Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 8812282, 11 pages
https://doi.org/10.1155/2021/8812282

Research Article

Associating Multivariate Traits with Genetic Variants Using Collapsing and Kernel Methods with Pedigree- or Population-Based Studies

Li-Chu Chien

Center for Fundamental Science, Kaohsiung Medical University, Kaohsiung, Taiwan

Correspondence should be addressed to Li-Chu Chien; lcchien@kmu.edu.tw

Received 6 September 2020; Revised 2 January 2021; Accepted 8 January 2021; Published 10 February 2021

Academic Editor: Giuseppe Pontrelli

In genetic association analysis, several relevant phenotypes or multivariate traits with different types of components are usually collected to study complex or multifactorial diseases. Over the past few years, jointly testing for association between multivariate traits and multiple genetic variants has become more popular because it can increase statistical power to identify causal genes in pedigree- or population-based studies. However, most of the existing methods mainly focus on testing genetic variants associated with multiple continuous phenotypes. In this investigation, we develop a framework for identifying the pleiotropic effects of genetic variants on multivariate traits by using collapsing and kernel methods with pedigree- or population-structured data. The proposed framework is applicable to the burden test, the kernel test, and the omnibus test for autosomes and the X chromosome. The proposed multivariate trait association methods can accommodate continuous phenotypes or binary phenotypes and further can adjust for covariates. Simulation studies show that the performance of our methods is satisfactory with respect to the empirical type I error rates and power rates in comparison with the existing methods.

1. Introduction

Genome-wide association studies (GWAS) intend to find genetic variants such as single nucleotide polymorphisms (SNPs) associated with common traits or with complex diseases [1, 2]. Association studies, where the correlation relationship between a genetic variant and a trait is evaluated, are helpful for mapping genes influencing complex diseases [3]. In the study of complex diseases, data on several correlated phenotypes or a multivariate phenotype with several components are often collected to get a better understanding of the disease [1, 3, 4]. Multivariate correlated traits are influenced through multiple variants simultaneously. Therefore, by a suitable joint or multivariate analysis framework of multivariate traits, we can not only gain more statistical power to identify pleiotropic effects of genetic variants on multivariate traits [3, 5–12] but also can further understand the genetic architecture of the disease of interest [5, 13]. Thus, recently, the joint analysis of multivariate traits has become popular because it can increase statistical power over analyzing only one trait at a time [1, 4].

Several statistical methods have been developed to identify the association between multivariate traits and a genetic variant [1, 5]. Current multivariate methods can be classified into three groups [1, 2, 5]: regression methods [14–16], variable reduction methods [11, 13, 17, 18], and combining analysis [9, 19–23]. However, many of the existing methods for multivariate association analysis cannot be straightaway extended to rare variant analyses, due to their enormous numbers causing the problems of multiple comparison or multiple testing and their low minor allele frequencies [2, 5, 24]. Moreover, sparsity of data could lead to problems on estimating regression parameters and fitting regression models [2]. Hence, it is necessary for proposing statistical methods for identifying the association between multivariate traits and multiple genetic variants (common and/or rare variants) [5]. In recent years, various statistical techniques have been proposed for this purpose in GWAS [8, 17, 25–27]. Furthermore, several approaches have been extendedly developed for the investigation of rare variants associated with multivariate traits [2, 28–38].

Although these new developments keep many benefits, existing methods have some potential limitations [39]. Most current methods are constructed under some specific assumptions about the effects of genetic variants on multivariate traits [39]. These current approaches suffer a severe loss in power once the model assumptions are violated [26, 39].

In this investigation, we develop the statistical methods for identifying pleiotropic effects of genetic variants on multivariate traits using collapsing and kernel methods with pedigree- or population-structured data. The proposed multivariate trait association method is able to handle binary phenotypes or continuous phenotypes and further can adjust for covariates. Moreover, the proposed multivariate trait association method not only can leverage the dependence on the phenotypes but also can account for the sample relatedness in the pedigree-based or population-based structured data.

The rest of the article is organized as follows. In the materials and methods section, we construct the multivariate effect model using the joint GEE model formulation (JGEE) [40]. We apply the JGEE to pedigree- or population-structured data and introduce a retrospective framework for analyzing multivariate traits in genetic association studies. The proposed framework is applicable to the burden test, the kernel test, and the omnibus test for autosomes and the X chromosome. In the simulation studies, we examine the finite sample size performance of the proposed multivariate association methods and evaluate the comparison results with the existing method, Multi-SKAT [39]. Concluding remarks and future possibilities for continuity are given in the conclusion section and the limitation section.

2. Materials and Methods

2.1. Notations. To describe the proposed multivariate trait association method based on the pedigree- or population-based structured data, we suppose that there are N independent pedigrees and each pedigree has n_i subjects. We assume that the n_i subjects have been sequenced in a genetic region of interest (e.g., a gene) that contains p variants. Let $\mathbf{y}_{ik} = (y_{i1k}, y_{i2k}, \cdots, y_{\mathrm{in}_i k})^T$ be the $n_i \times 1$ phenotype vector for the k^{th} phenotype of the i^{th} pedigree. Let $\mathbf{y}_i = (\mathbf{y}_{i1}, \mathbf{y}_{i2}, \cdots, \mathbf{y}_{iK})$ be the $(n_i \times K) \times 1$ response vector for the K phenotypes that we are interested in. Let $\mathbf{x}_{im} = (x_{i1m}, x_{i2m}, \cdots, x_{\mathrm{in}_i m})^T$ be the $n_i \times 1$ vector for the m^{th} covariate of the i^{th} pedigree. Let $\mathbf{x}_i = (\mathbf{x}_{i0}, \mathbf{x}_{i1}, \cdots, \mathbf{x}_{iq})$ be the $n_i \times (q+1)$ covariate matrix for the $(q+1)$ nongenetic covariates that we want to adjust for. Let $\boldsymbol{\alpha}_k = (\alpha_{0k}, \alpha_{1k}, \cdots, \alpha_{qk})^T$ be the $(q+1) \times 1$ vector of regression coefficients of the $(q+1)$ nongenetic covariates with the element α_{mk} being the effect of the m^{th} covariate on the k^{th} trait. Let $\mathbf{g}_i = (\mathbf{g}_{i1}, \mathbf{g}_{i2}, \cdots, \mathbf{g}_{ip})$ be the $n_i \times p$ genetic matrix for p genetic variants in a target region of interest where $\mathbf{g}_{il} = (\mathrm{g}_{i1l}, \mathrm{g}_{i2l}, \cdots, \mathrm{g}_{\mathrm{in}_i l})^T$ is the $n_i \times 1$ vector for a genetic variant l ($\mathrm{g}_{ijl} = 0$, 1, or 2 for 0, 1, or 2 copies of the minor allele, respectively). Let $\boldsymbol{\beta}_k = (\beta_{1k}, \beta_{2k}, \cdots, \beta_{pk})^T$ be the $p \times 1$ vector of regression coefficients of the p genetic variants with the element β_{lk} being the effect of the l^{th} genetic variant on the k^{th} trait.

2.2. Multitrait Regression-Based Tests for Pedigree Data. We let $\mathbf{X}_i = \mathbf{I}_K \otimes \mathbf{x}_i$ be the $(n_i \times K) \times ((q+1) \times K)$ covariate matrix and $\mathbf{G}_i = \mathbf{I}_K \otimes \mathbf{g}_i$ be the $(n_i \times K) \times (p \times K)$ genotype matrix for the i^{th} pedigree where $\mathbf{I}_K$ is an identity matrix of dimension $K \times K$ and $\otimes$ stands for the Kronecker product. According to the generalized linear model [41], we assume that the marginal density of y_{ijk} is $f(y_{ijk}) = \exp[\{y_{ijk}\theta_{ijk} - a(\theta_{ijk}) + b(y_{ijk})\}/\phi]$ with two moments, $\mu_{ijk} = E(y_{ijk}) = \partial a(\theta_{ijk})/\partial\theta_{ijk}$ and $\mathrm{Var}(y_{ijk}) = (\partial\mu_{ijk}/\partial\theta_{ijk})\phi$, where ϕ is a scale parameter. Let $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}^T, \boldsymbol{\theta}_{i2}^T, \cdots, \boldsymbol{\theta}_{iK}^T)^T$ be the $(n_i \times K) \times 1$ vector with the components $\boldsymbol{\theta}_{ik} = (\theta_{i1k}, \theta_{i2k}, \cdots, \theta_{\mathrm{in}_i k})^T$, $k = 1, 2, \cdots, K$ and $\boldsymbol{\eta}_i = (\boldsymbol{\eta}_{i1}^T, \boldsymbol{\eta}_{i2}^T, \cdots, \boldsymbol{\eta}_{iK}^T)^T$ be the $(n_i \times K) \times 1$ vector with the components $\boldsymbol{\eta}_{ik} = (\eta_{i1k}, \eta_{i2k}, \cdots, \eta_{\mathrm{in}_i k})^T = \mathbf{x}_i\boldsymbol{\alpha}_k + \mathbf{g}_i\boldsymbol{\beta}_k$, $k = 1, 2, \cdots, K$ for the k^{th} trait of the i^{th} pedigree.

Based on the joint GEE model formulation [40], we construct the multivariate linear model for describing the association relationship between K correlated traits and genetic variants, which is given as follows:

$$\boldsymbol{\mu}_i = g(\boldsymbol{\eta}_i) \text{ and } \boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{G}_i\boldsymbol{\beta}, \tag{1}$$

where $g^{-1}(\bullet)$ is the inverse function of $g(\bullet)$ and is a response-specific link function [40], $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_{i1}^T, \boldsymbol{\mu}_{i2}^T, \cdots, \boldsymbol{\mu}_{iK}^T)^T$ is the $(n_i \times K) \times 1$ vector of the expected mean of the multivariate traits $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \mathbf{y}_{i2}^T, \cdots, \mathbf{y}_{iK}^T)^T$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \cdots, \boldsymbol{\alpha}_K^T)^T$ is the $((q+1) \times K) \times 1$ vector of regression coefficients of the $(q+1)$ nongenetic covariates for the K correlated traits, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \cdots, \boldsymbol{\beta}_K^T)^T$ is the $(p \times K) \times 1$ vector of regression coefficients of the p genetic variants for the K correlated traits.

Let $\mathbf{R}_{n_i}(\boldsymbol{\varphi})$ and $\mathbf{R}_K(\boldsymbol{\gamma})$ be the $n_i \times n_i$ within-in cluster correlation matrix and the $K \times K$ multivariate-response cluster correlation matrix, which depend on a vector of parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\gamma}$, respectively. The $(n_i \times K) \times (n_i \times K)$ working (or approximate) covariance matrix of $\mathbf{y}_i$ is given by [40].

$$\mathbf{V}_i = \mathbf{A}_i^{1/2}\left(\mathbf{R}_K(\boldsymbol{\gamma}) \otimes \mathbf{R}_{n_i}(\boldsymbol{\varphi})\right)\mathbf{A}_i^{1/2}\phi, \tag{2}$$

where $\boldsymbol{A}_i = \mathrm{diag}(\boldsymbol{A}_{i1}, \boldsymbol{A}_{i2}, \cdots, \boldsymbol{A}_{iK})$ is a $(n_i \times K) \times (n_i \times K)$ block diagonal matrix with the components $\boldsymbol{A}_{ik} = \mathrm{diag}(\partial\mu_{i1k}/\partial\theta_{i1k}, \partial\mu_{i2k}/\partial\theta_{i2k}, \cdots, \partial\mu_{\mathrm{in}_i k}/\partial\theta_{\mathrm{in}_i k})$, $k = 1, 2, \cdots, K$ being the $n_i \times n_i$ diagonal matrices. According to equation (1), under the null hypothesis of no association between genotypes and phenotypes, we propose the multivariate association methods including the homogeneous kernel statistic (HoK), the heterogeneous kernel statistic (HeK), and burden test (BT). Moreover, we propose the homogeneous omnibus test (HoO) and heterogeneous omnibus test (HeO) by combining the HoK with the BT and by combining the HeK with the BT, respectively.

2.2.1. Kernel Statistic. We let $\boldsymbol{H}$ be a $p \times p$ correlation matrix of genotype scores with element $H_{ll'}$ for markers l and l'. Let

m_l denote the minor allele frequency (MAF) of the l^{th} marker. Let $\boldsymbol{S}_i = \widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_i) = (\boldsymbol{S}_{i1}^T, \boldsymbol{S}_{i2}^T, \cdots, \boldsymbol{S}_{iK}^T)^T$ be the $((n_i \times K) \times 1)$ vector of the standard residuals with components $\boldsymbol{S}_{ik} = (S_{i1k}, S_{i2k}, \cdots, S_{\text{in}_i k})^T, k = 1, 2, \cdots, K$, where $\widehat{\mathbf{V}}_i^{-1}$ is the inverse matrix of $\widehat{\mathbf{V}}_i$. Here, $\widehat{\mathbf{V}}_i$ and $\widehat{\boldsymbol{\mu}}_i$ are the estimates of $\boldsymbol{V}_i$ and $\boldsymbol{\mu}_i$. Here and henceforth, all estimates are calculated based on the null hypothesis of the genetic effects $\boldsymbol{\beta}$ equal to zero. All unknown parameters and the working within-in and multivariate-response cluster correlation matrices are estimated by the R package JGEE [42].

(1) Homogeneous Kernel Statistic. We suppose that w_l is a marker-specific weight of the l^{th} variant and assume that the genetic effects on the K different phenotypes are homogeneous (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_K$). Based on the JGEE model with the genotype as random variables considered, we propose the homogeneous quadratic (kernel) association statistic (HoK) as follows:

$$\begin{aligned}\kappa_{\text{Ho}} &= \sum_{l=1}^{p}\left[w_l \sum_{k=1}^{K}\sum_{i=1}^{N} \mathbf{g}_{il}^T \Delta\wedge_{ik} \boldsymbol{A}\wedge_{ik} \boldsymbol{S}_{ik}\right]^2 = \sum_{l=1}^{p}\left[w_l \sum_{k=1}^{K} Z_{lk}\right]^2 \\ &= \sum_{l=1}^{p}[w_l Z_l]^2 = \sum_{l=1}^{p}\tilde{Z}_l^2,\end{aligned} \tag{3}$$

where $\tilde{Z}_l = w_l Z_l$, $Z_l = \sum_{k=1}^{K} Z_{lk}$, $Z_{lk} = \sum_{i=1}^{N} \mathbf{g}_{il}^T \widehat{\Delta}_{ik} \widehat{\boldsymbol{A}}_{ik} \boldsymbol{S}_{ik}$, $\widehat{\boldsymbol{A}}_{ik}$ is the estimate of $\boldsymbol{A}_{ik}$, and $\widehat{\Delta}_{ik}$ is the estimate of $\Delta_{ik} = \text{diag}\,(\partial\theta_{i1k}/\partial\eta_{i1k}, \partial\theta_{i2k}/\partial\eta_{i2k}, \cdots, \partial\theta_{\text{in}_i k}/\partial\eta_{\text{in}_i k})$ that is a $n_i \times n_i$ diagonal matrix for the k^{th} phenotype of the i^{th} pedigree. The null distribution of κ_{Ho} asymptotically follows a mixture chi-square distribution $\sum_{l=1}^{p} \lambda_l \chi_{l,1}^2$, where $\chi_{l,1}^2$s are independent random variables following a chi-square distribution with one degree of freedom and $(\lambda_1, \lambda_2, \cdots, \lambda_p)$ are nonzero eigenvalues of the null covariate matrix of $\text{Cov}_0(\tilde{Z}_l, \tilde{Z}_{l'}) = 2C_{\text{Ho}} w_l w_{l'} H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}$ where $C_{\text{Ho}} = \sum_{i=1}^{N}[(\sum_{k=1}^{K} \boldsymbol{S}_{ik}^T \widehat{\boldsymbol{A}}_{ik} \widehat{\Delta}_{ik}) \Omega_i (\sum_{k=1}^{K} \widehat{\Delta}_{ik} \widehat{\boldsymbol{A}}_{ik} \boldsymbol{S}_{ik})]$ and Ω_i is a $n_i \times n_i$ matrix of genetic correlations for all n_i individuals in the i^{th} pedigree, which has the same definition given by Schaid et al. [43] and can be calculated by the R package kinship2 [44]. When the genetic relationship between subjects j and j' in the i^{th} pedigree is unknown, the elements of the genetic correlation Ω_i can be estimated through genomic data [43, 45], and its estimate is given by [43]

$$\widehat{\Omega}_i = \frac{1}{p}\sum_{l=1}^{p} \frac{\left(g_{ijl} - 2m_l\right)\left(g_{ij'l} - 2m_l\right)}{2m_l(1-m_l)}. \tag{4}$$

(2) Heterogeneous Kernel Statistic. We assume that the genetic effects on the K different phenotypes are heterogeneous (i.e., $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \neq \cdots \neq \boldsymbol{\beta}_K$). The heterogeneous quadratic (kernel) association statistic (HeK) is defined by

$$\begin{aligned}\kappa_{\text{He}} &= \sum_{l=1}^{p}\sum_{k=1}^{K}\left[w_{lk}\sum_{i=1}^{N} \mathbf{g}_{il}^T \Delta\wedge_{ik} \boldsymbol{A}\wedge_{ik} \boldsymbol{S}_{ik}\right]^2 = \sum_{l=1}^{p}\sum_{k=1}^{K}[w_{lk} Z_{lk}]^2 \\ &= \sum_{l=1}^{p}\sum_{k=1}^{K}\tilde{Z}_{lk}^2,\end{aligned} \tag{5}$$

where $\tilde{Z}_{lk} = w_{lk} Z_{lk}$ and w_{lk} is a marker-specific weight of the l^{th} variant of the k^{th} trait. The null distribution of κ_{He} asymptotically follows a mixture chi-square distribution $\sum_{l=1}^{(p\times K)} \lambda_l \chi_{l,1}^2$, where $\chi_{l,1}^2$s are independent random variables following a chi-square distribution with one degree of freedom, and $(\lambda_1, \lambda_2, \cdots, \lambda_{(p\times K)})$ are nonzero eigenvalues of the null covariate matrix of $\text{Cov}_0(\tilde{Z}_{lk}, \tilde{Z}_{l'k'}) = 2C_{\text{He}} w_{lk} w_{l'k'} H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}$, where $C_{\text{He}} = \sum_{i=1}^{N}[\mathbf{S}_{ik}^T \widehat{\mathbf{A}}_{ik} \widehat{\Delta}_{ik} \Omega_i \widehat{\Delta}_{ik'} \widehat{\mathbf{A}}_{ik'} \mathbf{S}_{ik'}]$.

Theoretical p values of κ_{Ho} and κ_{He} are approximately calculated by Kuonen's saddlepoint method [46] and obtained by the R package pchisqsum. A theory for the derivation of the HoK test (κ_{Ho}) and the HeK test (κ_{He}) is shown in Appendix S1.

2.2.2. Burden Test. We let $\tilde{\mathbf{g}}_i^T = \sum_{l=1}^{p} w_l \mathbf{g}_{il}^T$ be a weighted average of genotype scores for the i^{th} pedigree. On the basis of the HoK test (κ_{Ho}) and the HeK test (κ_{He}) in equations (3) and (5) with the same marker-specific weight of the l^{th} variant for each trait k (i.e., $w_l = w_{lk}, k = 1, 2, \cdots, K$), we propose the burden test (BT) as follows:

$$\text{BT} = \frac{\left[\sum_{i=1}^{N}\left(\sum_{k=1}^{K} \boldsymbol{S}_{ik}^T \boldsymbol{A}\wedge_{ik} \Delta\wedge_{ik}\right)\tilde{\mathbf{g}}_i\right]^2}{\sum_{i=1}^{N}\left[\left(\sum_{k=1}^{K} \boldsymbol{S}_{ik}^T \widehat{\boldsymbol{A}}_{ik} \widehat{\Delta}_{ik}\right)\text{Cov}_0(\tilde{\mathbf{g}}_i)\left(\sum_{k=1}^{K} \widehat{\Delta}_{ik} \widehat{\boldsymbol{A}}_{ik} \boldsymbol{S}_{ik}\right)\right]}, \tag{6}$$

where the null covariance matrix of $\tilde{\mathbf{g}}_i$ is given by

$$\begin{aligned}\text{Cov}_0(\tilde{\mathbf{g}}_i) &= \text{Cov}_0\left(\sum_{l=1}^{p} w_l \mathbf{g}_{il}\right) = \sum_{l=1}^{p} w_l^2 \text{Cov}_0(\mathbf{g}_{il}, \mathbf{g}_{il}) \\ &\quad + 2\sum_{l=1}^{p}\sum_{l'=l+1}^{p} w_l w_{l'} \text{Cov}_0(\mathbf{g}_{il}, \mathbf{g}_{il'}) \\ &= \Omega_i \sum_{l=1}^{p}\sum_{l'=1}^{p} 2w_l w_{l'} H_{ll'} \sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}.\end{aligned} \tag{7}$$

Then,

$$\mathrm{BT} = \frac{\left[\sum_{i=1}^{N}\left(\sum_{k=1}^{K} \boldsymbol{S}_{ik}^{T}\boldsymbol{A}\wedge_{ik}\Delta\wedge_{ik}\right)\tilde{\mathbf{g}}_i\right]^2}{2\sum_{l=1}^{p}\sum_{l'=1}^{p} w_l w_{l'} H_{ll'}\sqrt{m_l(1-m_l)m_{l'}(1-m_{l'})}C_{\mathrm{Ho}}}. \tag{8}$$

The null distribution of BT asymptotically follows a chi-square distribution with one degree of freedom.

2.2.3. Omnibus Test. Let p_{Ho}, p_{He}, and p_{BT} denote the p values obtained by the HoK, HeK, and BT statistics. Based on the idea of the p value combination method through the Cauchy distribution [47–49], we propose the homogeneous omnibus test (HoO) and heterogeneous omnibus test (HeO).

(1) Homogeneous Omnibus Test. Combining the p_{Ho} with the p_{BT}, we construct the homogeneous omnibus test (HoO) as follows:

$$O_{\mathrm{Ho}} = -\frac{1}{2}\left[F_C^{-1}(p_{\mathrm{Ho}}) + F_C^{-1}(p_{\mathrm{BT}})\right], \tag{9}$$

where F_C^{-1} stands for the inverse cumulative distribution function of the standard Cauchy distribution.

(2) Heterogeneous Omnibus Test. Combining the p_{He} with the p_{BT}, we construct the heterogeneous omnibus test (HeO) as follows:

$$O_{\mathrm{He}} = -\frac{1}{2}\left[F_C^{-1}(p_{\mathrm{He}}) + F_C^{-1}(p_{\mathrm{BT}})\right]. \tag{10}$$

The null distributions of the O_{Ho} test and the O_{He} test asymptotically follow a standard Cauchy distribution [47–49]. The p values of the O_{Ho} test and the O_{He} test are calculated by the R package RNOmni [50].

The kernel statistic, the burden test, and the omnibus test are also applicable to the X chromosome. Additional technical information for extensions to the X chromosome is shown in Appendix S2.

3. Simulation Studies

We conduct the numerical simulation studies to assess the finite sample performance of the proposed methods and evaluate the comparison results with two existing methods, the minimum p value SKAT statistic (mPK), and the minimum p value burden statistic (mPB) [39]. The two existing methods are implemented by the R package Multi-SKAT [39]. Based on the similar simulation set-up as those usually considered from existing genetic association tests [39, 43, 51], we investigate the effect of the proposed methods, HoK, HeK, BT, HoO, and HeO, for identifying genetic variants that are associated with multiple traits. We simultaneously generate 10,000 European-like (EUR) and 10,000 admixed African American-like (AA) haplotypes of length 200 kb using a calibrated human demographic model through the COSI software [51, 52]. A 3 kb region is randomly selected in our numerical simulations. We generate a total of 10,000 databases for each simulation scenario in our studies.

3.1. Type I Error Rate and Power Simulations. In the heterogeneous population with nuclear family data considered, continuous and binary phenotypes for trait k for individual j in the i^{th} family are generated from the multivariate linear model in equation (1) with $K = 2$ and $n_i = 3$. More precisely, continuous and binary phenotypes are generated by the following linear and logit models, respectively:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{G}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i, \tag{11}$$

$$\mathrm{logit}(P(\mathbf{y}_i = 1)) = \mathbf{X}_i\boldsymbol{\alpha} + \mathbf{G}_i\boldsymbol{\beta}, \tag{12}$$

where $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \mathbf{y}_{i2}^T)^T$, $\mathbf{X}_i = \mathbf{I}_2 \otimes \mathbf{x}_i$, $\mathbf{G}_i = \mathbf{I}_2 \otimes \mathbf{g}_i$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i11}, \varepsilon_{i21}, \varepsilon_{i31}, \varepsilon_{i12}, \varepsilon_{i22}, \varepsilon_{i32})^T$. Here, the elements $\mathbf{x}_{i0} = (x_{i10}, x_{i20}, x_{i30})^T$ of the covariance matrix $\mathbf{x}_i = (\mathbf{x}_{i0}, \mathbf{x}_{i1}, \mathbf{x}_{i2})$ is a 3×1 vector of all ones. The elements $\mathbf{x}_{i1} = (x_{i11}, x_{i21}, x_{i31})^T$ of $\mathbf{x}_i$ are independently generated with an equal probability of being 0 or 1. The elements $\mathbf{x}_{i2} = (x_{i12}, x_{i22}, x_{i32})^T$ of $\mathbf{x}_i$ are generated from a multivariate normal distribution with a mean of 0.5 and a covariance matrix with diagonal entries of 1 and all off-diagonal entries of 0.1. The regression coefficients of the covariate matrix $\mathbf{x}_i$ for the k^{th} correlated trait are given by $\boldsymbol{\alpha}_k = (\alpha_{0k}, \alpha_{1k}, \alpha_{2k})^T = (0.01, 0.1, 0.1)^T$ and $\boldsymbol{\alpha}_k = (\alpha_{0k}, \alpha_{1k}, \alpha_{2k})^T = (-1.4, 0.1, 0.1)^T$, respectively, for continuous traits and binary traits for $k = 1, 2$.

For continuous traits, the error terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{i11}, \varepsilon_{i21}, \varepsilon_{i31}, \varepsilon_{i12}, \varepsilon_{i22}, \varepsilon_{i32})^T$ in equation (11) follow a multivariate normal distribution having a mean of zero, a within-in cluster correlation matrix (i.e., $\mathrm{Cor}(\varepsilon_{ijk}, \varepsilon_{ij'k})$) with diagonal entries of 1 and all off-diagonal entries of 0.2 and a subject-across-response correlation matrix (i.e., $\mathrm{Cor}(\varepsilon_{ijk}, \varepsilon_{ij'k'})$) with diagonal entries of 0.3 and all off-diagonal entries of 0.1. Similarly, binary traits $\mathbf{y}_i$ in equation (12) are generated with the same within-in cluster correlation matrix (i.e., $\mathrm{Cor}(y_{ijk}, y_{ij'k})$) and the same subject-across-response correlation matrix (i.e., $\mathrm{Cor}(y_{ijk}, y_{ij'k'})$) as the continuous traits $\mathbf{y}_i$ in equation (11). These correlated phenotypes are generated by the R package BinNor [53].

For type I error simulations, the regression coefficients of genetic variants, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, in equations (11) and (12) are equal to zero under the null hypothesis. For power simulations, under the alternative hypothesis, we simulate that 35% of low variants with the MAF < 0.03 are causal. For each setting, either all causal SNPs have a positive effect, or 80% of causal SNPs are positive, and 20% of causal SNPs are negative. The regression coefficients of genetic variants, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, are set by $0.095 \times |\log_{10}(m_l)|$ or $0.095 \times \log_{10}(m_l)$ corresponding to the risk or protective variant $l, l = 1, 2, \cdots, p$ [51]. Under the assumption that the genetic effects on the two different phenotypes are heterogeneous (i.e., $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$), the genetic effects $\boldsymbol{\beta}_1$ for the first traits $\mathbf{y}_{i1}$ are set as described

above, while the genetic effects $\boldsymbol{\beta}_2$ for the second traits $\mathbf{y}_{i2}$ are set by zero. On the other hand, under the assumption that the genetic effects on the two different phenotypes are homogeneous (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$), the genetic effects $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ for the first and second traits have the same settings as described above.

We simulate 1,400 nuclear families with 800 nuclear families from the European samples and 600 nuclear families from African-American samples. The marker-specific weight w_l for variant l is considered as the beta density function $w_l = \text{Beta}(m_l, \lambda_1, \lambda_2)$ with shape parameters $\lambda_1 > 0$ and $\lambda_2 > 0$ [51]. To study the effect of the marker-specific weight w_l of variant l on the phenotypes, we consider the unweighted marker-specific weight with $w_l = \text{Beta}(m_l, 1, 1) = 1$ and the weighted marker-specific weight with $w_l = \text{Beta}(m_l, 1, 25)$ [51]. The empirical type I error rates based on fifty thousand replicates and the empirical power rates based on two thousand replicates are reported for all simulation results. The "exchangeable" and "unstructured" structures are considered for the working within-cluster and multivariate-response correlation matrices for the proposed methods, HoK, HeK, and BT, respectively.

4. Results

4.1. Empirical Type I Error Rates. Table 1 reports the results of a simulation comparison on empirical type I error rates when the phenotypes are considered to be continuous. Table 1 displays that the proposed methods, HoK, HoO, HeK, HeO, and BT, well control the empirical type I error rates regardless of the weight of the marker-specific weight. Similarly, the existing methods, mPK and mPB, have good performance on controlling the empirical type I error rates. Our simulation results show that the seven competing methods, HoK, HoO, HeK, HeO, BT, mPK, and mPB, reasonably control the empirical type I error rates for autosome analyses with continuous traits. The seven competing approaches display similar performance in terms of the empirical type I error rates for the X chromosome analyses with continuous traits (Appendix S3: Table S1).

Table 2 reports the empirical type I error rates based on the proposed methods, HoK, HeK, BT, HoO, and HeO, for the binary data. The two existing methods, mPK and mPB, aren't included for comparison. This reason is that implementing the two existing methods, mPK and mPB, via the R package Multi-SKAT [39], the MPMM (multiple phenotype mixed model) function in the R package PHENIX [54–56] is a necessary tool for this process. However, the MPMM function is suitable for the continuous phenotypes [56] or is suitable for the binary phenotypes with the condition that the number of cases is sufficiently large [39]. In other words, in some sense, the two existing methods, mPK and mPB, are limited to continuous phenotypes [39].

Table 2 shows that the proposed methods appropriately control the type I error rates when the marker-specific weight is considered for $w_l = \text{Beta}(m_l, 1, 1)$ or $w_l = \text{Beta}(m_l, 1, 25)$ for variant l for binary traits. On the other hand, the empirical type I error rates of the proposed methods for X chromosome analyses with binary traits are depicted in Table S2 in Appendix S3. These empirical type I error rates show similar results as that for autosome analyses.

In summary, our simulation results show that the proposed multivariate trait association methods, HoK, HoO, HeK, HeO, and BT, have reasonable control of type I error rates for continuous traits or binary traits whether the marker is X chromosomal or autosomal. On the other hand, the existing methods, mPK and mPB, yield well-controlled type I error rates for the autosome analyses or the X chromosome analyses with continuous traits (Table 1 or Table S1), regardless of the weight of the marker-specific weight.

4.2. Empirical Power. Figure 1 exhibits the comparison results of the empirical power rates for the autosome analyses with continuous traits, when the working within-cluster and multivariate-response correlation matrices of the proposed methods, HoK, HeK, and BT, are considered to be exchangeable. As expected, the empirical power rates of the seven competing methods with a weighted marker-specific weight of $w_l = \text{Beta}(m_l, 1, 25)$ are higher than that with an unweighted marker-specific weight of $w_l = \text{Beta}(m_l, 1, 1) = 1$. The heterogeneous kernel statistic (HeK) has slightly greater empirical power rates than other methods, when the genetic effects on the different phenotypes are heterogeneous (i.e., $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$), and causal SNPs have positive effects or negative effects on phenotypes. On the other hand, the existing method, mPB, has bigger empirical power rates, when the genetic effects on the different phenotypes are heterogeneous (i.e., $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$), and all causal SNPs have a positive association on phenotypes. Moreover, the empirical power rates of the homogeneous omnibus test (HoO) are larger than that of the other six competing methods, when the genetic effects on the different phenotypes are homogeneous (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$). Evidently, the seven competing methods have their respective advantages in identifying the association between genetic effects and multiple continuous traits for autosome analyses.

Similar empirical power rates are obtained from the working within-cluster and multivariate-response correlation matrices of the proposed methods, HoK, HeK, and BT, considered to be unstructured. Hence, these empirical power rates are not shown in order to save space. On the other hand, the seven competing approaches display a similar performance in testing for the X chromosome analyses with continuous traits (Appendix S3: Figure S1).

Figure 2 exhibits the comparison results of empirical power rates for the autosome analyses with binary traits when the working within-cluster and multivariate-response correlation matrices of the proposed methods, HoK, HeK, and BT, are considered to be exchangeable. As a similar reason for investigating the empirical type I error rates with binary traits, the two existing methods, mPK and mPB, aren't included for power comparison.

Figure 2 shows that the heterogeneous kernel statistic (HeK) and the heterogeneous omnibus test (HeO) outperform over other methods in terms of the empirical power rates, when the genetic effects on the different phenotypes are heterogeneous (i.e., $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$). On the other hand, the empirical power rates of the homogeneous omnibus test (HoO) are bigger than that of the other competing methods,

Table 1: Empirical type I errors of the seven competing methods with continuous traits.

Marker-specific weight (w_l)	Nominal level	Working correlation	Method						
			HoK[3]	HoO	HeK	HeO	BT	mPK[4]	mPB
Unweighted marker-specific weight[1]	0.05	U/U[2]	0.04876	0.04960	0.05036	0.05228	0.04914	0.04352	0.04692
		E/E	0.04866	0.04994	0.05016	0.05216	0.04914		
	0.01	U/U	0.00918	0.01012	0.01016	0.01030	0.01034	0.00854	0.01036
		E/E	0.00924	0.00994	0.01008	0.01022	0.01028		
	0.001	U/U	0.00078	0.00082	0.00086	0.00070	0.00084	0.00084	0.00088
		E/E	0.00080	0.00078	0.00084	0.00070	0.00082		
	0.0001	U/U	0.00008	0.00002	0.00006	0.00008	0.00008	0.00006	0.00014
		E/E	0.00006	0.00002	0.00006	0.00008	0.00008		
Weighted marker-specific weight	0.05	U/U	0.05030	0.04998	0.05158	0.05134	0.04696	0.04604	0.04536
		E/E	0.05054	0.05010	0.05176	0.05122	0.04714		
	0.01	U/U	0.00992	0.00942	0.01080	0.00972	0.00888	0.00978	0.01008
		E/E	0.00992	0.00944	0.01088	0.00978	0.00886		
	0.001	U/U	0.00078	0.00086	0.00126	0.00098	0.00082	0.00124	0.00134
		E/E	0.00076	0.00088	0.00122	0.00102	0.00080		
	0.0001	U/U	0.00006	0.00008	0.00006	0.00006	0.00010	0.00002	0.00010
		E/E	0.00006	0.00008	0.00008	0.00006	0.00010		

[1]The unweighted marker-specific weight is given by $w_l = \text{Beta}(m_l, 1, 1) = 1$; the weighted marker-specific weight is given by $w_l = \text{Beta}(m_l, 1, 25)$. [2]U/U represents the structures of the working within-cluster and multivariate-response correlation matrices considered by the unstructured structures; E/E represents the structures of the working within-cluster and multivariate-response correlation matrices considered by the exchangeable structures. [3]HoK, HoO, HeK, HeO, and BT are our proposed methods. [4]mPK and mPB are executed by the R package Multi-SKAT [39].

Table 2: Empirical type I errors of the five competing methods with binary traits.

Marker-specific weight (w_l)	Nominal level	Working correlation	Method				
			HoK[3]	HoO	HeK	HeO	BT
Unweighted marker-specific weight[1]	0.05	U/U[2]	0.04944	0.05154	0.05086	0.05280	0.04952
		E/E	0.04930	0.05144	0.05068	0.05318	0.04946
	0.01	U/U	0.00974	0.00994	0.00982	0.01026	0.01000
		E/E	0.00974	0.00998	0.00984	0.01028	0.00998
	0.001	U/U	0.00068	0.00084	0.00100	0.00098	0.00106
		E/E	0.00066	0.00084	0.00102	0.00094	0.00104
	0.0001	U/U	0.00008	0.00002	0.00012	0.00010	0.00000
		E/E	0.00008	0.00002	0.00012	0.00010	0.00002
Weighted marker-specific weight	0.05	U/U	0.05170	0.04900	0.05256	0.04922	0.04576
		E/E	0.05168	0.04920	0.05232	0.04930	0.04556
	0.01	U/U	0.01028	0.00976	0.00996	0.00972	0.00886
		E/E	0.01024	0.00982	0.00986	0.00976	0.00884
	0.001	U/U	0.00110	0.00080	0.00096	0.00090	0.00088
		E/E	0.00112	0.00076	0.00096	0.00088	0.00090
	0.0001	U/U	0.00004	0.00008	0.00010	0.00012	0.00006
		E/E	0.00006	0.00008	0.00010	0.00012	0.00008

[1]The unweighted marker-specific weight is given by $w_l = \text{Beta}(m_l, 1, 1) = 1$; the weighted marker-specific weight is given by $w_l = \text{Beta}(m_l, 1, 25)$. [2]U/U represents the structures of the working within-cluster and multivariate-response correlation matrices considered by the unstructured structures; E/E represents the structures of the working within-cluster and multivariate-response correlation matrices considered by the exchangeable structures. [3]HoK, HoO, HeK, HeO, and BT are our proposed methods.

when the genetic effects on the different phenotypes are homogeneous (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$). As expected, in general, the heterogeneous kernel statistic (HeK) is more powerful than the homogeneous kernel statistic (HoK), when the genetic effects on the different phenotypes are heterogeneous (i.e., $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$). On the other hand, the homogeneous kernel statistic (HoK) is more powerful than the heterogeneous kernel statistic (HeK), when the genetic effects on the different phenotypes

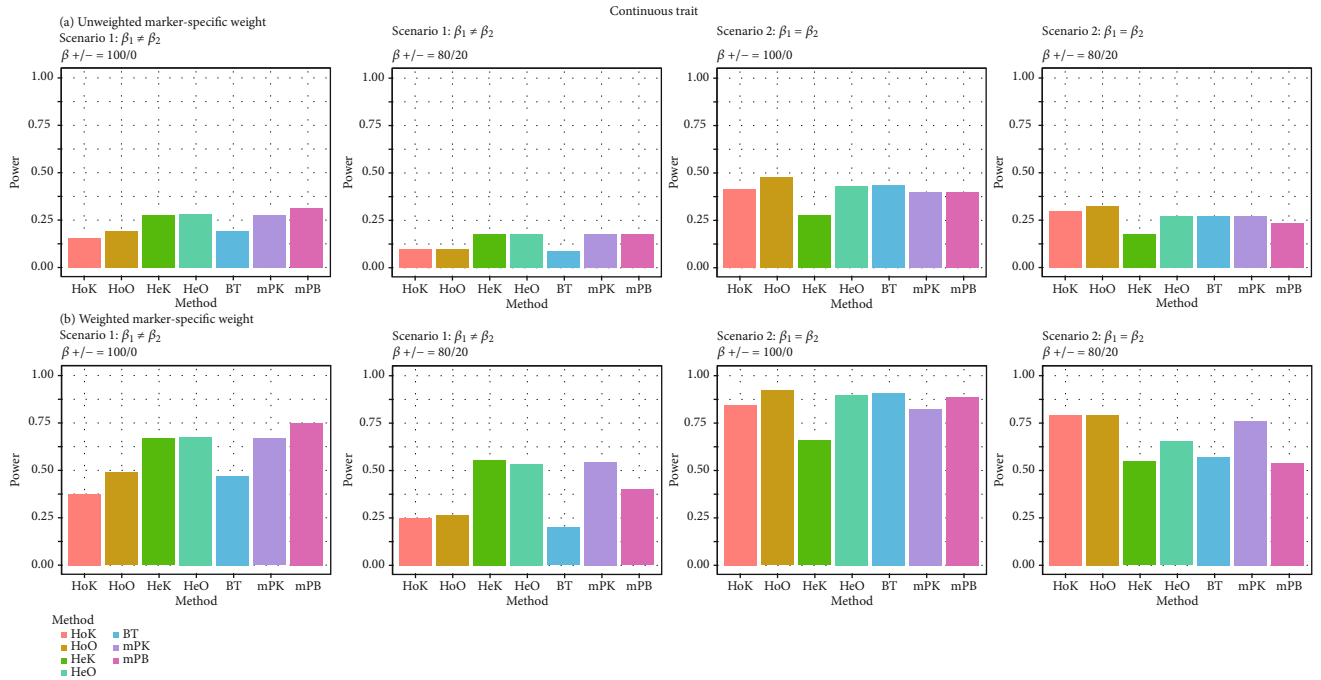

FIGURE 1: Power comparisons of the seven competing methods with continuous traits for each scenario at the nominal level of 0.001. (a) Unweighted marker-specific weight: $w_l = \text{Beta}(m_l, 1, 1) = 1$. (b) Weighted marker-specific weight: $w_l = \text{Beta}(m_l, 1, 25)$.

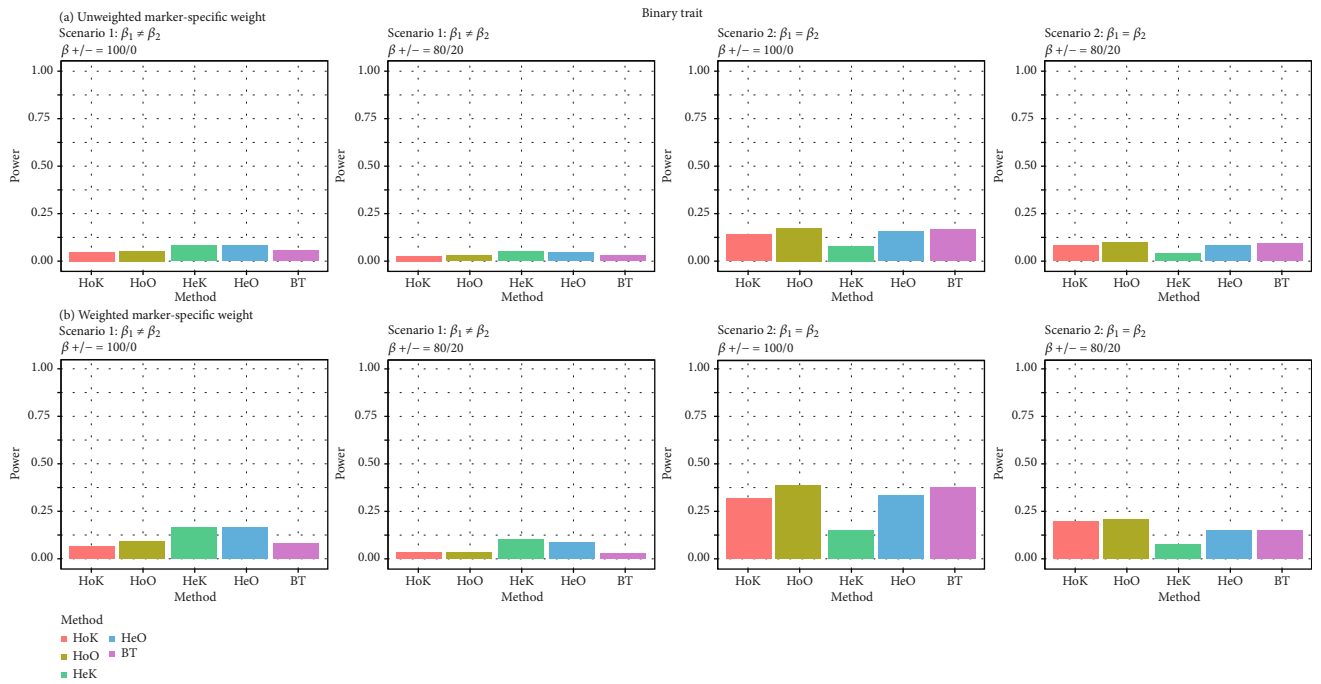

FIGURE 2: Power comparisons of the five competing methods with binary traits for each scenario at the nominal level of 0.001. (a) Unweighted marker-specific weight: $w_l = \text{Beta}(m_l, 1, 1) = 1$. (b) Weighted marker-specific weight: $w_l = \text{Beta}(m_l, 1, 25)$.

are homogeneous (i.e., $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$). In a word, the proposed methods, HoK, HoO, HeK, HeO, and BT, have their respective merits in examining the association between genetic effects and multiple binary traits for autosome analyses.

Similarly, when the working within-cluster and multivariate-response correlation matrices of the proposed methods, HoK, HeK, and BT, are considered to be unstructured, the empirical power rates have similar results and thus

they are omitted. On the other hand, the empirical power rates of the proposed methods for X chromosome analyses with binary traits are presented in Figure S2 in Appendix S3. These empirical power rates show similar results as that discussed in Figure 2.

In summary, the seven competing methods, HoK, HoO, HeK, HeO, BT, mPK, and mPB, have their respective merits in diagnosing whether genetic effects are associated with multiple continuous traits for autosome analyses or the X chromosome analyses. Similarly, the proposed methods, HoK, HoO, HeK, HeO, and BT, have their respective advantages in examining whether there are associations between genetic effects and multiple binary traits for autosome analyses or the X chromosome analyses.

To furthermore examine the performance of the proposed methods, additional simulation studies for continuous traits and binary traits are presented in Appendix S4 and Appendix S5 with higher correlations of phenotypes and higher dimensions of phenotypes considered, respectively. In general, these competing methods based on higher correlations of phenotypes or higher dimensions of phenotypes can provide a bigger empirical power rate for the analysis of continuous traits or binary traits. However, we note that these competing methods based on higher correlations of phenotypes or higher dimensions of phenotypes more easily have empirical type I error rate inflation at a smaller nominal level, especially for binary data analysis (Appendix S5: Tables S5-S6 and Appendix S6: Table S7), in comparison with these methods based on lower correlations of phenotypes or lower dimensions of phenotypes. A detailed discussion of these additional simulation results is given in Appendixes S4 and S5.

However, we note that the proposed methods have a high computational cost, especially for binary data. Under our simulation setting and framework, we carry out a single simulated data set by using a computer based on one CPU core at 2.1 GHz. The average computational times of the homogeneous and heterogeneous tests with a weighted marker-specific weight $w_l = \text{Beta}(m_l, 1, 25)$ under the alternative hypothesis for continuous data are 0.83 and 0.91 minutes, respectively, while that for binary data are 4.77 and 4.80 minutes, respectively. Therefore, in the current version, such a framework algorithm implementation is unsatisfactory for analyzing a large-scale high-dimensional data set in practice.

5. Conclusion

In this investigation, we develop a retrospective framework for identifying the pleiotropic effects of genetic variants on multivariate traits by using collapsing and kernel methods with pedigree- or population-structured data. The proposed framework, corresponding to the burden test, the kernel test, and the omnibus test, provides a sound basis for genetic association analyses for autosomes and the X chromosome. The proposed multivariate trait association methods based on the JGEE model can flexibly accommodate continuous phenotypes or binary phenotypes and further can adjust for covariates.

One critical advantage of the proposed methods is that the homogeneous kernel statistic (HoK), the heterogeneous kernel statistic (HeK), and the burden test (BT) retain all of the benefits of the retrospective tests proposed by Schaid et al. [43] who treated the genotype data as random variables by conditioning the phenotypes as constants. On the other hand, the homogeneous omnibus test (HoO) and the heterogeneous kernel statistic (HeO) keep the advantages of the Cauchy combination tests proposed by Liu and Xie [48] who showed that the Cauchy combination tests are robust to model misspecification and robustly protect the type I error rates [49].

Another important benefit of the proposed method is that the HoK test, the HeK test, and the BT test keep the benefits of the JGEE model that validly account for complex correlations between subjects within the cluster (within-cluster correlations) and between different phenotypes from the same subjects (multivariate-response correlations). Moreover, the proposed test statistics, HoK, HeK, and BT, based on the JGEE model can efficaciously account for covariate adjustment whether the phenotypes are continuous or binary.

Our simulation studies show that an unweighted marker-specific weight $w_l = \text{Beta}(m_l, 1, 1) = 1$ and an exchangeable structure of the working within-cluster and multivariate-response correlations are recommended for the practical data analysis if the data cannot sufficiently provide valid information for estimating the structures of the working within-cluster and multivariate-response correlations before the start of the data analysis. Moreover, the homogeneous kernel statistic (HoK) is more robust than the heterogeneous statistic (HeK) in controlling the empirical type I errors, because the null distribution of the HeK statistic asymptotically follows a mixture chi-square distribution with a larger degree of freedom, in comparison with the null distribution of the HoK statistic. However, the HeK statistic is more powerful than the HoK statistic when the genetic effects on the different phenotypes are heterogeneous.

On the other hand, our simulation results show that for the autosome analyses or the X chromosome analyses with continuous traits, the seven competing methods, HoK, HoO, HeK, HeO, BT, mPK, and mPB, show good performance with well-controlled type I errors, while the seven competing methods have their respective merits for identifying the association between the genetic effects and multiple continuous traits. In addition, our simulation results show that for the autosome analyses or the X chromosome analyses with binary traits, the proposed methods, HoK, HoO, HeK, HeO, and BT, can control empirical type I errors with lower correlations of phenotypes or with lower dimensions of phenotypes (Table *2* and Table S2), while these proposed methods have their respective advantages for identifying the genetic variants associated with multiple binary traits. However, we observe that the proposed methods, HoK, HoO, HeK, HeO, and BT, with higher correlations of phenotypes or with higher dimensions of phenotypes, more easily have the infection of empirical type I errors at a smaller nominal level (Appendix S5: Tables S5-S6 and Appendix S6: Table S7), although these method under such situations have higher empirical power rates.

6. Limitation

The proposed multivariate trait association methods have their limitations. First, these proposed methods cannot simultaneously include the continuous traits and binary traits in analysis. Thus, future studies are needed to extend the idea of the proposed multivariate trait association methods for simultaneously considering continuous traits and binary traits in analysis. Second, the multivariate trait association methods, based on higher correlations of phenotypes or higher dimensions of phenotypes, easily suffer from the problem of the inflated type I errors, especially when the binary traits are considered (Appendix S5: Tables S5-S6 and Appendix S6: Table S7). Although the JGEE model provides an efficient algorithm for estimating the structure of the working within-cluster and multivariate-response correlations, a large-scale pedigree study always suffers from a more complex and high-dimensional structure of the within-cluster and multivariate-response correlations in pedigree database analysis. Hence, in the future, a more effective algorithm for estimating the complicated and high-dimensional (or higher correlational) structure of the working within-cluster and multivariate-response correlations is necessary to be proposed, especially when the analysis focuses on the binary traits. Third, in comparison with the null distribution of the homogeneous kernel statistic, the null distribution of the heterogeneous kernel statistic follows a larger degree of freedom test, which easily causes such a heterogeneous test to suffer from the problem of the type I error inflation. Therefore, overcoming the problem of the type I error inflation from the heterogeneous test is an essential part of the future work. Fourth, the proposed methods, which have a high computational cost especially for binary data, are inappropriate for analyzing large-scale high-dimensional data in practice. Thus, a more effective algorithm for reducing computational cost is needed to be proposed in further research. Moreover, the software of the proposed methods is computationally inconvenient and particularly inadequate for the mass GWAS data in practice. Therefore, the software of the proposed methods, which is convenient to be used, is a further work in the future. Fifth, our current work focuses mainly on the low- and common-frequency variants. Extension of the proposed methods to the rare variants deserves further works.

Data Availability

The data supporting the findings of this study are available within the article and its supplementary materials.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to thank the editor and the referees for their constructive comments, which significantly improve the presentation of the article. This work is supported by grant MOST 108-2118-M-037-001-MY2 of Ministry of Science and Technology, Taiwan, R.O.C.

Supplementary Materials

Appendix S1: the null distribution of the kernel statistic. Appendix S2: extension to the X chromosome. Appendix S3: simulation results based on the X chromosome. Appendix S4: additional simulation studies for continuous traits. Appendix S5: additional simulation studies for binary traits. Appendix S6: limitation *(Supplementary Materials)*

References

[1] H. Zhu, S. Zhang, and Q. Sha, "A novel method to test associations between a weighted combination of phenotypes and genetic variants," *PLoS One*, vol. 13, no. 1, article e0190788, 2018.

[2] S. Lee, S. Won, Y. J. Kim et al., "Rare variant association test with multiple phenotypes," *Genetic Epidemiology*, vol. 41, no. 3, pp. 198–209, 2017.

[3] Q. Yang and Y. Wang, "Methods for analyzing multivariate phenotypes in genetic association studies," *Journal of probability and statistics*, vol. 2012, Article ID 652569, 13 pages, 2012.

[4] X. Liang, Q. Sha, and S. Zhang, "Joint analysis of multiple phenotypes in association studies using allele-based clustering approach for non-normal distributions," *Annals of Human Genetics*, vol. 82, no. 6, pp. 389–395, 2018.

[5] Z. Wang, Q. Sha, S. Fang, K. Zhang, and S. Zhang, "Testing an optimally weighted combination of common and/or rare variants with multiple traits," *PLoS One*, vol. 13, no. 7, article e0201186, 2018.

[6] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, "Pleiotropy in complex traits: challenges and strategies," *Nature Reviews Genetics*, vol. 14, no. 7, pp. 483–495, 2013.

[7] M. Stephens, "A unified framework for association analysis with multiple related phenotypes," *PLoS One*, vol. 8, no. 7, article e65245, 2013.

[8] X. Zhou and M. Stephens, "Efficient multivariate linear mixed model algorithms for genome-wide association studies," *Nature Methods*, vol. 11, no. 4, pp. 407–409, 2014.

[9] X. Liang, Z. Wang, Q. Sha, and S. Zhang, "An adaptive Fisher's combination method for joint analysis of multiple phenotypes in association studies," *Scientific Reports*, vol. 6, no. 1, article 34323, 2016.

[10] Z. Wang, X. Wang, Q. Sha, and S. Zhang, "Joint analysis of multiple traits in rare variant association studies," *Annals of Human Genetics*, vol. 80, no. 3, pp. 162–171, 2016.

[11] Z. Wang, Q. Sha, and Z. Zhang, "Joint analysis of multiple traits using "optimal" maximum heritability test," *PLoS One*, vol. 11, no. 3, article e0150975, 2016.

[12] H. Zhu, S. Zhang, and Q. Sha, "Power comparisons of methods for joint association analysis of multiple phenotypes," *Human heredity*, vol. 80, no. 3, pp. 144–152, 2015.

[13] H. Aschard, B. Vilhjálmsson, N. Greliche, P. Morange, D. Trégouët, and P. Kraft, "Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies," *The American Journal of Human Genetics*, vol. 94, no. 5, pp. 662–676, 2014.

[14] A. Korte, B. J. Vilhjálmsson, V. Segura, A. Platt, Q. Long, and M. Nordborg, "A mixed-model approach for genome-wide association studies of correlated traits in structured populations," *Nature Genetics*, vol. 44, no. 9, pp. 1066–1071, 2012.

[15] P. F. O'Reilly, C. J. Hoggart, Y. Pomyen et al., "MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS," *PLoS One*, vol. 7, no. 5, article e34861, 2012.

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

[17] M. A. R. Ferreira and S. M. Purcell, "A multivariate test of association," *Bioinformatics*, vol. 25, no. 1, pp. 132-133, 2009.

[18] L. Klei, D. Luca, B. Devlin, and K. Roeder, "Pleiotropy and principal components of heritability combine to increase power for association analysis," *Genetic Epidemiology*, vol. 32, no. 1, pp. 9–19, 2008.

[19] P. C. O'Brien, "Procedures for comparing samples with multiple endpoints," *Biometrics*, vol. 40, no. 4, pp. 1079–1087, 1984.

[20] Q. Yang, H. Wu, C.-Y. Guo, and C. S. Fox, "Analyze multivariate phenotypes in genetic association studies by combining univariate association tests," *Genetic Epidemiology*, vol. 34, no. 5, pp. 444–454, 2010.

[21] S. van der Sluis, D. Posthuma, and C. V. Dolan, "TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies," *PLoS Genet*, vol. 9, no. 1, article e1003235, 2013.

[22] J. Kim, Y. Bai, and W. Pan, "An adaptive association test for multiple phenotypes with GWAS summary statistics," *Genetic Epidemiology*, vol. 39, no. 8, pp. 651–663, 2015.

[23] X. Zhu, T. Feng, B. O. Tayo et al., "Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension," *The American Journal of Human Genetics*, vol. 96, no. 1, pp. 21–36, 2015.

[24] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *The American Journal of Human Genetics*, vol. 83, no. 3, pp. 311–321, 2008.

[25] J. Huang, A. Johnson, and C. O'Donnell, "PRIMe: a method for characterization and evaluation of pleiotropic regions from multiple genome-wide association studies," *Bioinformatics*, vol. 27, no. 9, pp. 1201–1206, 2011.

[26] D. Ray, J. Pankow, and S. Basu, "USAT: a unified score-based association test for multiple phenotype-genotype analysis," *Genetic Epidemiology*, vol. 40, no. 1, pp. 20–34, 2016.

[27] J. Ried, A. Döring, K. Oexle et al., "PSEA: phenotype set enrichment analysis-a new method for analysis of multiple phenotypes," *Genetic Epidemiology*, vol. 36, no. 3, pp. 244–252, 2012.

[28] Q. Yan, D. Weeks, J. Celedón et al., "Associating multivariate quantitative phenotypes with genetic variants in family samples with a novel kernel machine regression method," *Genetics*, vol. 201, no. 4, pp. 1329–1339, 2015.

[29] X. Zhan, N. Zhao, A. Plantinga et al., "Powerful genetic association analysis for common or rare variants with high-dimensional structured traits," *Genetics*, vol. 206, no. 4, pp. 1779–1790, 2017.

[30] A. Maity, P. Sullivan, and J. Tzeng, "Multivariate phenotype association analysis by marker-set kernel machine regression," *Genetic Epidemiology*, vol. 36, no. 7, pp. 686–695, 2012.

[31] UK10K Consortium, J. Sun, K. Oualkacha et al., "A method for analyzing multiple continuous phenotypes in rare variant association studies allowing for flexible correlations in variant effects," *European Journal of Human Genetics*, vol. 24, no. 9, pp. 1344–1351, 2016.

[32] Y. Wang, A. Liu, J. L. Mills et al., "Pleiotropy analysis of quantitative traits at gene level by multivariate functional linear models," *Genetic Epidemiology*, vol. 39, no. 4, pp. 259–275, 2015.

[33] F. P. Casale, B. Rakitsch, C. Lippert, and O. Stegle, "Efficient set tests for the genetic analysis of correlated traits," *Nature Methods*, vol. 12, no. 8, pp. 755–758, 2015.

[34] K. A. Broadaway, D. J. Cutler, R. Duncan et al., "A statistical approach for testing cross-phenotype effects of rare variants," *The American Journal of Human Genetics*, vol. 98, no. 3, pp. 525–540, 2016.

[35] A. Cichonska, J. Rousu, P. Marttinen et al., "metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis," *Bioinformatics*, vol. 32, no. 13, pp. 1981–1989, 2016.

[36] J. Lin, R. Tabassum, S. Ripatti, and M. Pirinen, "MetaPhat: detecting and decomposing multivariate associations from univariate genome-wide association statistics," *Frontiers in Genetics*, vol. 11, 2020.

[37] S. Bhattacharjee, P. Rajaraman, K. B. Jacobs et al., "A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits," *The American Journal of Human Genetics*, vol. 90, no. 5, pp. 821–835, 2012.

[38] 23andMe Research Team, Social Science Genetic Association Consortium, P. Turley et al., "Multi-trait analysis of genome-wide association summary statistics using MTAG," *Nature Genetics*, vol. 50, no. 2, pp. 229–237, 2018.

[39] D. Dutta, L. Scott, M. Boehnke, and S. Lee, "Multi-SKAT: general framework to test for rare-variant association with multiple phenotypes," *Genetic Epidemiology*, vol. 43, no. 1, pp. 4–23, 2019.

[40] G. Inan and R. Yucel, "Joint GEEs for multivariate correlated data with incomplete binary outcomes," *Journal of Applied Statistics*, vol. 44, no. 11, pp. 1920–1937, 2017.

[41] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, no. 1, pp. 13–22, 1986.

[42] G. Inan, "JGEE: joint generalized estimating equation solver," 2015, *R package version 1.1*.

[43] D. J. Schaid, S. K. McDonnell, J. P. Sinnwell, and S. N. Thibodeau, "Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data," *Genetic Epidemiology*, vol. 37, no. 5, pp. 409–418, 2013.

[44] J. Sinnwell, T. Therneau, D. Schaid, E. Atkinson, and C. Mester, "kinship2," 2020, *R package version 1.8.5*.

[45] T. Thornton and M. S. McPeek, "ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure," *The American Journal of Human Genetics*, vol. 86, no. 2, pp. 172–184, 2010.

[46] H. Chen, J. B. Meigs, and J. Dupuis, "Sequence kernel association test for quantitative traits in family samples," *Genetic Epidemiology*, vol. 37, no. 2, pp. 196–204, 2013.

[47] Y. Liu, S. Chen, Z. Li, A. C. Morrison, E. Boerwinkle, and X. Lin, "ACAT: a fast and powerful p value combination

method for rare-variant analysis in sequencing studies," *The American Journal of Human Genetics*, vol. 104, no. 3, pp. 410–421, 2019.

[48] Y. Liu and J. Xie, "Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures," *Journal of the American Statistical Association*, vol. 115, no. 529, pp. 393–402, 2020.

[49] Z. R. McCaw, J. M. Lane, R. Saxena, S. Redline, and X. Lin, "Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies," *Biometrics*, vol. 76, no. 4, pp. 1262–1272, 2020.

[50] Z. R. McCaw, "Rank normal transformation omnibus test," 2019, *R package version 0.7.1*.

[51] S. Lee, T. M. Teslovich, M. Boehnke, and X. Lin, "General framework for meta-analysis of rare variants in sequencing association studies," *The American Journal of Human Genetics*, vol. 93, no. 1, pp. 42–53, 2013.

[52] S. F. Schaffner, C. Foo, S. Gabriel, D. Reich, M. J. Daly, and D. Altshuler, "Calibrating a coalescent simulation of human genome sequence variation," *Genome Research*, vol. 15, no. 11, pp. 1576–1583, 2005.

[53] A. Amatya, H. Demirtas, and R. Gao, "Simultaneous generation of multivariate binary and normal variates," 2020, *R package version 2.3.2*.

[54] R. Torices and A. J. Muñoz-Pajares, "Phenotypic integration index," 2017, *R package version 1.3.1*.

[55] R. Torices and A. J. Muñoz-Pajares, "PHENIX: an R package to estimate a size-controlled phenotypic integration index," *Applications in Plant Sciences*, vol. 3, no. 5, article 1400104, 2015.

[56] A. Dah, V. Hore, V. Iotchkova, and J. Marchini, "Network inference in matrix-variate Gaussian models with non-independent noise," 2013, https://arxiv.org/abs/1312.1622.