $\boldsymbol{P}$ evaluated at H_0'. The score function of τ_ℓ, evaluated at H_0' is

$$S_{\tau_\ell, n} = (\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}_0^{-1}\boldsymbol{K}_\ell^*\boldsymbol{V}_0^{-1}(\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})/2 - \text{trace}(\boldsymbol{K}_\ell^*\boldsymbol{P}_0),$$

where $\boldsymbol{K}_\ell^*$ is a block diagonal matrix with $\boldsymbol{K}_\ell$ as the ℓth block and zero otherwise, $\widehat{\boldsymbol{\beta}}$ is estimated under null and $\boldsymbol{V}_0$ denotes $\boldsymbol{V}$ evaluated under null. Similar to Tzeng et al. [2011], we use the first term as the test statistic, and to test for $H_0' : \tau_1 = \cdots = \tau_p = 0$, we now propose to use the combined score type test statistic

$$T_n = \sum_{\ell=1}^{p} S_{\tau_\ell, n} = (\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}\boldsymbol{V}_0^{-1}\boldsymbol{K}\boldsymbol{V}_0^{-1}(\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}).$$

To derive the null distribution of the test statistic T_n, we first compute the eigenvalue decomposition of $\boldsymbol{K} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathrm{T}}$ and observe that $T_n = \boldsymbol{r}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{r}$ is a quadratic form where $\boldsymbol{r} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{V}_0^{-1}(\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}})$. Note that under H_0', $\boldsymbol{r}$ follows a Gaussian distribution with mean zero and covariance matrix $\boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}_0\boldsymbol{U}$, and therefore the distribution of T_n is a mixture of chi-squared random variables with weights being the

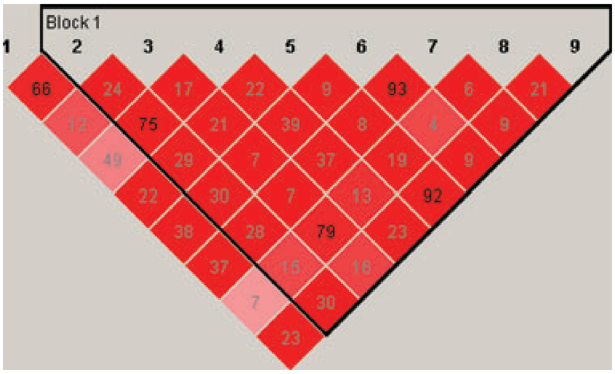

Fig. 1. LD of gene *SLC17A1*.

TABLE I. Frequency of multimarker genotypes as observed in *SLC17A1* of CATIE schizophrenia samples. Displayed are the genotypes appearing ≥ 1%

Genotype	Frequency (in real data)	Relative frequency (used in simulation)
221212211[a]	113	0.191
222202220	103	0.174
112111121	72	0.121
111121112	46	0.078
220222202	41	0.069
112112221	38	0.064
111122212	28	0.047
211122212	26	0.044
111222212	24	0.041
002021122	20	0.034
212112221	15	0.025
002020022	10	0.017
102122222	10	0.017
102022222	9	0.015
002121122	8	0.013
002122222	8	0.013
220222211	8	0.013
102021122	7	0.012
202022222	7	0.012

[a] The string represents the multimarker genotype of the nine SNPs, with each digit showing the minor allele count of a locus.

diagonal elements of $\boldsymbol{D}$. One can approximate the distribution of T_n by moment matching [e.g., Duchesne and Lafaye [2010]] or by the empirical approach as described below. First, we generate independent and identically distributed random vectors $\boldsymbol{w}_1^*, \ldots, \boldsymbol{w}_B^*$ from multivariate normal distribution with mean zero and identity covariance matrix for a large number B, and compute the realizations of $\boldsymbol{r}$ under H_0' as $\boldsymbol{r}_b^* = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{P}_0^{1/2}\boldsymbol{w}_b^*$ for $b = 1, \ldots, B$. Then realizations of T_n can be generated as $T_{n,b}^* = \boldsymbol{r}_b^{*\mathrm{T}}\boldsymbol{D}\boldsymbol{r}_b^*$. Finally, we compute the P-value as

$$P\text{-value} = \sum_{b=1}^{B} 1(T_n < T_{n,b}^*)/B,$$

where $1(\cdot)$ denotes the indicator function.

Note that one needs to provide a working covariance matrix $\boldsymbol{\Sigma}$ in order to perform this test. We propose to first use a working independence structure, that is assuming $\boldsymbol{\Sigma}$ is a diagonal matrix with unknown variances, fit the null model $\boldsymbol{Y} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and compute the residuals $\widehat{e} = \boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$. Then we estimate the covariance matrix under null using these residuals. The final estimated covariance matrix can then be used to conduct the test.

SIMULATION STUDY

We demonstrate the performance of our proposed testing procedures through simulation study. For $i = 1, \ldots, n$ and $k = 1, \ldots, p$, we generated data from the following model:

$$Y_{ki} = \boldsymbol{X}_i^{\mathrm{T}}\boldsymbol{\beta}_k + h_k(\boldsymbol{Z}_i) + \epsilon_{ki},$$

where $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{im})^{\mathrm{T}}$, $\boldsymbol{X}_i = (X_{i1}, X_{i2})^{\mathrm{T}}$, and $(\epsilon_{1i}, \ldots, \epsilon_{pi})^{\mathrm{T}} = \text{Normal}(0, \boldsymbol{\Sigma}_{\text{true}})$. We generated $\boldsymbol{X}_i$ from a bivariate standard normal distribution and set the true value of $\boldsymbol{\beta}_k = (0.2, 0.4)^{\mathrm{T}}$ for $k = 1, \ldots, p$. We simulated the m-SNP genotype data $\boldsymbol{Z}_i$ based on the first gene (*SLC17A1*) in the CATIE antibody study that contained nine SNPs. The structure of linkage disequilibrium (LD) is given in Figure 1, and the multimarker genotype distribution is shown in Table I. For simulation purposes, we only take genotypes with ≥ 7 occurrences (i.e., 1% of sample size in data example, $n = 690$). We considered simulation scenarios: (1) $n = 100$, $m = 9$, (2) $n = 200$, $m = 9$, and (3) $n = 200$, $m = 30$. In scenarios (1) and (2), $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{im})^{\mathrm{T}}$ were generated directly from *SLC17A1* by Table I, and in scenario (3), we added additional 21 nuisance SNPs where each Z_{im} took value 0, 1, or 2 with probability 0.3, 0.5, and 0.2, respectively. We considered the following choices of the functions h: (a) sparse effect, where $h_1 = \delta(z_1 + z_2 + z_3 + z_1z_4z_5 - z_6/3 - z_7z_8/2 + (1 - z_9))$ for $\delta = 0, 0.04, \ldots, 0.16$ and $h_2 = \cdots = h_p = 0$, and (b) common effect, where $h_1^* = h_1 + \delta z_3$ and $h_2 = \cdots = h_p = \delta z_3$ for $\delta = 0, 0.04, \ldots, 0.16$. For the dimension of phenotypes, we considered $p = 3$ and $p = 10$. The case with $p = 3$ was in accordance to the CATIE data, and the case $p = 10$ represented the case of larger number of phenotypes. For $p = 3$, we used three choices of $\boldsymbol{\Sigma}_{\text{true}}$:

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.95 & 0 & 0. \\ 0 & 0.86 & 0 \\ 0 & 0 & 0.89 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.95 & 0.07 & 0.23 \\ 0.07 & 0.86 & 0.24 \\ 0.23 & 0.24 & 0.89 \end{pmatrix},$$

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.95 & 0.57 & 0.43 \\ 0.57 & 0.86 & 0.24 \\ 0.43 & 0.24 & 0.89 \end{pmatrix}$$

The second covariance matrix $\boldsymbol{\Sigma}_2$ was the estimated covariance matrix from the CATIE data under the null model, where the correlations among phenotypes ranged from 0.11 to 0.28. The first choice $\boldsymbol{\Sigma}_1$ was just the diagonal part of $\boldsymbol{\Sigma}_2$ and was used to evaluate the performance of our procedure where the responses are actually independent. The third choice $\boldsymbol{\Sigma}_3$ introduced higher correlation between Y_1 and (Y_2, Y_3) (i.e., correlation coefficient $r = (0.63, 0.62)$), and was used to demonstrate the impact of correlations on the power of our testing procedure. For $p = 10$, we added seven more error variables independent to others, each with variance 1.

TABLE II. Simulation results for $p = 3$ and 10. Displayed are the type I error rates in percent of our test (MV-KM) along with the univariate KM-test with Bonferroni correction (UV-KM) and principal component based KM tests (PC1-KM and PC*k*-KM), as described in the simulation section, for $\alpha = 5\%$ and 0.5%

		$\boldsymbol{\Sigma}_1$		$\boldsymbol{\Sigma}_2$		$\boldsymbol{\Sigma}_3$	
(n, m)	α :	5%	0.5%	5%	0.5%	5%	0.5%
Case : $p = 3$							
	MV-KM	5.11	0.45	5.10	0.46	5.16	0.47
	PC1-KM	4.52	0.41	4.62	0.41	4.54	0.43
(100, 9)	PC*k*-KM	4.41	0.36	4.45	0.41	4.55	0.35
	UV-KM	4.56	0.35	4.34	0.36	3.99	0.37
	MV-KM	5.01	0.47	5.05	0.47	4.91	0.43
	PC1-KM	4.55	0.42	4.52	0.41	4.53	0.43
(200, 9)	PC*k*-KM	4.49	0.39	4.45	0.43	4.21	0.37
	UV-KM	4.66	0.38	4.54	0.40	4.03	0.34
	MV-KM	5.10	0.46	5.07	0.48	5.13	0.45
	PC1-KM	4.51	0.43	4.71	0.46	4.55	0.43
(200, 30)	PC*k*-KM	4.44	0.38	4.49	0.46	4.51	0.41
	UV-KM	4.50	0.33	4.41	0.37	3.97	0.37
Case : $p = 10$							
	MV-KM	4.71	0.42	4.70	0.43	5.20	0.43
	PC1-KM	4.82	0.47	4.66	0.42	4.64	0.48
(100, 9)	PC*k*-KM	4.43	0.40	4.15	0.41	4.12	0.32
	UV-KM	4.10	0.38	4.43	0.37	4.18	0.37
	MV-KM	4.59	0.47	5.10	0.44	4.20	0.51
	PC1-KM	3.62	0.32	1.12	0.38	4.00	0.35
(200, 9)	PC*k*-KM	3.61	0.32	3.71	0.33	3.80	0.30
	UV-KM	3.68	0.37	3.70	0.33	3.50	0.30
	MV-KM	4.88	0.41	4.67	0.39	4.92	0.40
	PC1-KM	3.88	0.30	4.26	0.34	4.00	0.32
(200, 30)	PC*k*-KM	3.03	0.31	3.03	0.28	3.27	0.24
	UV-KM	2.91	0.27	2.91	0.25	2.90	0.23

We compared the proposed MV-KM test with (1) UV-KM test and (2) principal component-based tests. In UV-KM, we performed p separate UV-KM regressions using each of the phenotypes as response and corrected for multiple testing using Bonferroni method. In PC-based tests, we first performed a PCA on the phenotypes to obtain the p principal components. We then conducted kernel machine test either using the first PC as response (referred to as PC1-KM) or using the top k PCs that retained 90% of the variations with the Bonferroni correction (referred to as PC*k*-KM). We used the IBS kernel for this comparison. When fitting the MV-KM model, we first fitted a working independence model and estimated $\boldsymbol{\Sigma}$ using residual covariance matrix, and then used the estimated $\boldsymbol{\Sigma}$ as the working covariance matrix in the testing procedure. For each scenario, we generated 20,000 simulated data sets for type I error evaluation and 1,000 data sets for power comparison. The P-values were calculated based on $B = 10{,}000$ resampled statistics.

We provide the type I error analysis in Table II. Significance levels of $\alpha = 0.05$ and $\alpha = 0.005$ were considered. Roughly speaking, MV-KM retained the nominal type I error rate in all the scenarios, UV-KM were conservative especially in high correlation setting and when $\alpha = 0.005$, PC1-KM and PC*k*-KM had their type I error rates between the two. When phenotypes were correlated, UV-KM appeared to be more conservative than PC1-KM and PC*k*-KM when $p = 3$, but was about the same or more conservative when $p = 10$. The conservative results of UV-KM and PC*k*-KM can be attributed to the correlation among the test statistics. Even in the case of independent phenotypes (i.e., Σ_1), different test statistics were correlated because they were based on the same predictor values from the same subjects.

Results for power analysis are displayed in Tables III and IV. For power analysis, we first focus on $p = 3$, case (a): sparse effect, where only Y_1 was associated with the gene. We see that UV-KM had the best power when the phenotypes were independent ($\boldsymbol{\Sigma}_1$) or in low correlation ($\boldsymbol{\Sigma}_2$), and MV-KM had the best power when correlation was high (e.g., $\boldsymbol{\Sigma}_3$). PC1-KM had inferior power compared to UV-KM and MV-KM. However, PC*k*-KM had similar power to MV-KM in the zero or low correlation cases but had inferior power in the high correlation case. For $p = 3$ case with case (b) (common effect), where all the responses shared some common genetic effects, MV-KM had similar power as UV-KM for zero or low correlation, and had better power than UV-KM for high correlation. In this case, PC1-KM and PC*k*-KM showed similar pattern as in the sparse case.

For $p = 10$ case with sparse effect, we observed a similar pattern as in $p = 3$ except that the gap between UV-KM and MV-KM become larger for case (a) with zero ($\boldsymbol{\Sigma}_1$) or low correlation ($\boldsymbol{\Sigma}_2$). This is not too surprising because in case (a), there was only one phenotype (Y_1) exhibiting association, and the correlations among the phenotypes were mostly 0 (i.e., in $\boldsymbol{\Sigma}_1$, all Ys had 0 correlations with each other; in $\boldsymbol{\Sigma}_2$, all Ys had 0 correlation except that $r = 0.11 \sim 0.28$ between Y_1, Y_2, and Y_3). The results suggested that when signal is sparse among Ys and Ys also have little correlations, multiple correction may be more efficient than multivariate modeling. Nevertheless, once some correlations became moderate-even only among a few Ys (e.g., in $\boldsymbol{\Sigma}_3$, all Ys had 0 correlation except that $r = 0.28 \sim 0.63$ between Y_1, Y_2, and Y_3), we saw MV-KM can still utilize the information from other phenotypes and gave significant power gain over UV-KM and PC-based tests. The results of $p = 10$ with common effect were also similar to $p = 3$ case (b). When correlation was higher, MV-KM had the best power as seen earlier. In cases (a) and (b), both PC-based tests had worse power than both MV-KM and UV-KM across all scenarios.

We also investigated the effect of choosing different kernel functions on type I error and power of the proposed MV-KM test. We consider the setting with $n = 100$, $m = 9$, and $p = 3$ with sparse and common effects as above. We consider three different kernel functions to summarize the multi-SNP information: (a) IBS kernel, (b) quadratic kernel, i.e., the second-order polynomial kernel, and (c) linear kernel, i.e., the first-order polynomial kernel. Both IBS kernel and quadratic kernel model epistatic and nonlinear SNP effects, but unlike IBS kernel that modeled such effect implicitly, the quadratic kernel uses specific forms, i.e., the linear and quadratic main effects plus the pairwise interactions. In contrast, the linear kernel only consider the linear additive SNP effects. The type I errors at 5% and 0.5% level are given in Table V (based on 20, 000 simulated data sets), and the power analysis at $\alpha = 0.05$ is displayed in Table VI (based on 1,000 simulated data sets). It is evident that in terms of both type I error and power, the three kernel behave similarly for the $h(\cdot)$ functions considered in the simulation study.

TABLE III. Simulation results for $p = 3$. Displayed are the power of our test (MV-KM) along with the univariate KM test with Bonferroni correction (UV-KM) and principal component based KM tests (PC1-KM and PC*k*-KM), as described in the simulation section, for $\alpha = 0.05$

		Σ_1				Σ_2				Σ_3			
(n, m)		0.04	0.08	0.12	0.16	0.04	0.08	0.12	0.16	0.04	0.08	0.12	0.16
Case: (a) Sparse effect													
	MV-KM	0.085	0.229	0.467	0.701	0.089	0.235	0.467	0.72	0.202	0.594	0.902	0.99
	PC1-KM	0.068	0.143	0.281	0.421	0.055	0.094	0.178	0.299	0.058	0.097	0.181	0.300
(100, 9)	PC*k*-KM	0.079	0.211	0.462	0.711	0.076	0.226	0.482	0.749	0.067	0.246	0.535	0.770
	UV-KM	0.082	0.278	0.58	0.85	0.08	0.278	0.585	0.854	0.077	0.274	0.581	0.857
	MV-KM	0.122	0.457	0.843	0.987	0.131	0.466	0.853	0.985	0.341	0.895	0.998	1
	PC1-KM	0.105	0.271	0.525	0.744	0.083	0.173	0.356	0.547	0.083	0.134	0.347	0.569
(200, 9)	PC*k*-KM	0.125	0.421	0.815	0.971	0.124	0.448	0.825	0.973	0.128	0.414	0.734	0.922
	UV-KM	0.141	0.542	0.915	0.995	0.136	0.54	0.916	0.996	0.13	0.536	0.923	0.997
	MV-KM	0.084	0.282	0.647	0.905	0.097	0.304	0.669	0.91	0.235	0.784	0.99	1
	PC1-KM	0.074	0.169	0.359	0.573	0.060	0.117	0.232	0.363	0.057	0.119	0.218	0.368
(200, 30)	PC*k*-KM	0.082	0.257	0.636	0.912	0.077	0.279	0.677	0.920	0.073	0.324	0.668	0.896
	UV-KM	0.084	0.381	0.83	0.982	0.088	0.377	0.822	0.98	0.08	0.364	0.828	0.975
Case: (b) Common effect													
	MV-KM	0.078	0.198	0.405	0.669	0.089	0.222	0.445	0.715	0.215	0.635	0.93	0.993
	PC1-KM	0.057	0.095	0.164	0.252	0.046	0.074	0.107	0.146	0.049	0.068	0.111	0.162
(100, 9)	PC*k*-KM	0.073	0.198	0.419	0.676	0.076	0.211	0.455	0.724	0.066	0.249	0.570	0.823
	UV-KM	0.07	0.217	0.482	0.728	0.07	0.218	0.488	0.717	0.063	0.212	0.465	0.729
	MV-KM	0.109	0.403	0.79	0.957	0.127	0.432	0.815	0.976	0.365	0.916	0.998	1
	PC1-KM	0.078	0.145	0.293	0.440	0.060	0.106	0.179	0.291	0.058	0.094	0.191	0.289
(200, 9)	PC*k*-KM	0.110	0.388	0.754	0.943	0.114	0.425	0.809	0.968	0.122	0.443	0.735	0.923
	UV-KM	0.125	0.398	0.804	0.974	0.122	0.407	0.809	0.971	0.107	0.415	0.824	0.981
	MV-KM	0.083	0.232	0.579	0.864	0.093	0.265	0.636	0.899	0.248	0.817	0.99	1
	PC1-KM	0.050	0.095	0.180	0.284	0.044	0.067	0.117	0.186	0.054	0.069	0.136	0.201
(200, 30)	PC*k*-KM	0.066	0.220	0.569	0.862	0.066	0.249	0.643	0.913	0.066	0.356	0.709	0.931
	UV-KM	0.065	0.264	0.655	0.914	0.072	0.26	0.657	0.918	0.057	0.249	0.627	0.926

DATA ANALYSIS

The CATIE antibody study is motivated by the availability of genome-wide association SNP data and antibody level quantification for three neurotrophic herpesviruses in schizophrenia cases from the CATIE studies [Lieberman et al., 2005; Sullivan et al., 2008; Yolken et al., 2011]. Given that (1) genetic variation in the MHC has emerged as a robust and replicable risk factor for schizophrenia [International Schizophrenia Consortium, 2009; Shi et al., 2009; Stefansson et al., 2009], (2) the known role of genetic variation in this region in responses to infectious agents, and (3) the epidemiological and clinical associations of exposure to these infectious agents on risk of schizophrenia [Yolken et al., 2011], one major goal is to understand the association between antibody responses to the neurotrophic herpesviruses and genes that are related to schizophrenia and located around the MHC region.

The original CATIE study examined whether atypical antipsychotics can reduce morbidity and resource use compared to a conventional antipsychotic drug for patients suffering from chronic schizophrenia. About 51% of the 1,460 CATIE participants provided DNA samples, and genotype data were available for 492K SNPs [Sullivan et al., 2008]. Focused on the CATIE schizophrenia samples, Yolken et al. [2011] measured the IgG class antibodies to three herpesviruses that are capable of establishing persistent infection within the human central nervous system: Herpes Simplex Virus type 1 (HSV-1), Herpes Simplex Virus type 2 (HSV-2), and Cytomegalovirus (CMV). They found that exposure to these neurotrophic infectious agents were associated with cognitive deficits in schizophrenia samples. In our analysis, we examined the association between antibody levels and those genes that were located on chromosome 6p22.1 and were reported to be associated with schizophrenia. There were 12 gene regions selected for evaluation (Table VII). We note that in a complete gene-based study, the analysis should include two steps: first to detect genes that exhibited global association with the traits, and second to evaluate variant-specific effects within the associated genes using refined approaches, such as single-SNP analysis or haplotype penalized regression [Tzeng and Bondell, 2009]. In this data application, we focus on the illustration of how MV-KM can be used as a tool for the first detection step.

For each of the gene regions, we consider three kernel functions to summarize the multi-SNP information: the IBS kernel, the quadratic kernel, and the linear kernel. We applied four approaches to evaluate the gene-level effect: the proposed MV-KM method, PC1-KM, PC*k*-KM, and UV-KM. The P-values were obtained using $B = 100{,}000$ resampled statistics. Each analysis adjusted for age and sex. The correlation between the antibody levels were estimated as 0.011 for HSV-1 and HSV-2, 0.203 for HSV-1 and CMV, and 0.280 for HSV-2 and CMV, which indicated a weak dependence

TABLE IV. Simulation results for $p = 10$. Displayed are the power of our test (MV-KM) along with the univariate KM test with Bonferroni correction (UV-KM) and principal component based KM tests (PC1-KM and PC*k*-KM), as described in the simulation section, for $\alpha = 0.05$

		Σ_1				Σ_2				Σ_3			
(n, m)		0.04	0.08	0.12	0.16	0.04	0.08	0.12	0.16	0.04	0.08	0.12	0.16
Case: (a) Sparse effect													
	MV-KM	0.066	0.122	0.204	0.321	0.066	0.136	0.251	0.38	0.158	0.45	0.781	0.945
	PC1-KM	0.069	0.095	0.124	0.171	0.067	0.083	0.128	0.174	0.060	0.094	0.139	0.208
(100, 9)	PC*k*-KM	0.055	0.097	0.191	0.251	0.055	0.102	0.186	0.273	0.056	0.067	0.115	0.278
	UV-KM	0.069	0.176	0.425	0.724	0.064	0.176	0.427	0.717	0.056	0.169	0.427	0.721
	MV-KM	0.083	0.244	0.566	0.858	0.097	0.298	0.647	0.881	0.28	0.857	0.993	1
	PC1-KM	0.050	0.095	0.170	0.273	0.047	0.117	0.175	0.273	0.067	0.134	0.234	0.359
(200, 9)	PC*k*-KM	0.053	0.145	0.435	0.737	0.061	0.153	0.390	0.691	0.045	0.078	0.156	0.406
	UV-KM	0.074	0.376	0.838	0.983	0.074	0.366	0.829	0.982	0.079	0.36	0.829	0.988
	MV-KM	0.062	0.141	0.316	0.548	0.071	0.167	0.394	0.619	0.166	0.67	0.957	0.999
	PC1-KM	0.034	0.061	0.121	0.174	0.039	0.061	0.120	0.187	0.047	0.072	0.142	0.237
(200, 30)	PC*k*-KM	037	0.106	0.319	0.467	0.025	0.078	0.209	0.496	0.019	0.042	0.097	0.272
	UV-KM	0.052	0.249	0.688	0.951	0.055	0.247	0.691	0.95	0.052	0.233	0.684	0.953
Case: (b) Common effect													
	MV-KM	0.064	0.126	0.245	0.404	0.064	0.14	0.298	0.477	0.168	0.492	0.841	0.964
	PC1-KM	0.076	0.124	0.247	0.525	0.066	0.104	0.211	0.366	0.057	0.083	0.139	0.246
(100, 9)	PC*k*-KM	0.058	0.114	0.233	0.384	0.062	0.109	0.239	0.382	0.058	0.081	0.155	0.259
	UV-KM	0.068	0.143	0.322	0.556	0.066	0.142	0.326	0.573	0.054	0.127	0.309	0.587
	MV-KM	0.093	0.287	0.708	0.944	0.094	0.367	0.754	0.972	0.334	0.9	0.997	1
	PC1-KM	0.066	0.216	0.495	0.817	0.077	0.180	0.429	0.714	0.055	0.141	0.279	0.516
(200, 9)	PC*k*-KM	0.060	0.181	0.480	0.847	0.058	0.160	0.454	0.797	0.050	0.095	0.224	0.526
	UV-KM	0.066	0.271	0.665	0.947	0.058	0.298	0.667	0.953	0.06	0.288	0.696	0.95
	MV-KM	0.064	0.172	0.41	0.701	0.06	0.188	0.475	0.774	0.196	0.722	0.984	1
	PC1-KM	0.077	0.154	0.369	0.607	0.060	0.106	0.385	0.503	0.047	0.076	0.185	0.346
(200, 30)	PC*k*-KM	0.026	0.103	0.374	0.627	0.033	0.070	0.333	0.629	0.024	0.052	0.148	0.358
	UV-KM	0.031	0.162	0.513	0.863	0.035	0.168	0.516	0.857	0.043	0.176	0.524	0.87

TABLE V. Simulation results for $(n, m, p) = (100, 9, 3)$. Displayed are the type I rates (in percent) of our test (MV-KM) for different choices of kernels: linear, quadratic, and IBS, for $\alpha = 5\%$ and 0.5%

	Σ_1		Σ_2		Σ_3	
α :	5%	0.5%	5%	0.5%	5%	0.5%
Linear	5.11	0.41	5.30	0.40	5.3	0.44
Quadratic	5.13	0.40	5.23	0.41	5.26	0.44
IBS	5.11	0.45	5.10	0.46	5.16	0.47

among different antibody levels. The proportion of the variance explained by the top k PCs was 59%, 87%, and 100% for $k = 1, 2$, and 3, respectively. So we reported the results of PC1-KM and PC2-KM. The significance level for the 12-gene analysis was 0.05/12= 0.0042.

The *P*-values based on IBS kernel are given in Table VII. The *P*-values reported were adjusted for the multiphenotype tests for UV-KM and PC2-KM. MV-KM identified one significant gene region, i.e., MHC (*P*-value = 0.0006). The UV-KM method identified three significant regions, including MHC (*P*-value = 0.0006) and two additional genes, BTN2A1 (*P*-value 0.0036) and POM121L2 (*P*-value = 0.0009). This data set consisted of small-dimensional phenotype data with low correlations, which is an ideal scenario for UV-KM according to the simulation findings. In this regard, it is not surprising that MV-KM did not identify as many signals as UV-KM, but it has exhibited ability to identify the most biologically meaningful region (i.e., MHC). PC1-KM did not identify any regions. PC2-KM identified BTN2A1 (*P*-value 0.0038) to be significant.

The *P*-values based on quadratic and linear kernels are shown in Table VIII. Comparing to Table VII, we observed that using quadratic kernels yielded more significant results than IBS kernel and linear kernel regardless of univariate, PC-based, or multivariate analyses, especially for genes BTN2A1 and POM121L2. This suggested that there might be some nonadditive effect among SNPs within those genes, and the quadratic function was more efficient in modeling such effects for the CATIE antibody data.

DISCUSSION

In this work, we present a MV-KM testing procedure for studying the joint effect of multiple markers on multiple phenotypes simultaneously. The kernel machine serves as a powerful dimension-reduction tool to capture interaction and nonlinear effects among markers. The multivariate framework incorporates the potentially correlated multidimensional phenotypic information, accommodates common or different environmental covariates for each trait, and detects genetic effects that affect single or multiple traits. We derived a score-based test to assess the

TABLE VI. Simulation results for (n, m, p) = (100, 9, 3). Displayed are the power of our test (MV-KM) for different choices of kernels: linear, quadratic, and IBS, for $\alpha = 0.05$

	Σ_1				Σ_2				Σ_3			
c :	0.04	0.08	0.12	0.16	0.04	0.08	0.12	0.16	0.04	0.08	0.12	0.16
Case: (a) Sparse effect												
Linear	0.085	0.235	0.458	0.703	0.091	0.239	0.450	0.729	0.194	0.612	0.909	0.983
Quadratic	0.081	0.225	0.480	0.711	0.090	0.236	0.484	0.745	0.188	0.610	0.915	0.994
IBS	0.085	0.229	0.467	0.701	0.089	0.235	0.467	0.720	0.202	0.594	0.902	0.990
Case: (b) Common effect												
Linear	0.074	0.205	0.405	0.642	0.089	0.226	0.454	0.736	0.203	0.653	0.936	0.993
Quadratic	0.070	0.191	0.419	0.669	0.083	0.213	0.453	0.716	0.200	0.649	0.933	0.996
IBS	0.078	0.198	0.405	0.669	0.089	0.222	0.445	0.715	0.215	0.635	0.930	0.993

TABLE VII. Results of CATIE antibody analysis using IBS kernel. The genes are ordered according to the genomic regions. The significant threshold for the *P*-value is based on Bonferroni correction for 12 genes analysis, i.e., 0.05/12 = 0.0042. Values given in bold indicate p-values that are less than the significant threshold

			P-values of IBS kernel			
Gene	No. of subjects	No. of SNPs	MV-KM	PC1-KM	PC2-KM	UV-KM
SLC17A1	690	9	0.0060	0.1224	0.0217	0.0090
SLC17A3	404	13	0.3235	0.0048	0.0096	0.0152
BTN3A2	666	5	0.0271	0.0116	0.0231	0.0605
BTN2A2	437	5	0.6566	0.9489	0.8658	0.7934
BTN2A1	693	4	0.0129	0.0235	**0.0038**	**0.0036**
HIST1H2AG	436	2	0.7133	0.8254	0.5476	0.7181
HIST1H2BJ	678	2	0.0501	0.5375	0.0231	0.0259
PRSS16	425	1	0.9826	0.7031	0.9119	0.9749
POM121L2	676	4	0.0042	0.0758	0.0048	**0.0009**
ZNF184	406	8	0.0605	0.0662	0.0840	0.0367
NOTCH4	618	24	0.4977	0.0725	0.1397	0.2588
MHC	516	787	**0.0006**	0.0287	0.0185	**0.0006**

TABLE VIII. Results of CATIE antibody analysis using quadratic and linear kernels. The genes are ordered according to the genomic regions. The significant threshold for the *P*-value is based on Bonferroni correction for 12 genes analysis, i.e., 0.05/12 = 0.0042. Values given in bold indicate p-values that are less than the significant threshold

	P-values of quadratic kernel				*P*-values of linear kernel			
Gene	MV-KM	PC1-KM	PC2-KM	UV-KM	MV-KM	PC1-KM	PC2-KM	UV-KM
SLC17A1	**0.0015**	0.0563	0.0066	**0.0006**	**0.0039**	0.0567	0.0118	**0.0036**
SLC17A3	0.2853	0.0044	0.0088	0.0146	0.2827	0.0075	0.0149	0.0155
BTN3A2	0.0362	0.0358	0.0703	0.0452	0.0189	0.0177	0.0351	0.0794
BTN2A2	0.5100	0.8943	0.6911	0.5848	0.5775	0.9316	0.8093	0.7464
BTN2A1	**0.0014**	0.0094	**0.0018**	**0.0006**	0.0046	0.0131	**0.0014**	**0.0027**
HIST1H2AG	0.8610	0.7581	0.7949	0.8819	0.7693	0.6806	0.8174	0.8217
HIST1H2BJ	0.0328	0.6816	0.0118	0.0161	0.0393	0.6641	0.0266	0.0170
PRSS16	0.9826	0.6999	0.9099	0.9728	0.9826	0.7071	0.9142	0.9695
POM121L2	**0.0015**	0.0351	0.0042	$\mathbf{<10^{-5}}$	**0.0038**	0.1180	0.0102	**0.0006**
ZNF184	0.0643	0.8879	0.0529	0.0432	0.0782	0.6922	0.0834	0.1238
NOTCH4	0.4311	0.0909	0.1735	0.2683	0.4172	0.0577	0.1121	0.1633
MHC	**0.0005**	0.0220	0.0155	**0.0003**	**0.0005**	0.0245	0.0201	**0.0003**

multimarker effects on the multiple traits collectively. We conducted simulations to evaluate the performance of the proposed MV-KM testing procedure. We also studied the performance of the commonly adapted strategies for kernel machine analysis on multiple phenotypes, i.e., UV-KM, PC1-KM, and PC*k*-KM. The results indicated that none of these approaches has the uniformly best power: UV-KM gave the highest power when the phenotypes are independent or have weak correlations. MV-KM had the highest power once the correlation became stronger. Finally, PC-based tests tended to have similar or less power than MV-KM for zero or low correlation cases but had inferior power for high correlation cases. Although the optimal test depends on the magnitude of the phenotype correlation and the effect patterns, MV-KM can still serve as a reasonable tool for multiple phenotype analysis—it often yields comparable power to UV-KM even when phenotypes have none or mild correlations, and performs the best when phenotypes have increased correlation or share common genetic mechanisms.

We note that despite the fact that the methods presented here aimed toward testing the effect of SNP sets, one could also employ similar techniques to include environmental factors and interaction terms in the model, and develop tests on the overall effect of gene and gene-environment interactions. In addition, the proposed methodology can be readily generalized to accommodate other types of predictors such as those encountered in the analysis of expression, methylation, or metabolism. One would need to use different kernels suitable for each data type while keeping the methodology same. In that sense, we have presented here a general framework to perform set analysis on multivariate phenotypes for different "omic" data types.

As noted in the data analysis, a complete gene-level analysis should include two aspects: (1) to detect the global association between gene and traits, and (2) to comprehend the signal identified at gene level. The proposed MV-KM method can serve as a screening tool for step (1), and then follow-up analysis could be performed to dissect the gene-level signals identified. In this regard, it is desired to perform phenotype-specific tests to identify the source of the global signals and understand the effect patterns. These inference procedures require estimating the genetic effects for each trait and we are currently working on the extensions along this direction.

Finally, in this work we illustrated how MV-KM can be used to screen for promising genes using common SNPs through a chromosome-wide search. The proposed method is directly applicable to other scenarios: The set aggregation can be performed at the level of exons, LD blocks, genes, pathways, or networks; the analyzed variants can be common or rare; and the search can be extended to exome-wide or genome-wide. MV-KM is computationally efficient due to its permutation-free features. For the case with $(n, m, p) = (100, 9, 3)$, $B = 10{,}000$ and using IBS kernel, it takes 1.2 sec, on an average, to run one MV-KM test on an Intel Xeon 3.33 GHz machine with 12 Gb RAM (using only one processing core). To focus on rare variants, one can incorporate weights into the kernel function based on minor allele frequencies, functionality, or estimated effect size to better target variants of interest.

ACKNOWLEDGMENTS

The authors thank Dr. Robert Yolken at Johns Hopkins University for providing antibody data. This work was supported by National Institutes of Health grants R00 ES017744 (to A.M.), R01 MH074027 (to J.Y.T. and P.F.S.), and P01 CA142538 (to J.Y.T.).

REFERENCES

Cristianini N, Shawe-Taylor J. 2000. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge: Cambridge University Press.

Duchesne P, 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.

Gottesman II, Gould TD. 2003. The endophenotype concept in psychiatry: etymology and strategic intentions. Am J Psychiatry 160:636–645.

Harville DA. 1977. Maximum likelihood approaches to variance component estimation and to related problems. J Am Stat Assoc 72:320–338.

International Schizophrenia Consortium. 2009. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460:748–752.

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

Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. 2008. A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 82:386–397.

Lange C, John C, Whittaker JC, Macgregor AJ. 2002. Generalized estimating equations: a hybrid approach for mean parameters in multivariate regression models. Stat Modelling 2:163–181.

Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, Keefe RS, Davis SM, Davis CE, Lebowitz BD, Severe J, Hsiao JK. 2005. Effectiveness of antipsychotic drugs in patients with chronic schizophrenia. N Engl J Med 22:1209–1223.

Liu J, Pei Y, Papasian CJ, Deng HW. 2009. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. Genet Epidemiol 33:217–727.

Maity A, Lin X. 2011. Powerful tests for detecting a gene effect in the presence of possible gene-gene interactions using garrote kernel machines. Biometrics 67:1271–1284.

Shi J, Levinson DF, Duan J, Sanders AR, Zheng Y, Pe'er I, Dudbridge F, Holmans PA, Whittemore AS, Mowry BJ, Olincy A, Amin F, Cloninger CR, Silverman JM, Buccola NG, Byerley WF, Black DW, Crowe RR, Oksenberg JR, Mirel DB, Kendler KS, Freedman R, Gejman PV. 2009. Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460:753–757.

Stefansson H, Ophoff RA, Steinberg S, Andreassen OA, Cichon S, Rujescu D, Werge T, Pietiläinen OP, Mors O, Mortensen PB, Sigurdsson E, Gustafsson O, Nyegaard M, Tuulio-Henriksson A, Ingason A, Hansen T, Suvisaari J, Lonnqvist J, Paunio T, Børglum AD, Hartmann A, Fink-Jensen A, Nordentoft M, Hougaard D, Norgaard-Pedersen B, Böttcher Y, Olesen J, Breuer R, Möller HJ, Giegling I, Rasmussen HB, Timm S, Mattheisen M, Bitter I, Réthelyi JM, Magnusdottir BB, Sigmundsson T, Olason P, Masson G, Gulcher JR, Haraldsson M, Fossdal R, Thorgeirsson TE, Thorsteinsdottir U, Ruggeri M, Tosato S, Franke B, Strengman E, Kiemeney LA; Genetic Risk and Outcome in Psychosis (GROUP), Melle I, Djurovic S, Abramova L, Kaleda V, Sanjuan J, de Frutos R, Bramon E, Vassos E, Fraser G, Ettinger U, Picchioni M, Walker N, Toulopoulou T, Need AC, Ge D, Yoon JL, Shianna KV, Freimer NB, Cantor RM, Murray R, Kong A, Golimbet V, Carracedo A, Arango C, Costas J, Jönsson EG, Terenius L, Agartz I, Petursson H, Nöthen MM, Rietschel M, Matthews PM, Muglia P, Peltonen L, St Clair D, Goldstein DB, Stefansson K, Collier DA. 2009. Common variants conferring risk of schizophrenia. Nature 460:744–747.

Suhre K, Shin SY, Petersen AK, Mohney RP, Meredith D, Wägele B, Altmaier E; CARDIoGRAM, Deloukas P, Erdmann J, Grundberg E, Hammond CJ, de Angelis MH, Kastenmüller G, Köttgen A, Kronenberg F, Mangino M, Meisinger C, Meitinger T, Mewes HW, Milburn MV, Prehn C, Raffler J, Ried JS, Römisch-Margl W, Samani NJ, Small KS, Wichmann HE, Zhai G, Illig T, Spector TD, Adamski J, Soranzo N, Gieger C. 2011. Human metabolic individuality in biomedical and pharmaceutical research. Nature 477:54–60.

Sullivan PF, Lin D, Tzeng JY, van den Oord E, Perkins D, Stroup TS, Wagner M, Lee S, Wright FA, Zou F, Liu W, Downing AM, Lieberman J, Close SL. 2008. Genomewide association for schizophrenia in the CATIE study: results of stage 1. Mol Psychiatry 13:570–584.

Tzeng JY, Bondell H. 2010. A comprehensive approach to haplotype specific analysis via penalized likelihood. Eur J Hum Genet 18:95–103.

Tzeng JY, Zhang D. 2007. Haplotype-based association analysis via variance component score test. Am J Hum Genet 81:927–938.

Tzeng JY, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, Worrall BB, Hsu FC, Thomas DC, Sullivan PF. 2011. Detecting gene and gene-environment effects of common and uncommon variants on quantitative traits: a marker-set approach using gene-trait similarity regression. Am J Hum Genet 89:277–288.

Verzilli CJ, Stallard N, Whittaker JC. 2005. Bayesian modelling of multivariate quantitative traits using seemingly unrelated regressions. Genet Epidemiol 28:313–325.

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X. 2010. Powerful SNP set analysis for case-control genome wide association studies. Am J Hum Genet 86:929–942.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82–93.

Yolken RH, Torrey EF, Lieberman JA, Yang S, Dickerson FB. 2011. Serological evidence of exposure to Herpes Simplex Virus type 1 is associated with cognitive deficits in the CATIE schizophrenia sample. Schizophr Res 128:61–65.

Zapala MA, Schork NJ. 2006. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci USA 103:19430–19435.

Zhang H, Liu CT, Wang X. 2010. An association test for multiple traits based on the generalized Kendall's Tau. J Am Stat Assoc 105:473–481.

Zhu W, Zhang H. 2009. Why do we test multiple traits in genetic association studies? J Korean Stat Soc 38:1–10.